UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

ADRIANO QUILIÃO DE OLIVEIRA

Síntese de Fotografias e Vídeos com Depth-Image-Based Rendering

Tese apresentada como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação

Orientador: Prof. Dr. Marcelo Walter
Co-orientador: Prof. Dr. Cláudio Rosito Jung

Porto Alegre
2019

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

de Oliveira, Adriano Quilião

Síntese de Fotografias e Vídeos com Depth-Image-Based Rendering / Adriano Quilião de Oliveira. – Porto Alegre: PPGC da UFRGS, 2019.

146 f.: il.

Tese (doutorado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2019. Orientador: Marcelo Walter; Co-orientador: Cláudio Rosito Jung.

1. DIBR. 2. Síntese de vistas. 3. Inpainting. 4. Informação temporal. I. Walter, Marcelo. II. Jung, Cláudio Rosito. III. Título.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. Rui Vicente Oppermann
Vice-Reitora: Profa. Jane Fraga Tutikian
Pró-Reitor de Pós-Graduação: Prof. Celso Giannetti Loureiro Chaves
Diretor do Instituto de Informática: Prof. Carla Maria Dal Sasso Freitas
Coordenador do PPGC: Prof[a]. Luciana Salete Buriol
Bibliotecária-chefe do Instituto de Informática: Beatriz Regina Bastos Haro

“The pessimist sees difficulty in every opportunity.
The optimist sees the opportunity in every difficulty.”
— WINSTON CHURCHILL

AGRADECIMENTOS

Primeiramente, gostaria de agradecer aos meus orientadores Marcelo Walter e Cláudio Jung. A disponibilidade, apoio, confiança e ótima orientação foram imprescindíveis para o desenvolvimento desta tese. Agradeço também por acreditarem em mim oito anos atrás, esta oportunidade me proporcionou uma experiência de aprendizado e crescimento pessoal imensurável, não tenho palavras para agradecer.

Aos colegas de laboratório compartilhado que estiveram ao meu lado durante este desafio, apoiando, auxiliando e dando sempre ótimas dicas. Neste ponto, deixo um agradecimento especial ao meu amigo Thiago, que mesmo muito ocupado, sempre encontrou tempo para me ajudar com revisões e dar conselhos sobre a pesquisa. Aqui, já estendo meus agradecimentos aos colegas de apartamento, Uillian, Gustavo e, em especial, ao irmão que me acompanha desde os tempos do Alegrete, Arthur, pela parceria e ótimo ambiente de convivência.

Ao PPGC pela excelente estrutura e seus professores pelos preciosos ensinamentos. Juntamente, às agências governamentais CAPES e CNPq, pelo suporte financeiro prestado durante o desenvolvimento deste trabalho.

Gostaria de agradecer aos amigos de Cachoeira, por toda preocupação, reza, conselhos, brincadeiras, churrascos, cervejadas (desde a produção até o consumo), isso me ajudou a manter à calma, mesmo nos momentos mais complicados. Aqui, deixo um agradecimento especial ao Marcel e ao Jefferson, que me acompanharam desde às caronas para Alegrete até o dia da defesa, o apoio de vocês foi imprescindível. Aos amigos do Alegrete, Rafael, Robson e Marcelo, que mesmo com a distância sempre deram um jeito de se fazer presentes apoiando.

Por fim, agradeço ao mais importante, minha família. Aos meus pais, minha fortaleza e motivo de tanta vontade de lutar para alcançar objetivos como este, obrigado por todo amor e apoio incondicional. Infelizmente, não posso agradecer pessoalmente ao meu irmão, mas sei, que mesmo longe, ele está torcendo e me ajudado de alguma forma a alcançar os meu sonhos. À minha namorada Andriza, por me aguentar, apoiar e acreditar, mesmo quando eu não acreditei, me dando força e ajudando todos os dias em tudo que fosse possível. Também, gostaria de agradecer ao restante da minha família, tios, tias, primos e avós, por tudo, vocês são muito importantes. Encerro com um agradecimento especial aos meus tios Nilton e Paula, que juntamente com meus pais, me incentivaram a estudar, antes mesmo de eu ter ciência do quanto isso seria importante na minha vida.

RESUMO

O processo de síntese de vistas com *Depth-Image-Based Rendering* (DIBR) se apresenta como um meio promissor para viabilizar aplicações como TV3D, *Free Viewpoint Video*, e outras relacionadas com Realidade Virtual e Realidade Aumentada. DIBR permite que sejam produzidos inúmeros pontos de vista virtuais da mesma cena utilizando apenas uma imagem de referência e seu respectivo mapa de profundidades. Contudo, artefatos (*cracks* e *ghosts*) e regiões sem informação (*holes*) são formados no processo de síntese, os quais precisam ser tratados ou preenchidos. Nesta tese, são apresentadas duas abordagens para a síntese de vistas com DIBR: ATA e DHS. A abordagem ATA foi desenvolvida a partir de estudos aprofundados acerca da geração de fotografias sintéticas. Esta identifica *cracks* vazios e translúcidos e reconstrói as regiões afetadas com um algoritmo especializado. *Ghosts* são identificados e reprojetados para as posições corretas de acordo com um processo de avaliação. As regiões sem informação restantes são preenchidas com uma adaptação de um popular algoritmo de *inpainting*, que emprega *patches* com tamanho dinâmico, copiados da imagem de referência, e que se ajusta a diferentes tipos de *hole*. Já a abordagem DHS utiliza os avanços produzidos com ATA, apresentando um método de reconstrução dos *holes* ainda mais robusto e confiável, ciente da estrutura e composição da imagem. *Ghosts* são tratados antes da geração da vista virtual. Para os *holes*, utiliza-se um algoritmo de *inpainting* baseado em *superpixels* hierárquicos, que reconstrói as regiões vazias com base na composição de sua vizinhança, copiando conteúdo da imagem de referência. Adicionalmente, propõe-se um método robusto para a geração de um modelo de *background* incremental para vídeos, que pode ser incorporado em qualquer abordagem DIBR. Como exemplo, detalha-se sua integração ao DHS, que apresentou melhores resultados na avaliação quadro a quadro realizada. Exaustivos testes comprovam que as abordagens propostas apresentam melhores resultados quantitativos e qualitativos quando comparados com diversos métodos recentes e competitivos, tanto na geração de fotografias como de vídeos sintéticos, em testes com mapas de profundidade *ground truth* e reais. Como contribuição adicional desta tese, avaliou-se o impacto do uso de mapas de profundidade reais produzidos com casamento estéreo no processo de síntese com DIBR, e analisou-se a relação entre métricas de qualidade empregadas em ambos problemas.

Palavras-chave: DIBR. síntese de vistas. inpainting. informação temporal.

Synthesis of Still Images and Videos with Depth-Image-Based Rendering

ABSTRACT

The view synthesis process with Depth-Image-Based Rendering (DIBR) is presented as a promising way to enable applications like TV3D, Free Viewpoint Video, and others related to Virtual Reality and Augmented Reality. DIBR allows numerous virtual views of the same scene to be produced using only a single reference image and its depth map. However, artifacts (cracks and ghosts) and regions without information (holes) are formed in the synthesis process, which need to be treated or filled. In this thesis, we present two approaches for view synthesis with DIBR: ATA and DHS. We developed the ATA approach from in-depth studies on the generation of synthetic still images. This approach identifies empty and translucent cracks and reconstructs the affected regions with a specialized algorithm. Then, ghosts are identified through an evaluation process and warped to the correct positions. Finally, the remaining empty regions are filled with an adaptation of a popular inpainting algorithm that employs dynamically sized patches copied from the reference image and fits different hole types. The DHS approach, on the other hand, uses the advances produced with ATA, presenting an even more robust and reliable hole reconstruction method, aware of the image structure and composition. Ghosts are treated before the virtual view generation. An inpainting algorithm based on hierarchical superpixels is used for hole filling, which reconstructs empty regions based on their neighborhood composition by copying the contents of the reference image. Additionally, we propose a robust method for generating an incremental background model for videos that can be incorporated into any DIBR approach. As an example, we detailed its integration with DHS, which presented better results in the frame-by-frame evaluation. Exhaustive testing proves that the proposed approaches yield better quantitative and qualitative results when compared to several recent and competitive methods, both in the generation of synthetic still images as in videos, in tests with ground truth and real depth maps. As an additional contribution of this thesis, we evaluated the impact of using real depth maps produced with stereo matching in the synthesis process with DIBR and analyzed the relationship between quality metrics employed in both problems.

Keywords: DIBR, view synthesis, inpainting, temporal information.

LISTA DE ABREVIATURAS E SIGLAS

AR	*Augmented Reality*
BGE	*Background Extension*
CRF	*Conditional Random Field*
d-D	d-Dimensional
DIBR	*Depth-Image-Based Rendering*
FDC	*Foreground Depth Correlation*
FVV	*Free Viewpoint Video*
GMM	*Gaussian Mixture Model*
FDC	*Foreground Depth Correlation*
HHF	*Hierarchical Hole-Filling*
LDI	*Layered Depth Image*
MRF	*Markov Random Field*
MST	*Minimum Spanning Tree*
MVD	*Multiview Video-Plus-Depth*
MSE	*Mean Square Error*
MW-MSE	*Multi-scale Wavelet Mean Squared Error*
MW-PSNR	*Morphological-Wavelet PSNR*
OOFAs	*Out-of-Field Areas*
PSNR	*Peak Signal-to-Noise Ratio*
RANSAC	*Random Sample Consensus*
RGB	*Red, Green, Blue*
RMS	*Root Mean Square*
STRRED	*Spatio-Temporal-Reduced Reference Entropic Differences*
SSD	*Sum of Squared Differences*

SSIM	*Structural Similarity Index*
TV3D	Televisão 3D
VR	*Virtual Reality*
VSRS	*View Synthesis Reference Software*

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Motivação

Nos últimos anos, diversos avanços ocorreram tanto em tecnologias para visualização de vídeos e imagens em 3D quanto na produção de conteúdo utilizando múltiplas câmeras e sensores para captação de profundidade em cenas. Aparelhos celulares lançados recentemente, como o iPhone Xs[1] da Apple e o Galaxy S9+[2] da Samsung, possuem câmera embutida de visão estéreo de fácil uso. Esses aparelhos permitem que o usuário final capture fotografias em 3D (que pode ser obtida pelo uso sincronizado das duas câmeras) e ganham destaque como uma forma promissora para gravação, armazenamento e compartilhamento de pontos de vista, tanto para fotografias como vídeos. Em paralelo, diversos trabalhos estão sendo propostos na literatura, com o objetivo de viabilizar aplicações como Televisão em 3D (TV3D) e *Free Viewpoint Video* (FVV) (LUO et al., 2016; KELLNHOFER et al., 2017; RAHAMAN; PAUL, 2018), que visam prover uma experiência mais realista (com a percepção de profundidade) e interativa (com a livre navegação em cena) para o usuário. Simultaneamente, novos modelos de interatividade estão emergindo, como é o caso do Facebook 3D Photos[3], que fornece para o usuário a sensação de profundidade e movimento, para uma única imagem estática.

Para exibição de conteúdo em 3D, *displays* estereoscópicos clássicos requerem um par de imagens estéreo para cada quadro de vídeo. Com uma tecnologia diferente, *displays* autoestereoscópicos (ou automultiscópicos) de múltiplas vistas proporcionam uma experiência 3D imersiva sem o uso de óculos, exibindo imagens diferentes de acordo a posição do visualizador, sendo considerados como o futuro da televisão e do cinema (DU et al., 2014). Contudo, para cada quadro de vídeo, esses monitores fornecem até 128 vistas separadas (SCHWARZ; OLSSON; SJöSTRöM, 2013), e imagens correspondentes são necessárias. Da mesma forma, sistemas para FVV requerem um grande número de pontos de vista, mesmo quando permitem uma pequena liberdade de navegação (RAHAMAN; PAUL, 2018). Tornar isso possível representa um desafio, pois capturar e transmitir pontos de vista arbitrários exigiria muitas câmeras, codificação complexa e uma grande largura de banda (LUO; ZHU, 2017). Neste contexto, destaca-se a síntese de vistas com

[1] Mais detalhes sobre o iPhone Xs em: <https://www.apple.com/lae/iphone-xs/>.

[2] Mais detalhes sobre o Galaxy S9+ em: <https://www.samsung.com/br/smartphones/galaxy-s9/>.

[3] Mais detalhes sobre a aplicação estão disponíveis em: <https://facebook360.fb.com/2018/10/11/3d-photos-now-rolling-out-on-facebook-and-in-vr/>.

o modelo *Depth-Image-Based Rendering* (DIBR) (FEHN, 2004) como uma boa solução para este problema, uma vez que permite a geração de múltiplos pontos de vista "virtuais" da mesma cena utilizando apenas uma imagem (ou quadro de vídeo) e seu respectivo mapa de profundidades.

No modelo DIBR, novas vistas virtuais (ou sintéticas) são produzidas por meio da equação de 3D *warping* (MARK; MCMILLAN; BISHOP, 1997), utilizada para projetar cada *pixel* da imagem real para um ponto de vista estipulado com base em seu valor de profundidade e parâmetros de câmera. As vistas sintéticas podem ser empregadas tanto na produção de pontos de vista não disponíveis para FVV quanto nos diferentes formatos para exibição em 3D ou, ainda, em qualquer aplicação que necessite de duas ou múltiplas imagens de uma mesma cena. Contudo, diferentes tipos de artefatos aparecem na vista sintética após a projeção, como *cracks*[4], *ghosts*, *disocclusions* e *out-of-field areas* (OOFAs) (MUDDALA; SJÖSTRÖM; OLSSON, 2016), os quais precisam ser tratados.

Os *cracks* podem ocorrer devido a erros de arredondamento na estimativa das coordenadas com 3D *warping* (AHN; KIM, 2013) e/ou por descontinuidades de profundidade em regiões com textura homogênea na imagem de referência (MORI et al., 2008). Como normalmente os valores de profundidade não são estimados para cada *pixel* de maneira isolada, quando um ponto da imagem de origem é projetado para a vista virtual, o erro faz com que uma linha longa e fina sem informação se forme na imagem sintética. A Figura 1.1(a) destaca em vermelho diversas ocorrências deste artefato. Como variação, os *cracks* podem se apresentar na forma translúcida, como pode ser visto na Figura 1.1(b), onde informação do *background*[5] foi projetada no interior do artefato (como destacado por setas amarelas), preenchendo-o com conteúdo inadequado (MUDDALA; SJÖSTRÖM; OLSSON, 2016). Após a detecção das ocorrências de ambas as formas deste artefato, faz-se necessário remover o conteúdo (quando houver) das regiões identificadas e reconstruí-las.

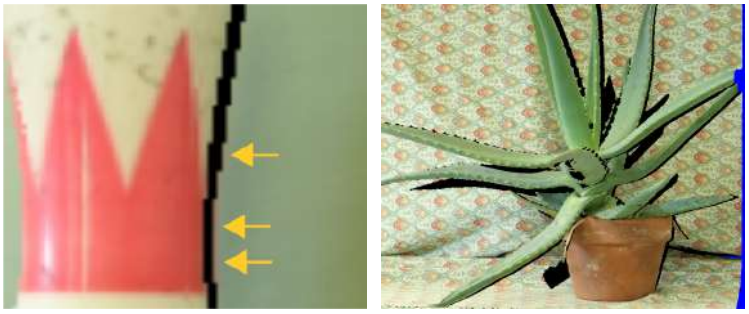
Regiões oclusas no ponto de vista utilizado para captura da imagem de referência podem se tornar visíveis na posição escolhida para projeção (PURICA et al., 2015). Neste caso, formam-se as *disocclusions*, que são regiões sem informação de projeção no ponto de vista virtual, tipicamente maiores do que os *cracks* e com formato irregular. Essas regiões correspondem a partes do *background* que foram expostas, as quais estavam sendo

[4]Neste trabalho, optou-se por manter os termos em inglês associados a nomenclatura empregada para o modelo DIBR na literatura.

[5]Nesta tese, quando usados os termos *background* e *foreground*, estes fazem referência a dois objetos vizinhos em uma região delimitada da imagem, onde o primeiro está em uma camada de profundidade mais ao fundo em relação ao segundo.

Figura 1.1: Artefatos encontrados no processo de geração de vistas sintéticas com o modelo DIBR. (a) *Cracks* vazios sinalizados em vermelho; (b) *Crack* translúcido em azul, indicado por setas amarelas; (c) *Disocclusions* preenchidas com a cor verde. (d) *Ghost* apontado por setas amarelas na borda de objeto no *background* (em verde); (e) OOFAs destacadas pela cor azul na extremidade direita da imagem.

(a) (b) (c)



(d) (e)

Fonte: As imagens de (a)-(d) foram retiradas de (OLIVEIRA, 2016). (e) O autor. Originalmente, todas as imagens foram adaptadas de (HIRSCHMULLER; SCHARSTEIN, 2007).

ocluídas por algum objeto do *foreground* na imagem de referência (LUO et al., 2016). Na Figura 1.1(c), são exibidas ocorrências deste artefato em verde, onde, por exemplo, pode ser observado que o pincel (no *foreground*) sobrepõe o busto (no *background)*. Em relação a este artefato, o desafio consiste em realizar seu preenchimento de maneira coerente, considerando que para áreas perdidas ou danificadas, só se pode esperar produzir algo plausível, ao invés de uma reconstrução exata (OLIVEIRA et al., 2001).

As bordas das *disocclusions* podem conter artefatos, denominados *ghosts*, que ocorrem quando as descontinuidades de profundidade não são suficientemente nítidas no domínio da imagem (OLIVEIRA et al., 2015). Neste caso, ao definir-se o valor de profundidade associado a região de transição entre dois objetos em camadas diferentes na cena, atribui-se equivocadamente o mesmo valor de profundidade do *background* para a borda do objeto no *foreground*. Então, em razão deste erro na definição do valor de profundidade, a silhueta do objeto no *foreground* aparece na borda do *background* da *disocclusion* após a projeção. A Figura 1.1(d) destaca um exemplo deste artefato com

setas amarelas. Essas regiões devem ser removidas ou tratadas, de modo que o efeito causado pelo artefato seja suprimido.

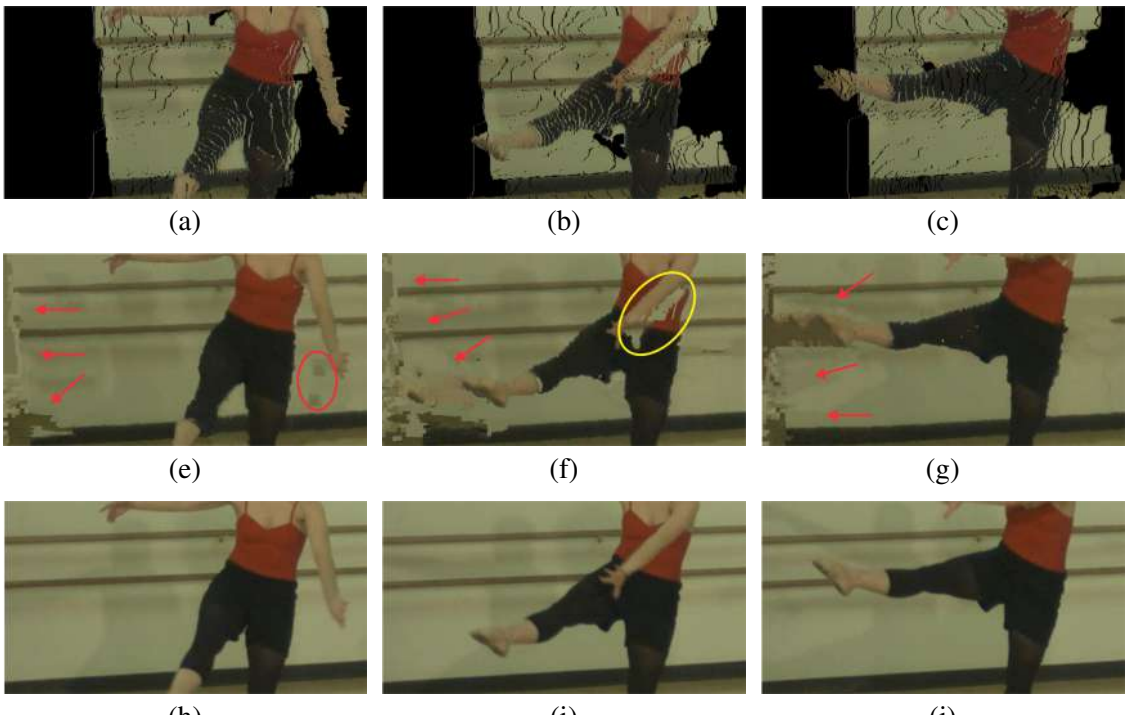
Por fim, as OOFAs ocorrem porque o ponto de vista virtual extrapola os limites da vista de referência, criando regiões sem informação nas extremidades da imagem sintética (MUDDALA; SJÖSTRÖM; OLSSON, 2016). Essas regiões são geralmente grandes e podem ser localizadas de acordo com a direção adotada na projeção. Por exemplo, se houver apenas deslocamento horizontal para a esquerda em relação ao ponto de vista real, as OOFAs serão identificadas na borda direita da imagem, como ocorre na Figura 1.1(e). Assim como as *disocclusions*, essas regiões precisam ser preenchidas e, inclusive, algumas abordagens tratam ambos genericamente como *holes*, apesar de não apresentarem as mesmas características de vizinhança e composição.

Para a aplicação em vídeos, o modelo DIBR demanda o uso de mecanismos de controle de informação temporal. Erros no mapa de profundidades podem fazer com que pequenos artefatos apareçam e sumam durante a sucessão de quadros, possivelmente causando *flickering*, o que prejudica a experiência do usuário (FICKEL, 2015). O mesmo pode ocorrer por incoerências no preenchimento de *disocclusions* e OOFAs, pois padrões de reconstrução distintos para regiões iguais em quadros sucessivos fazem com que a estrutura da cena mude ao longo do tempo. A segunda linha da Figura 1.2 destaca com setas vermelhas a variação da reconstrução para três quadros consecutivos, utilizando a abordagem de (LUO et al., 2016).

O modelo DIBR apresenta-se como peça-chave para a viabilização de diversas tecnologias (LUO; ZHU, 2017). Entretanto, mesmo diante de diversas abordagens propostas recentemente na literatura, fica evidente que muito esforço ainda precisa ser empreendido para tornar viável sua aplicação real, como destaca a Figura 1.2. Em geral, trabalhos correlatos priorizam a reconstrução das *disocclusions* (KöPPEL; MüLLER; WIEGAND, 2016; RAHAMAN; PAUL, 2018; LIE; HSIEH; LIN, 2018; LUO et al., 2019), e concentram seus esforços somente na utilização adequada de informação do *background* no preenchimento, desconsiderando a manutenção do formato dos objetos da cena – como destaca a elipse amarela na Figura 1.2(e). Outro ponto a ser observado trata da necessidade de recuperação da textura nas *disocclusions*, negligenciada, algumas vezes, como no caso de (SOLH; ALREGIB, 2012b), que apenas recria cor nestas regiões por meio de sucessivas médias de intensidade da vizinhança. Ainda, destaca-se a ausência de uma abordagem que utilize informação temporal não somente como fonte de conteúdo para o preenchimento de *disocclusions* em quadros futuros, como em (SCHMEING; JIANG,

2015; LUO; ZHU, 2017), mas também na identificação e supressão de artefatos (como *flickering*) e erros de preenchimento, que podem provocar desconforto no espectador. Por fim, observa-se que as abordagens propostas na literatura não apresentam em seus resultados experimentais uma avaliação em cenário real, com mapas de profundidade estimados com técnicas de casamento estéreo, por exemplo. Desta forma, mesmo com a comprovação de bons resultados utilizando mapas de profundidade *ground truth*, não existem garantias da efetividade dos métodos em um cenário de aplicação concreto.

Figura 1.2: Resultados obtidos pelo método de (LUO et al., 2016) nos quadros 6, 7 e 8 (apresentados em cada coluna, respectivamente) do *dataset* Ballet, com a projeção da vista 4 (real) para a 1 (virtual). A primeira linha exibe as imagens projetadas com os *holes* em preto, a seguinte o resultado obtido com a aplicação do método proposto por (LUO et al., 2016), e a última o *ground truth* (imagens reais da vista 1) de cada quadro. Ainda, nas imagens (e), (f) e (g), destaca-se por setas e uma elipse vermelha artefatos decorrentes de erros no preenchimento dos *holes* e, por uma elipse amarela em (f), um equívoco, onde parte do corpo da bailarina foi preenchido durante a projeção e aplicação do método de *inpainting* por informação de *background*.



Fonte: Imagens cedidas por Luo et al. (2016) e adaptadas pelo autor. Imagens originais retiradas do *dataset* Ballet de (ZITNICK et al., 2004).

1.2 Objetivos

1.2.1 Objetivo Geral

Este trabalho tem como objetivo principal propor uma abordagem para síntese de vistas com o modelo DIBR, com soluções para cada um dos seus problemas de pesquisa – que se resumem principalmente a remoção de artefatos e preenchimento dos *holes*. Com a abordagem, deseja-se produzir imagens sintéticas de boa qualidade, sem artefatos que possam prejudicar a experiência do usuário, e com coerência temporal para aplicação em vídeos.

A abordagem proposta deve empregar o padrão de entrada de dados definido para o modelo DIBR, contando apenas com uma imagem colorida e seu respetivo mapa de profundidades (seja *ground truth* ou real – gerado por uma técnica de casamento estéreo, por exemplo). O uso deste padrão se justifica pela fácil adaptação a diferentes cenários de transmissão de vídeo 2D existentes, acrescentando menos de 10-20% no consumo de banda em relação à imagem colorida (FEHN, 2004), permitindo que aplicações como FVV, 3DTV, transmissão de vídeos 3D em geral, entre outras, sejam viabilizadas.

Durante o processo de síntese de vistas, deseja-se que todos os artefatos sejam tratados adequadamente, de acordo com suas características. Ainda, almeja-se que as regiões sem informação de projeção na imagem sintética sejam reconstruídas com o uso de conteúdo confiável, com o auxílio de informação de segmentação e/ou semântica e/ou contexto, inibindo a inserção de artefatos visuais na vista sintética. Para vídeos, deseja-se utilizar informação temporal para o preenchimento parcial de *disocclusions* e OOFAs. No mesmo sentido, pretende-se incorporar mecanismos para evitar problemas relacionados à coesão temporal, que atenuem efeitos indesejáveis para o usuário. Ao final, espera-se que a abordagem proposta apresente resultados quantitativos e qualitativos competitivos, no comparativo com as técnicas que compõem o estado da arte para geração de vistas sintéticas com DIBR.

1.2.2 Objetivos Específicos

Abaixo, são apresentados os objetivos específicos relacionados com esta tese:

- propor métodos para remoção e preenchimento de *cracks* vazios e translúcidos.

- propor uma técnica que trate ou evite as ocorrências de *ghosts*, sem a remoção de conteúdo da imagem.
- propor uma abordagem para manutenção da coesão temporal e reconstrução parcial de *disocclusions* para vídeos, que utilize conteúdo de quadros predecessores.
- investigar mecanismos que permitam utilizar informação de segmentação/semântica/contexto no preenchimento de *disocclusions* e OOFAs, para a manutenção do formato dos objetos.
- propor um algoritmo de *inpainting* que reconstrua as regiões sem informação de projeção adequadamente, e que preveja a manutenção do formato dos objetos da cena, classificando e preenchendo cada *hole* com o conteúdo apropriado.

1.3 Contribuições da Tese

Nesta tese, propõe-se uma nova abordagem para síntese de vistas com o modelo DIBR, com etapas de pré-processamento que removem cada um dos tipos de artefatos e, por fim, preenche os diferentes tipos de *hole* de acordo com suas especificidades tanto para fotografias como vídeos. A abordagem provê também um mecanismo que permite empregar informação temporal na reconstrução parcial dos *holes* em vídeos. As contribuições mais importantes deste trabalho são:

I. uma abordagem para a detecção e preenchimento de *cracks* (vazios e translúcidos) de maneira rápida e eficiente.

II. uma abordagem para a identificação e remoção de *ghosts* sem a remoção de conteúdo da imagem, por meio de um pós-processamento do mapa de profundidades.

III. um algoritmo de preenchimento para os *holes* que se baseia em suas especificidades, e visa preservar a estrutura da cena.

IV. um método de construção de um modelo de *background* incremental para vídeos que permite que abordagens DIBR possam preencher parcialmente *holes* com informação de quadros predecessores.

V. uma avaliação e discussão sobre o impacto dos algoritmos que compõem o estado da arte em casamento estéreo no contexto de síntese de vistas com o modelo DIBR.

1.4 Organização dos Capítulos

O restante do texto desta tese está organizado como segue. No Capítulo 2, apresentam-se conceitos fundamentais para a compreensão deste trabalho, e a revisão do estado da arte para métodos de geração de vistas sintéticas com o modelo DIBR. O Capítulo 3 descreve uma abordagem completa para a síntese de vistas com o modelo DIBR, focada principalmente em fotografias, que apresenta métodos específicos para cada tipo de artefato e *hole*, desenvolvidos com base em suas características individuais. No Capítulo 4, propõe-se uma nova abordagem DIBR, para síntese de fotografias e vídeos, que remove *ghosts* antes mesmo da projeção da imagem de referência para o ponto de vista virtual, e guia-se pela informação estrutural definida por *superpixels* hierárquicos para preencher os *holes*. No Capítulo 5, apresenta-se um novo método para a construção de um modelo de *background* incremental, empregado no preenchimento parcial de *holes* por abordagens DIBR. O Capítulo 6 exibe os resultados experimentais dos métodos propostos, com avaliação qualitativa e quantitativa em diferentes *datasets*, em um comparativo com o estado da arte. Ainda neste capítulo, avalia-se a aplicação de métodos para a geração de vistas sintéticas com DIBR em um cenário real utilizando mapas de profundidade produzidos por técnicas de casamento estéreo. O Capítulo 7 discute as considerações finais desta tese, apresentando algumas conclusões parciais e indicativo de trabalhos futuros. Além disso, na última seção deste capítulo, são enumeradas as contribuições acadêmicas produzidas durante o desenvolvimento deste trabalho.

2 FUNDAMENTAÇÃO TEÓRICA E REVISÃO BIBLIOGRÁFICA

Neste capítulo, apresentam-se conceitos fundamentais para uma melhor compreensão do tema abordado nesta tese e, juntamente, o estado da arte para a geração de imagens sintéticas com o modelo DIBR. Na Seção 2.1, descreve-se detalhadamente o modelo DIBR especificando suas características e problemas relacionados com a sua viabilidade. Após, na Seção 2.2, são sumarizados os principais trabalhos relacionados com o tema abordado neste estudo.

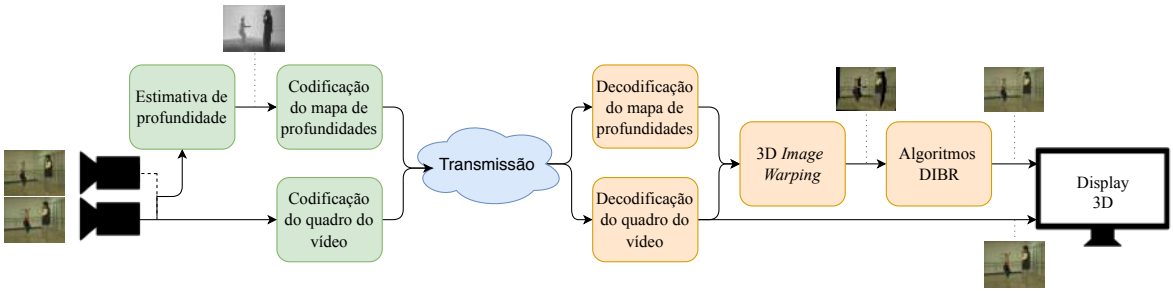
2.1 Modelo DIBR

Nos últimos anos, inúmeras pesquisas foram desenvolvidas, tanto por parte da academia quanto da indústria, com o objetivo de viabilizar o consumo de vídeos 3D sob demanda. Essas pesquisas estão ligadas a diferentes problemas inerentes a geração, transmissão e visualização deste tipo de conteúdo. Neste contexto, foi proposto por Fehn (2004) o modelo DIBR, como parte de um sistema para TV3D, adequado à infraestrutura de transmissão televisiva existente na Europa.

No *pipeline* proposto originalmente para o modelo DIBR em (FEHN, 2004), ilustrado na Figura 2.1, cada quadro de vídeo é associado a um mapa de profundidades, representado por uma imagem em escala de cinza. Ambas as imagens e metadados adicionais são codificados e enviados por meio da infraestrutura de transmissão existente (que comporta o acréscimo de dados), seguindo o formato europeu de TV digital. Então, no lado cliente da aplicação, um *set-top box* recebe e decodifica os dados referentes a cada quadro. Com a informação decodificada, utiliza-se 3D *image warping* para projetar a imagem 2D colorida para um ponto de vista virtual, com base no valor de profundidade associado a cada *pixel*. A definição do local apropriado para o ponto de vista virtual depende do tipo de *display* utilizado (e do sistema de projeção), para o qual será gerada a vista complementar que irá compor o par estéreo juntamente com o quadro do vídeo. Após esta etapa, diversos problemas associados especificamente à geração de imagens sintéticas com o modelo DIBR – detalhados na Subseção 2.1.2 – são solucionados, como a remoção de artefatos e o preenchimento de regiões sem informação de projeção. Por fim, as imagens sintéticas são exibidas em um *display* 3D, fornecendo ao usuário a percepção de profundidade nos vídeos.

Após o surgimento em 2004, o modelo DIBR passou a ser explorado por diversas

Figura 2.1: Diagrama de blocos de um sistema de TV3D com DIBR. Os blocos em verde destacam os processos necessários para a transmissão de conteúdo no formato adotado pelo modelo DIBR (uma imagem colorida – quadro do vídeo – e uma em escala de cinza correspondendo ao mapa de profundidades). Após a transmissão, nos blocos em amarelo, são destacados os processos que são executado em um *set-top box* no lado usuário, para a produção do conteúdo que será exibido no *display* 3D.



Fonte: O autor.

pesquisas, inclusive para outras aplicações além de TV3D. Mori et al. (2008) e Zinger, Do and With (2010), por exemplo, apresentam abordagens para viabilizar FVV, gerando imagens sintéticas complementares aos pontos de vista reais, que permitem a navegação na cena. Entretanto, DIBR não se apresenta como solução somente para FVV, mas também como uma forma promissora para sintetizar vistas virtuais em muitas outras aplicações populares e recentes de multimídia imersiva, como *Virtual Reality* (VR), *Augmented Reality* (AR), *light field multiview videos*, etc. (TIAN et al., 2019). Isso demonstra que, de um modo geral, o modelo se destina a quaisquer aplicações onde se faça necessária a geração de diferentes pontos de vista para a mesma cena, para os quais não se tenha imagens reais ou não exista capacidade de transmissão ou armazenamento das mesmas.

Como limitante à expansão do modelo, destaca-se a necessidade de um mapa de profundidades denso associado à imagem colorida. Contudo, trabalhos recentes como (WANG et al., 2018) e (GUO et al., 2018) demonstram que se pode inferir informação de profundidade com base em uma única imagem. Sendo assim, utilizando soluções como estas, quaisquer fotografias/vídeos existentes poderiam ser adaptados para aplicações baseadas no modelo DIBR.

2.1.1 Geração da Vista Sintética

Para gerar um novo ponto de vista com o modelo DIBR, faz-se necessário extrair informação de profundidade para cada *pixel* da imagem colorida. Esta informação

é essencial para que se possa fazer a projeção da imagem de referência para o ponto de vista virtual. Por este motivo, apresenta-se na próxima subseção uma discussão acerca do processo de extração de profundidade. Em seguida, na Subseção 2.1.1.2, descreve-se o processo de projeção com 3D *image warping*.

2.1.1.1 Extração de profundidade

A informação de profundidade para imagens 2D pode ser obtida de diversas formas, como: por câmeras RGB-D (como Kinect, Google Tango, etc.) (ZOLLHÖFER et al., 2018); por pares de câmera estéreo (BARNARD; FISCHLER, 1982) via técnicas de casamento estéreo (ZHANG et al., 2015; MOZEROV; WEIJER, 2015; TANIAI et al., 2018); e até mesmo por métodos de predição de profundidade com o uso de Redes Neurais Convolucionais (*Convolutional Neural Networks* – CNNs) (EIGEN; FERGUS, 2015; LIU et al., 2016; WANG et al., 2018). Técnicas de casamento estéreo, por exemplo, permitem extrair de maneira confiável um mapa denso de profundidades a partir de duas ou mais imagens. Essas técnicas identificam *pixels* correspondentes nas imagens e convertem suas posições no 2D em informação de profundidade no 3D (SZELISKI, 2010).

A Figura 2.2 ilustra um *setup* estéreo simples, com um par de câmeras que capturam simultaneamente o mesmo objeto no 3D, sob diferentes perspectivas. No exemplo, a mudança de perspectiva se refere somente ao deslocamento horizontal entre o centro das duas câmeras (*CAM1* e *CAM2*), denominado *baseline* (referenciado por b). As técnicas de casamento estéreo tem como objetivo identificar a correspondência de pontos, como no caso do ponto não ocluso P localizado no objeto 3D da cena, projetado no plano das duas imagens em diferentes posições.

Considerando que as imagens capturadas por *CAM1* e *CAM2* estão retificadas, a identificação de correspondências se resume a uma busca no 1D (no eixo x) (SHIBAHARA et al., 2007). A diferença da posição de projeção nas duas imagens perspectivas do mesmo ponto 3D, se denomina disparidade (MUDDALA, 2015), dada por

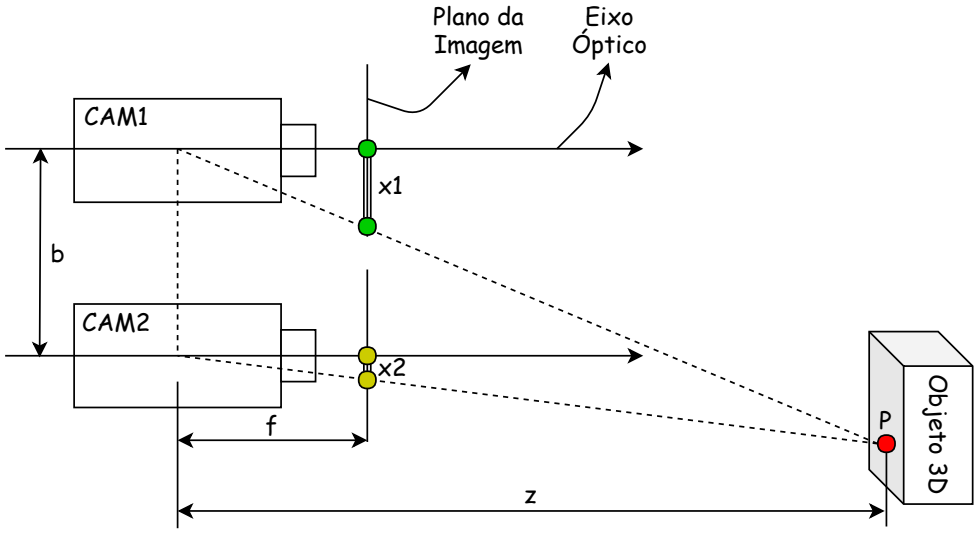
$$d = |x_1 - x_2|. \tag{2.1}$$

A disparidade possui uma relação inversa com a profundidade z no 3D, e pode ser obtida por meio da equação

$$z = \frac{f \cdot b}{d}, \tag{2.2}$$

onde f denota a distância focal, que corresponde à distância entre o centro da câmera e

Figura 2.2: Ilustração de *setup* estéreo utilizado para aquisição de profundidade a partir de técnicas de casamento estéreo.



Fonte: O autor, com base em (FICKEL, 2015; MUDDALA, 2015).

o plano da imagem, medida em *pixels*. Neste mesmo contexto, o *baseline* b se dá em coordenadas de mundo e a disparidade em *pixels*. Observa-se que os os parâmetros f e b são fixos para quaisquer pontos da imagem (por serem inerentes do *setup* estéreo), e apenas o valor de disparidade pode variar de acordo com cada *pixel* (com base no conteúdo representado). Com isso, fica claro como a identificação de pontos correspondentes em imagens perspectivas permite que seja obtido o valor real de profundidade associado aos *pixels* no 3D (a Figura 2.2 expressa esta relação).

A Figura 2.3 apresenta um exemplo de correspondência de pontos, onde são exibidas duas imagens – da esquerda e direita – retificadas, capturadas sob diferentes perspectivas. No exemplo, os círculos em vermelho identificados por x_1 na imagem da esquerda, correspondem a x_2 na imagem da direita, pois as linhas são coincidentes. Estes pontos estão em coordenadas x diferentes com relação ao plano das imagens, as quais podem ser utilizadas para determinar o valor de disparidade em cada uma das três correspondências dadas como exemplo. Sendo assim, o processo de estimativa do mapa de disparidades completo consiste na identificação da correspondência para cada coordenada (x, y) associada a cada *pixel* em ambas imagens (SZELISKI, 2010).

2.1.1.2 3D Image Warping

O modelo DIBR permite gerar novos pontos de vista virtuais para uma imagem colorida 2D a partir de parâmetros de câmera e da informação de profundidade associada

Figura 2.3: Ilustração da correspondência de pontos em duas imagens retificadas, capturadas sob diferentes perspectivas.

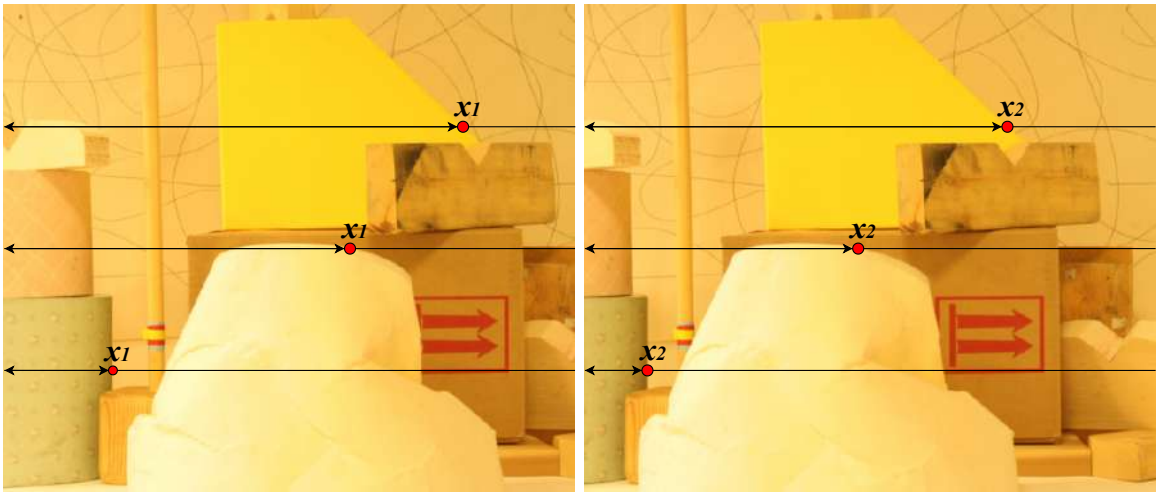


Imagem da Esquerda (CAM1) Imagem da Direita (CAM2)

Fonte: O autor, com base em (SCHARSTEIN; SZELISKI, 2002; FICKEL, 2015), utilizando imagens de (HIRSCHMULLER; SCHARSTEIN, 2007).

a cada *pixel*. Para tal, normalmente, utiliza-se a equação de 3D *image warping*, proposta por McMillan (1997), que se baseia no modelo de câmera *pinhole*. Neste caso, como definido em (FEHN, 2004), realiza-se o processo de geração de novas vistas sintéticas com a execução basicamente de duas etapas: na primeira, os pontos da imagem de referência são reprojetados em coordenadas 3D de mundo, com base na informação de profundidade; após, na segunda, esses pontos no espaço 3D são projetados no plano da imagem da câmera virtual.

Além desta abordagem genérica, outras opções também podem ser exploradas para a formação da vista sintética. Como discutido na Subseção 2.1.1.1, o valor de disparidade equivale ao deslocamento horizontal entre quaisquer dois *pixels* correspondentes em um par de imagens estéreo. Desta forma, com base no exemplo dado na Figura 2.3, pode-se concluir que a imagem capturada por $CAM1$ pode ser utilizada para gerar o ponto de vista da $CAM2$ (e vice-versa), com base somente nos valores de disparidade associados a cada *pixel*. Considerando que as coordenadas (x_1, y_1) do espaço de disparidade são coincidentes com as coordenadas de *pixel* da imagem de referência, a relação de correspondência entre as duas pode ser dada por:

$$x_2 = x_1 + s \cdot d(x_1, y_1), \tag{2.3}$$

$$y_2 = y_1, \tag{2.4}$$

onde $s = \pm 1$ determina o sentido de projeção para a vista virtual (SCHARSTEIN; SZELISKI, 2002). Se a imagem de referência for projetada para o lado direito, atribui-se -1, e para o sentido inverso 1. Com base nesta equação, pode-se perceber que o processo de projeção pode ser realizado sem o uso da informação de profundidade. Como detalhado, o valor de disparidade e o sentido de projeção podem ser suficientes, pois permitem estimar a posição de cada *pixel* da imagem de referência no ponto de vista virtual. De outro modo, quando se faz necessário utilizar a informação de profundidade para a projeção, se os valores de distância focal e do *baseline* estiverem disponíveis, pode-se converter os valores de disparidade para profundidade utilizando a Equação 2.2. A forma como a projeção para o ponto de vista virtual é definida, depende do *setup* utilizado para a captação das imagens e, dentro das limitações, das possibilidades que serão exploradas pela aplicação alvo.

A Figura 2.4(a) exibe o mapa de disparidades[1] *ground truth* associado à imagem obtida pela $CAM1$ da Figura 2.3, estimado com relação a $CAM2$. Utilizando os valores de disparidade associados a cada *pixel* da imagem de referência (imagem da esquerda), pode-se produzir a vista sintética correspondente à capturada pela $CAM2$, por meio da Equação 2.3. O resultado obtido por este processo pode ser visualizado na Figura 2.4(b). A imagem sintética produzida normalmente possui artefatos visuais e contém regiões sem informação de projeção (destacadas em azul na figura), devido a oclusões ou até mesmo por erros do processo de estimativa do mapa de disparidades. Estes problemas são inerentes do processo de geração de vistas sintéticas com o modelo DIBR e precisam ser tratados por abordagens específicas.

O processo de projeção da vista de referência pode ser realizado de duas maneiras distintas: com o *forward warping*, onde a vista de referência e o mapa de disparidades são projetados diretamente para o ponto de vista desejado; ou com *backward warping*, onde projeta-se primeiramente o mapa de disparidades para o ponto de vista virtual, preenchem-se as regiões sem informação, e então utiliza-se seu conteúdo para localizar os *pixels* da vista sintética de maneira inversa na imagem de referência (MUDDALA, 2015). Esta segunda maneira reduz a quantidade de conteúdo a ser estimado na vista sintética. Entretanto, não existem garantias de que a informação de disparidade estimada leve a localização do *pixel* adequado ao preenchimento na vista de referência.

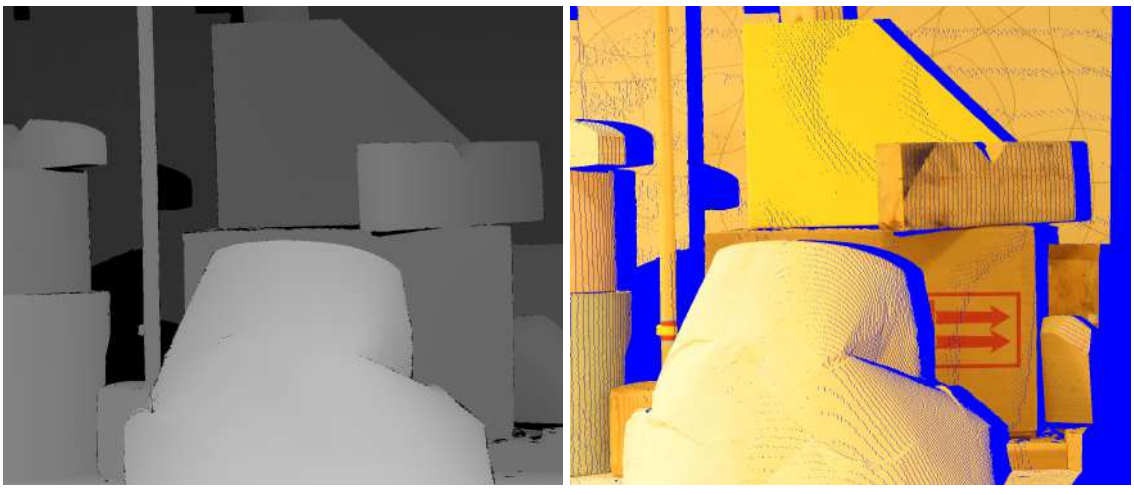
Com 3D *image warping*, mais de um *pixel* pode ser mapeado no mesmo ponto da

[1] Na sequência do texto, os problemas e soluções propostas estão descritos em termos de disparidade. Como detalhado anteriormente, os valores de disparidade podem ser convertidos para profundidade, quando for necessário.

imagem sintética. Para resolver este problema, deve-se selecionar o ponto mais próximo da câmera (de menor profundidade), pois este sobrepõe os objetos de *background* na cena (MORI et al., 2008).

Para a geração de vistas sintéticas, duas abordagens diferentes são encontradas na literatura, no que se refere ao modelo de entrada de dados e a formulação da imagem sintética após o 3D *image warping*. O modelo V+D, definido originalmente em (FEHN, 2004) para o modelo DIBR, consiste em uma vista de referência (V) e o mapa de disparidades correspondente (D). Com relação a D, dependendo do *dataset* utilizado, as soluções DIBR utilizam informação de profundidade ou de disparidade. Comumente, são fornecidos os valores de disparidade e os parâmetros definidos no *setup* estéreo. Por isso, adotou-se como padrão no texto disparidade, ao invés de profundidade. Para formar a imagem sintética, o modelo DIBR utiliza extrapolação de vistas (MUDDALA, 2015). Nesta situação, quando se projeta a imagem de referência no ponto de vista virtual, são excedidos os limites físicos estabelecidos no momento da captação, e grandes regiões sem informação são formadas em sua lateral. O outro modelo, 2V+2D ou nV+nD, é composto por dois ou n pontos de vista de referência, cada um com seu mapa de disparidade. Para formar a imagem sintética, utiliza-se o processo de interpolação de vistas, onde as 2 ou n vistas são projetadas com base nos respectivos mapas D e então combinadas (MUD-

Figura 2.4: Exemplo de projeção com 3D *image Warping*. Em (a), apresenta-se o mapa de disparidades *ground truth* para a imagem à esquerda, exibida na Figura 2.3, estimado com relação a da direita (exibida na mesma figura). Em (b), exibe-se o resultado da projeção da imagem da $CAM1$, para o ponto de vista da $CAM2$, com os buracos sem informação destacados em azul.



(a) (b)

Fonte: O Autor, com imagens adaptadas de (HIRSCHMULLER; SCHARSTEIN, 2007).

DALA, 2015). Neste processo, as imagens de referência são projetadas para o ponto de vista virtual e combinadas, assim como ocorre com os mapas D, restando pequenas regiões sem informação.

2.1.2 Artefatos e *Holes*

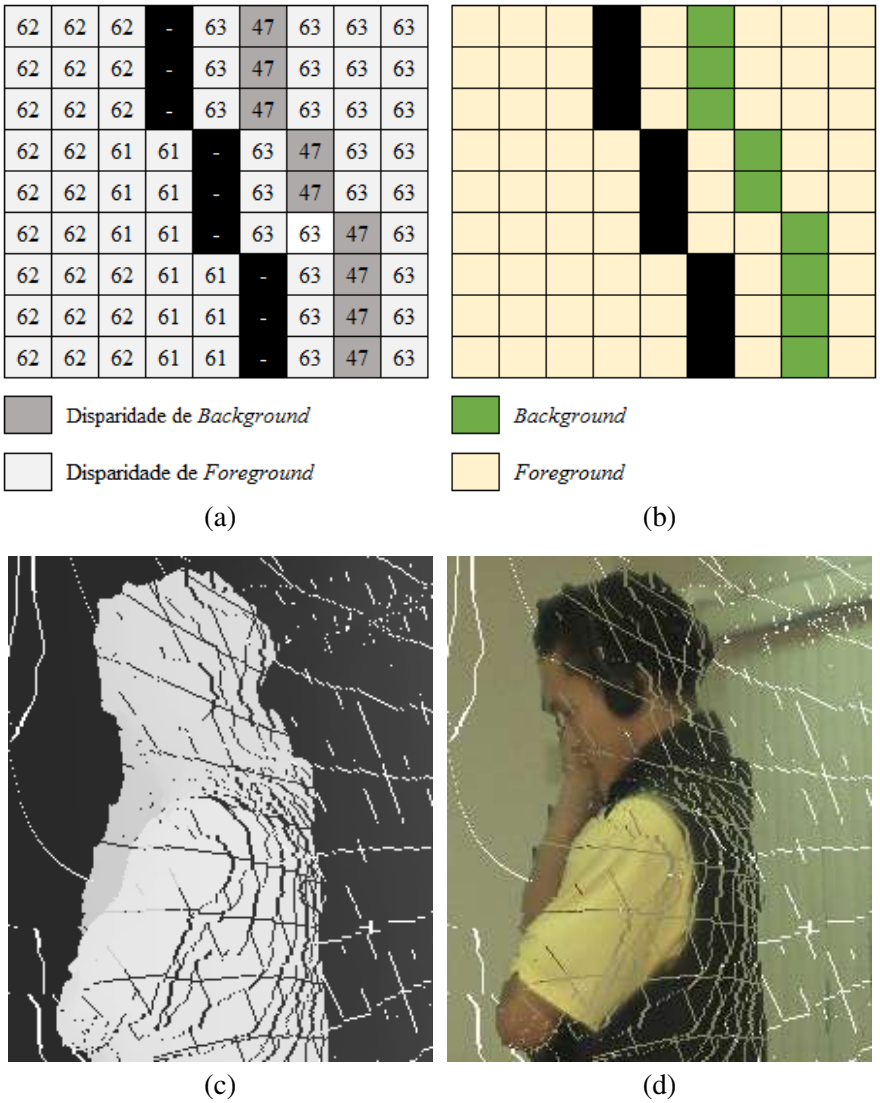
Após a projeção, diferentes tipos de artefatos e inúmeras regiões sem informação de projeção – denominadas genericamente de *holes* – se formam na imagem sintética. Artefatos são gerados por diferentes razões como, por exemplo, erros no mapa de disparidades, como é o caso dos *cracks* e *ghosts*. Já os *holes* são resultantes do processo de geração de vistas sintéticas empregado pelo modelo DIBR. Estes podem corresponder a regiões oclusas que passam a ser expostas no ponto de vista virtual, classificadas como *disocclusions*, ou ser áreas que não fazem parte da imagem de referência, por extrapolar os limites físicos utilizados para sua captação, denominadas OOFAs. Nas próximas subseções, encontram-se detalhados os diferentes tipos de artefatos e *holes*, especificando suas causas, composição tanto na imagem como no mapa disparidades, e indicação de solução ideal.

2.1.2.1 Cracks Vazios e Translúcidos

Cracks são regiões vazias com um ou dois *pixels* de largura na imagem projetada (MUDDALA; SJÖSTRÖM; OLSSON, 2016), que possuem o formato de uma rachadura longa e fina. Além de serem causados por erros nos mapas D, sua justificativa mais comum tem origem no arredondamento de valores em ponto flutuante – obtidos com a Equação 2.3, por exemplo – para inteiros, utilizado para definir as coordenadas de projeção de cada *pixel* no domínio da imagem do ponto de vista virtual (MORI et al., 2008; AHN; KIM, 2013). Por este motivo, os *cracks* são encontrados em regiões com disparidade aproximadamente homogênea, que correspondem normalmente ao interior dos objetos. Nas Figuras 2.5(a) e (b), apresenta-se uma ilustração da ocorrência deste artefato na imagem e no mapa de disparidades, destacada em preto. Por não possuir disparidade nestas regiões, a figura ilustra seu valor como "-". Por definição, quando o artefato não possui conteúdo, recebe o nome de *crack* vazio.

Em outros casos, o objeto no *background* da imagem é projetado na mesma posição do artefato. Assim, sua textura e valores de disparidade preenchem esta região

Figura 2.5: Ocorrência de *cracks* nos mapas de disparidades (no lado esquerdo) e nas imagens coloridas (no lado direito), com ilustrações na linha de cima e exemplos reais na de baixo. No mapa de disparidades (a), a forma vazia do artefato é identificada pela cor preta. Já a translúcida, é destacada pela cor verde em (b) e pelo valor "47" em (a). Nos exemplos reais, podem ser notados *cracks* na forma vazia em braco, tanto em (c) como em (d). A forma translúcida pode ser identificada em (c) pela tonalidade mais escura nas disparidades que representam o corpo do rapaz, que corresponde a cor da parede em (d).



Fonte: O Autor, com imagens adaptadas do *dataset Ballet* de (ZITNICK et al., 2004).

incorretamente, gerando o *crack* na forma translúcida (MUDDALA; SJÖSTRÖM; OLSSON, 2016). As Figuras 2.5(a) e (b) apresentam um exemplo da ocorrência deste artefato, identificado na imagem como uma linha vertical com forma irregular de cor verde, e no mapa de disparidades pelo valor "47", inferior ao da vizinhança. Após a detecção, o conteúdo dos *cracks* translúcidos precisa ser removido, e então ambas as formas devem ser preenchidas.

Exemplos reais da ocorrência de ambos os tipos de *crack* podem ser visualizados na Figura 2.5(d). Nesta imagem, a forma translúcida do artefato pode ser identificada no corpo do ator, onde parte da parede ao fundo foi projetada. Na Figura 2.5(c), exibe-se o mapa de disparidades correspondente a imagem (d), na qual as mesmas regiões apresentam menor intensidade, associada a parede. Em ambas figuras, diversas ocorrências de *cracks* vazios podem ser visualizadas em branco, espalhadas pela imagem.

2.1.2.2 Ghosts

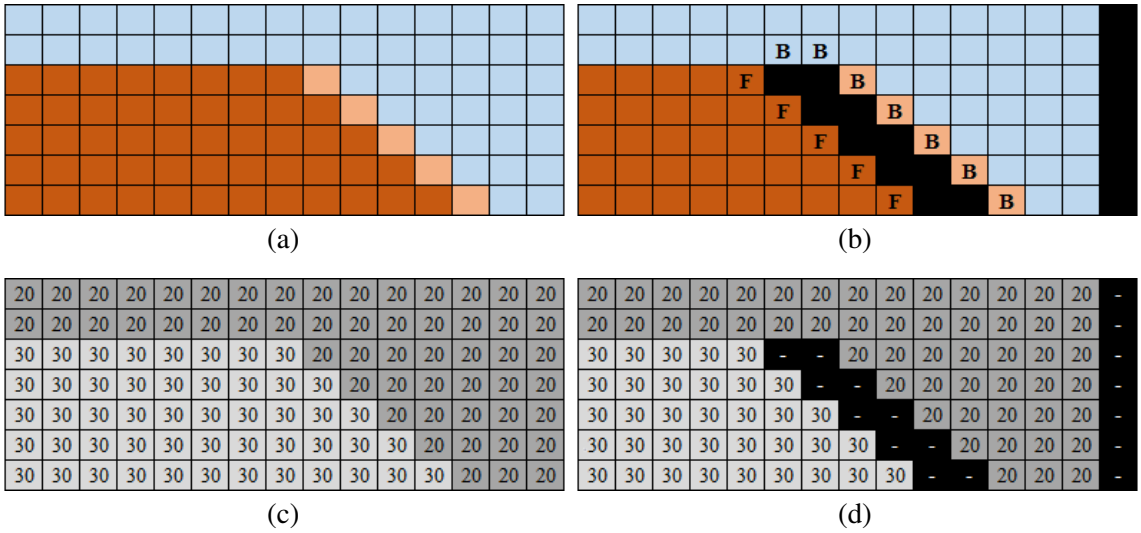
Os *Ghosts* aparecem na borda da região de *background* das *disocclusions*, como uma representação da silhueta dos objetos de *foreground*. Em sua composição, o artefato apresenta uma mistura de cor correspondente à região de transição de objetos vizinhos na imagem de referência (MUDDALA; SJÖSTRÖM; OLSSON, 2016). Isto ocorre, pois, geralmente, a borda entre dois objetos não é dada por uma mudança brusca de valores de cor, mas sim um gradiente, que pode ser mais similar ao elemento no *foreground* ou *background*, e que só pode receber o valor de disparidade de um dos dois objetos (SCHMEING; JIANG, 2011). Desta forma, no momento da definição do valor de disparidade da borda do objeto de *foreground*, atribui-se erradamente o mesmo valor associado ao *background*. Como consequência, os objetos são separados com o processo de projeção, devido à diferença dos valores de addisparidade, e o artefato se forma na borda de *background* das *disocclusions*.

A Figura 2.6 exemplifica a ocorrência de um *ghost*, na qual o valor de disparidade associado a transição dos dois objetos na imagem (em laranja claro) possui o mesmo valor do *background* (B) “20”. Neste caso, o valor deveria ser o mesmo do *foreground* (F) “30”. Como resultado, a região de transição é projetada para o lado errado da *disocclusion*, produzindo o artefato.

Estes artefatos impactam no processo de preenchimento das *disocclusions*, uma vez que os algoritmos de *inpainting* são tipicamente guiados pela borda do *background* (OLIVEIRA et al., 2015; MUDDALA; SJÖSTRÖM; OLSSON, 2016). Desta forma, se o artefato não for tratado, informação de *foreground* pode ser incorretamente propagada nas *disocclusions*.

Como evidenciado, o artefato se forma por um erro na estimativa do mapa de disparidades, onde se deveria atribuir a regiões de transição de objetos sempre o valor associado ao *foreground* e não ao *background*. Intuitivamente, esse artefato poderia ser evitado ainda no processo de estimativa dos valores de disparidade. Contudo, inúmeras aborda-

Figura 2.6: Ilustração da geração de um *ghost*. Na primeira linha são exibidas as imagens coloridas e abaixo os mapas de disparidade correspondentes. Em (a), apresenta-se a imagem e o mapa antes da projeção. Ao lado, em (b), exibe-se o resultado do processo de projeção, onde pode ser identificado um *ghost* em laranja claro, e uma *disocclusion* em preto, cuja borda está segmentada em região de *background* (B) e *foreground* (F).



Fonte: O Autor.

gens com diferentes tecnologias para a estimativa de disparidade já foram apresentados na literatura, sem mecanismos específicos para evitar o artefato, as quais precisariam ser reavaliadas individualmente, para tornar isso possível. Desta forma, uma solução ideal para este problema poderia ser produzida por um algoritmo de pós-processamento, que se adaptasse a qualquer abordagem existente.

2.1.2.3 Disocclusions e OOFAs

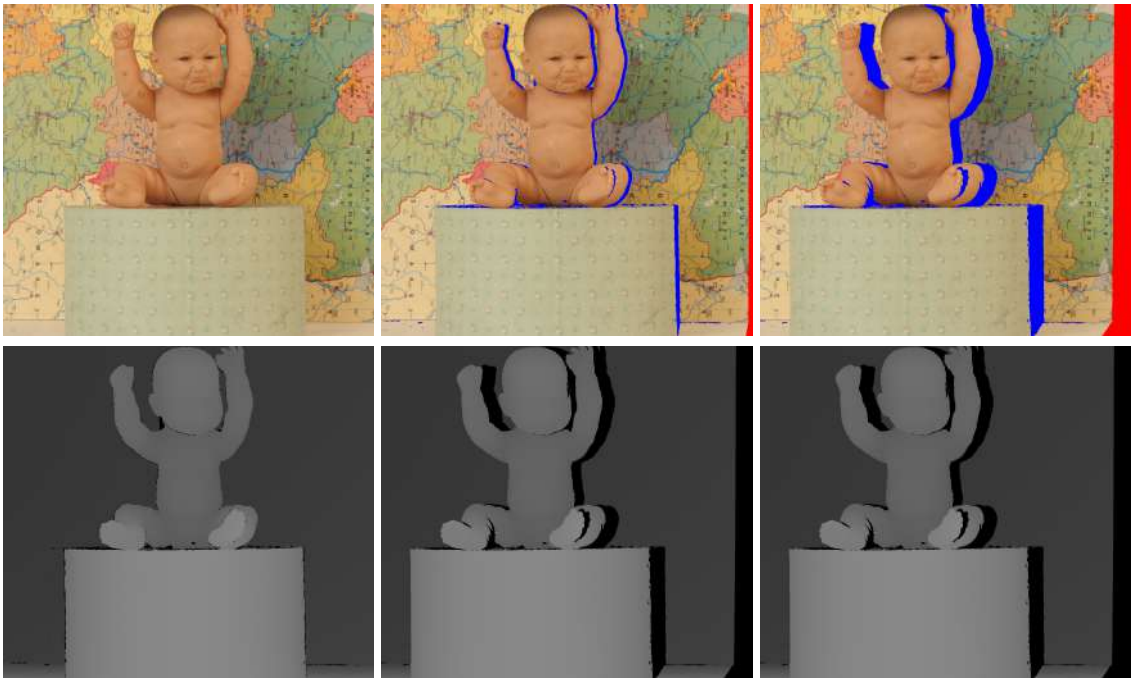
Os *holes* se formam em razão do processo de projeção e correspondem a regiões que não podem ser visualizadas no ponto de vista utilizado para captação da imagem de referência, e que são expostas no ponto de vista virtual. Definir uma abordagem para tratar estas novas regiões expostas se destaca como o problema mais significativo dentre os desafios impostos pelo modelo DIBR (ZHANG; VAZQUEZ; KNORR, 2011).

As *disocclusions* ocorrem em descontinuidades de disparidade, próximas às bordas dos objetos (MUDDALA, 2015). Estes *holes* são produzidos na vista sintética quando parte do *background* ocluso por objetos de *foreground*, e consequentemente não capturado, é exposto no ponto de vista virtual (LUO; ZHU, 2017). Por este motivo, estas regiões devem ser reconstruídas com informação somente de *background*. O tamanho e formato destes *holes* podem variar, aumentando proporcionalmente à distância definida

entre o ponto de vista de referência e o virtual. Na Figura 2.7, são apresentados exemplos de projeção para o *dataset* “Baby1” de (HIRSCHMULLER; SCHARSTEIN, 2007), com diversas ocorrências de *disocclusions* destacas em azul. Como pode ser visto no exemplo, quando se projeta a vista 1 (referência) para um ponto de vista mais próximo como o da vista 2 (com $b = 40mm$), as regiões sem informação são bem menores que as apresentadas na imagem ao lado, para a vista 5 que está mais distante. Para que se tenha uma ideia, esta imagem foi gerada utilizando um *baseline* quatro vezes maior ($b = 160mm$) que o definido para a vista 2. Estas regiões são normalmente grandes tanto em largura como altura e preenchê-las adequadamente é um desafio (OLIVEIRA; WALTER; JUNG, 2018).

As *OOFAs* são regiões sem informação localizadas nas extremidades da vista sintética. Este tipo de *hole* ocorre porque o ponto definido para a vista virtual extrapola os limites físicos de captura do utilizado para obter a imagem de referência, o que faz com que não exista informação de projeção para as bordas da imagem sintética (MUDDALA; SJÖSTRÖM; OLSSON, 2016). Assim como no caso das *disocclusions*, este tipo de *hole*

Figura 2.7: Exemplo da ocorrência de *disocclusions* e OOFAs após o processo de projeção. Na primeira linha são apresentadas as imagens coloridas, e abaixo os respectivos mapas de disparidade. Em (a) apresenta-se a imagem de referência (vista 1), e ao lado nas colunas (b) e (c) o resultado da projeção para as vistas 2 e 5, respectivamente. Em azul, destacam-se as *disocclusions* e em vermelho as OOFAs.



(a) (b) (c)

Fonte: O Autor, com imagens adaptadas de (HIRSCHMULLER; SCHARSTEIN, 2007).

geralmente é grande e varia de tamanho de acordo com o *baseline* estabelecido para a projeção da imagem de referência.

As Figuras 2.7(b) e (c) destacam no canto direito as OOFAs em vermelho. A identificação dessas regiões pode ser realizada com uma busca a partir da extremidade inversa ao sentido de projeção da imagem, seguindo até que conteúdo válido seja localizado. Mais especificamente, se a imagem de referência for projetada para a esquerda, como no exemplo da Figura 2.7(c), as OOFAs serão formadas no lado direito (e vice-versa), que será definido como ponto de partida para a busca. O mesmo vale para projeções verticais, pois se a imagem for projetada para baixo, os *holes* deste tipo serão identificados no topo da vista sintética.

Por ocorrerem nas bordas da imagem, não existem vestígios concretos sobre qual informação deve ser empregada na reconstrução das OOFAs, diferentemente das *disocclusions*, que precisam ser preenchidas com conteúdo de *background*. Este tipo de *hole* possui informação de vizinhança apenas em uma das extremidades, com objetos de diferentes camadas de disparidade na borda, que se apresentam como a única fonte de informação para guiar o processo de reconstrução. Na prática, não se tem o mínimo de informação sobre quais elementos poderiam estar fora dos limites de captação da câmera de referência, e novos objetos podem aparecer na cena, o que aumenta mais este desafio.

2.2 Trabalhos Relacionados

Nesta seção, são abordados trabalhos relacionados diretamente ao tema desta tese. Inicialmente, na Subseção 2.2.1, descrevem-se algumas das principais abordagens relativas a síntese de vistas com o modelo DIBR. Na subseção seguinte, aborda-se a estimativa e uso de informação de *background* por parte de abordagens DIBR. Este tema está diretamente relacionado ao método proposto para a geração de modelos de *background*, apresentado no Capítulo 5. Por fim, na Subseção 2.2.3, discutem-se estudos relacionados com a avaliação da geração de imagens sintéticas com o uso de mapas de disparidade reais, vinculados com a análise apresentada mais a frente no Capítulo 6.

2.2.1 Abordagens para Síntese de Vistas com o Modelo DIBR

Após a definição do modelo DIBR, inúmeras abordagens foram propostas na literatura, tanto por parte da comunidade acadêmica quanto da indústria, concentrando esforços na solução dos diferentes problemas relativos a sua viabilidade. Nestas abordagens, são apresentados métodos para o tratamento de artefatos e preenchimento de *holes* utilizando como entrada o modelo V+D ou 2V+2D, com foco em diferentes aplicações. A seguir, encontram-se sumarizados – em ordem cronológica – alguns dos principais trabalhos relacionados a geração de vistas sintéticas com o modelo DIBR divididos em duas subseções. Na primeira, descrevem-se trabalhos que não empregam informação temporal e, na segunda, os que utilizam este recurso no processo de síntese de vistas.

2.2.1.1 Abordagens que não empregam informação temporal

Mori et al. (2008) propõem um método de geração de vistas sintéticas para FVV utilizando o modelo 2V+2D. Inicialmente, os dois mapas de profundidade mais próximos da vista sintética são projetados para o ponto alvo, nos quais aplica-se o filtro da mediana e uma filtragem bilateral, para preencher *cracks* e remover irregularidades, respectivamente. Após, as *disocclusions* são dilatadas para evitar possíveis *ghosts*, e os mapas são utilizados para projetar as respectivas imagens coloridas. Por fim, as duas vistas são combinadas, e os *holes* preenchidos pelo algoritmo de *inpainting* de (TELEA, 2004). Segundo Tian et al. (2019), este método foi adotado pelo *MPEG 3D video Group* e passou a ser conhecido como *View Synthesis Reference Software* (VSRS), que foi especificado em (TANIMOTO et al., 2008). Nesta versão, além da proposta original com o modelo 2V+2D (VSRS2), passou-se a contar com uma variante que emprega V+D como entrada (VSRS1) utilizando a extrapolação de vistas como processo base.

Oh, Yea and Ho (2009), apresentam uma abordagem semelhante à desenvolvida por (MORI et al., 2008), a qual incorpora informação de profundidade no algoritmo de *inpainting*. Nesta proposta, substitui-se temporariamente a informação de *foreground* vizinha a cada *hole* pela de *background*, para induzir o algoritmo de (TELEA, 2004) a propagar somente o conteúdo da borda que contém os pontos com maior profundidade. Com esta estratégia, soluciona-se o problema da inserção de cores de objetos de *foreground* nas *disocclusions*, mas não da reconstrução de textura nessas regiões, devido a natureza do algoritmo de *inpainting* adotado.

Utilizando o mesmo modelo de entrada empregado pelas abordagens anteriores,

Zinger, Do and With (2010) propõem um *pipeline* completo para a geração de vistas sintéticas para FVV. Inicialmente, projetam-se as imagens e mapas mais próximos para o ponto de vista virtual, ignorando *pixels* em regiões com alta descontinuidade de profundidade, maiores que um limiar pré-determinado, para evitar *ghosts*. Em seguida, aplica-se o filtro da mediana nos mapas de profundidade para preencher os *cracks*. Com base nos valores de profundidade estimados, preenchem-se os *cracks* na imagem colorida, seguindo o conceito definido como *inverse warping* (que corresponde ao *backward warping*). Por fim, as imagens resultantes são combinadas, e as *disocclusions* remanescentes são preenchidas com um algoritmo de interpolação ponderada, adequado para estimativa de pequenos *holes*. Todavia, assimetrias fotométricas (de cor, por exemplo) não são verificadas nesta e em nenhuma das abordagens descritas anteriormente, as quais devem ser consideradas, uma vez que as imagens são captadas por diferentes sensores.

Daribo and Saito (2011) propõem utilizar o modelo V+D com informação residual e *inpainting*, visando gerar imagens para TV3D. Inicialmente, projeta-se a imagem de referência para o ponto de vista virtual. Após, boa parte dos *holes* são preenchidos com informação "residual", oriunda de outra vista real vizinha a virtual. Em seguida, as regiões sem informação restantes são recuperadas por uma adaptação do algoritmo de *inpainting* de (CRIMINISI; PEREZ; TOYAMA, 2004). Com o objetivo de selecionar primeiramente *patches* com menor variação de profundidade, foi incluído um termo de regularidade no cálculo de prioridades, de modo a preencher primeiramente regiões homogêneas da imagem. Ainda, na estimativa de similaridade, introduziu-se a distância SSD (*Sum of Squared Differences*) de profundidade de forma ponderada. No entanto, observa-se que não existe uma correlação entre a diferença de profundidade e a semelhança de cor/textura, o que pode comprometer a escolha pelo melhor candidato.

Solh and AlRegib (2012b) apresentam uma abordagem diferente para o preenchimento dos *holes*, denominada *Hierarchical Hole-Filling* (HHF), avaliada com o modelo de entrada V+D para FVV e vídeos 3D. Nessa abordagem, produzem-se estimativas de baixa resolução em uma representação de pirâmide da imagem projetada, computadas pela média em blocos de 5×5 de conteúdo válido, que são propagadas para escalas maiores abaixo. Dentro de algumas escalas em multirresolução, obtêm-se uma estimativa de alta resolução da vista sintética sem buracos, utilizada para preencher os *holes*. Os autores apresentam também uma variação deste algoritmo em (SOLH; ALREGIB, 2012a), denominada *depth adaptive* HHF, que prioriza *pixels* que possuem maior valor de profundidade. Como pode ser visto nos resultados experimentais apresentados em

ambos trabalhos, este algoritmo reconstrói pequenas regiões adequadamente, entretanto, para *holes* maiores, ambas as abordagens produzem um resultado visual semelhante a um borramento.

Ahn and Kim (2013) propõem um *pipeline* completo para FVV também com o modelo V+D. Neste trabalho, regiões que podem conter *ghosts* são identificadas no mapa de profundidades original por meio de uma dilatação morfológica, seguida da aplicação de um limiar. Após a projeção, essas regiões são removidas da vista sintética quando localizadas no *background*, para suprimir possíveis ocorrências do artefato. Em seguida, aplica-se o filtro da mediana para remover os *cracks* tanto na imagem como no mapa de profundidades. Após, uma interpolação ponderada preenche regiões com até 100 *pixels* de área. Para os demais *holes*, utiliza-se uma adaptação do algoritmo de *inpainting* de (CRIMINISI; PEREZ; TOYAMA, 2004). Na adaptação, calcula-se o termo de dados utilizando uma matriz Hessiana, e na inicialização do termo de confiança, indica-se com 1 apenas regiões classificadas como *background*. Além disso, para a busca pelo melhor *patch*, considera-se uma região delimitada com 160×80 *pixels*, centrada no *patch* a ser preenchido, realizada apenas na região segmentada como *background*, para evitar que conteúdo do *foreground* seja copiado. Para determinar a correspondência entre os *patches*, além da distância SSD no espaço de cores, considera-se no cálculo a distância entre as linhas e a homogeneidade de profundidade. Assim como no caso de (DARIBO; SAITO, 2011), esta abordagem mistura o conceito de similaridade por cor/textura com a homogeneidade de profundidades, mesmo não existindo correlação.

Na abordagem apresentada em (OLIVEIRA et al., 2015), são propostos métodos para o tratamento de artefatos e preenchimento de *disocclusions*, destinadas ao modelo de entrada V+D. Para detectar os *cracks* vazios, aplica-se uma operação de abertura morfológica em um mapa binário com as regiões sem informação de projeção identificadas, seguida do cálculo de complemento absoluto, para detectar as ocorrências do artefato pelo formato. Para preencher as regiões identificadas, aplica-se o algoritmo *fast inpainting* (OLIVEIRA et al., 2001). Em seguida, pontos isolados – considerados *outliers* – são removidos do interior das *disocclusions* por meio de uma operação de abertura morfológica. Após, regiões que podem conter *ghosts* são detectadas de acordo com o sentido de projeção da vista virtual, e os *pixels* candidatos são verificados individualmente em relação a sua similaridade com ambas as extremidades (*foreground* e *background*) da respectiva *disocclusion*. Se um candidato for mais similar a sua vizinhança no *foreground*, move-se o *pixel* para esta extremidade do *hole*. Por fim, as *disocclusions* são preenchi-

das com uma adaptação do algoritmo de (CRIMINISI; PEREZ; TOYAMA, 2004), que substitui o termo de dados por um novo de profundidades (que prioriza candidatos no *background*), e os *patches* são comparados no espaço de cores RGB (*red, green, blue*). Esta abordagem não considera as OOFAs em sua avaliação.

Luo and Zhu (2017) propõem preencher *holes* com o uso de um modelo de *background* de cenário. Inicialmente, faz-se um pré-processamento no mapa de profundidades suavizado com o intuito de colocar as bordas dos objetos sobre os elementos de *foreground*. Após, estas bordas são identificadas por meio do detector de Canny (CANNY, 1986). Então, partindo do mapa de bordas, aplica-se um algoritmo que iterativamente remove conteúdo das regiões vizinhas, que contenham profundidade similar aos pontos identificados, até que os objetos do *foreground* sejam inteiramente removidos da imagem. Nesta abordagem, produz-se um prolongamento da imagem denominado *background extension* (BGE), que tem como objetivo extrapolar os limites físicos do modelo de *background*, para produzir conteúdo para as OOFAs. Para reconstruir os *holes* (incluindo o BGE), utiliza-se uma adaptação do algoritmo de Criminisi (CRIMINISI; PEREZ; TOYAMA, 2004), que busca por candidatos somente em regiões com profundidade similar a do *patch* a ser preenchido. Em seguida, para formular a vista virtual, realiza-se a projeção tanto do modelo quanto da imagem de referência. Por fim, são preenchidas as regiões sem informação de projeção com o modelo de *background* estendido, e pequenos *holes* gerados por descontinuidades de profundidade com o algoritmo de *inpainting* adaptado. Além de depender de um processo de segmentação preciso, métodos que reconstroem modelos de *background* completos, removendo elementos de *foreground*, demandam por mais processamento que o necessário, por reconstruir regiões que não são expostas em momento algum dos vídeos.

Muddala, Sjöström and Olsson (2016) apresentam um *pipeline* completo para a geração de vistas sintéticas com DIBR, que baseia-se no uso de uma imagem auxiliar denominada *layered depth image* (LDI). Inicialmente, formula-se a LDI removendo da imagem de referência *pixels* que sobrepõem regiões do *background* na vista projetada, com base no processo de 3D *warping* e na análise de descontinuidades de profundidade. Para evitar *ghosts*, são removidos dois *pixels* ao longo da borda de *background* dos *holes*. Então, utiliza-se uma adaptação do algoritmo de (CRIMINISI; PEREZ; TOYAMA, 2004) para reconstruir a LDI. Na adaptação, os termos de prioridade foram alterados, para que a confiança seja contabilizada apenas em *pixels* classificados como *background*, e o termo de dados foi substituído pelo proposto em (MUDDALA; OLSSON; SJÖSTRÖM,

2013), com o objetivo de detectar detalhes da estrutura da imagem apenas no contexto de vizinhança. Adicionalmente, realiza-se a busca pelo melhor *patch* para o preenchimento apenas em regiões classificadas como *background*, adicionando a diferença de profundidade ponderada no cálculo de similaridade. Após, preenchem-se as *disocclusions* na vista sintética com conteúdo da LDI projetada. Por fim, estima-se o conteúdo dos *cracks* com uma propagação de *background* simples, e os *holes* restantes são preenchidos com o algoritmo de *inpainting* adaptado, sem considerar informação de profundidade. Como destacado anteriormente, a inclusão de profundidade no cálculo de estimativa de similaridade pode levar a escolha de *patches* incoerentes em relação a cor/textura.

Cho et al. (2017) propuseram uma variação do algoritmo de (CRIMINISI; PEREZ; TOYAMA, 2004) para o preenchimento de *disocclusions*. Nesta abordagem, adiciona-se ao cálculo de prioridades o termo de regularidade de (DARIBO; SAITO, 2011), e um novo termo, que considera concomitantemente a maior média de profundidade e a probabilidade do ponto pertencer ao *background*. Já para a busca pelo melhor *patch* para o preenchimento, soma-se a distância SSD no espaço de cores com um termo que mede a regularidade de profundidade do *patch* candidato, visando selecionar regiões homogêneas. Quando o *patch* a ser preenchido está em uma borda, realiza-se adicionalmente um comparativo ponderado dos pontos de *foreground* e *background*, com maior peso para as diferenças deste segundo grupo. Contudo, salienta-se que a regularidade de profundidade não garante a similaridade entre *patches*, o que pode vir a prejudicar o comparativo.

Em uma abordagem diferente, Dai and Nguyen (2017) propõem preencher os *holes* com um algoritmo baseado em *hierarchical clustering*, empregando o modelo 2V+2D. Inicialmente, para identificar regiões que podem conter *ghosts*, aplica-se o detector de bordas de Canny no mapa de profundidades original. Então, as bordas detectadas são dilatadas, e os *pixels* classificados como pertencentes ao *background* são substituídos pelo filtro da mediana na imagem projetada. Após, as imagens e os mapas de profundidade são combinados, em um processo que pondera as intensidades de cada *pixel* de acordo com a proximidade entre as câmeras reais e a virtual. Para cada *hole*, define-se a quantidade de planos de profundidade no seu entorno, em uma abordagem gulosa de agrupamento hierárquico. Com base nesta informação, preenche-se o mapa de profundidades, partindo da borda do plano mais ao fundo do *hole*, atribuindo gradualmente o valor da mediana. Em seguida, preenche-se a imagem colorida com uma filtragem bilateral, guiada pelo mapa de profundidades. Por fim, as bordas são suavizadas com o filtro da mediana para evitar o efeito serrilhado nesta região. Nesta abordagem, o algoritmo de preenchimento considera

que o plano de profundidade mais ao fundo deve seguir como guia para a reconstrução do *hole* inteiro. No entanto, quando existem mais de dois objetos na vizinhança desta área, esta premissa pode ser facilmente refutada, pois a região a ser reconstruída pode não pertencer somente a um destes objetos.

2.2.1.2 Abordagens que empregam informação temporal

Yao et al. (2014) propõem em sua abordagem preencher *disocclusions* em vídeos sintéticos com informação de uma imagem de *background* estável temporalmente, seguindo o modelo V+D. Primeiramente, constrói-se um modelo de *background* com um processo *offline*. Neste processo, gera-se um modelo estável utilizando um *Gaussian Mixture Model* (GMM) a nível de *pixel*. Após, refina-se este modelo com um processo denominado *Foreground Depth Correlation* (FDC), que tem como objetivo remover *pixels* de *foreground* estáveis. No FDC, divide-se cada imagem em duas partes, com base no valor de profundidade, utilizando *k-means*. Este processo é repetido quadro a quadro, com o objetivo de ir adicionando conteúdo classificado como *background* com FDC. Por fim, as *disocclusions* são identificadas com base em um limiar de profundidade e preenchidas com o conteúdo do modelo de *background*. Os demais *holes* são preenchidos com o algoritmo de *inpainting* de Criminisi, Perez and Toyama (2004). Apesar de estimar adequadamente o conteúdo de *background* para os vídeos, esta abordagem requer um processo *offline* para computar o modelo de *background*, que se apresenta inviável para aplicações como TV3D.

Schmeing and Jiang (2015) propõem um método de *inpainting* baseado em *superpixels* para o preenchimento de *disocclusions*, com aplicação em vídeos, empregando o modelo de entrada V+D. Inicialmente, são computados os *superpixels* com SLIC (ACHANTA et al., 2010) na imagem projetada. Após, aplica-se uma erosão em cada segmento, com o objetivo de evitar que *ghosts* sejam propagados no processo de reconstrução. Então, para determinar a ordem de preenchimento de cada *disocclusion*, seleciona-se na borda de *background* o ponto que possui maior média de profundidade e mais informação válida em sua vizinhança. Considerando o entorno do ponto selecionado, analisam-se *superpixels* candidatos tanto no domínio espacial (em região delimitada) quanto temporal (de quadros passados) do vídeo, para determinar o melhor candidato ao preenchimento, utilizando como critério a similaridade por cor e profundidade. Então, copia-se o conteúdo do *superpixel* que minimiza a função de custo para o interior da *disocclusion*. Após, os termos de prioridade são atualizados com a informação do segmento, e repete-se o pro-

cesso até preencher completamente a imagem. Em sua essência, *superpixels* apresentam cor aproximadamente homogênea e geralmente não extrapolam bordas. Portanto, nesta abordagem, quando o candidato ao preenchimento estiver na intersecção de dois objetos, apenas conteúdo de um poderá ser copiado, comprometendo a reconstrução das bordas no interior dos *holes*.

Köppel, Müller and Wiegand (2016) apresentam um método híbrido para o preenchimento de *disocclusions*. Primeiramente, separa-se a estrutura da textura na vista de referência por meio do filtro de (XU et al., 2012). Então, projeta-se a imagem resultante e o mapa de profundidades para o ponto de vista virtual, e inicializa-se o conteúdo dos *holes* com a abordagem proposta por (NDJIKI-NYA et al., 2011). Após, faz-se a reconstrução da estrutura que separa diferentes regiões de textura no interior das *disocclusions*, por meio de uma adaptação do algoritmo de (CRIMINISI; PEREZ; TOYAMA, 2004). Nesta adaptação, computa-se a prioridade com um termo de dados modificado, apenas na região de *background*. Além disso, para determinar o melhor candidato para o preenchimento, calcula-se a diferença no espaço de cores RGB e, de forma ponderada, a similaridade na região inicializada do *hole* e também entre os *patches* no quadro anterior. Após refazer a estrutura, utiliza-se um modelo espacial auto regressivo para reconstruir as regiões homogêneas de textura. Então, se o preenchimento for considerado como não satisfatório, com base em um limiar de intensidades, aplica-se o algoritmo baseado em *patches* no *hole*, mas sem adaptações. Por fim, aplica-se um filtro gaussiano nas bordas de transição entre o *foreground* e as regiões reconstruídas, de modo a evitar a aparência serrilhada. Contudo, salienta-se que nesta abordagem o correto preenchimento dos *disocclusions* está fragilmente atrelado a identificação e recuperação exata das bordas dominantes.

Luo et al. (2016) propõem uma abordagem similar a apresentada em (LUO; ZHU, 2017), para preencher *holes* em vídeos. Para construir o modelo, aplica-se o algoritmo de segmentação Random Walks (GRADY, 2006) no mapa de profundidades suavizado, com o objetivo de identificar e remover elementos de *foreground*. Nesta etapa, definem-se as sementes iniciais para o algoritmo no *foreground* e *background* por meio de operações morfológicas de erosão e dilatação, e do detector de bordas de Canny. Com base nas sementes, o algoritmo gera um mapa de probabilidades, que permite estimar e remover os pontos de *foreground* da imagem, para formar o modelo de *background* limpo. Para vídeos com câmeras não estacionárias, considerando quadros consecutivos, estima-se a matriz de homografia por meio do casamento de pontos-chave detectados com SURF (BAY; TUYTELAARS; GOOL, 2006) combinado com *Random sample consensus* (RANSAC)

(FISCHLER; BOLLES, 1981), que permite a projeção do modelo do instante $t-1$ para t via transformação perspectiva. Como elemento central da abordagem, utiliza-se um GMM para modelar a parte estática de *background*, o qual permite controlar e atualizar o modelo com informação temporal em cada iteração. Para iniciar o processo de síntese, projeta-se o modelo de *background* parcialmente preenchido por informação temporal para a vista virtual, e os *holes* referentes aos objetos de *forground* removidos e as OOFAs são preenchidos pelo algoritmo de *inpainting* de (CRIMINISI; PEREZ; TOYAMA, 2004). Finalmente, para produzir a vista sintética, realiza-se a projeção da imagem de referência para o ponto de vista virtual, e utiliza-se o modelo de *background* para preencher as regiões sem informação de projeção. Este método recai nos mesmos problemas relacionados a outra abordagem proposta pelos autores em (LUO; ZHU, 2017).

Lie, Hsieh and Lin (2018) propõem preencher os *holes* em vídeos utilizando um *background sprite model* incremental, atualizado com informação espacial e temporal. Após inicializar o modelo incremental com o primeiro quadro do vídeo, determina-se a correspondência de pontos em cada transição, se houver movimento de câmera, utilizando *Motion Vectors* em um processo guiado pela correspondência de blocos. Então, remove-se a região de *foreground* do quadro corrente, com base em seus valores de profundidade com relação ao *background sprite model* e, após, com um refinamento baseado no crescimento de regiões. Os valores de *background* do quadro corrente são então registrados no modelo, ponderando suas intensidades de acordo com um fator predeterminado. Para formar a vista sintética, tanto o *background spite model* como o quadro corrente são projetados para o ponto de vista virtual. Então, os *cracks* vazios são preenchidos em ambos, por uma interpolação espacial. Por fim, para gerar a vista sintética, as *disocclusions* são preenchidas no quadro projetado com o conteúdo do modelo, e se ainda restarem *holes*, aplica-se novamente a interpolação espacial. Esta abordagem permite armazenar qualquer conteúdo visível e útil para o preenchimento no *background sprite model*, no entanto, a interpolação adotada para o preenchimento dos *holes* restantes, não reconstrói textura nestas regiões.

Luo et al. (2019) propõem um *framework* composto por quatro módulos para o preenchimento de *disocclusions*. O primeiro módulo, de extração de *foreground*, permite que elementos possam ser removidos tanto da vista virtual como da vista de referência. Para segmentar os elementos de *foreground*, sementes são determinadas no mapa de profundidades pré-processado por uma filtragem bilateral cruzada com o uso de operações morfológicas e do detector de bordas de Canny. Após, aplica-se o algoritmo de segmenta-

ção Random Walks fornecendo estas sementes como entrada, para separar as duas regiões. O segundo módulo tem como objetivo compensar o movimento de câmera entre quadros consecutivos. Para isso, inicialmente, pontos-chave SURF são combinados e refinados com RANSAC. Após, estes são utilizados com *Perspective-n-Point* para encontrar as projeções de imagem correspondentes, para estimar a matriz de transformação de câmera. No módulo de modelo de *background* modificado, são apresentadas três propostas de modelos diferentes, baseados no filtro da mediana temporal, *Loopy Belief Propagation* e GMM. O último módulo tem como objetivo preencher os *holes*, com um método convencional de *inpainting*, como (CRIMINISI; PEREZ; TOYAMA, 2004) ou (KOMODAKIS; TZIRITAS, 2007), aplicado no modelo de *background* projetado, utilizado para completar a vista sintética. Apesar de o *framework* apresentado ser bastante flexível, a construção do modelo sempre vai depender da correta segmentação dos elementos de *foreground*, pois se alguma parte for mantida, os algoritmos de *inpainting* podem ser induzidos a reconstruções erradas.

2.2.2 Estimativa de Conteúdo de *Background* para DIBR

A subtração de *background* de vídeos capturados por câmeras estáticas tem sido um dos tópicos de pesquisa mas ativos em visão computacional nas últimas duas décadas, em razão do seu grande número de aplicações, como monitoramento de tráfego e vigilância inteligente de atividades humanas em espaços públicos (BOUWMANS et al., 2019). Como parte deste processo, produz-se um modelo matemático de *background* estático, empregado no comparativo com cada quadro da sequência de vídeo (JUNG, 2009; VISWANATH et al., 2015). Neste comparativo, identificam-se *pixels* do quadro analisado que não possuem correspondência com o modelo, os quais são indicados como candidatos a pertencer ao *foreground* (MOHAMED; TAHIR; ADNAN, 2010). Com isso, torna-se possível identificar objetos de *foreground* e separá-los do *background* em sequências de vídeo.

No contexto de DIBR, não se tem como objetivo fim identificar ou remover objetos de *foreground*, mas sim utilizar a informação estimada no modelo de *background* estático para preencher parcial ou totalmente *holes* em imagens sintéticas, como ocorre em (YANG, 2015; LUO et al., 2016; LIE; HSIEH; LIN, 2018). Essas abordagens baseiam-se na premissa de que alguns dos *holes* formados após a projeção (mais especificamente, as *disocclusions*), correspondem a regiões do *background* que estavam cobertas por elemen-

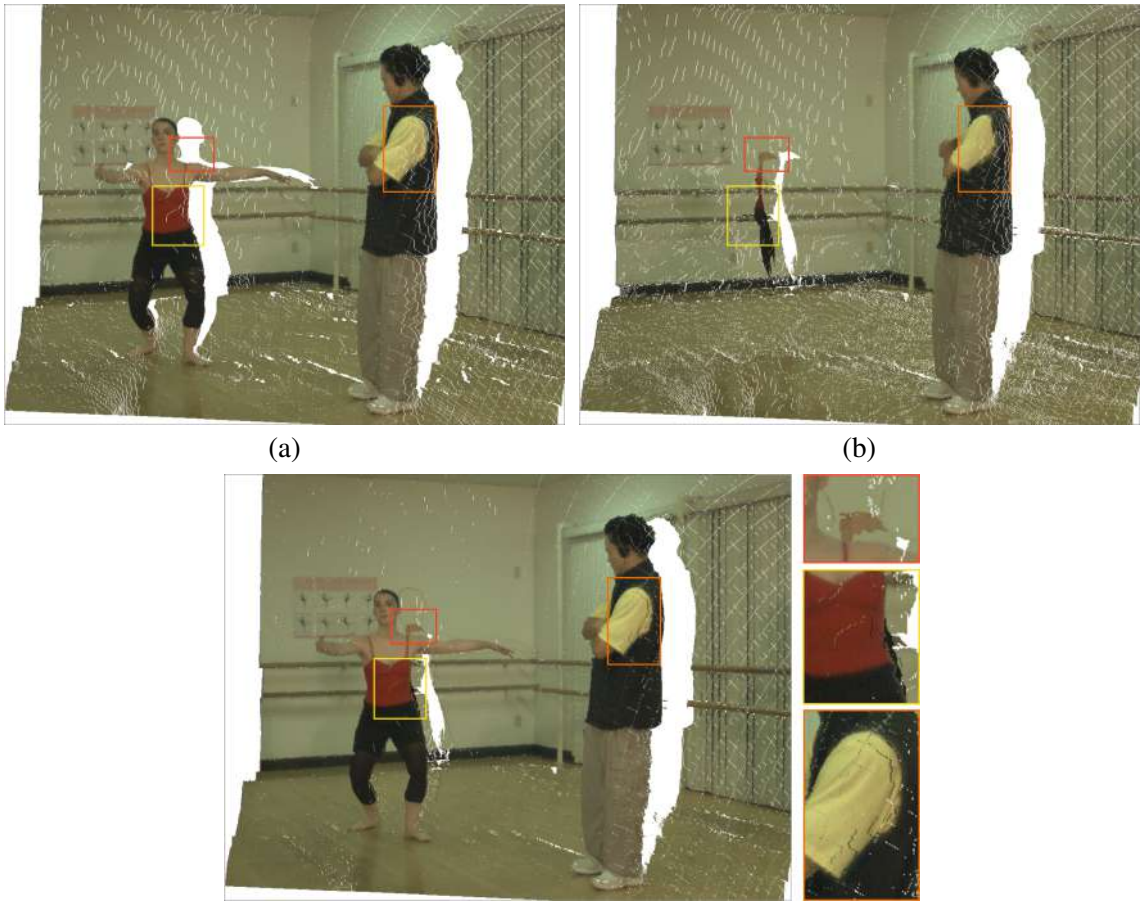
tos no *foreground* que foram expostas. Neste caso, se o modelo de *background* estático – projetado para o ponto de vista virtual – possuir conteúdo disponível onde ocorrem as *disocclusions*, basta copiar a informação para a imagem sintética.

De fato, modelos de *background* estático permitem que *disocclusions* sejam preenchidas de forma eficiente e confiável, evitando processamento desnecessário de algoritmos de *inpainting*. Contudo, como observa Xu et al. (2016), métodos para a estimativa de modelos de *background* enfrentam diferentes desafios, tais como: mudança de iluminação; *background* dinâmico (por movimento de árvores, semáforos, etc.); sombras de objetos no *foreground*; ruído no vídeo (devido a ruído do sensor, artefatos de compressão, etc.). Estes problemas tornam o processo de estimativa ainda mais desafiador. Por isso, o método de formulação do modelo não deve levar em conta somente a identificação dos elementos de *background*, mas também alterações decorrentes do processo de captação do vídeo.

Métodos para a geração de modelos de *background* contam, normalmente, somente com o conteúdo do vídeo. Mas com DIBR, tem-se sempre informação adicional de disparidade associada a cada *pixel* de cada um dos quadros da sequência de vídeo. Neste caso, além de poder estimar padrões de intensidade nas imagens da sequência de vídeo, as abordagens podem analisar o posicionamento no cenário de *pixels* e/ou segmentos, para definir qual parte do conteúdo pertence efetivamente ao *background*. Entretanto, mapas de disparidade podem conter inconsistências temporais, decorrentes do modo como foram estimados (YAO et al., 2014). Portanto, validações adicionais são requeridas quando este tipo de informação for utilizada no processo de geração do modelo de *background*.

Abordagens DIBR normalmente produzem modelos de *background* com base em informação de cor e disparidade (ou profundidade), processando um ou mais quadros de uma sequência de vídeo. Por exemplo, Yao et al. (2014) baseiam-se na observação de que a maioria das regiões oclusas em uma cena pertencem ao *background* que está coberto por objetos de *foreground*, e essas regiões oclusas podem se tornar visíveis em outros quadros, devido ao movimento do *foreground*. Desta forma, se os quadros do vídeo forem analisados em sequência, e o conteúdo associado ao maior valor de profundidade for acumulado no modelo, ao final será formada uma imagem com todo o conteúdo de *background* visível no vídeo. Contudo, se esta estratégia for adotada, partes do *background* podem não ser exibidas no vídeo, fazendo com que objetos no *foreground* permaneçam como parte do modelo. Uma estratégia diferente foi adotada por Luo et al. (2019), que utiliza apenas um quadro da sequência de vídeo para formular o modelo de *background*, no qual objetos

Figura 2.8: Preenchimento de vista sintética com informação de um modelo de *background*. Em (a) e (b) são exibidos a imagem de referência e o modelo de *background*, respectivamente, após a projeção. A imagem (c) corresponde a (a) preenchida com o conteúdo disponível em (b). Na imagem (c), destacam-se erros de preenchimento em laranja, vermelho e amarelo (marcados nas mesmas posições em (a) e (b)), produzidos pela cópia de informação inadequada de (b), exibidas em detalhe ao lado.



(a) (b)

(c)

Fonte: O autor, com imagens adaptadas do *dataset Ballet* de (ZITNICK et al., 2004).

de *foreground* são segmentados, removidos e as regiões correspondentes reconstruídas com um algoritmo de *inpainting*. Apesar da abordagem precisar de apenas um quadro do vídeo para construir o modelo, sua formulação depende da adequada detecção e reconstrução das regiões ocupadas pelos objetos de *foreground*. Neste caso, um erro em algum dos processos pode fazer com que seja mantido um modelo incoerente para todo o vídeo. Com base no exposto, pode-se perceber que falhas em etapas do processo ou a falta de informação real de *background* podem levar à construção de um modelo de *background* com erros ou conteúdo de *foreground*. Por esta razão, não se pode copiar diretamente a informação do modelo para a vista sintética, e validações adicionais são requeridas, de modo a evitar que possíveis erros de estimativa possam ser replicados.

A Figura 2.8 exibe um exemplo do preenchimento direto de uma imagem sintética com conteúdo de um modelo de *background*. No exemplo, a imagem (a) foi preenchida com conteúdo da (b), produzindo (c). Como pode ser observado em detalhe no lado direito da Figura 2.8(c), a cópia direta de informação pode produzir erros. Na região destacada em vermelho, pode-se observar que parte do corpo da bailarina preservado no modelo de *background* foi utilizado para preencher parcialmente a *disocclusion*. Neste caso, se houver uma verificação baseada em disparidade ou coerência com os elementos de *background* vizinhos da *disocclusion*, a informação incorreta pode ser identificada e descartada. Abaixo, em amarelo e laranja, *cracks* translúcidos podem ser visualizados, decorrentes do preenchimento inadequado de ocorrências do artefato na forma vazia com conteúdo retirado do modelo. Ainda, fica evidente no exemplo que modelos produzidos com câmeras estáticas não são capazes de produzir conteúdo para o preenchimento de OOFAs, por não conseguirem capturar esta região. Sendo assim, após analisar os possíveis erros de preenchimento, fica evidente que modelos de *background* não podem ser utilizados para preencher *cracks* e OOFAs. De outro modo, se o conteúdo a ser copiado for verificado adequadamente, parte significativa das *disocclusions* pode ser preenchida com informação confiável, sem o uso de algoritmos de *inpainting*.

2.2.3 Avaliações de Abordagens para Síntese de Vistas em Cenário Real

Após uma vasta busca por material associado a síntese de vistas com mapas de disparidade reais na literatura, foram encontrados apenas dois trabalhos relacionados, os quais estão sumarizados nesta subseção.

Lu e seus colegas (LU; YANG; LAFRUIT, 2009) descobriram que a medida de erro *root mean square* (RMS) de mapas de disparidade estimados pode não se correlacionar com a qualidade de vistas interpoladas. Estas conclusões são tiradas de experimentos com diferentes algoritmos de casamento estéreo, usando apenas um método de interpolação de vistas. Levando em conta os problemas comumente encontrados em *pipelines* para interpolação de vistas, os autores propõem uma nova métrica para classificar os algoritmos de casamento estéreo. Na prática, os métodos de interpolação de vistas e de DIBR não sofrem exatamente dos mesmos problemas, de modo que sua métrica não se aplica ao contexto deste trabalho. Vale ressaltar que os métodos para interpolação de vistas requerem pelo menos duas imagens de cor e de disparidade ($2V + 2D$), que são combinadas para formar a vista sintética, diferindo da abordagem DIBR clássica que utiliza apenas

uma imagem de cor e um mapa de disparidades ($V + D$).

Da mesma forma, Fuhr et al. (2013) compararam diferentes algoritmos de casamento estéreo tendo como aplicação interpolação de vistas. Os autores consideraram diferentes algoritmos de casamento estéreo, escolhidos para esta finalidade, e apenas um método de interpolação de vistas, assim como em (LU; YANG; LAFRUIT, 2009). Além disso, uma única métrica foi usada para classificar algoritmos de casamento estéreo, e as novas vistas são avaliadas por métricas de qualidade de imagem de propósito geral. Eles concluíram que "*number of bad pixels*" nos mapas de disparidade estimados, comumente utilizada para avaliar métodos de casamento estéreo, está fracamente correlacionada com o PSNR e SSIM de vistas sintéticas.

Não foram encontrados trabalhos que tenham como objetivo central analisar o impacto dos mapas de disparidades estimados com casamento estéreo em abordagens destinadas a síntese de vistas (interpoladas ou DIBR).

2.3 Conclusões do Capítulo

Neste capítulo foram apresentados os principais conceitos relacionados com a geração de vistas sintéticas com o modelo DIBR. Inicialmente, discutiram-se os problemas relacionados à aquisição de disparidade e/ou profundidade, essencial para a geração de novos pontos de vista virtuais via 3D *image warping*. Como descrito, após a projeção, artefatos e *holes* são produzidos na imagem sintética. Dentre os artefatos, destacam-se os *cracks* que devem ser identificados em ambas as formas, removidos e preenchidos adequadamente, e principalmente os *ghosts*, que além de prejudicar a qualidade visual da vista sintética, comprometem o processo de preenchimento das *disocclusions*. Estas, por sua vez, apresentam-se como principal entrave para viabilidade do modelo DIBR, pois são normalmente grandes (tanto em largura como altura). Além disso, para uma reconstrução adequada das *disocclusions*, deve-se considerar apenas informação de *background*, o que requer normalmente um processo preciso de classificação do conteúdo da imagem (completa ou de uma parte). Da mesma forma, preencher as OOFAs (quando necessário), impõe-se como um desafio, não só pelo seu tamanho, mas também devido a pouca disponibilidade de informação de vizinhança para sua reconstrução.

Na Tabela 2.1 são sumarizados os trabalhos relacionados com o tema desta tese, destacando o modelo de entrada empregado, tipos de artefatos tratados e *holes* preenchidos. No resumo, pode-se observar que a grande maioria dos trabalhos não apresenta

abordagens dedicadas a identificação e preenchimento de todos os artefatos, com destaque para os *cracks* translúcidos e *ghosts*. Além disso, as técnicas empregadas no tratamento dos *ghosts*, geralmente, removem conteúdo válido da imagem, mas deveriam apenas realocar os *pixels* que formam o artefato para o local correto. Em sua grande maioria, os trabalhos propostos ao longo dos anos apresentaram como foco principal a reconstrução das *disocclusions*, visando utilizar adequadamente conteúdo de *background*. Nestes trabalhos, na maioria dos casos, utiliza-se incoerentemente a abordagem projetada para as *disocclusions* para preencher *cracks* e OOFAs, que possuem características próprias e necessitam de métodos específicos. Como diferencial, nas publicações mais recentes, nota-se o emprego de informação temporal nos algoritmos de *inpainting*. Contudo, poucas abordagens visam controlar a coerência em vídeos, seja para evitar efeitos como *flickering* e/ou erros de preenchimento.

Tabela 2.1: Visão geral dos trabalhos que compõem o estado da arte para a geração de imagens sintéticas com o modelo DIBR, relacionando modelo de entrada empregado, artefatos tratados, tipos de *holes* preenchidos e uso de informação temporal. Na tabela, abordagens que apresentam métodos específicos para os itens representados nas colunas são indicados por "x". Em alguns trabalhos, os autores preenchem *cracks* e OOFAs com a mesma abordagem proposta para as *disocclusions*, neste caso os itens são simbolizados com "*". As abordagens que utilizam o modelo de entrada 2V+2D apresentam "–" na coluna das OOFAs, devido ao fato de não precisarem preencher este tipo de *hole*. Além disso, nas duas últimas colunas, "Preenc." indica o uso de informação temporal para a reconstrução parcial dos *holes* e "Coesão" faz referência a manutenção da coerência na transição de quadros para vídeos.

Métodos	Entrada		Artefatos			*Holes*		Inf. Temporal	
	V+D	2V+2D	*Cracks* Vazios	*Cracks* Transl.	*Ghosts*	*Disocclusions*	OOFAs	Preenc.	Coesão
Mori et al. (2008), Tanimoto et al. (2008) – VSRS2		x	x		x	x	-		
Tanimoto et al. (2008) – VSRS1	x		x		x	x	*		
Oh, Yea and Ho (2009)		x	x		x	x	-		
Zinger, Do and With (2010)		x	x		x	x	-		
Daribo and Saito (2011)	x[1]		*			x	x		
Solh and AlRegib (2012b)	x		*			x	*		
Ahn and Kim (2013)	x		x	x	x	x	*		
Yao et al. (2014)	x		*			x	*	x	x[2]
Oliveira et al. (2015)	x		x		x	x			
Schmeing and Jiang (2015)	x		*			x	*	x	
Köppel, Müller and Wiegand (2016)	x		*			x	*		x
Luo et al. (2016)	x		*			x	*	x	x[2]
Muddala, Sjöström and Olsson (2016)	x		x	x	x	x	x		
Luo and Zhu (2017)	x		*			x	x		
Cho et al. (2017)	x		*			x	*		
Dai and Nguyen (2017)		x	x		x	x	-		
Lie, Hsieh and Lin (2018)	x		x			x	*	x	x[2]
Luo et al. (2019)	x		*			x	*	x	x[2]
Capítulo 3	x		x	x	x	x	x		
Capítulo 4	x		x	x	x	x	x	x[3]	x[2,3]

[1]Método utiliza como base o modelo de entrada V+D, mas necessita de informação complementar (dita residual) de outra vista.

[2]Esta abordagem utiliza um modelo de *background* como fonte de informação temporal, o que permite controle da coesão em parte dos *holes*.

[3]Esta abordagem utiliza o modelo de *background* incremental proposto no Capítulo 5.

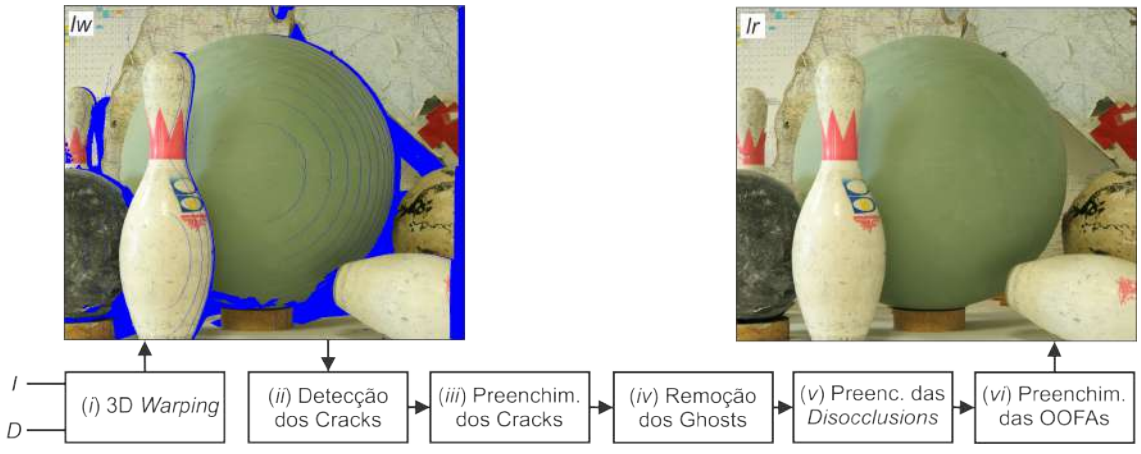
3 UM MÉTODO DIBR CIENTE DO TIPO DE ARTEFATO PARA SÍNTESE DE VISTAS

Neste capítulo, descreve-se a abordagem denominada "*An Artifact-type Aware DIBR Method for View Synthesis*" (ATA), publicada em (OLIVEIRA; WALTER; JUNG, 2018), que corresponde a avanços e extensões produzidos a partir de (OLIVEIRA, 2016). Mais especificamente, na Seção 3.1, apresenta-se uma visão geral da abordagem, com o passo a passo que compõe o *pipeline* proposto. Na seção seguinte, detalham-se os algoritmos empregados na detecção e remoção dos *cracks* vazios e translúcidos. Na Seção 3.3, descreve-se o algoritmo proposto para o tratamento dos *ghosts*. Por fim, na Seção 3.4, apresenta-se o algoritmo de *inpainting* empregado no preenchimento das *disocclusions* e OOFAs.

3.1 Visão Geral da Abordagem

A abordagem ATA divide-se em seis etapas, cada uma com um objetivo específico, executadas de modo serial. Na Figura 3.1, ilustra-se o passo a passo adotado pela abordagem, onde as tarefas são identificadas por blocos. Como pode ser visto no *pipeline*, inici-

Figura 3.1: Diagrama de blocos com o passo a passo adotado pela abordagem ATA. Como entrada para o processo, utiliza-se uma imagem de referência I e o respectivo mapa de disparidades D. Inicialmente, projeta-se I para o ponto de vista virtual com *3D image warping*, gerando I_w (com *holes* e artefatos destacados em azul). Então, após a execução de todas as etapas que compõem o *pipeline*, produz-se a imagem sintética I_r.



Fonte: O Autor, com imagens adaptadas do *dataset* Bowling1 de (HIRSCHMULLER; SCHARSTEIN, 2007).

almente projetam-se a imagem de referência I e seu respectivo mapa de disparidades D para o ponto de vista virtual com 3D *image warping*, gerando I_w e D_w, respectivamente. Após este processo, artefatos e *holes* são revelados na imagem sintética (destacados pela cor azul em I_w), e então dá-se início ao tratamento dos problemas inerentes ao modelo DIBR.

Na etapa (ii), *cracks* vazios e translúcidos são detectados simultaneamente com um processo de filtragem, que permite, ao mesmo tempo, estimar valores de disparidade para estas regiões em D_w. Na etapa seguinte, são removidos de I_w os *pixels* correspondentes ao artefato, que são preenchidos com o algoritmo HHF. No passo seguinte, etapa (iv), regiões que podem conter *ghosts* são identificadas e avaliadas de acordo com sua similaridade em relação a ambas extremidades das *disocclusions*, e *pixels* classificados como artefato são projetados para o *foreground*.

Para finalizar a vista sintética, faz-se necessário reconstruir adequadamente os *holes*. Deste modo, na etapa (v), preenchem-se as *disocclusions* em I_w e D_w concomitantemente, com uma adaptação do algoritmo de *inpainting* de (CRIMINISI; PEREZ; TOYAMA, 2004), que visa reconstruir essas regiões apenas com informação de *background* copiada da imagem de referência I. Então, no último passo, utiliza-se outra variação deste mesmo algoritmo de *inpainting* para preencher as OOFAs, com base nas suas especificidades. Por fim, têm-se como resultado a imagem sintética I_r, sem artefatos e com os *holes* totalmente preenchidos. A seguir explica-se cada uma dessas etapas.

3.2 Detecção e Preenchimento dos Cracks

3.2.1 Detecção Simultânea das Formas Vazia e Translúcida

Os *cracks* são normalmente encontrados no interior de objetos, circundados por valores uniformes de disparidade. Estes possuem forma aproximadamente linear, podendo ser horizontais, verticais ou diagonais (YANG et al., 2011), de acordo com a direção adotada no momento da projeção, como pode ser visto na Figura 3.2(a). Na forma vazia, as regiões afetadas não possuem valor de disparidade associado. Já na translúcida, o artefato possui maior valor de disparidade – que pertence ao objeto no *background* – que a sua vizinhança, composta pelo elemento de *foreground*. A Figura 3.2(b) exibe diversas ocorrências deste artefato no mapa de disparidades, em um tom de cinza escuro na face da bailarina. Sendo assim, para que os *cracks* vazios assumam exatamente o mesmo padrão

Figura 3.2: Exemplo da identificação de *cracks*. (a) Imagem de entrada com *cracks* vazios e translúcidos. (b) Mapa de disparidades associado a (a). (c) *Cracks* vazios e translúcidos identificados em azul e vermelho, respectivamente.

(a) (b) (c)

Fonte: O Autor, com imagens adaptadas do *dataset* Ballet de (ZITNICK et al., 2004).

dos translúcidos, basta que estes assumam valor de disparidade inferior ao da vizinhança. No caso ilustrado na Figura 3.2(b), os *pixels* em branco assumiriam uma cor ainda mais escura que a sua vizinhança (como preto).

O Algoritmo 1 apresenta o pseudocódigo com o passo a passo da abordagem proposta para a detecção de ambas as formas do artefato. Inicialmente, na Linha 2, para que os *cracks* vazios assumam as características dos translúcidos, atribui-se em D_w o valor -1 em todos os pontos que não possuem informação de projeção. Com isso, todas as ocorrências do artefato passam a apresentar valor de disparidade inferior à vizinhança, correspondente ao objeto no *foreground*.

Para identificar as regiões afetadas pelo artefato, basta seguir o processo proposto anteriormente para a detecção dos *cracks* translúcidos em (OLIVEIRA, 2016), mas com um elemento estruturante de tamanho diferente. Como exibido na Linha 4 do Algoritmo 1, aplica-se uma operação de fechamento morfológico com um elemento estruturante em forma de linha $Hl_4 = [1\ 1\ 1\ 1]^\top$ para detectar *cracks* verticais. O tamanho do elemento estruturante foi definido de acordo com a largura dos *cracks*, que como especificado anteriormente possuem 1 ou 2 *pixels* de largura. Esta filtragem substitui cada valor de disparidade em D_w pelo maior valor dentro da área de operação de Hl_4. Com o processo de filtragem, todos os *pixels* relacionados ao artefato mudam de valor significativamente, recebendo o valor de intensidade de um vizinho no *foreground*.

No próximo passo, Linha 5 do Algoritmo 1, calcula-se a diferença entre o mapa filtrado e o original. Por fim, para formular a imagem binária com os *cracks* identificados VT, como definido a partir da Linha 6, verificam-se quais são as ocorrências reais do artefato com o uso de um limiar predeterminado λ. Neste caso, se λ for pequeno, mínimos

locais de D_w relativos a geometria de objetos podem ser equivocadamente considerados como artefato. Porém, se λ for muito grande, ocorrências reais de *cracks* podem não ser identificadas. Com base em experimentos, definiu-se $\lambda = 5$, que demonstrou um bom compromisso entre estes dois extremos.

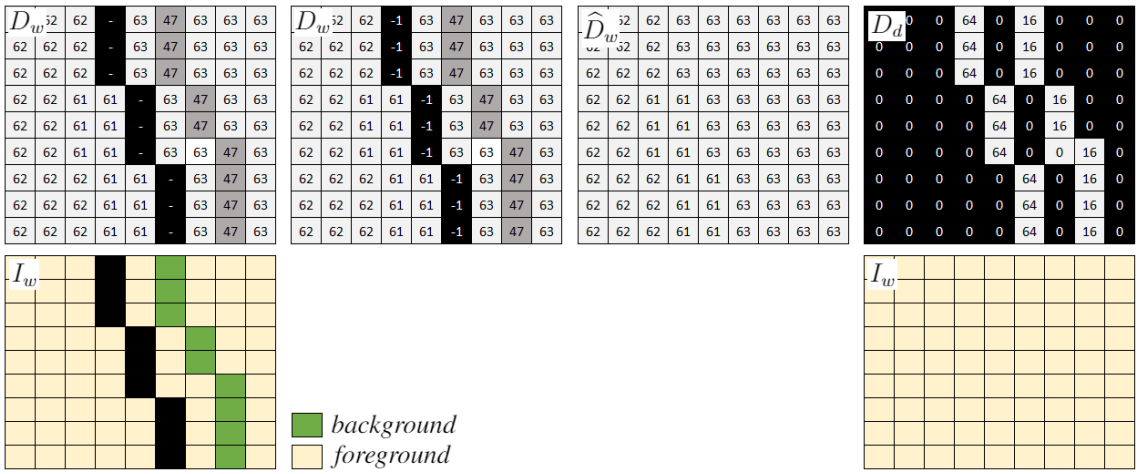
Algoritmo 1: Pseudocódigo com a abordagem de detecção simultânea de *cracks* vazios e translúcidos.

Entrada: D_w mapa de disparidades projetado
Hl_4 elemento estruturante linear
Saída: VT imagem binária com *cracks* identificados

início
  Inicializa D_w com -1 em todos os *pixels* sem informação de projeção;
  Inicializa VT com 0 em todos os *pixels*;
  $\hat{D}_w \leftarrow$ imagem resultante do fechamento morfológico de D_w com Hl_4;
  $D_d \leftarrow \hat{D}_w - D_w$;
  para todo $\boldsymbol{p} \in D_d$ **faça**
    se $D_d(\boldsymbol{p}) \geq \lambda$ **então**
      $VT(\boldsymbol{p}) \leftarrow 1$;
    fim
  fim
fim

A Figura 3.3 exibe na primeira linha o resultado de cada uma das operações executadas para a detecção dos *cracks*. Para detectar *cracks* horizontais (caso exista projeção vertical), basta executar o Algoritmo 1 novamente com ${Hl_4}^\top$ e unir as imagens binárias

Figura 3.3: Processo de identificação dos *cracks* vazios e translúcidos. Na linha superior, apresenta-se o resultado de cada uma das operações empregadas no processo de identificação. Abaixo, no lado esquerdo, exibi-se a identificação do *foreground* e *background* em I_w antes da remoção do artefato, e na outra extremidade o resultado após o preenchimento.



Fonte: O Autor.

produzidas. Os pontos identificados são removidos tanto de I_w como de D_w, para que seja realizado o processo de reconstrução.

3.2.2 Análise da Representatividade e Distribuição dos *Cracks*

Na Tabela 2.1, discutida no capítulo anterior, foram sumarizados 17 trabalhos diferentes, que contam com técnicas destinadas principalmente ao preenchimento de *disocclusions*. Como pode ser observado na tabela, dentre essas abordagens, 9 não empregam mecanismos específicos para o preenchimento dos *cracks* vazios. Considerando os 8 trabalhos restantes, que tratam esta forma do artefato de modo exclusivo, somente 2 removem a forma translúcida. Então, considerando a pouca atenção dada ao tratamento dos *cracks*, decidiu-se analisar a representatividade em termos do número de *pixels* onde o artefato ocorre (em ambas as formas) em relação ao conteúdo total da imagem. Como complemento, examinou-se a distribuição do artefato nas imagens.

A Tabela 3.1 apresenta uma análise referente à representatividade e localização dos *cracks* em vistas sintéticas. Estes dados foram computados com o uso do algoritmo de detecção proposto, aplicado em diferentes *datasets* (detalhados mais a frente, no Capítulo 6) amplamente empregados na avaliação de algoritmos de síntese de vistas com DIBR.

Como destacado na coluna TO (taxa de ocorrência) da tabela, dada em razão de *pixels*, *cracks* podem corresponder a $8,43\%$ do conteúdo total das vistas sintéticas de um vídeo (com 100 quadros), com um desvio padrão médio de $0,43\%$. Ainda, pode ser observado que as vistas sintéticas contém em média $3,6\%$ de seu conteúdo coberto pelo artefato, com um desvio padrão médio de $0,16\%$ (indicando um pequeno distanciamento da média). Como a avaliação indica, o artefato representa uma porção significativa das vistas sintéticas. Isto fica evidente na Figura 3.5(a), que apresenta o resultado produzido pelo algoritmo de detecção proposto em diferentes casos de teste.

Quando a imagem de uma cena é apresentada na tela do computador, os observadores olham primeiramente para o centro da cena (BINDEMANN, 2010). Por este motivo, artefatos visuais – como *cracks* – quando concentrados no centro da imagem, tendem a ser mais facilmente percebidos por usuários do que quando estão bem distribuídos ou acumulados em regiões periféricas. Com base nisso, foi definida uma métrica para estimar a distância média dos *pixels* detectados como ocorrências do artefato com relação ao centro da imagem. A coluna TC (tendência central) apresenta esta estimativa

Tabela 3.1: Taxa de ocorrência e análise da localização dos *cracks* na avaliação de diferentes *datasets*. A tabela exibe o resultado de testes com *datasets* de fotografias (F) e vídeos (V), indicados na coluna Tipo. Para os casos de teste com vídeos foi calculada a média de ocorrência nos primeiros 100 quadros. Na coluna Projeção, por exemplo, "$1 \rightarrow 2$" indica que o ponto de vista 1 (de referência) foi projetado para o 2 (virtual). A coluna TO se refere a taxa de ocorrência do artefato, TC a tendência central – computada com base na média da distância dos *pixels* correspondentes a *cracks* com relação ao centro da imagem – e σ o desvio padrão para ambos. Destaca-se em negrito a média e o maior valor encontrados para TO e TC nas imagens e vídeos analisados.

Dataset	Tipo	Projeção	TO	σ_{TO}	TC	σ_{TC}
11 *Datasets* de Middleburry	F	$1 \rightarrow 2$	0,75%	0,17%	0,4882	0,0522
Ballet	V	$4 \rightarrow 1$	6,27%	0,16%	0,4577	0,0068
Ballet	V	$4 \rightarrow 3$	2,98%	0,07%	0,4199	0,0053
Ballet	V	$5 \rightarrow 2$	4,94%	0,16%	0,4631	0,0081
Ballet	V	$5 \rightarrow 4$	2,81%	0,07%	0,4249	0,0057
Breakdancer	V	$4 \rightarrow 1$	**8,43%**	0,42%	0,4147	0,0090
Breakdancer	V	$4 \rightarrow 3$	3,51%	0,16%	0,3919	0,0087
Breakdancer	V	$5 \rightarrow 2$	8,39%	0,41%	0,4048	0,0078
Breakdancer	V	$5 \rightarrow 4$	4,23%	0,15%	0,4218	0,0066
PoznanHall2	V	$5 \rightarrow 7$	1,73%	0,16%	**0,5459**	0,0349
PoznanHall2	V	$6 \rightarrow 7$	1,06%	0,16%	0,5372	0,0279
PoznanStreet	V	$3 \rightarrow 4$	2,59%	0,10%	0,4684	0,0056
PoznanStreet	V	$3 \rightarrow 5$	4,35%	0,15%	0,4877	0,0047
Dancer	V	$1 \rightarrow 5$	0,97%	0,03%	0,5189	0,0195
Dancer	V	$5 \rightarrow 9$	1,01%	0,05%	0,4977	0,0315
Média			**3,60%**	0,16%	**0,4629**	0,0156

Fonte: O Autor.

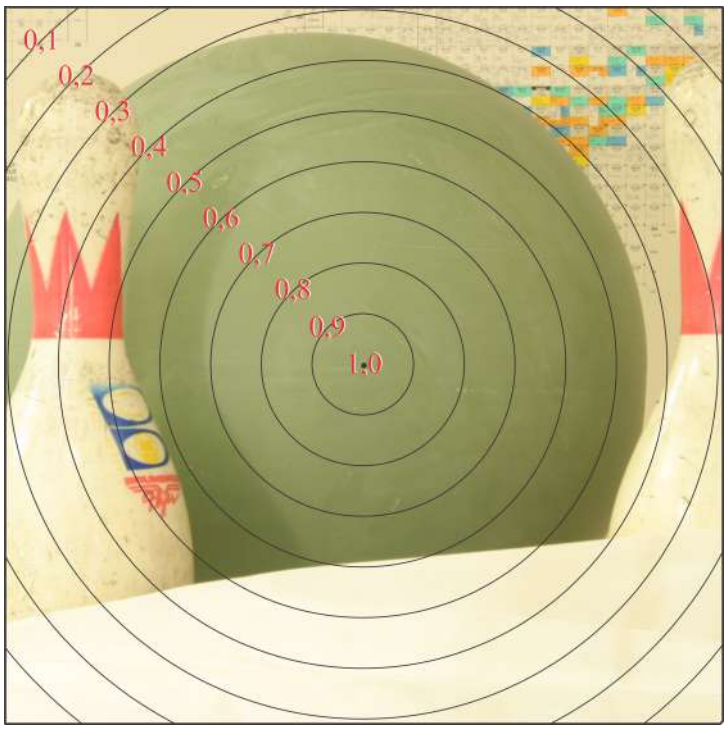
no intervalo $[0, 1]$. O cômputo de TC foi definido de acordo com a estratégia ilustrada na Figura 3.4. Como pode ser constatado, um valor próximo de 1 indica uma alta concentração do artefato no centro da imagem e um valor baixo aponta que as ocorrências tendem a ser encontradas nas regiões periféricas. A estimativa de TC é dada individualmente para cada *pixel* – cuja coordenada é (x, y) – por meio da seguinte equação:

$$TC(x, y) = 1 - \frac{\sqrt{(ma_c - x)^2 + (nl_c - y)^2}}{TC_{MAX}}, \tag{3.1}$$

$$TC_{MAX} = \frac{\sqrt{{ma_c}^2 + {nl_c}^2}}{2}, \tag{3.2}$$

onde ma corresponde a altura da imagem e nl a largura, que são utilizados para identificar a linha e a coluna centrais da imagem $ma_c = (ma/2 + 0,5)$ e $nl_c = (nl/2 + 0,5)$, respectivamente. Na Tabela 3.1, apresenta-se o valor médio de TC para cada um dos *datasets*, considerando todos os quadros do vídeo ou todas as fotografias.

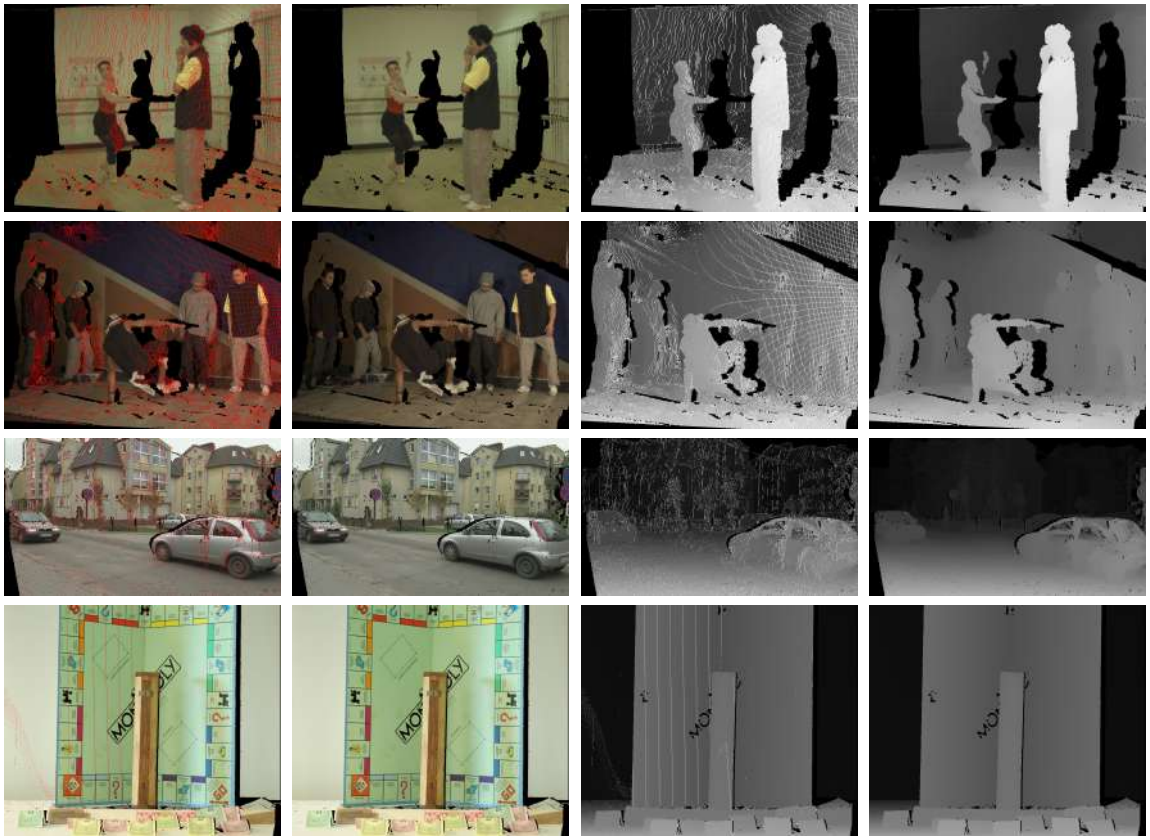
Figura 3.4: Ilustração que exemplifica a abordagem empregada na estimativa da distribuição dos *cracks*. A métrica proposta tem como objetivo detectar o quão próximos do centro da imagem estão os artefatos, por meio do computo de TC (tendência central) que é definido no intervalo $[0, 1]$. Como pode ser visto, quanto menor for a distância com relação ao centro da imagem, maior será o valor computado por TC, e vice-versa.



Fonte: O Autor com uma imagem adaptada de (HIRSCHMULLER; SCHARSTEIN, 2007).

Os valores exibidos na Tabela 3.1 representam a distribuição dos *pixels* indicados como *cracks* na imagem. As estimativas fornecem uma noção de maior concentração do artefato em uma região intermediária, entre o extremo e o centro da imagem. O maior valor médio encontrado para TC foi $0,5459$, que está muito próximo da média geral $0,4629$ e que, por sua vez, está próximo do meio do intervalo. Estes valores poderiam indicar uma distribuição equivalente no centro e no extremo da imagem, sem ocorrências na região intermediária. No entanto, o baixo desvio padrão médio $(0,0156)$ comprova o contrário. Então, com base nesta estimativa, pode-se concluir que os *cracks* se concentram principalmente em uma região intermediária entre o extremo da imagem e o centro.

Figura 3.5: Resultados produzidos com as abordagens propostas para a detecção e preenchimento dos *cracks* vazios e translúcidos em diferentes *datasets*. Em (a), exibe-se I_w com os *cracks* vazios e translúcidos detectados em vermelho. Ao lado, em (b), o resultado produzido pelo HHF no preenchimento das regiões correspondentes ao artefato. Em (c), são exibidas as ocorrências do artefato em branco e, em (d), o resultado produzido pela abordagem de preenchimento proposta.



(a) (b) (c) (d)

Fonte: O Autor, com imagens adaptadas de (ZITNICK et al., 2004; SCHWARZ; MARPE; WIEGAND, 2010; HIRSCHMULLER; SCHARSTEIN, 2007).

3.2.3 Preenchimento dos *Cracks*

O preenchimento dos *cracks* detectados VT em D_w não requer o uso de qualquer algoritmo ou processamento adicional, pois as intensidades de disparidade para estas regiões foram previamente estimadas pela operação de fechamento morfológico, na fase de detecção, como pode ser visto em $\hat{D}_w$ na Figura 3.3. Portanto, neste caso, a tarefa consiste basicamente em copiar de $\hat{D}_w$ os valores dos pontos identificados para D_w, que correspondem ao objeto de *foreground*, compatível com a região a ser reconstruída. Na Figura 3.5(d) exibe-se o resultado produzido pela técnica de preenchimento proposta em diferentes casos de testes, considerando as ocorrências do artefato em branco no mapa de

disparidades D_w exibido na imagem ao lado.

Como discutido na seção anterior, o preenchimento adequado dos *cracks* (destacados na Figura 3.2(c)) em I_w, demanda o uso de um algoritmo específico, embasado em suas características. Por padrão, os *cracks* são finos e possuem informação confiável na vizinhança, a qual pode ser empregada na sua reconstrução. Com base nisso, optou-se por realizar o preenchimento destas regiões com o algoritmo HHF (SOLH; ALREGIB, 2012b) – descrito previamente na Seção 2.2. Este algoritmo se destaca pela sua capacidade de reconstruir precisamente pequenas áreas, similares às produzidas pelo artefato. Além disso, a escolha fundamenta-se na análise apresentada em (OLIVEIRA, 2016), que comprova que esta abordagem obtém melhores resultados no preenchimento de *cracks*, quando comparado com diversos outros algoritmos apresentados na literatura, como (OLIVEIRA et al., 2001) ou (CRIMINISI; PEREZ; TOYAMA, 2004). A Figura 3.5(b) exibe o resultado final produzido com a abordagem proposta para o tratamento de ambas as formas do artefato em I_w.

3.3 Remoção dos *Ghosts*

Como definido anteriormente, os *ghosts* são normalmente compostos por uma mistura de cor de dois objetos vizinhos na cena (um no *foreground* e outro no *background*), correspondente a região de transição entre eles. Em razão do processo de estimativa do mapa de disparidades, estas regiões erroneamente recebem o valor associado ao objeto no *background*. Como efeito desta falha, após a etapa de projeção da vista de referência, esta região de transição passa a ser visualizada no lado de *background* das *disocclusions*. Visualmente, este artefato corresponde à silhueta dos objetos de *foreground* na região de *background* da imagem. Para resolver este problema, propõe-se identificar ocorrências do artefato e projetá-las para os locais apropriados (extremidade de *foreground* da *disocclusion*).

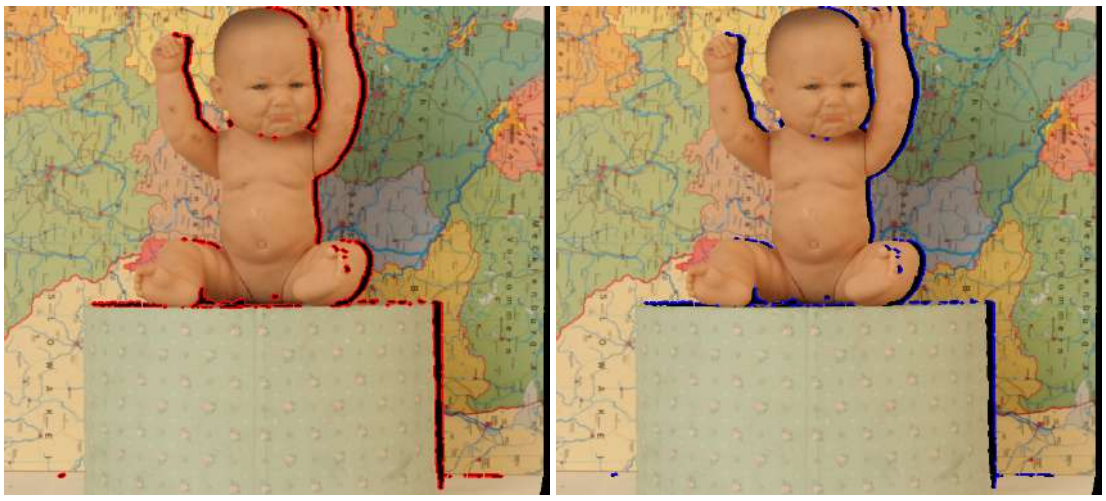
Na abordagem proposta, inicialmente são selecionadas as regiões candidatas a *ghost* na vista sintética, seguindo o método definido no Algoritmo 2. Para tanto, como exibido na Linha 2, cria-se uma máscara binária G com todas as regiões sem informação identificadas por 1. Como este tipo de artefato pode ser encontrado somente nas bordas de *disocclusions*, as OOFAs precisam ser descartadas. Por isso, como detalhado na Linha 3, as regiões correspondentes a este tipo de *hole* são identificadas e removidas. Mais especificamente, o valor 0 é atribuído nas posições correspondentes de G. As OOFAs

ocorrem somente nas extremidades da vista sintética, e podem ser identificadas com base na(s) direção(ões) adotada(s) para a projeção de I. Por exemplo, se a vista de referência estiver à esquerda da sintética, essas regiões deverão aparecer no lado direito de I_w (e vice-versa). Nesse caso, verifica-se cada Linha de I_w, começando pela borda mais à direita, parando quando se encontra conteúdo, ou seja, o fim da OOFA. Para o caso oposto, inicia-se pela borda mais à esquerda. Este processo segue o mesmo procedimento quando existem deslocamentos verticais, porém com a verificação das extremidades inferior e superior.

Com as OOFAs removidas, inicia-se a identificação das regiões que podem conter o artefato. Para tal, como detalhado na Linha 4 do Algoritmo 2, aplica-se uma dilatação morfológica em G assim como em (OLIVEIRA, 2016), mas com um elemento estruturante planar na forma de diamante – devido a forma irregular das bordas – de tamanho 2 para expandir as *disocclusions*, gerando a imagem $\hat{G}$. Após, são separados os candidatos $G_{\delta\Omega}$ por meio do operador de complemento absoluto $\setminus$, como pode ser visto na Linha 5. Com isso, é selecionada apenas a região expandida das *disocclusions*, como pode ser visto na Figura 3.6(a), que apresenta o resultado desta operação.

Os *ghosts* ocorrem somente na região de *background* das *disocclusions* e, portanto, faz-se necessário remover os pontos de *foreground* de $G_{\delta\Omega}$. Para uma imagem com múltiplas camadas de profundidade e muitos objetos, definir globalmente quais regiões pertencem ao *foreground* ou *background*, como em (LUO et al., 2016), pode ser uma

Figura 3.6: Exemplo da seleção de pontos candidatos a *ghosts*. Em (a), destaca-se $G_{\delta\Omega}$ em vermelho. Ao lado, em (b), apresenta-se em azul os candidatos após a aplicação do extrator de *background*.



(a) (b)

Fonte: O Autor, com imagens adaptadas de (HIRSCHMULLER; SCHARSTEIN, 2007).

tarefa complicada até mesmo para humanos. Entretanto, pode ser simples determinar localmente esta separação para objetos vizinhos de uma *disocclusion*. Para tanto, utiliza-se um extrator *foreground-background* que deve ser aplicado em cada *disocclusion* individualmente, proposto anteriormente em (OLIVEIRA, 2016), para a definição da prioridade de preenchimento dos *holes* no algoritmo de *inpainting*.

O extrator utilizado decompõe uma região delimitada – próxima a região alvo – em *foreground* e *background* com a simples aplicação de um limiar determinado dinamicamente. O pseudocódigo correspondente ao extrator está definido a partir da Linha 6 do Algoritmo 2. Como pode ser visto nas Linhas 7 e 8, estima-se um limiar de separação T_Ω para cada *disocclusion* Ω por meio do computo da média truncada (*trimmed mean*) dos valores de disparidade correspondentes a borda do *hole* definida em $G_{\delta\Omega}$. Para computar T_Ω, calcula-se a média de uma amostra dada como entrada, removendo os $\alpha\%$ maiores e menores valores da amostra. Estes valores são removidos para inibir possíveis *outliers* que possam comprometer a estimativa da média. Para evitar erros e manter uma margem de segurança, utilizou-se $\alpha = 10\%$. Como pode ser visto na Linha 9 do algoritmo, repete-se o processo de estimativa em cada *disocclusion* individualmente aplicando o limiar, de modo que todos os pontos com disparidade menor do que T_Ω (região do *foreground*) sejam removidos, restando apenas os candidatos pertencentes ao *background* em $G_{\delta\Omega}$. Na Figura 3.6(b), indica-se em azul o resultado da aplicação deste extrator nos pontos destacados em vermelho da imagem ao lado.

Algoritmo 2: Pseudocódigo com a abordagem para a identificação de possíveis ocorrências de *ghosts*.

Entrada: D_w mapa de disparidades projetado
Hd_2 elemento estruturante em forma de diamante
α percentual a ser removido da média truncada
Saída: $G_{\delta\Omega}$ imagem binária com regiões candidatas a *ghost*

1 **início**
2 Inicializa G com 1 nas regiões sem informação em D_w e 0 nas demais;
3 Atribuí 0 na região correspondente às OOFAs em G;
4 $\hat{G} \leftarrow$ imagem resultante da dilatação morfológica de G com Hd_2;
5 $G_{\delta\Omega} \leftarrow \hat{G} \setminus G$;
`// Extrator foreground-background`
6 **para todo** $\Omega \in G$ **faça**
7 $\delta\Omega \leftarrow$ contorno de Ω definido em $G_{\delta\Omega}$;
8 $T_\Omega \leftarrow$ valor da média truncada computada em $D_w(\delta\Omega)$ com α;
9 $G_{\delta\Omega}(D_w(\delta\Omega) < T_\Omega) \leftarrow 0$;
10 **fim**
11 **fim**

O último passo para a remoção do artefato, trata da checagem da consistência local de cada *pixel* candidato $\boldsymbol{p} \in G_{\delta\Omega}$. Para isso, utiliza-se uma abordagem similar a desenvolvida em (OLIVEIRA et al., 2015), que baseia-se na similaridade de $\boldsymbol{p}$ com seus vizinhos em ambos os lados da *disocclusion* (*background* e *foreground*). A abordagem adotada está detalhada em forma de pseudocódigo no Algoritmo 3.

Inicialmente, estimam-se os valores de disparidade associados a $G_{\delta\Omega}$ no *foreground*. Para isso, aplica-se uma dilatação morfológica em D, utilizando Hl_4, e então repete-se a operação com ${Hl_4}^{\top}$, produzindo um mapa de disparidades com as áreas de *foreground* estendidas, denominado D_{FG} (Linha 2). Então, inicia-se o processo comparação individual de cada candidato com seus vizinhos no *background* (definidos a partir de $\boldsymbol{p}$) e no *foreground* (estipulados de acordo com o ponto estimado $\boldsymbol{p}_{FG}$). Para definir o ponto no *foreground*, inicialmente (na Linha 4), determina-se o ponto de origem $\boldsymbol{p}'$, correspondente a $\boldsymbol{p}$ na imagem de referência, por meio do *inverse warping* computado com o valor de disparidade $D_w(\boldsymbol{p})$. Após, estima-se $\boldsymbol{p}_{FG}$ – na Linha 5 – com a projeção de $\boldsymbol{p}'$ para o ponto de vista virtual, utilizando o valor de disparidade $D_{FG}(\boldsymbol{p}')$. Em seguida, como definido nas Linhas 6 e 7 do algoritmo, calcula-se a média em *patches* de 9×9 centrados em $\boldsymbol{p}$ (corresponde ao *background*) e $\boldsymbol{p}_{FG}$ (*foreground*), considerando somente *pixels* válidos e não candidatos, gerando μ_{BG} e μ_{FG}, respectivamente. Então, conforme especificado nas Linhas 8 e 9, para verificar a consistência de $\boldsymbol{p}$, computa-se a diferença absoluta entre a sua intensidade e a de μ_{BG} e μ_{FG}, gerando d_{BG} e d_{FG}, respectivamente. Se $d_{BG} \geq d_{FG}$ ou $d_{BG} > \gamma$ (Linha 10), com γ sendo um limiar de similaridade, classifica-se $\boldsymbol{p}$ como um *ghost*. Esta verificação leva em conta tanto a similaridade entre o ponto e o *foreground* como grandes diferenças de intensidade com relação aos vizinhos no *background*.

Por fim, como definido a partir da Linha 11 do Algoritmo 3, os pontos classificados como *ghost* são reposicionados na extremidade correta da *disocclusion*, mais especificamente, no ponto $\boldsymbol{p}_{FG}$. Como em (OLIVEIRA et al., 2015), definiu-se $\gamma = 11$ em todos os experimentos. Destaca-se que esta abordagem não remove os *pixels* correspondentes ao artefato, pois não se trata de um "defeito" na imagem, mas apenas de inconsistências produzidas na estimativa do mapa de disparidades, as quais precisam ser corrigidas.

3.4 Preenchimento de *Disocclusions* e OOFAs

O preenchimento de *disocclusions* e OOFAs se apresenta como o maior desafio do processo de síntese de vistas. Na reconstrução dessas áreas, faz-se necessário atribuir

Algoritmo 3: Pseudocódigo com a abordagem para a avaliação de candidatos a *ghost*.

Entrada: I_w imagem colorida projetada
D mapa de disparidades
D_w mapa de disparidades projetado
$G_{\delta\Omega}$ imagem binária com regiões candidatas a *ghost*
Hl_4 elemento estruturante linear
γ limiar de similaridade
Saída: I_w imagem colorida projetada sem artefatos
D_w mapa de disparidades projetado sem artefatos

início
  $D_{FG} \leftarrow$ imagem resultante da dilatação morfológica de D com Hl_4 e ${Hl_4}^{\top}$;
  para todo $\boldsymbol{p} \in G_{\delta\Omega}$ **faça**
    $\boldsymbol{p}' \leftarrow$ ponto de origem de $\boldsymbol{p}$ determinado pelo computo do *inverse warping* com D_w;
    $\boldsymbol{p}_{FG} \leftarrow$ ponto estimado no *foreground* com a projeção de $\boldsymbol{p}'$ para o ponto de vista virtual, utilizando D_{FG};
    $\mu_{FG} \leftarrow$ resultado da média de um *patch* de 9×9 centrado em $\boldsymbol{p}_{FG}$ de I_w (computada nos *pixels* válidos sem considerar candidatos a *ghost*);
    $\mu_{BG} \leftarrow$ resultado da média de um *patch* de 9×9 centrado em $\boldsymbol{p}$ de I_w (computada nos *pixels* válidos sem considerar candidatos a *ghost*);
    $d_{FG} \leftarrow \mu_{FG} - I_w(\boldsymbol{p}_{FG})$;
    $d_{BG} \leftarrow \mu_{BG} - I_w(\boldsymbol{p})$;
    se $d_{BG} \geq d_{FG} \vee d_{BG} > \gamma$ **então**
      $I_w(\boldsymbol{p}_{FG}) \leftarrow I_w(\boldsymbol{p})$;
      $I_w(\boldsymbol{p}) \leftarrow 0$;
      $D_w(\boldsymbol{p}_{FG}) \leftarrow D_w(\boldsymbol{p})$;
      $I_w(\boldsymbol{p}) \leftarrow 0$;
    fim
  fim
fim

textura adequada e prolongar corretamente as estruturas dos objetos na cena, para que o resultado final seja o mais natural possível. Para esta tarefa, foi desenvolvida uma extensão do popular método de *inpainting* baseado em *patches* de Criminisi, Perez and Toyama (2004).

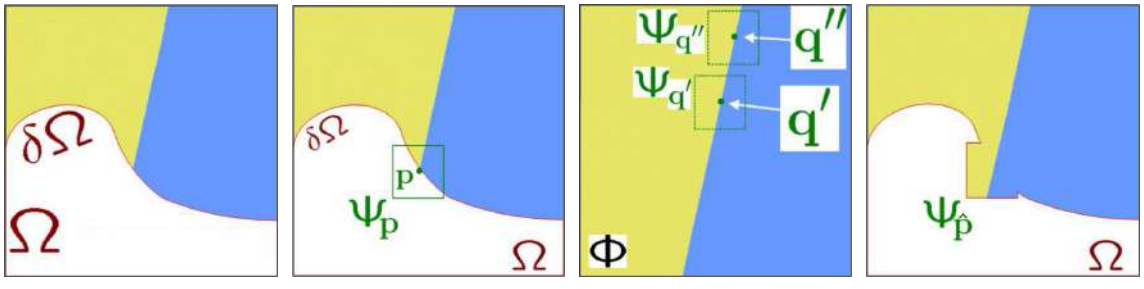
No algoritmo de *inpainting* original, de modo genérico, parte-se de uma imagem com uma região sem informação – *hole* – Ω, que precisa ser preenchida. Para que isso seja realizado da maneira apropriada, o autor segue dois processos fundamentais: **computo de prioridades**; **seleção do melhor *patch* para o preenchimento**. O computo das prioridades define a ordem na qual Ω é preenchido, por meio da análise dos pontos na

borda do *hole* $\delta\Omega$. Nesta etapa, são delimitados *patches* centrados em cada um dos pontos de $\delta\Omega$, que são utilizados para computar dois termos de prioridade (confiança e dados). Estes termos visam preencher primeiramente o ponto que possui mais informação confiável no seu redor, e dar continuidade para a borda dos objetos que estão na extremidade de Ω. Com o resultado do produto dado pelos dois termos, determina-se qual *patch* deve ser preenchido primeiramente. Após selecionar o candidato com maior prioridade, parte-se para a etapa de seleção do melhor *patch* para o preenchimento. Neste caso, o *patch* mais similar em termos de cor ao candidato é escolhido, por meio de um processo que analisa toda a região válida da imagem. Estes dois processos são repetidos iterativamente, até que Ω esteja completamente preenchido. Se houver mais de um *hole* na imagem, todos serão considerados como Ω e processados conjuntamente.

Na abordagem proposta, cujo processo de preenchimento encontra-se ilustrado na Figura 3.7, preenche-se individualmente cada *hole* Ω da vista sintética I_w, utilizando informação da imagem de referência I. Com o uso de I, evita-se que artefatos decorrentes de erros de projeção possam ser copiados durante o preenchimento e, além disso, mais conteúdo fica disponível para pesquisa e cópia de informação durante o processo de reconstrução.

Assim como no algoritmo original, em cada iteração do algoritmo, calcula-se a prioridade de preenchimento de todos os *patches* Ψ_p, centrados em cada ponto $\boldsymbol{p}$ da borda

Figura 3.7: Propagação de estrutura por síntese de textura baseada em *patches*. (a) Imagem a ser reconstruída, com a região alvo Ω e seu contorno $\delta\Omega$. (b) Cômputo da prioridade para um dado *patch* Ψ_p, centrado no ponto $\boldsymbol{p} \in \delta\Omega$, posteriormente escolhido para o preenchimento. (c) Possíveis candidatos ($\Psi_{q'}$ ou $\Psi_{q''}$) ao preenchimento do *patch* alvo, distribuídos na região de busca Φ em I. (d) Preenchimento da parte vazia de $\Psi_{\hat{p}}$ com o conteúdo do candidato que apresentou maior similaridade no processo de busca. Observa-se que tanto textura como estrutura (linha de separação dos objetos) foram propagados inicialmente dentro do *hole*. Com o preenchimento parcial de Ω, $\delta\Omega$ passou a assumir um formato diferente.



(a) (b) (c) (d)

Fonte: Adaptada de (CRIMINISI; PEREZ; TOYAMA, 2004).

$\delta\Omega$ (como exemplificado na Figura 3.7(b)), para garantir que a região alvo Ω seja reconstruída na ordem correta. Após, seleciona-se o *patch* $\Psi_{\hat{p}}$ com o maior valor de prioridade para ser preenchido. Para preencher os pontos vazios de $\Psi_{\hat{p}}$, faz-se a busca pelo *patch* mais adequado em uma região de interesse $\Phi \in I$. Na busca, todos os *patches* de Φ são comparados com $\Psi_{\hat{p}}$ e, dentre os candidatos (como $\Psi_{q'}$ e $\Psi_{q''}$, na Figura 3.7(c)), seleciona-se o mais similar com relação a cor $\Psi_{\hat{q}}$. Por fim, são preenchidos os *pixels* vazios de $\Psi_{\hat{p}}$ com o conteúdo de $\Psi_{\hat{q}}$, como pode ser visto na Figura 3.7(d). Neste algoritmo, repete-se iterativamente o processo de cálculo de prioridades e busca pelo melhor *patch* até que o *hole* inteiro seja preenchido.

3.4.1 Estimativa da Prioridade de Preenchimento

Na abordagem proposta, foram estabelecidas duas estratégias diferentes para a estimativa de prioridades para determinar a ordem adequada a cada tipo de *hole*: (i) para as OOFAs, segue-se o cálculo padrão, proposto originalmente no algoritmo, que prioriza a continuação de bordas fortes (estrutura de transição entre objetos), cercada por *pixels* de alta confiança; (ii) para as *disocclusions*, propõe-se selecionar o *patch* com a maior média de profundidade, cercado pela maior quantidade de *pixels* classificados como *background*, para preencher primeiramente esta região.

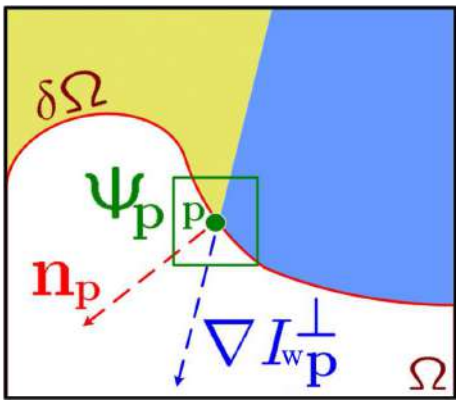
Com base nesta estratégia, dado um *patch* Ψ_p (veja a Figura 3.8), define-se sua prioridade em cada iteração calculando:

$$P(\boldsymbol{p}) = \begin{cases} C(\boldsymbol{p}) \cdot D(\boldsymbol{p}) & \text{se } \boldsymbol{p} \in \delta\Omega \text{ de uma OOFA} \\ B(\boldsymbol{p}) \cdot E(\boldsymbol{p}) & \text{caso contrário} \end{cases}, \tag{3.3}$$

onde $P(\boldsymbol{p})$ se refere a prioridade para um dado *pixel* no ponto $\boldsymbol{p} \in \delta\Omega$. $C(\boldsymbol{p})$ e $D(\boldsymbol{p})$ são os termos de confiança e dados, respectivamente, propostos em (CRIMINISI; PEREZ; TOYAMA, 2004). Além disso, $E(\boldsymbol{p})$ representa o termo de profundidades, proposto previamente em (OLIVEIRA et al., 2015), e $B(\boldsymbol{p})$ o de *background*, descrito em (OLIVEIRA, 2016). Estes termos são computados como segue:

$$C(\boldsymbol{p}) = \frac{\sum_{\boldsymbol{q} \in \Psi_p \cap (I_w - \Omega)} C(\boldsymbol{q})}{|\Psi_p|}, \tag{3.4}$$

Figura 3.8: Diagrama com a notação adotada pelo algoritmo de *inpainting*. Dado um *patch* Ψ_p, $\boldsymbol{n}_p$ refere-se a normal do contorno $\delta\Omega$ de uma região alvo Ω e $\nabla I_{w_p}^{\perp}$ indica o *isophote* – direção e intensidade – no ponto $\boldsymbol{p}$.



Fonte: Adaptada de (CRIMINISI; PEREZ; TOYAMA, 2004).

$$D(\boldsymbol{p}) = \frac{|\nabla I_{w_p}^{\perp} \cdot \boldsymbol{n}_p|}{\rho}, \tag{3.5}$$

$$E(\boldsymbol{p}) = \frac{\sum_{\boldsymbol{q} \in \Psi_p \cap (I_w - \Omega)} D_w(\boldsymbol{q})}{|\Psi_p|}, \tag{3.6}$$

$$B(\boldsymbol{p}) = \frac{\sum_{\boldsymbol{q} \in \Psi_p \cap (I_w - \Omega)} M_{BG}(\boldsymbol{q})}{|\Psi_p|}, \tag{3.7}$$

onde $|\Psi_p|$ indica o número de *pixels* não vazios em Ψ_p, e ρ é o fator de normalização (por exemplo, $\rho = 255$ para uma imagem em escala de cinza), $\boldsymbol{n}_p$ é um vetor unitário e ortogonal a $\delta\Omega$ no ponto $\boldsymbol{p}$, ∇I_{wp} representa o valor máximo do gradiente da imagem em $\Psi_p \cap I_w$ e $\perp$ denota o operador ortogonal, utilizados para indicar a direção das estruturas lineares (chamadas de *isophotes* na literatura, no contexto de *inpainting*). Além disso, M_{BG} corresponde a um mapa binário de *background* gerado pelo extrator de *foreground-background* (descrito na Seção 3.3) para cada Ω individualmente. Antes de calcular as prioridades, faz-se necessário inicializar C com 0 em todos os pontos pertencentes a Ω e 1 nos demais. Para definir a prioridade, computa-se $P(\boldsymbol{p})$ em cada *patch* na borda $\delta\Omega$ e, por fim, seleciona-se para o preenchimento o ponto $\hat{\boldsymbol{p}}$ que maximiza P.

O termo de confiança $C(\boldsymbol{p})$ tem como objetivo mensurar a quantidade de informação confiável ao redor de um dado ponto $\boldsymbol{p}$. Neste termo, tem-se como intenção preencher primeiramente *patches* que têm mais informação confiável para comparativo durante a busca pelo melhor candidato ao preenchimento. Além disso, a atualização gradativa

dos valores deste termo faz com que os *holes* sejam preenchidos de maneira concêntrica, como detalhado mais a frente. Já o termo de dados $D(\boldsymbol{p})$, tem como objetivo medir a força dos *isophotes* que atingem a borda $\delta\Omega$ em cada iteração. Neste caso, busca-se dar maior prioridade ao *patch* que contém a borda que incide no *hole*, para estimular primeiramente a síntese de estruturas lineares que dão forma a região a ser reconstruída. Esta combinação permite que regiões totalmente desconhecidas possam ser reconstruídas da maneira mais adequada possível, enquadrando-se no caso das OOFAs.

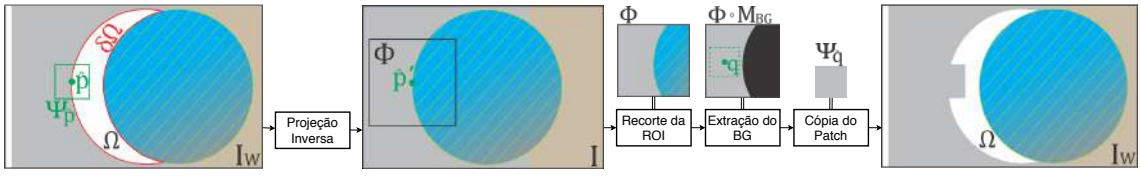
Como evidenciado, as *disocclusions* devem ser reconstruídas somente com conteúdo de *background*. Portanto, para uma reconstrução adequada dessas regiões, deve-se evitar que $\Psi_{\hat{p}}$ seja definido na borda de *foreground*. Por este motivo, utiliza-se tanto o termo de profundidades quanto o de *background*. Com a estimativa de $E(\boldsymbol{p})$, obtém-se o *patch* que contém o menor valor médio de disparidade e, consequentemente, maior de profundidade. Já $B(\boldsymbol{p})$, desempenha duplo papel, pois permite controlar tanto a confiança ao redor de um dado ponto $\boldsymbol{p}$, como a sua condição de pertencer ao *background*.

3.4.2 Busca pelo Melhor *Patch* para o Preenchimento

Geralmente, utiliza-se a imagem projetada I_w como fonte de cor/textura no processo de busca de informação para o preenchimento, como em (DARIBO; SAITO, 2011; OLIVEIRA et al., 2015). Entretanto, erros de projeção em I_w podem ser inseridos no interior das regiões reconstruídas. Por isso, nesta abordagem, realiza-se a busca na imagem de referência I, seguindo o esquemático ilustrado na Figura 3.9, proposto anteriormente em (OLIVEIRA, 2016).

O primeiro passo para a realização da busca, trata da delimitação da região de inte-

Figura 3.9: Processo de busca pelo melhor *patch*. Primeiro, reprojeta-se o centro $\hat{\boldsymbol{p}} \in \delta\Omega$ do *patch* a ser preenchido para seu ponto de origem $\hat{\boldsymbol{p}}'$ na imagem de referência I. Então, uma região de busca Φ centrada em $\hat{\boldsymbol{p}}'$ é definida e retirada de I. Se Ω corresponder a uma *disocclusion*, retira-se o *background* de Φ. Por fim, encontra-se o *patch* mais similar $\Psi_{\hat{q}}$, e sua informação é utilizada para preencher os *pixels* vazios de $\Psi_{\hat{p}}$ em I_w.



Fonte: Adaptado de (OLIVEIRA; WALTER; JUNG, 2018).

resse Φ em I. A busca não pode ser realizada na imagem completa, pois seria impossível determinar com precisão quais objetos pertencem ao *background* ou *foreground*, e informação incorreta poderia ser copiada. Como $\hat{\boldsymbol{p}}$ se refere a posição do centro do *patch* em I_w, faz-se necessário reprojetá-lo para seu ponto de origem $\hat{\boldsymbol{p}}'$ em I, antes da delimitação da região de busca. Nesta etapa, determina-se $\hat{\boldsymbol{p}}'$ por meio do *inverse warping* do valor de disparidade de $\hat{\boldsymbol{p}}$ em D_w. Sabendo que a região próxima a $\hat{\boldsymbol{p}}'$ deve conter conteúdo similar ao do *patch* a ser preenchido, com base no conceito de localidade espacial de textura (KAWAI; SATO; YOKOYA, 2009), define-se uma região de busca Φ com tamanho $N \times N$ centrada em $\hat{\boldsymbol{p}}'$. Neste trabalho, definiu-se $N = 69$ empiricamente, o qual resultou em uma região de busca de tamanho confiável quanto a classificação de *background* na análise de diferentes *datasets*. Se N for definido com um valor muito pequeno, o espaço de busca será muito limitado, o que pode levar a seleção de *patches* com pouca similaridade por falta de conteúdo para pesquisa. De outro modo, se o valor for muito grande, o processo de classificação do *background* pode falhar, devido a introdução de novos elementos em Φ.

Para prevenir que conteúdo do *foreground* seja copiado no interior das *disocclusions*, refina-se o espaço de busca com base na informação do mapa binário M_{BG} – gerado na estimativa de prioridades. O mapa contém a definição de quais *pixels* pertencem ao *background* na região vizinha a Ω, imprescindível para evitar que conteúdo inapropriado seja utilizado no preenchimento das *disocclusions*. Para garantir que pontos de *foreground* vizinhos sejam removidos do espaço de busca, primeiramente, aplica-se uma operação de erosão morfológica em M_{BG} com um elemento estruturante em forma de diamante, com tamanho 7 (definido experimentalmente). Então, calcula-se o Hadamard *product* (PETERSEN; PEDERSEN et al., 2008) entre M_{BG} e Φ, que consiste na multiplicação ponto a ponto entre duas matrizes, para delimitar a região de busca. Observa-se que OOFAs são compostas também por informação de *foreground* e, portanto, este processo não se aplica ao seu preenchimento. Finalmente, define-se qual *patch* $\Psi_{\hat{q}}$ tem maior similaridade com $\Psi_{\hat{p}}$ em Φ, da seguinte maneira:

$$\Psi_{\hat{q}} = \arg \min_{\Psi_q \in \Phi} s(\Psi_{\hat{p}}, \Psi_q), \tag{3.8}$$

$$s(\Psi_{\hat{p}}, \Psi_q) = \sum_{\boldsymbol{x} \in \Omega_v(\Psi_{\hat{p}})} \|\Psi_{\hat{p}}(\boldsymbol{x}) - \Psi_q(\boldsymbol{x})\|, \tag{3.9}$$

onde $\Omega_v(\Psi_{\hat{p}})$ denota os *pixels* dentro de $\Psi_{\hat{p}}$ com informação válida, e $\Psi_{\hat{p}}(\boldsymbol{x})$ se refere

ao vetor RGB relacionado a cor do *pixel* no ponto $\boldsymbol{x}$. Com a minimização da soma das diferenças absolutas no espaço de cores, garante-se que $\Psi_{\hat{q}}$ seja o *patch* mais similar a $\Psi_{\hat{p}}$ em Φ.

Para garantir que o mínimo de erro seja copiado a cada iteração do algoritmo, aplica-se o processo de busca com *patches* de tamanho adaptativo. A busca inicia com *patches* de 9×9. Então, se o valor mínimo obtido em $s(\Psi_{\hat{p}}, \Psi_q)$ exceder um limiar β, reduz-se $\Psi_{\hat{p}}$ e Ψ_q em dois *pixels* nas duas dimensões, e reinicia-se o processo de busca. No caso extremo, quando o tamanho 3×3 for alcançado, aceita-se o candidato mesmo se este exceder β. Objetiva-se com esta operação copiar o mínimo possível de informação quando não existe conteúdo confiável o suficiente. Se β for alto, um *patch* grande e com baixa similaridade pode ser copiado. No caso oposto, um valor baixo pode fazer com que o algoritmo propague sucessivamente *patches* pequenos. Nesta etapa, definiu-se $\beta = 35$ assim como em (OLIVEIRA, 2016).

No final de cada iteração, preenche-se a parte vazia do *patch* $\Psi_{\hat{p}}$ em Ω com o conteúdo de $\Psi_{\hat{q}}$ em I. Da mesma forma, copia-se o conteúdo do *patch* em D para D_w. Além disso, atualiza-se o mapa M_{BG}, atribuindo-se o valor 1 nos pontos preenchidos. Em C, atribui-se o valor de confiança calculado para $\hat{\boldsymbol{p}}$ em todos os pontos vazios de $\Psi_{\hat{p}}$. Repete-se todo o processo em cada região Ω individualmente, até que não haja pontos vazios na vista sintética.

3.5 Conclusões do Capítulo

Neste capítulo, apresentou-se uma abordagem completa para a geração de vistas sintéticas com o modelo DIBR. Inicialmente, foi descrito o método proposto para a detecção dos *cracks*, que possibilita a identificação tanto da forma translúcida como da vazia em uma única etapa. Em seguida, foi exibida uma avaliação, que permite concluir que o artefato pode compreender até 8,4% do conteúdo de uma vista sintética. Tal avaliação justifica o uso de técnicas especializadas para reconstrução destas regiões e, para tanto, foi proposto o uso do algoritmo HHF, que realiza um preenchimento local consistente. Para o tratamento dos *ghosts*, propôs-se uma análise de regiões candidatas (localizadas na borda de *background* das *disocclusions*), que visa reprojetar os pontos identificados para o local adequado, de acordo com um valor de disparidade estimado. Por fim, preenche-se as *disocclusions* e as OOFAs com uma adaptação do algoritmo de (CRIMINISI; PEREZ; TOYAMA, 2004), que se ajusta as especificidades de cada um destes *holes*, e utiliza

patches de tamanho adaptativo extraídos da imagem de referência. Observa-se que esta abordagem destina-se principalmente a casos estáticos, por não possuir mecanismos para controle e reconstrução baseada em informação temporal. Entretanto, isso não inviabiliza seu uso em vídeos.

4 SÍNTESE DE VISTAS COM DIBR BASEADA EM SUPERPIXELS HIERÁRQUICOS

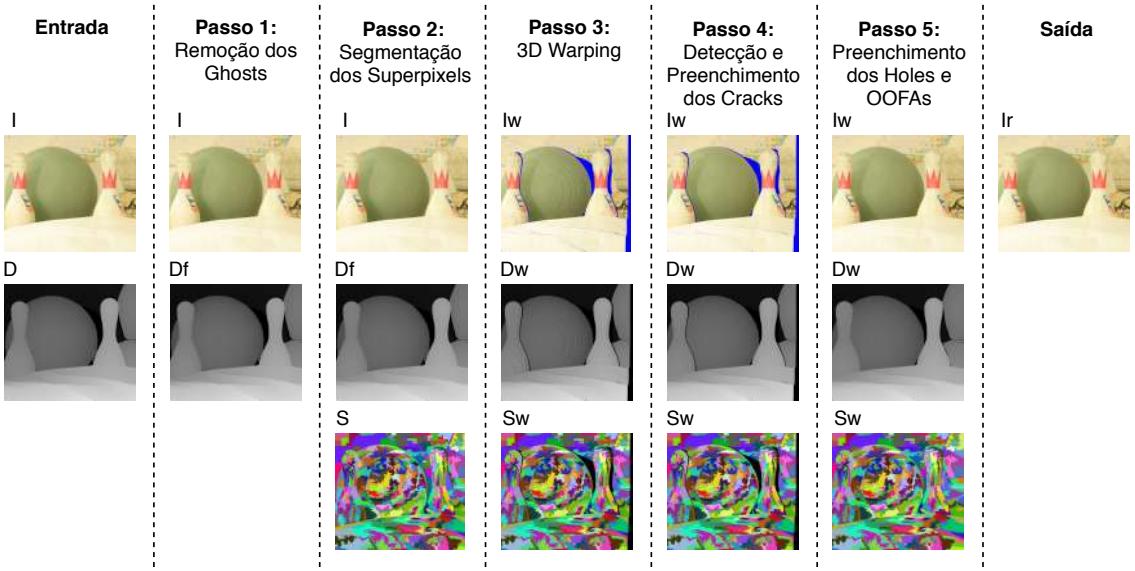
Neste capítulo, apresenta-se uma nova abordagem destinada a geração de vistas sintéticas com DIBR, que baseia-se na estrutura definida por *superpixels* hierárquicos para a reconstrução de *holes*, com foco tanto em fotografias como vídeos. Na próxima seção, detalha-se o *pipeline* proposto, definindo o papel de cada um dos passos necessários para a produção de imagens sintéticas. Após, na Seção 4.2, descreve-se um novo método para a remoção dos *ghosts*, que tem como finalidade suprimir o artefato antes mesmo da projeção da imagem de referência para o ponto de vista virtual. Na Seção 4.3, apresenta-se o algoritmo empregado na segmentação dos *superpixels* e, em seguida, na Seção 4.4, são descritos os processos de síntese de vistas e de detecção e preenchimento dos *cracks*. Por fim, na Seção 4.5, detalha-se um novo algoritmo de *inpainting* desenvolvido para o preenchimento dos *holes*, guiado pelas características individuais destas regiões sem informação, com base na interação dada pela estrutura formada pelos *superpixels* hierárquicos.

4.1 Visão Geral da Abordagem

Nesta abordagem, emprega-se o modelo de entrada V+D, assim como na ATA (descrita anteriormente no Capítulo 3). Entretanto, a ordem em que são solucionados os problemas e a metodologia adotada nesta abordagem são predominantemente diferentes. Na Figura 4.1, apresenta-se o passo a passo associado a esta abordagem, com a ilustração do resultado obtido em cada uma das etapas que compõem o *pipeline* proposto.

No primeiro passo, antes da projeção da imagem de referência I e do mapa de disparidades D para o ponto de vista virtual, realiza-se a remoção dos *ghosts*, evitando que o artefato seja formado. Nesta etapa, *pixels* de I que correspondem ao artefato têm seu valor de disparidade em D substituído por um valor associado ao objeto de *foreground* vizinho. Este processo repara incoerências produzidas na estimativa de D, produzindo um mapa D_f que não permite a formação de *ghosts*. Após, utilizam-se I e D_f para produzir uma segmentação de *superpixels* hierárquicos S, com *pixels* agrupados por critérios de cor e disparidade. Então, com base no mapa de disparidades reparado D_f, gera-se a imagem sintética I_w com 3D *image warping* e, juntamente, D_w e S_w. No passo seguinte,

Figura 4.1: Ilustração do *pipeline* desenvolvido para a geração de vistas sintéticas com a abordagem baseada em *superpixels* hierárquicos. Abaixo da descrição de cada passo, apresenta-se o resultado gerado pela solução adotada na imagem colorida, no mapa de disparidades e na segmentação dos *superpixels* (após o passo 2), respectivamente.



Fonte: O Autor, com imagens adaptadas do *dataset* Bowling1 de (HIRSCHMULLER; SCHARSTEIN, 2007).

para tratar os *cracks*, utiliza-se a solução proposta na abordagem ATA (na Seção 3.2), com a adição de uma filtragem para reconstrução das regiões afetadas pelo artefato em S_w. Por fim, aplica-se o algoritmo de *inpainting* desenvolvido, o qual segue as bases definidas em (CRIMINISI; PEREZ; TOYAMA, 2004), incorporando propriedades criadas para o processo de busca da abordagem ATA, e baseia-se em *superpixels* hierárquicos, que permitem maior controle do processo de reconstrução dos *holes*. Este algoritmo apresenta diversas possibilidades para extensão em relação a inclusão de informação temporal, que podem ser exploradas tanto na reconstrução dos *holes* quanto no controle de coerência na sucessão de quadros. A seguir, estão detalhados cada um destes passos.

4.2 Remoção de Ghosts por Refinamento do Mapa de Disparidades

Nesta abordagem, removem-se inicialmente os *ghosts*, para que estes não prejudiquem o algoritmo que segmenta os *superpixels*. O processo de segmentação depende diretamente da correspondência precisa entre disparidade e cor. Estes artefatos são inicialmente identificados e, após, tem seu valor de disparidade corrigido. O processo de identificação dos *ghosts* se divide em duas etapas: seleção e análise de *pixels* candidatos.

Este artefato ocorre em regiões onde existe uma mudança abrupta do valor de disparidade, que corresponde a transição entre dois objetos na cena. Desta forma, se este padrão for detectado em D, automaticamente são identificadas regiões que podem potencialmente conter *ghosts*. O Algoritmo 4 detalha a abordagem proposta para segmentação de regiões candidatas a *ghost*.

Com o objetivo de detectar estas variações características, aplica-se uma operação de dilatação morfológica em D, com um elemento estruturante em forma de diamante – devido a não uniformidade da borda dos objetos – com tamanho 1, produzindo $\tilde{D}$. Esta operação está definida na Linha 3 do algoritmo, e faz com que objetos no *foreground* sejam estendidos sobre seus respectivos vizinhos no *background*, proporcionalmente ao tamanho do elemento estruturante, por possuírem maior valor de disparidade. Neste caso, as áreas expandidas correspondem exatamente a *pixels* que podem pertencer ao *foreground*, mas receberam valor de disparidade de *background*, ou seja, *ghosts* em potencial. Se um *pixel* candidato não corresponder ao artefato, o processo de análise tratará de manter seu conteúdo inalterado.

Após a filtragem, quaisquer mudanças de disparidade em D são representadas por novos valores em $\tilde{D}$, mesmo que sejam relativas a geometria dos objetos. Entretanto, deseja-se selecionar apenas variações significativas de disparidade, que possam produzir *disocclusions* após o processo de projeção. Desta forma, para selecionar somente reais candidatos, avalia-se individualmente cada ponto $\boldsymbol{p}_c \in D$ quanto a sua variação de intensidade em $\tilde{D}$ para formular o conjunto de potenciais *ghosts* G_c. Como pode

Algoritmo 4: Pseudocódigo com a abordagem para a identificação de regiões onde podem ocorrer *ghosts*.

Entrada: D mapa de disparidades
Hd_1 elemento estruturante em forma de diamante
λ limiar que indica a variação máxima de disparidade
Saída: G_c imagem binária com regiões candidatas a *ghost*
$\tilde{D}$ imagem resultante da dilatação morfológica de D

```
1 início
2     Inicializa G_c com 0 em todos os pixels;
3     D̃ ← imagem resultante da dilatação morfológica de D com Hd_1;
4     para todo p_c ∈ D faça
5         se (D̃(p_c) − D(p_c)) ≥ λ então
6             G_c(p_c) ← 1;
7         fim
8     fim
9 fim
```

ser visto na Linha 5 do Algoritmo 4, o processo de avaliação consiste em verificar se $(\tilde{D}(\boldsymbol{p}_c) - D(\boldsymbol{p}_c)) \geq \lambda$. Neste caso, λ corresponde ao limiar definido na Seção 3.2, proposto originalmente para identificar mudanças que podem representar *cracks* translúcidos. Sendo assim, se esta verificação for válida, assinala-se $\boldsymbol{p}_c$ em G_c, como definido na Linha 6. Na Figura 4.2(b), destaca-se G_c em branco na imagem I, produzido com a técnica proposta a partir do mapa de disparidades exibido ao lado.

Após determinar o conjunto de pontos G_c, faz-se necessário analisar o conteúdo de cada *pixel* candidato em relação a vizinhança, e identificar quais equivalem realmente a uma ocorrência de *ghost*. Nesta etapa, utiliza-se um processo similar ao proposto na abordagem ATA (Seção 3.3), que baseia-se no comparativo de cor com as regiões de *foreground* e *background* vizinhas de cada ponto. Este processo está definido em forma de pseudocódigo no Algoritmo 5.

Primeiramente, como pode ser visto na Linha 2, inicializa-se o mapa de disparidades refinado D_f com os valores de D. Como definido na Linha 4, para determinar a consistência de um dado ponto candidato $\boldsymbol{p}_g \in G_c$, delimita-se um *patch* Ψ_{p_g} de 5×5, com tamanho definido experimentalmente, centrado em $\boldsymbol{p}_g$ na imagem I. Em Ψ_{p_g}, existem *pixels* de *background* e *foreground* e, portanto, faz-se necessário dividir o *patch* em duas partes. Para tanto, como definido na Linha 5 do algoritmo, aplica-se o extrator *foreground-background* (definido no Algoritmo 2 da Seção 3.3), computado nos valores de

Figura 4.2: Exemplo de seleção e verificação de *pixels* candidatos a *ghost*. Na imagem (a), apresenta-se um recorte do mapa de disparidade utilizado para a identificação de pontos candidatos a *ghost*. Ao lado, em (b), destaca-se na cor branca os pontos candidatos G_c na imagem I correspondente e, em (c), são exibidos apenas os pontos classificados como *ghost*.

(a) (b) (c)

Fonte: O Autor, com imagens adaptadas do *dataset* Baby1 de (HIRSCHMULLER; SCHARSTEIN, 2007).

disparidade do *patch*, sem considerar pontos candidatos. Este extrator permite dividir Ψ_{p_g} em dois conjuntos de *pixels*, FG (*foreground*) e BG (*background*). Após, calcula-se a média de intensidade de cor de FG e BG (para cada um dos canais RGB), obtendo μFG e μBG, respectivamente, nas Linhas 6 e 7. Então, para computar a similaridade entre cada *pixel* candidato $I(\boldsymbol{p}_g)$ e sua vizinhança em FG e BG, calculam-se as respectivas diferenças d_{FG} e d_{BG}, como definido nas Linhas 8 e 9 do algoritmo. Por fim, classifica-se o ponto candidato $\boldsymbol{p}_g$ com base no limiar de similaridade γ, descrito na Seção 3.3, seguindo a avaliação definida na Linha 10 do algoritmo.

À medida que pontos correspondentes a *ghost* vão sendo identificados, D_f vai sendo refinado. Neste caso, sempre que um candidato $\boldsymbol{p}_g$ for classificado como *ghost*, têm-se a confirmação de que o valor $D_f(\boldsymbol{p}_g)$ não corresponde ao objeto representado. Por isso, como pode ser visto na linha 11 do Algoritmo 5, atribui-se em $D_f(\boldsymbol{p}_g)$ o valor de $\tilde{D}(\boldsymbol{p}_g)$, que corresponde a disparidade do objeto no *foreground*.

Algoritmo 5: Pseudocódigo com a abordagem para a avaliação de candidatos a *ghost*.

Entrada: I imagem colorida
D mapa de disparidades
G_c imagem binária com regiões candidatas a *ghost*
$\tilde{D}$ imagem resultante da dilatação morfológica de D
γ limiar de similaridade
Saída: D_f mapa de disparidade refinado

1 **início**
2 $D_f \leftarrow D$;
3 **para todo** $\boldsymbol{p}_g \in G_c$ **faça**
4 $\Psi_{p_g} \leftarrow$ *patch* de 5×5 centrado em $\boldsymbol{p}_g$, na imagem I, desconsiderando *pixels* identificados em G_c;
5 $T_\Omega \leftarrow$ limiar obtido com o extrator *foreground-background* aplicado em $D(\Psi_{p_g})$;
6 $\mu_{FG} \leftarrow$ resultado da média dos *pixels* em $I(\Psi_{p_g})$ onde $D(\Psi_{p_g}) \geq T_\Omega$;
7 $\mu_{BG} \leftarrow$ resultado da média dos *pixels* em $I(\Psi_{p_g})$ onde $D(\Psi_{p_g}) < T_\Omega$;
8 $d_{FG} \leftarrow |\mu_{FG} - I(\boldsymbol{p}_g)|$;
9 $d_{BG} \leftarrow |\mu_{BG} - I(\boldsymbol{p}_g)|$;
10 **se** $d_{BG} \geq d_{FG} \vee d_{BG} > \gamma$ **então**
11 $D_f(\boldsymbol{p}_g) \leftarrow \tilde{D}(\boldsymbol{p}_g)$;
12 **fim**
13 **fim**
14 **fim**

O processo completo – seleção e análise de *pixels* candidatos – remove apenas ocorrências de *ghosts* com um *pixel* de largura. Por isso, faz-se necessário repetir o processo duas vezes, para que todas as ocorrências do artefato sejam identificadas, conside-

rando o padrão de até 2 *pixels* de largura, definido na literatura. Assim, após o término da primeira iteração, os pontos candidatos são novamente detectados, com base apenas em D_f, desconsiderando D. Nesta etapa, analisa-se o contorno dos objetos em duas etapas por existir uma dependência entre *pixels* vizinhos que correspondem ao artefato. Ou seja, se um *pixel* adjacente ao objeto no *foreground* não corresponder a um *ghost*, seu vizinho – ao lado – também não corresponderá.

Ao fim do processo, D_f corresponderá a um mapa de disparidade sem possíveis ocorrências de *ghost*. Como consequência, após a projeção, os *pixels* correspondentes ao artefato serão exibidos na extremidade de *foreground* da *disocclusion*, e não junto ao *background*, evitando a formação dos *ghosts*.

4.3 Segmentação de *Superpixels*

Após a projeção para o ponto de vista virtual, inúmeros *holes* são produzidos na imagem sintética, correspondendo a *cracks* vazios (ou translúcidos removidos), *disocclusions* e OOFAs. Esses, possuem características particulares que indicam qual tipo de conteúdo pode ser explorado para realizar sua reconstrução. Portanto, ter conhecimento sobre informações específicas e precisas sobre o entorno de cada uma das regiões a serem preenchidas se torna essencial para que seja produzida uma reconstrução adequada. Neste caso, em termos de informação, pode-se considerar desde cor até informação semântica dos objetos.

Na Figura 4.3(b), apresenta-se um exemplo de mapa semântico correspondente a

Figura 4.3: Exemplo da segmentação semântica de uma imagem. Em (a), apresenta-se a imagem colorida e em (b) a segmentação semântica correspondente, com cores indicando as classes dos objetos. Ao lado, em (c) e (d), exibe-se o resultado da projeção da imagem e do mapa semântico, respectivamente, com os *holes* destacados em branco.

(a) (b) (c) (d)

Fonte: O Autor, com imagens adaptadas do *dataset* Bowling1 de (HIRSCHMULLER; SCHARSTEIN, 2007).

fotografia exibida em (a), com cores indicando as classes dos objetos. Com a projeção destas imagens para um ponto de vista virtual, *holes* – destacados em branco – são gerados naturalmente em ambas, como pode ser visto nas Figuras 4.3(c) e (d). Em uma solução convencional, informação de cor e/ou disparidade seria utilizada para a reconstrução destas regiões na fotografia. Entretanto, neste caso, existe informação semântica – definindo a classe do objeto – associada às regiões vizinhas de cada *hole*. Desta forma, para tornar o processo de reconstrução mais robusto, seria plausível procurar informação para o preenchimento destas regiões somente em áreas que contém o mesmo rótulo semântico. Por exemplo, se o *hole* alvo está na bola, basta que seja realizada uma busca por informação para preenchimento somente onde existe este mesmo objeto, neste caso, nas regiões de cor amarela na Figura 4.3(b). Uma solução perfeita de segmentação e identificação de rótulos semânticos permitiria que esta busca fosse expandida para outras imagens. Desta forma, se o *hole* está na bola, quaisquer outras bolas identificadas em milhares de outras imagens poderiam ser consultadas e empregadas no preenchimento desta região.

No entanto, a segmentação e a classificação semântica ainda representam um desafio para a comunidade acadêmica, o que tem incentivado o desenvolvimento de inúmeras novas abordagens ao longo dos anos, como (LONG; SHELHAMER; DARRELL, 2015; CHEN et al., 2018; AHN; KWAK, 2018). Em alguns casos, este processo representa um desafio até mesmo para humanos, devido a oclusões na cena, ou até mesmo porque o objeto foge das suas características padrão. Mesmo assim, trabalhos recentes conseguem obter taxas de acerto satisfatórias nesta tarefa, no entanto, limitam-se a identificar uma quantidade pequena de classes de objetos. Neste caso, se um algoritmo de preenchimento fosse alimentado pelos resultados produzidos por alguma destas abordagens, este estaria limitado a reconstruir *holes* somente em alguns tipos específicos de objeto.

Diante desta realidade, os *superpixels* se apresentam como uma boa alternativa para segmentação e representação estrutural da imagem. *Superpixels* descrevem um grupo de *pixels* similares em cor ou outras propriedades de baixo nível (STUTZ, 2015), que revelam características locais que podem auxiliar na reconstrução dos *holes*. Diversos métodos foram produzidos para este fim nos últimos anos, como (ACHANTA et al., 2012; BERGH et al., 2012; ACHANTA; SüSSTRUNK, 2017; JAMPANI et al., 2018), com características distintas. Portanto, diferentes fatores devem ser analisados quando se deseja escolher um algoritmo para segmentação de *superpixels*. Como definido por Stutz, Hermans and Leibe (2018), alguns dos principais requisitos para a segmentação de *superpixels* são:

Partição – *superpixels* devem definir uma partição da imagem, pois devem ser disjuntos e atribuir um rótulo único para cada *pixel*.

Conectividade – os *superpixels* devem representar um conjunto de *pixels* conectados.

Aderência a borda – *superpixels* devem preservar bordas de imagens.

Compacidade – se refere a regularidade e suavidade, de modo que na ausência de bordas na imagem, os *superpixels* devem ser compactos, posicionados regularmente e exibir bordas suaves.

Eficiência – *superpixels* devem ser gerados eficientemente.

Número controlável – o número de *superpixels* deve ser controlável.

Alguns destes fatores estão relacionados com a aplicação alvo, como aderência a borda e eficiência, enquanto que outros são desejáveis em quaisquer casos, como partição e conectividade. No contexto da abordagem proposta, deseja-se ter uma segmentação de *superpixels* que represente a estrutura da imagem. Deste modo, o único fator listado que não impacta diretamente no resultado desejado se refere a compacidade, por não se fazer necessária.

Após testar algumas opções como (ACHANTA et al., 2012; ACHANTA; SüSSTRUNK, 2017), optou-se por utilizar o algoritmo Superpixels Hierarchy (SH), proposto por Wei et al. (2018). Este algoritmo produz segmentos hierárquicos, com aderência às bordas, partição e conectividade, mantendo consistentemente os relacionamentos de vizinhança em diversos níveis de granularidade. Com SH, pode-se gerar quaisquer níveis de granularidade de segmentação *on the fly* (demonstrando eficiência e controle na quantidade de *superpixels*). Além disso, eles têm a capacidade de representar claramente a estrutura dos objetos na cena, o que fica evidente no exemplo apresentado na Figura 4.4. A seguir, descreve-se o algoritmo detalhadamente, com uma adaptação desenvolvida, que incorpora na segmentação informação de disparidade para produzir *superpixels* coerentes com a aplicação alvo.

Figura 4.4: Exemplo de uso do algoritmo de segmentação *Superpixels Hierarchy*. Em (a), apresenta-se a imagem original e, em (b), são exibidas segmentações com 16, 256, 4096, e 65536 *superpixels*, respectivamente.

(a) (b)

Fonte: Retirado de (WEI et al., 2018).

4.3.1 Algoritmo *Superpixel Hierarchy* Adaptado

4.3.1.1 Algoritmo Superpixel Hierarchy Original

Para manter restrições de hierarquia, o algoritmo SH emprega um grafo não direcionado $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ como estrutura de dados base, composta por n vértices $v \in \mathcal{V}$ e m arestas $e \in \mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Visando preservar a conectividade intrínseca das imagens, os autores formularam $\mathcal{G}$ de modo que cada *pixel* seja representado por um vértice, conectado aos seus 4 vizinhos na horizontal e vertical (ou seja, 4-conectado). Neste caso, as arestas $e_{ij} = (v_i, v_j)$ representam as ligações com os *pixels* (vértices) vizinhos, as quais possuem um peso $w((v_i, v_j))$ associado, que corresponde a uma medida de dissimilaridade entre dois vértices v_i e v_j. Ao final da inicialização do grafo, tem-se uma estrutura que representa a imagem e a relação de conectividade dada por *pixels* vizinhos, que pode ser explorada no processo de segmentação.

Antes de inicializar o processo de segmentação, faz-se necessário definir a quantidade k de *superpixels* hierárquicos que devem ser extraídos da imagem ou, mais precisamente, segmentados no grafo. Com base nesta definição, pode-se dizer que uma segmentação $\mathcal{S}$ de $\mathcal{G}$ consiste de uma partição de $\mathcal{V}$ em k componentes disjuntos. Deste modo, cada componente $\mathcal{C} \in \mathcal{S}$ corresponde a um subgrafo conectado $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$, onde $\mathcal{V}' \subseteq \mathcal{V}$ e $\mathcal{E}' \subseteq \mathcal{E}$.

No algoritmo SH, realiza-se a segmentação com o método de Borůvka (WEST, 1996), em uma abordagem que se baseia no crescimento de regiões. O algoritmo de Borůvka é uma abordagem gulosa, que originalmente tem como finalidade computar uma árvore geradora de custo mínimo (ou *minimum spanning tree* – MST). Na modelagem

adotada no SH, interpreta-se o grafo $\mathcal{G}$ como um floresta composta por n árvores (neste caso, cada vértice/*pixel* corresponde a uma árvore). Para criar os *superpixels*, utiliza-se o processo de crescimento de regiões no qual cada árvore é combinada com o seu vizinho mais próximo, indicado pela aresta de menor valor. Neste processo, se o componente (ou árvore) $\mathcal{C}_2$ for o vizinho mais próximo de $\mathcal{C}_1$, não necessariamente $\mathcal{C}_1$ também será o vizinho mais próximo de $\mathcal{C}_2$. A distância entre duas árvores $\mathcal{C}_1$ e $\mathcal{C}_2$, conectadas por uma ou mais aresta (v_i, v_j), é dada por:

$$F(\mathcal{C}_1, \mathcal{C}_2) = \min_{v_i \in \mathcal{C}_1, v_j \in \mathcal{C}_2, (v_i, v_j) \in \mathcal{E}} w((v_i, v_j)). \tag{4.1}$$

Após realizar o processo de busca pelo vizinho mais próximo para cada árvore, formula-se um grafo auxiliar, onde cada vértice representa um *cluster* e cada aresta corresponde a aresta de menor peso escolhida. Caso existam vizinhos mais próximos mútuos, indica-se a relação entre os vértices por arestas duplicadas. Nos outros casos as arestas são distintivas. Para encontrar os componentes conectados, emprega-se a busca em profundidade. O algoritmo de Borůvka repete este processo de combinação das árvores até que reste apenas uma árvore. Na adaptação do algoritmo proposta para o SH foi adicionada uma funcionalidade, que tem como objetivo armazenar a ordem na qual as arestas são adicionadas na MST. No momento em que uma aresta é adicionada na MST, observa-se que o número de árvores na floresta é reduzido em razão de um. Desta forma, para produzir a segmentação de k *superpixels*, basta conectar os vértices das primeiras $n - k$ arestas do grafo (onde n corresponde a quantidade total de vértices), já memorizadas na estimativa da MST. O resultado final consiste de k componentes conectados, cada um com apenas um rótulo numérico sequencial, correspondente ao identificador do *superpixel*.

Durante o processo de segmentação, cada árvore é combinada diversas vezes com seu vizinho mais próximo, com base em uma estimativa de distância. Por se tratar da segmentação de uma imagem, diferentes critérios de similaridade podem ser empregados nesta estimativa. Considerando uma abordagem simples, a diferença absoluta em um dado espaço de cores poderia determinar se dois componentes $\mathcal{C}_1$ e $\mathcal{C}_2$ seriam ou não combinados.

A função que estima a similaridade entre *pixels* e, posteriormente, entre *clusters*, tem grande impacto sobre o resultado produzido por este algoritmo. Desta forma, para melhorar a precisão da segmentação, além de atributos da imagem, a abordagem permite integrar informação de confiança estimada por um detector de bordas na função de custo. Esta informação tem por finalidade não permitir que segmentos possam extrapolar bordas

de objetos da imagem, adicionando como custo a confiança dada pelo mapa de bordas. Deste modo, calcula-se a distância entre $\mathcal{C}_1$ e $\mathcal{C}_2$ da seguinte forma:

$$F(\mathcal{C}_1, \mathcal{C}_2) = d_c \cdot d_e, \tag{4.2}$$

onde d_c e d_e correspondem à distância de cor e de borda, respectivamente. Para estimar a distância de cor, mensura-se a diferença absoluta entre as médias de $\mathcal{C}_1$ e $\mathcal{C}_2$ no espaço de cores CIELab. Após a quarta iteração, para se adaptar ao crescimento dos *superpixels*, d_c passa a ser estimada pela distância χ^2 dos histogramas de cor (divididos igualmente em 20 classes). Para estimar o valor de d_e, calcula-se a média de confiança dada entre duas regiões.

4.3.1.2 Adaptação Proposta

A abordagem original de Wei et al. (2018) emprega a confiança de bordas dada por algoritmos como (HARIHARAN et al., 2011; DOLLAR; ZITNICK, 2013) no cômputo de d_e. Entretanto, a estimativa de confiança requer processamento adicional, e falsas detecções de borda podem levar a *oversegmentation* dos objetos, devido a textura e/ou ruído. Por ambos esses motivos, e também para utilizar uma estimativa de borda mais robusta, propomos agregar informação de disparidade na estimativa de d_e. Deste modo, ao invés de computar o mapa de bordas e utilizar o valor de confiança, que determina a probabilidade de existir uma borda, propomos empregar a diferença direta das disparidades. Considerando a modificação proposta, a equação de distância passa a ser computada da seguinte forma:

$$F(\mathcal{C}_1, \mathcal{C}_2) = d_c \cdot d_d, \tag{4.3}$$

$$d_d(\mathcal{C}_1, \mathcal{C}_2) = |D_f(\mathcal{C}_1) - D_f(\mathcal{C}_2)|, \tag{4.4}$$

onde D_f se refere ao mapa de disparidades com os *ghosts* removidos. Neste caso, quanto menor for a diferença de cor e disparidade, maior a probabilidade de $\mathcal{C}_1$ e $\mathcal{C}_2$ se unirem.

Antes de produzir a imagem sintética, faz-se necessário computar o mapa de *superpixels* correspondente S, estimado no conteúdo de I e D_f. Considerando as especificidades envolvidas no algoritmo de preenchimento dos *holes* desenvolvido (descrito mais a frente), optou-se por gerar um mapa com $k = 350$. Este valor produziu uma boa

representatividade da estrutura da imagem em diferentes casos de teste. Se o valor for muito alto, a imagem ficará muito subdividida e não irá representar a estrutura da imagem adequadamente. No caso oposto, com k muito baixo, o algoritmo acaba sendo forçado a misturar regiões com pouca similaridade.

Na Figura 4.5 são apresentados alguns exemplos de segmentação produzidos com a adaptação do algoritmo SH proposta. Nas imagens, pode-se observar que a granularidade definida permite representar grandes regiões homogêneas com um mesmo rótulo. Da mesma forma, regiões menores, com textura e cor específicas, recebem rótulos específicos. Isto faz com que a estrutura da imagem, interpretada pela informação de cor e disparidade, sejam segmentadas adequadamente.

Figura 4.5: Resultado produzido pela adaptação proposta do algoritmo *Superpixels Hierarchy*. Em (a) são exibidas as imagens e em (b) o mapa de *superpixels* correspondente, com $k = 350$.

(a) (b)

Fonte: O Autor, com imagens adaptadas de (HIRSCHMULLER; SCHARSTEIN, 2007; SCHWARZ; MARPE; WIEGAND, 2010).

4.4 Geração da Imagem Sintética e Tratamento dos *Cracks*

Após tratar os *ghosts* e segmentar os *superpixels*, inicia-se o processo de formação da imagem sintética. Deste modo, com base em D_f projeta-se I, D_f e S para o ponto de vista virtual, formando I_w, D_w e S_w, respectivamente. Com este processo, *cracks* vazios e translúcidos são produzidos nas imagens projetadas, os quais precisam ser tratados.

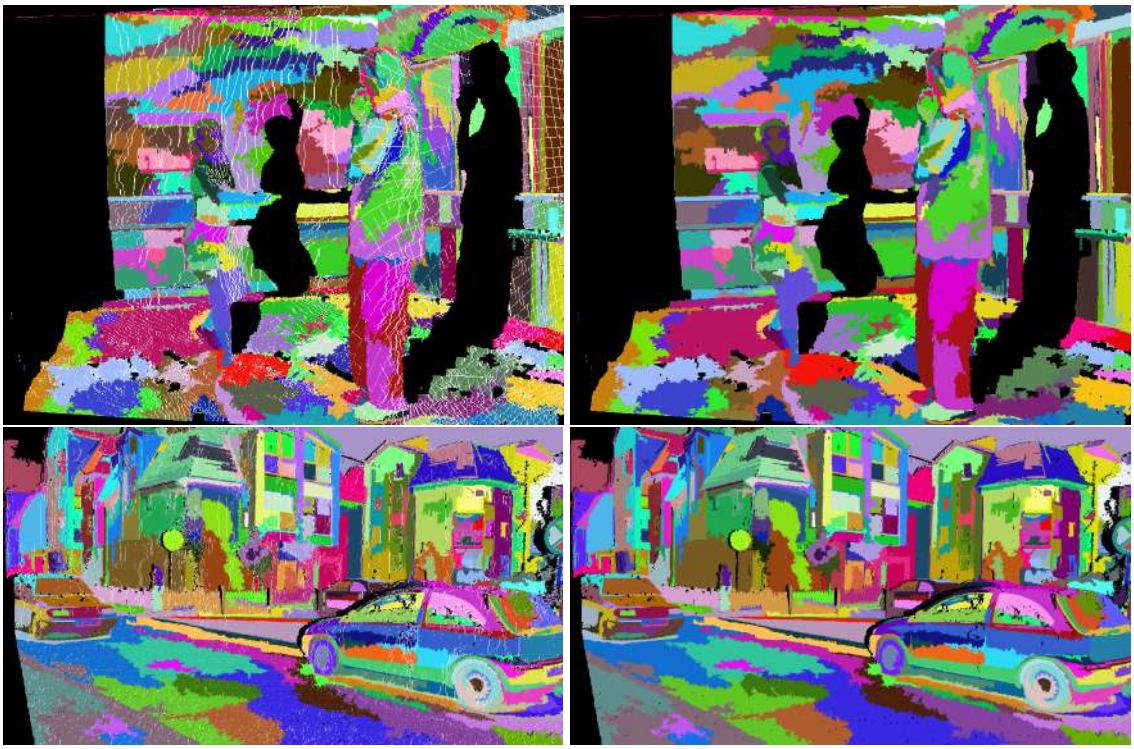
Neste caso, o artefato apresenta o mesmo padrão tratado pela técnica desenvolvida para a abordagem ATA (descrita na Seção 3.2). Portanto, não se faz necessário desenvolver uma nova técnica relativa a esta abordagem, uma vez que a solução proposta anteriormente permite detectar e preencher os *cracks* de ambas as formas em I_w e D_w. Contudo, nesta abordagem, faz-se necessário reconstruir o mapa de *superpixels* projetado.

Para preencher em S_w as regiões correspondentes aos *cracks*, propõe-se usar um processo de filtragem. Como as regiões a serem reconstruídas são pequenas, geralmente com o mesmo rótulo atribuído em ambas as extremidades da área formada pelo artefato, que corresponde praticamente ao mesmo caso encontrado no mapa de disparidades. Por isso, optou-se por utilizar uma abordagem similar à empregada em D_w. Para identificar os rótulos correspondentes a essas regiões, aplica-se uma operação de fechamento morfológico com um elemento estruturante linear $Hl_3 = [1\ 1\ 1]^\top$ em S_w, produzindo $\hat{S}_w$. Do mesmo modo, se houver projeção vertical, repete-se a filtragem com ${Hl_3}^T$ em $\hat{S}_w$. O elemento estruturante Hl_3 foi definido com este tamanho para preencher *cracks* com até 2 *pixels* de largura, seguindo as proporções estipuladas na abordagem ATA na Seção 3.2. Ao final da filtragem, os rótulos correspondentes ao artefato em S_w apresentam o valor correspondente em $\hat{S}_w$. Portanto, basta copiar o conteúdo das regiões identificadas para reconstruir parte do mapa de *superpixels*. Na Figura 4.6, apresentam-se exemplos do resultado de preenchimento dos *cracks* em S_w, produzidos pela técnica proposta.

4.5 Preenchimento dos *Holes*

Para preencher adequadamente os *holes*, faz-se necessário ter conhecimento sobre seu entorno e qual tipo de conteúdo deve ser empregado na sua reconstrução. Com base nisso, foi desenvolvida uma nova metodologia para classificação dos *holes*, detalhada na próxima subseção. O resultado da classificação é utilizado como guia pelo algoritmo de *inpainting* proposto (descrito na Subseção 4.5.2), para fazer com que cada *hole* seja preenchido com o processo e conteúdo adequados.

Figura 4.6: Resultado produzido pela abordagem proposta para preenchimento dos *cracks* em S_w. Em (a), encontram-se destacados em branco os *cracks* vazios e translúcidos e, em (b), o resultado após o preenchimento.



(a) (b)

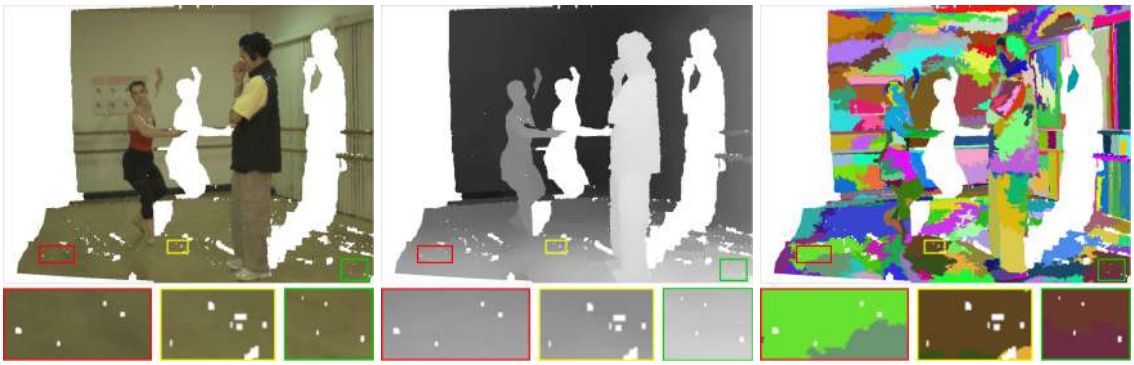
Fonte: O Autor, com imagens adaptadas de (ZITNICK et al., 2004; SCHWARZ; MARPE; WIEGAND, 2010).

4.5.1 Classificação dos *Holes*

Após o preenchimento dos *cracks*, restam predominantemente *holes* Ω de dois tipos: *disocclusions* e OOFAs. Essencialmente, as *disocclusions* são caracterizadas por apresentar duas ou mais camadas de profundidade na borda $\delta\Omega$, correspondentes aos objetos que fazem fronteira na imagem de referência. Já as OOFAs não possuem um padrão de composição de profundidade, mas sim de localização, e caracterizam-se por apresentar informação em apenas uma das extremidades de $\delta\Omega$.

Apesar de não existirem definições na literatura, uma simples inspeção visual permite observar um outro tipo de *hole* nas vistas sintéticas, que não pode ser caracterizado como OOFA ou *disocclusion*. Este apresenta um padrão de cor/textura homogêneo em $\delta\Omega$, com variações suaves de profundidade. Na Figura 4.7(a), são destacadas ocorrências deste tipo de *hole*, e as disparidades em escala de cinza (com intensidade inversamente proporcional à profundidade) correspondentes podem ser observadas ao lado em (b). Este

Figura 4.7: Imagens (a) I_w, (b) D_w e (c) S_w com os *cracks* preenchidos. São exibidos em detalhe abaixo das imagens alguns *holes* da classe FG, identificados por retângulos nas cores vermelho, amarelo e verde.



(a) (b) (c)

Fonte: O Autor, com imagens adaptadas de (ZITNICK et al., 2004).

tipo de *hole* ocorre no interior de elementos da cena (como objetos, paredes, etc.), devido a variações abruptas de profundidade em regiões específicas ou até mesmo por erros no processo de estimativa. As abordagens apresentadas na literatura, assim como ATA, costumam considerar estas áreas como *disocclusions*. Isto pode levar a inconsistências no preenchimento ou, no mínimo, a uma reconstrução menos precisa dessas regiões, devido ao não uso de parte do conteúdo equivocadamente considerado como *foreground*.

Com base nesta análise, foram criadas três classes genéricas para representação dos *holes*, as quais estão detalhadas a seguir:

***foreground* (FG) –** encontrados em regiões com cor aproximadamente homogênea, com transições de profundidade suaves em $\delta\Omega$;

***background* (BG) –** localizados nas bordas de I_w, com conteúdo somente em uma das extremidades. Correspondem quase em sua totalidade às OOFAs;

***foreground-background* (FB) –** possuem duas ou mais camadas de profundidade bem definidas em $\delta\Omega$. Equivalem às *disocclusions*.

Estas classes permitem classificar quaisquer tipos de *hole*, independentemente das inconsistências ou erros contidos nos mapas de profundidade (ou disparidade).

4.5.2 Algoritmo de Preenchimento

Algoritmos baseados em exemplares são explorados há bastante tempo por trabalhos descritos na literatura, com o objetivo de replicar textura e/ou reconstruir áreas sem informação em imagens (EFROS; LEUNG, 1999; ASHIKHMIN, 2001; CRIMINISI; PEREZ; TOYAMA, 2004; TONIETTO; WALTER; JUNG, 2005). Com base no processo de segmentação em *superpixels* hierárquicos, idealiza-se que este tipo de estrutura possa ser empregada como exemplar base para a reconstrução dos *holes*, assim como foi feito em (SCHMEING; JIANG, 2015). No entanto, as regiões a serem reconstruídas podem ser compostas por múltiplas texturas, separadas por "linhas estruturais" (bordas). Os *superpixels* representam em sua essência uma região com cor homogênea, disposta no interior de um elemento da cena (por não extrapolar bordas e planos de profundidade) e, portanto, não são capazes de representar a estrutura da imagem, demonstrando-se inadequados para desempenhar este papel.

Criminisi, Perez and Toyama (2004) estabeleceram um processo para reconstrução de textura, baseado em um esquemático definido pela ordem de preenchimento de regiões vazias e cópia de *patches* (baseado na similaridade de cor). Nos resultados apresentados por seu estudo, fica evidente a importância da prioridade de preenchimento dos *isophotes* e de regiões com informação confiável, para a propagação adequada de textura.

Contudo, o algoritmo de *inpainting* de (CRIMINISI; PEREZ; TOYAMA, 2004) apresenta fragilidades, como no caso do termo de dados empregado no cálculo de prioridades, estimado sobre o operador de gradiente, vulnerável a ruído (AHN; KIM, 2013). Além disso, como discutido anteriormente, os *holes* gerados com DIBR possuem características específicas e não podem ser reconstruídos de maneira genérica. Um bom exemplo disso são as *disocclusions*, que devem ser preenchidas somente com informação de *background*.

Com base nas premissas estabelecidas pelo algoritmo de (CRIMINISI; PEREZ; TOYAMA, 2004) e nos avanços produzidos por ATA, desenvolveu-se uma nova abordagem de preenchimento baseada em *patches* (capazes de representar ao mesmo tempo diferentes elementos, com textura/cor distintas). Esta, diferentemente das demais propostas na literatura, guiadas principalmente por informação do mapa de profundidades (GAUTIER; MEUR; GUILLEMOT, 2011; DARIBO; SAITO, 2011; LUO et al., 2016), baseia-se na informação estrutural da imagem, representada pela segmentação de *superpixels* hierárquicos. A interação entre os *superpixels* revela informações importantes, que

podem ser exploradas no processo de preenchimento. Por exemplo, se dois *superpixels* vizinhos fazem fronteira, isso é um indicativo de transição de textura ou mudança de plano de profundidade, ou até mesmo as duas coisas. Neste caso, uma região onde dois segmentos se intersectam representa um *isophote*, e esta informação pode ser utilizada para determinar a ordem de preenchimento.

Na abordagem de preenchimento proposta, preenche-se cada *hole* Ω individualmente seguindo a ordem dada por uma função de estimativa de prioridades. Para isto, utiliza-se uma equação de prioridades adequada a cada uma das classes de *hole*, estimada a partir de informação estrutural dos *superpixels* hierárquicos e de consistência da informação local, com base na análise de cada *patch* Ψ_p centrado em cada ponto $\boldsymbol{p}$ da borda $\delta\Omega$ de Ω. Após identificar o *patch* $\Psi_{\hat{p}}$ com maior prioridade, faz-se uma busca pelo candidato $\Psi_{\hat{q}}$ que melhor se adapta ao preenchimento de seus pontos vazios em uma região delimitada $\Phi \in I$. Para estimar a similaridade, emprega-se uma função que se adapta a cada classe de *hole*, considerando como critérios a semelhança de cor, rótulos e também a distribuição de profundidade em pontos vazios. O processo é repetido até que Ω esteja inteiramente preenchido. As próximas subseções detalham o algoritmo proposto.

4.5.2.1 Estimativa de Prioridade dos Candidatos

Considerando as especificidades de cada *hole*, que requerem processos de reconstrução próprios, foram desenvolvidas três equações para estimativa de ordem de preenchimento, baseadas nas caraterísticas de cada classe. Para identificar a classe de um dado Ω, utiliza-se apenas informação de $\delta\Omega$ em S_w e a localização do *hole*. Neste caso, classifica-se o *hole* como: FG, se $S_w(\delta\Omega)$ possuir apenas um rótulo, o que indica uma região com homogeneidade de cor e profundidade; BG, se estiver em contato com as bordas de I_w; FB, os demais *holes*, os quais possuem 2 ou mais rótulos em $S_w(\delta\Omega)$ e não estão nas extremidades de I_w.

Com base nas especificidades de cada classe de *hole*, definiu-se uma função de prioridade $P(\boldsymbol{p})$, que deve ser computada em cada *patch* Ψ_p centrado em cada ponto $\boldsymbol{p} \in \delta\Omega$ da seguinte maneira:

$$P(\boldsymbol{p}) = \begin{cases} C(\boldsymbol{p}) & \text{se } \Omega \in FG \\ C(\boldsymbol{p}) \cdot H(\boldsymbol{p}) \cdot R(\boldsymbol{p}) & \text{se } \Omega \in BG \\ B(\boldsymbol{p}) \cdot H(\boldsymbol{p}) \cdot R(\boldsymbol{p}) & \text{se } \Omega \in FB \end{cases}, \tag{4.5}$$

onde $C(\boldsymbol{p})$ se refere ao termo de confiança e $B(\boldsymbol{p})$ ao de *background*, definidos anteriormente na Seção 3.4.1. $H(\boldsymbol{p})$ e $R(\boldsymbol{p})$ se referem aos termos de heterogeneidade e regularidade, respectivamente, que são computados como segue:

$$H(\boldsymbol{p}) = \frac{\#u(S_w(\Psi_p))}{\#S}, \tag{4.6}$$

$$R(\boldsymbol{p}) = 1 - \frac{\sigma^2(h(S_w(\Psi_p)))}{\sigma^2_{MAX}} \tag{4.7}$$

onde u corresponde a uma função que retorna um vetor com os rótulos únicos de um conjunto de dados, considerando que estes são representados por valores inteiros em S_w, e o operador $\#$ indica a quantidade de elementos em um dado conjunto. Na Equação 4.7, σ^2 refere-se a variância, h corresponde a uma função que retorna um vetor com o somatório de elementos de cada rótulo dos pontos não vazios de S_w e σ^2_{MAX} indica a variância máxima que pode ocorrer no vetor retornado por h. Como a maior variância possível em um *patch* de 9×9 é dada entre os valores 1 e 79, desconsiderando $\boldsymbol{p}$ que é vazio, definiu-se $\sigma^2_{MAX} = 1521$. Após computar $P(\boldsymbol{p})$, seleciona-se o ponto $\hat{\boldsymbol{p}}$ que maximiza a função de prioridades, para preencher o *patch* correspondente em cada iteração.

O termo $H(\boldsymbol{p})$ têm uma funcionalidade similar a do termo de dados $D(\boldsymbol{p})$ (proposto por (CRIMINISI; PEREZ; TOYAMA, 2004)). A necessidade de desenvolvimento de $H(\boldsymbol{p})$ se fundamenta na vulnerabilidade a ruído apresentada pelo termo de dados (AHN; KIM, 2013), que pode levar a estimativas incorretas de prioridade. Quando subdivididos em uma quantidade adequada de segmentos, os *superpixels* hierárquicos representam idealmente a estrutura da imagem, por naturalmente não extrapolarem bordas e, no caso da adaptação proposta, por não misturarem conteúdo de objetos distintos. Isso ocorre devido a restrições que penalizam mudanças de profundidade e buscam formular segmentos com cor aproximadamente homogênea. Desta forma, se dois *superpixels* fazem fronteira em um dado *patch* Ψ_p, ou estes estão em planos de profundidade distintos, ou existe uma mudança abrupta de profundidade, ou ainda, ambos os casos. Portanto, esta região representa parte da estrutura da imagem, correspondente a intersecção de dois objetos ou a mudança de cor/textura, e deve ser preenchida anteriormente. Sendo assim, quanto mais rótulos diferentes forem encontrados em um *patch* Ψ_p, mais estruturas estão representadas nesta região e maior é a sua prioridade de preenchimento.

Como complemento a $H(\boldsymbol{p})$, foi desenvolvido o termo $R(\boldsymbol{p})$, que tem como objetivo selecionar o ponto candidato $\boldsymbol{p}$ que contém a maior regularidade na distribuição dos

rótulos em Ψ_p. Para tal, mede-se a frequência de ocorrência de cada rótulo no *patch*, com base no cálculo da variância, para determinar o quão próximo da média está a ocorrência de cada um dos rótulos em Ψ_p. Se a variância for baixa, existe uma regularidade na ocorrência dos rótulos e o valor de prioridade vai se aproximar de um. Diferentemente, se for apresentado um valor muito alto, mais próximo de zero estará a sua prioridade.

Com a introdução destes novos termos, as prioridades passam a ter um papel definido de acordo com a classe de cada *hole* e não mais somente ao tipo (*disocclusions* ou OOFAs). Desta forma, considerando as equações de prioridades, pode-se dizer que: se $\Omega \in BG$, faz-se necessário priorizar o ponto $\boldsymbol{p}$ que possui simultaneamente a maior concentração de pontos válidos (cofiabilidade) e que contém mais elementos estruturais (com base em $H(\boldsymbol{p})$ e $R(\boldsymbol{p})$); por fim, caso $\Omega \in BG$, além de dar prioridade para o ponto centrado em um *patch* com maior concentração de informação estrutural, faz-se necessário iniciar o preenchimento a partir da região que apresenta a maior acumulo de pontos no *background*, estimado com $B(\boldsymbol{p})$.

4.5.2.2 Busca pelo Melhor Patch para o Preenchimento

O processo de busca proposto deriva de diversas evoluções produzidas pela abordagem ATA (descritas na Subseção 3.4.2). Assim como em ATA, utiliza-se como fonte de cor/textura para o preenchimento dos *pixels* vazios de cada *patch* alvo $\Psi_{\hat{p}}$ uma região delimitada $\Phi \in I$, com tamanho $N \times N$, centrada no ponto de origem $\hat{\boldsymbol{p}}'$ (definido pela reprojeção de $\hat{\boldsymbol{p}}$). O uso de I permite que apenas informação confiável seja utilizada no processo de preenchimento, e a delimitação do espaço de busca permite explorar o conceito de localidade espacial de textura (KAWAI; SATO; YOKOYA, 2009).

Para evitar que conteúdo de *foreground* seja copiado nas *disocclusions*, ATA considera apenas uma parte segmentada como *background* de Φ para a busca por *patches* correspondentes. Contudo, este processo não se faz necessário, pois a própria estimativa de similaridade pode desempenhar este papel com maior eficiência, sem eliminar informação que poderia ser útil para o preenchimento. Para tanto, desenvolveu-se uma função de cálculo de similaridades, que não considera apenas similaridade por cor, que permitiria possivelmente a cópia de conteúdo inadequado, mas também a correspondência em termos de rótulos de identificação individual dos *superpixel*. Esta tem como objetivo adicional propagar a distribuição estrutural do *patch* a ser preenchido.

Deste modo, considerando a região de interesse $\Phi \in I$, realiza-se a busca pelo *patch* $\Psi_{\hat{q}}$ que mais se assemelha a $\Psi_{\hat{p}}$, considerando a classe de cada *hole*, da seguinte

maneira:

$$\Psi_{\hat{q}} = \arg\min_{\Psi_q \in \Phi} \begin{cases} s(\Psi_{\hat{p}}, \Psi_q) \cdot l(\Psi_{\hat{p}}, \Psi_q) & \text{se } \Omega \in FG \vee \Omega \in FB \\ s(\Psi_{\hat{p}}, \Psi_q) \cdot l(\Psi_{\hat{p}}, \Psi_q) + z(\Psi_q) & \text{se } \Omega \in BG \end{cases}, \quad (4.8)$$

$$l(\Psi_{\hat{p}}, \Psi_q) = 1 - \exp\left(-\frac{n(\Psi_{\hat{p}}, \Psi_q)}{\tau}\right),$$

$$z(\Psi_q) = \sum_{d \in D_w(\Psi_q \cup \Omega)} |d - T_\Omega|,$$

onde $s(\Psi_{\hat{p}}, \Psi_q)$ corresponde ao cálculo de similaridade no espaço de cores RGB (detalhada na Equação 3.8), $l(\Psi_{\hat{p}}, \Psi_q)$ a correspondência de rótulos e $z(\Psi_q)$ a diferença de disparidade nos pontos sem correspondência em $\Psi_{\hat{p}}$ (para identificar discrepâncias de profundidade). Nas equações, mais especificamente, $n(\Psi_{\hat{p}}, \Psi_q)$ corresponde a uma função que retorna o número de pontos com rótulos coincidentes nos dois *patches*, τ controla o decaimento da função exponencial (definido experimentalmente como 10) e T_Ω corresponde a média truncada dos valores de disparidade de $\delta\Omega$, calculado com $\alpha = 10\%$.

A estimativa de correspondência $l(\Psi_{\hat{p}}, \Psi_q)$ tem como objetivo computar quantos pontos possuem rótulos coincidentes nos dois *patches*. Neste caso, além de complementar a estimativa de similaridade por cor determinada por $s(\Psi_{\hat{p}}, \Psi_q)$, o cômputo da correspondência também tem como objetivo dar continuidade a estrutura da imagem. Mais especificamente, se os rótulos nos pontos não vazios coincidirem, a tendência é que os demais pontos sigam a mesma distribuição.

Os *holes* da classe BG possuem conteúdo em apenas uma de suas extremidades. Por isso, tende-se a ter *patches* com pouco conteúdo para a busca por similaridade. Desta forma, para complementar a estimativa de similaridade e evitar que conteúdo com profundidade muito discrepante seja copiado no interior de Ω, estima-se a distância entre a média truncada de $\delta\Omega \in D_w$ e de cada ponto sem correspondência em $\Psi_q \in D \cap \Phi$.

Do mesmo modo que em ATA, utiliza-se um processo de busca com *patches* de tamanho adaptativo em Φ. Ao fim de cada iteração, a região vazia de $\Psi_{\hat{p}}$ é preenchida com o conteúdo de $\Psi_{\hat{q}}$, copiado de I para I_w. O mesmo ocorre com D_w e S_w, que são preenchidos com informação de D e S, respectivamente. Os termos de prioridade C e B são atualizados conforme a definição dada na Subseção 3.4.2. Já os pontos a

serem preenchidos em $\Psi_{\hat{p}}$ nos termos H e R recebem valor relativo ao ponto $\hat{\boldsymbol{p}}$. Esse processo é repetido em cada região Ω individualmente, até que todos os pontos vazios sejam preenchidos.

4.6 Conclusões do Capítulo

Neste capítulo, foi descrita uma nova abordagem para síntese de vistas com o modelo DIBR. Para remover os *ghosts*, realiza-se a filtragem ponto a ponto de regiões que podem potencialmente corresponder ao artefato no mapa de disparidades original, de modo a evitar que o mesmo possa ser formado na vista sintética. Após, realiza-se a segmentação dos *superpixels* hierárquicos, com base na adaptação proposta do algoritmo de (WEI et al., 2018). Em seguida, com base no mapa de disparidades refinado, projeta-se a imagem de referência e os respectivos mapas (D_f e S) para o ponto de vista virtual. Então, ocorrências de *cracks* vazios e translúcidos são detectadas e preenchidas com a solução proposta na abordagem ATA em I_w e D_w e, adicionalmente, reconstruídas no mapa de *superpixels* projetado utilizando uma técnica baseada em morfologia matemática. Por fim, os *holes* são classificados de acordo com uma nova metodologia desenvolvida e, com base nisso, aplica-se um algoritmo de preenchimento baseado em *patches* que incorpora o esquemático proposto por (CRIMINISI; PEREZ; TOYAMA, 2004) e funcionalidades derivadas da abordagem ATA. Este algoritmo baseia-se na interação entre *superpixels* hierárquicos, tanto para o cálculo de prioridades como para a estimativa de similaridade adotada pelo processo de correspondências dos *patches*.

5 MODELO DE *BACKGROUND* INCREMENTAL

Neste capítulo, apresenta-se um novo método para a geração de modelos de *background* para vídeos, capturados com câmera estática, com informação de profundidade associada. Este, pode ser integrado a abordagens DIBR (como as descritas nos dois capítulos anteriores) como ferramenta auxiliar para o preenchimento de *disocclusions*. Deste modo, na próxima seção, expõe-se uma visão geral do método proposto. Após, na Seção 5.2, descreve-se o processo de construção do modelo inicial de *background*. Na Seção 5.3, detalha-se o processo de incremento e atualização deste modelo inicial, com base em informação de quadros futuros. Por fim, a Seção 5.4 apresenta o processo adotado para a inclusão de informação do modelo de *background* na abordagem DHS.

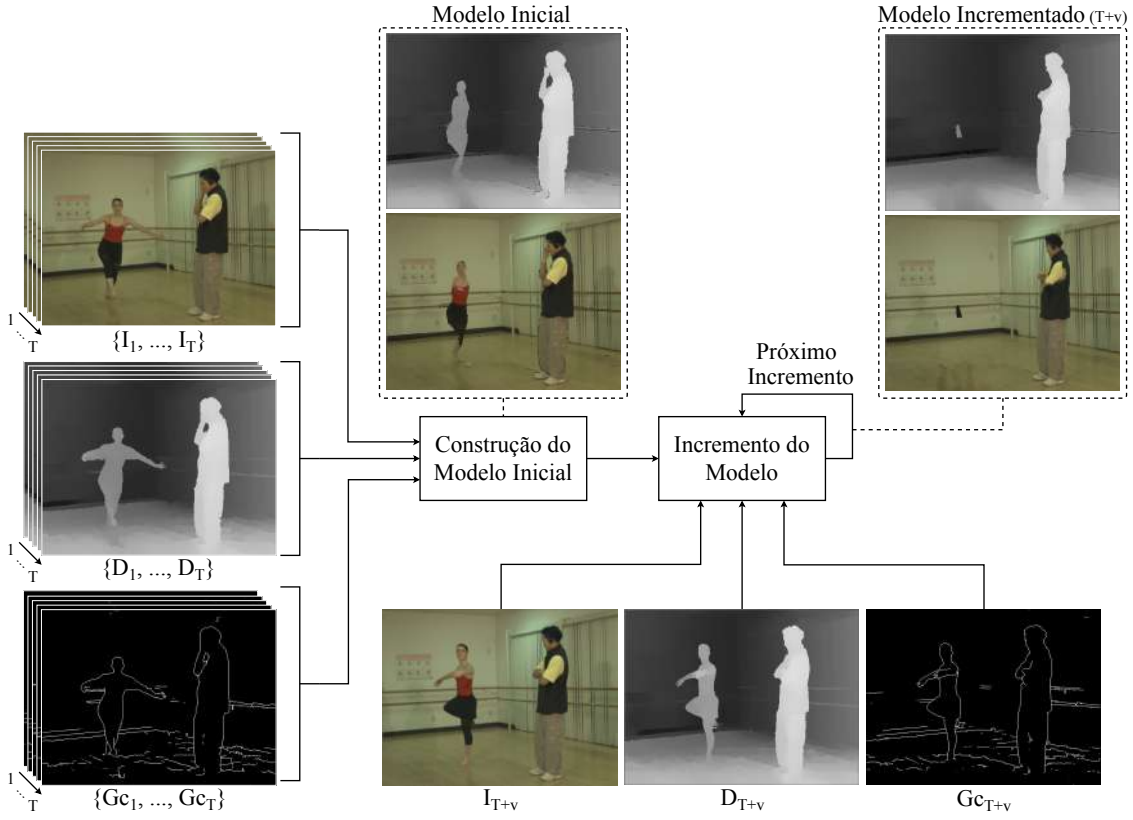
5.1 Visão Geral do Método Proposto

O processo de geração de modelos de *background* deve obedecer restrições e considerar especificidades inerentes da aplicação alvo. Abordagens DIBR tem como foco principal produzir conteúdo para TV3D e FVV. Estas aplicações, em especial, requerem que as imagens sejam geradas sob demanda e, portanto, não permitem que vídeos sejam processados de forma não causal – com informação de quadros futuros, por exemplo – para posterior renderização. Por este motivo, desenvolveu-se um método para a geração de modelos de *background* para sequências de vídeo com câmera estática, que se divide em duas fases: **construção**, onde um modelo inicial de *background* estático é computado; e **incremento**, que tem como objetivo atualizar o modelo à medida que novos quadros de vídeo são disponibilizados. Após a etapa de construção, que utiliza poucos quadros iniciais do vídeo, o modelo de *background* já pode ser empregado no *pipeline* DIBR. Testes efetuados com o método proposto revelaram que com apenas 5 quadros já se torna possível formular um modelo sem boa parte dos objetos de *foreground*. Isto permite que o conteúdo do modelo seja utilizado para o preenchimento de *disocclusions* (por exemplo), ao mesmo tempo em que é incrementado com dados do novo quadro utilizado como referência para a geração da vista sintética. Apesar do direcionamento dado a aplicações alvo do DIBR, observa-se que o método proposto pode ser empregado na solução de outros problemas de pesquisa que necessitem, por exemplo, da subtração de *background*, contanto que o vídeo disponha de informação de profundidade para cada *pixel*.

A Figura 5.1 apresenta um diagrama com uma visão geral do método proposto.

As operações realizadas em ambas fases do método proposto são guiadas por informação de cor e profundidade, o que permite uma identificação e controle mais preciso do conteúdo das imagens e do modelo. Como pode ser visto no diagrama, são utilizadas T imagens coloridas I, correspondentes aos quadros iniciais do vídeo, em conjunto com os respectivos mapas de disparidade D e de regiões candidatas a *ghost* G_c, para a construção do modelo de *background* inicial. Com base nesta entrada de dados, estima-se uma representação estável de informação de *background* (modelo inicial), que corresponde a uma imagem colorida M_I e seu respectivo mapa de disparidades M_D. Para formular M_I e M_D, faz-se uma análise do conteúdo de cada *pixel* individualmente, ao longo dos T quadros, em relação a imagem colorida, mapa de disparidades e de *ghosts*. Não existe

Figura 5.1: Diagrama que detalha o método proposto para a construção do modelo de *background*. Para a formulação do modelo inicial são utilizadas T imagens coloridas (I) – referentes aos quadros iniciais do vídeo – com os mapas de disparidade (D) correspondentes e, também, mapas com as regiões candidatas a *ghost* identificadas (G_c). Após a formulação deste modelo, inicia-se o processo de refinamento, com incrementos que utilizam informação dos quadros seguintes. No diagrama, apresenta-se o resultado do incremento do modelo de *background* utilizando um quadro no instante $T + v$, após v incrementos.



Fonte: O Autor, com imagens adaptadas do *dataset Ballet* de (ZITNICK et al., 2004).

dependência entre a estimativa dos *pixels*, por isso, este processo pode ser paralelizado. Antes de iniciar o processo de análise, ocorrências indicadas em G_c são descartadas para remover possíveis artefatos. Após, ocorrências de *foreground* são identificadas com base nos valores definidos em D e removidas. Nesta etapa, ao mesmo tempo, estima-se o valor de disparidade associado ao *pixel* em M_D. Em seguida, utilizando a amostra restante, elaborada somente com informação de *background* e sem possíveis artefatos, estima-se o conteúdo do *pixel* em M_I aplicando o método proposto por Jung (2009). O processo de construção de M_I e M_D utiliza somente conteúdo real das T imagens e mapas disponíveis. Portanto, elementos de *foreground* podem fazer parte do modelo inicial, quando se mantiverem fixos na cena. Por isso, faz-se necessário uma etapa de validação deste conteúdo por parte das abordagens DIBR que venham a empregá-lo no seu *pipeline*, para que conteúdo inadequado não seja copiado.

A segunda fase do método proposto começa após a estimativa do modelo inicial, e tem como objetivo refinar M_I e M_D com informação dos novos quadros, assim que estes são disponibilizados. Como observado por (YAO et al., 2014), e discutido anteriormente, a maior parte das regiões oclusas em um quadro de vídeo pertencem ao *background* que pode ser visível em outros instantes. Neste caso, pode-se utilizar o conteúdo revelado para incrementar o modelo de *background*. O mesmo pode ocorrer de forma inversa, e *pixels* de *background* podem ser cobertos por objetos de *foreground* em quadros consecutivos do vídeo. Contudo, logicamente, estas ocorrências não são úteis para o processo de incremento. Por outro lado, regiões sem alterações significativas em quadros posteriores do vídeo também podem ocorrer. Nesta situação, a informação mais recente pode ser útil para a atualização do modelo.

Com base nas observações anteriores, desenvolveu-se um processo de incremento para o modelo de *background* inicial, que se baseia na repetição do conteúdo de cada *pixel* ao longo do tempo. Neste processo, se um dado *pixel* no quadro $T + 1$, por exemplo, apresentar valor de disparidade inferior ao contido em M_D, este provavelmente corresponde a parte de um elemento do *background* que foi exposto. No entanto, se isso não se repetir nos quadros seguintes, a mudança de intensidade pode corresponder a um *outlier*, decorrente de um erro no processo de estimativa do mapa de disparidades. Por isso, na abordagem proposta, substitui-se o valor de um dado *pixel* em M_I e M_D somente quando este indicativo se repetir por um número predeterminado de vezes consecutivas, para que o modelo produzido seja o mais confiável possível. Além da substituição de conteúdo do modelo com informação de *background* revelada, parte do processo proposto consiste

da atualização de M_I e M_D, com informação dos novos quadros, em regiões estáveis. Na abordagem proposta, o conteúdo do modelo e do novo quadro são misturados, por meio de uma média ponderada, computada para cada *pixel* individualmente. A seguir, detalham-se as duas fases que compõem o método proposto.

5.2 Construção do Modelo de *Background* Inicial

A informação de *background* para vídeos pode ser formulada de diferentes maneiras, o que permite uma grande gama de possibilidades para a sua elaboração. Por esta razão, e pela relevância do tema, diversos trabalhos foram publicados na literatura abordando este problema, utilizando GMM, aprendizado de subespaço, Fuzzy, *Robust Principal Components Analysis*, entre outros (BOUWMANS, 2014). Durante os últimos anos, algumas abordagens clássicas, pertencentes a estas categorias, foram adaptadas para a formulação de modelos de *background* no contexto de DIBR, como GMM (YAO et al., 2014; LUO et al., 2016; LUO et al., 2019) e LBP (LUO et al., 2019), as quais apresentam como principal diferencial a inclusão de informação de profundidade no processo de estimativa. Estas, apesar de aprimorarem as abordagens originais, formulam os modelos sem considerar diretamente a possibilidade da existência de *outliers* e artefatos nas imagens e mapas de profundidade. Como discutido anteriormente, estes problemas são comuns no processo de síntese de vistas com DIBR, e precisam ser evitados. Com base nisso, desenvolveu-se um novo método para a construção de modelos de *background*, próprio para aplicação em DIBR, que se caracteriza por evitar a inclusão de artefatos e eliminar possíveis *outliers* do processo de estimativa.

No método proposto, a construção do modelo de *background* inicial se baseia na análise da informação individual de cada *pixel*. Denotaremos por $\boldsymbol{p} = (x, y)$ a coordenada de um *pixel* durante os primeiros T quadros do vídeo. Nesta análise, são verificados os padrões de intensidade de $\boldsymbol{p}$ nas imagens $\{I_1(\boldsymbol{p}), I_2(\boldsymbol{p}), ..., I_T(\boldsymbol{p})\}$, e nos mapas de disparidade $\{D_1(\boldsymbol{p}), D_2(\boldsymbol{p}), ..., D_T(\boldsymbol{p})\}$ e de regiões candidatas a *ghost* $\{G_{c1}(\boldsymbol{p}), G_{c2}(\boldsymbol{p}), ..., G_{cT}(\boldsymbol{p})\}$ correspondentes. As imagens e valores de disparidade são fornecidos como entrada para o método proposto, mas os mapas que indicam as regiões que podem conter *ghosts* precisam ser estimados. Como discutido anteriormente, este artefato ocorre em regiões onde existe uma mudança abrupta de profundidade (o que indica uma transição entre dois objetos). Sendo assim, formular um mapa G_c para cada quadro com regiões que potencialmente contém o artefato, corresponde a selecionar va-

riações abruptas de profundidade. Para esta tarefa, utilizou-se o método proposto para a seleção de candidatos a *ghost*, detalhado no Algoritmo 4. Este, executa uma operação de dilatação morfológica em D, seguida da aplicação de um limiar, para selecionar somente regiões que apresentem mudanças significativas de disparidade. No diagrama exibido na Figura 5.1 são exibidos alguns exemplos deste mapa. Estes mapas, juntamente com as imagens de entrada e a informação de disparidade, são utilizados para formular o modelo inicial, composto por uma imagem colorida M_I e pelo mapa de disparidades M_D correspondente.

Para estimar $M_I(\boldsymbol{p})$ e $M_D(\boldsymbol{p})$, são inicialmente removidas todas as ocorrências indicadas em $G_{c\{1,\ldots,T\}}(\boldsymbol{p})$ de $I_{\{1,\ldots,T\}}(\boldsymbol{p})$ e $D_{\{1,\ldots,T\}}(\boldsymbol{p})$, para evitar a inclusão de *ghosts* no modelo. Após, estima-se o valor de disparidade correspondente ao *background* em $\boldsymbol{p}$, com base nos valores de D não indicados em G_c ao longo dos T quadros. Teoricamente, a disparidade equivalente ao *background* em $\boldsymbol{p}$ deve ser indicada pelo menor valor encontrado na amostra restante, o qual representa o elemento mais ao fundo capturado neste ponto da cena. Contudo, sabe-se que além de artefatos o mapa de disparidades pode conter valores inconsistentes, decorrentes do seu processo de estimativa. Por este motivo, estima-se o valor mínimo de disparidade $D_{min}(\boldsymbol{p})$, eliminando $\alpha_{M_D}\%$ dos menores valores da amostra (de modo similar ao que ocorre no cálculo da média truncada), para evitar que possíveis *outliers* possam prejudicar a estimativa. Normalmente, o elemento do *background* é exibido em mais de um quadro e, portanto, mesmo que não existam *outliers*, seu valor de disparidade deverá ser preservado, desde que α_{M_D} seja relativamente baixo. Nos testes, definiu-se $\alpha_{M_D} = 10\%$, seguindo o padrão utilizado no cálculo da média truncada, empregada no extrator *foreground-background* (descrito na Seção 3.3). Esta estimativa permite que seja estabelecido um valor de referência para o *background* em $\boldsymbol{p}$ e, consequentemente, possibilita que seja determinado o valor de $M_D(\boldsymbol{p})$, pois assume-se que $M_D(\boldsymbol{p}) = D_{min}(\boldsymbol{p})$. Se um valor muito baixo for atribuído a $M_D(\boldsymbol{p})$ neste instante, em decorrência de algum *outlier*, e o mesmo corresponder a informação de cor de um objeto do *foreground*, o seu conteúdo permanecerá igual em todo o processo de incremento (descrito mais à frente). Como consequência disso, abordagens DIBR acabam sendo induzidas a propagar informação inconsistente para as vistas sintéticas, devido a cópia de conteúdo do *foreground* – classificado como *background* – para as *disocclusions*. Por este motivo, justifica-se esta precaução de remover uma fração α_{M_D} da amostra.

Nesta abordagem, assim como no trabalho de Yao et al. (2014), explora-se a correlação temporal de textura e informação de profundidade para gerar uma imagem co-

lorida de referência para o *background*, mas com algumas restrições relativas a possíveis inconsistências em ambos. Desta forma, com base no valor de referência de disparidade $D_{min}(\boldsymbol{p})$, são removidas de $I(\boldsymbol{p})$ as ocorrências que potencialmente correspondem a objetos do *foreground*. Para isso, aplica-se um limiar baseado em $D_{min}(\boldsymbol{p})$ e em uma margem de flutuação aceitável ν, visando descartar todas as ocorrências onde $D_{\{1,\ldots,T\}}(\boldsymbol{p}) > (D_{min}(\boldsymbol{p}) + \nu)$ ou $D_{\{1,\ldots,T\}}(\boldsymbol{p}) < (D_{min}(\boldsymbol{p}) - \nu)$. Se ν for definido como um valor muito alto, elementos de *foreground* podem permanecer na amostra. Por outro lado, um valor baixo poderá reduzir em exagero a amostra utilizada para definir a intensidade de cor associada a $\boldsymbol{p}$ no modelo, o que pode resultar em uma estimativa de cor menos confiável. Neste caso, para manter uma maior flexibilidade, definiu-se $\nu = \lambda$, que corresponde ao limiar que indica a máxima mudança aceitável de disparidade dentro de um mesmo objeto no processo de detecção de *cracks* translúcidos, descrito na Seção 3.2. Portanto, assim como estabelecido anteriormente para λ, definiu-se $\nu = 5$.

Após remover de $I(\boldsymbol{p})$ possíveis *ghosts* e elementos de *foreground*, pode-se iniciar o processo de estimativa de informação de cor para $M_I(\boldsymbol{p})$. Nesta etapa, emprega-se a abordagem proposta por Jung (2009) para a estimativa de modelos de *background*, utilizada em seu método de subtração de fundo. Esta abordagem foi seleciona por proporcionar a construção de um modelo de *background* robusto, de maneira simples e rápida, sem que seja necessário o uso de muitos quadros de vídeo. O trabalho original de Jung (2009) parte da premissa de que séries temporais de cada *pixel* de *background*, capturados por uma câmera estática, devem ser normalmente distribuídas e suas médias correspondem à imagem de *background*. Por isso, e considerando que objetos em movimento podem produzir *outliers* na distribuição, o autor optou pelo uso da média truncada baseada na similaridade com a mediana para a estimativa do modelo.

Na abordagem original, utilizam-se vídeos monocromáticos para a produção do modelo de *background*, e por este motivo, as imagens I são inicialmente convertidas para escala de cinza, gerando Ig. Após, computa-se uma imagem Ig_M que contém a mediana temporal de cada *pixel* $\boldsymbol{p}$ nas T amostras restantes de Ig. Então, para $0 \leq \alpha_{M_I} < 1$, a média truncada para cada *pixel* é obtida pelo cálculo da média temporal da amostra, desconsiderando os maiores $\lfloor \alpha_{M_I} \cdot T \rfloor$ desvios da mediana (onde, $\lfloor x \rfloor$ denota a função de maior inteiro, menor ou igual a um valor real x). Os desvios da mediana são dadas por $|Ig_t(\boldsymbol{p}) - Ig_M(\boldsymbol{p})|$, para cada item t da amostra, onde t se refere ao instante. Então, considerando que a função inteira f_{ord} – onde $1 \leq f_{ord}(\boldsymbol{p}, t) \leq T$ – retorna a posição de $|Ig_t(\boldsymbol{p}) - Ig_M(\boldsymbol{p})|$ em ordem crescente, o cálculo da média truncada que define o valor

de intensidade para a imagem do modelo pode ser dada por:

$$M_I(\boldsymbol{p}) = \frac{1}{T - \lfloor \alpha_{M_I} \cdot T \rfloor} \sum_{t \in A_{\alpha_{M_I}}(\boldsymbol{p})} I_t(\boldsymbol{p}), \tag{5.1}$$

onde $A_{\alpha_{M_I}}(\boldsymbol{p}) = \{t : f_{ord}(\boldsymbol{p}, t) \leq T - \lfloor \alpha_{M_I} \cdot T \rfloor\}$. Nesta etapa, se $\alpha_{M_I} = 0$, a função irá retornar exatamente a média (que é afetada pelos objetos em movimento). De outro modo, quando α_{M_I} se aproximar de um, a média truncada ficará próxima do valor da mediana (que corresponde a um instantâneo do *background* em um dado quadro). Assim como em (JUNG, 2009), definimos $\alpha_{M_I} = 0,3$. Na Equação 5.1, a imagem monocromática Ig foi utilizada para definir os elementos que devem compor a média na função f_{ord}. No entanto, ao contrário da abordagem original (que trabalha com imagens monocromáticas), $M_I(\boldsymbol{p})$ foi computada com base em I (colorida).

No método proposto, como parâmetro geral, um valor de T muito pequeno deve levar a construção de um modelo com mais conteúdo de *foreground* e uma menor confiabilidade. Por outro lado, uma quantidade exagerada de quadros, mesmo tendendo a produzir um modelo com mais conteúdo de *background*, faz com que muitas vistas sintéticas deixem de utilizar o conteúdo do modelo para o preenchimento das *disocclusions*. A Figura 5.2 exibe o resultado do método proposto, atribuindo diferentes valores para T.

5.3 Incremento do Modelo de *Background*

Após a construção do modelo de *background*, inicia-se o processo de incremento e atualização do seu conteúdo, com base em novos quadros do vídeo, à medida que es-

Figura 5.2: Resultado produzido pela abordagem proposta para a estimativa do modelo de *background* inicial, com T definido como 5 em (a), 10 em (b) e 20 em (c).

(a) (b) (c)

Fonte: O autor, com imagens adaptadas do *dataset Ballet* de (ZITNICK et al., 2004).

tes são disponibilizados. Nesta fase, assim como na anterior, para identificar diferenças entre o modelo pré-computado e um novo quadro, analisam-se as intensidades de cor (no espaço de cores RGB) e disparidade para cada *pixel* individualmente. Considerando que M_I e M_D contêm toda informação de *background* disponível no vídeo entre os instantes 1 e T, um *pixel* $\boldsymbol{p}$ em um novo quadro no instante $T+1$ pode consistir de: *(i)* uma região no *background* que foi exposta em razão da movimentação de um objeto no *foreground*, indicada por um valor maior de profundidade (menor de disparidade); *(ii)* parte do *foreground* que passou a ocluir o *background*, indicada com um valor menor de profundidade (maior de disparidade); *(iii)* uma região que não sofreu alterações significativas durante os $T+1$ quadros, permanecendo com características aproximadamente padronizadas de cor e profundidade (independentemente de pertencer ao *background* ou *foreground*). Sendo assim, com base nestas possibilidades, foram estabelecidas regras para a atualização do modelo de *background*, as quais têm como finalidade mantê-lo estável a menos que informação com indicativo de pertencer ao *background* seja identificada.

Além das alternativas detalhadas anteriormente, que levam em conta imagens e mapas de disparidades perfeitos, deve-se considerar também a possibilidade da existência de inconsistências tanto em I como D, que possam resultar na inclusão de artefatos e *outliers* no modelo. Por este motivo, ao receber um novo quadro no instante $T+1$ (ou em qualquer outro posterior), deve-se verificar primeiramente o seu mapa de regiões candidatas a *ghost*. Neste caso, se $G_{cT+1}(\boldsymbol{p}) = 1$, descarta-se imediatamente o conteúdo do *pixel* $\boldsymbol{p}$, para evitar a possibilidade da inserção de artefatos no modelo de *background*. Os *pixels* restantes são analisados, para garantir que um dado $\boldsymbol{p}$ seja: descartado, caso corresponda a um objeto no *foreground*, indicado por um alto valor de disparidade em relação a $M_D(\boldsymbol{p})$; inserido no modelo, caso seja comprovadamente uma parte do *background* revelada, o que é evidenciado por um baixo valor de disparidade em relação a $M_D(\boldsymbol{p})$; utilizado para atualizar o modelo, de modo a torná-lo mais confiável e estável, se este possuir uma mudança pouco significativa de disparidade com relação a $M_D(\boldsymbol{p})$.

Ao atualizar o modelo de *background*, deve-se ter em mente que este poderá ser empregado no preenchimento de *holes* de diversas vistas sintéticas. Por este motivo, não se pode permitir que erros sejam inseridos no processo de incremento, uma vez que estes podem ser replicados em muitos quadros do vídeo sintético, prejudicando a experiência do usuário. Desta forma, para determinar variações aceitáveis de mudança de disparidade para um dado *pixel*, foi definido um limiar σ_{M_D}. Com base neste limiar, pode-se identificar a necessidade de descarte, atualização ou substituição de um dado *pixel* $\boldsymbol{p}$ no modelo,

por meio da seguinte verificação:

- se $(M_D(\boldsymbol{p}) - D_{T+1}(\boldsymbol{p})) > \sigma_{M_D}$, indica-se para substituição o conteúdo de $\boldsymbol{p}$. Esta validação considera que apenas uma mudança significativa do valor de disparidade possa indicar a exibição de parte do *background* na posição do *pixel*. Para evitar que *outliers* comprometam o processo de incremento, estipula-se que esta verificação deva ser verdadeira por no mínimo Q_q vezes consecutivas. Neste caso, substitui-se o valor associado ao *pixel* no modelo pelo conteúdo de *background* que foi exibido nos Q_q últimos quadros. Um alto valor para Q_q pode fazer com que *pixels* do modelo não sejam substituídos, e um baixo pode permitir que um *outlier* em D_{T+1} habilite a substituição de $\boldsymbol{p}$. Nos testes realizados, $Q_q = 1$ (duas ocorrências seguidas) se fez suficiente para evitar este segundo caso.

- se $|M_D(\boldsymbol{p}) - D_{T+1}(\boldsymbol{p})| < \sigma_{M_D}$, atualiza-se o modelo de *background*, incluindo os valores de intensidade mais recentes. Neste caso, tanto M_I como M_D são redefinidos por meio do cômputo de uma média ponderada, que considera a quantidade de quadros que foram utilizados para sua estimativa atual.

- caso nenhuma das verificações anteriores seja satisfeita, descarta-se o conteúdo associado a $\boldsymbol{p}$ no quadro do instante $T + 1$. Este, pode corresponder a informação do *foreground* ou a algum *outlier* que foi exibido no *pixel* analisado.

Considerando que as operações são guiadas pelo valor de σ_{M_D}, deve-se levar em conta que, se um alto valor for atribuído a este limiar, conteúdo de *background* e *foreground* pode ser misturado, criando inconsistências no modelo. Por outro lado, a atribuição de um valor baixo pode inviabilizar a atualização do modelo, levando a substituição frequente do seu conteúdo. Visando manter uma boa margem de segurança neste processo, definiu-se $\sigma_{M_D} = 1$ em todos os testes.

Duas operações podem ser realizadas no modelo de *background*, substituição e atualização. Na substituição, quando os critérios são obedecidos, indica-se o valor de cor e disparidade do *pixel* pela média das intensidades dos Q_q quadros consecutivos. Já no caso da atualização, define-se o valor de $\boldsymbol{p}$ da seguinte maneira:

$$M_I(\boldsymbol{p}) = \frac{M_I(\boldsymbol{p}) \cdot M_Q(\boldsymbol{p}) + I(\boldsymbol{p})}{M_Q(\boldsymbol{p}) + 1}, \tag{5.2}$$

$$M_D(\boldsymbol{p}) = \frac{M_D(\boldsymbol{p}) \cdot M_Q(\boldsymbol{p}) + D(\boldsymbol{p})}{M_Q(\boldsymbol{p}) + 1}, \tag{5.3}$$

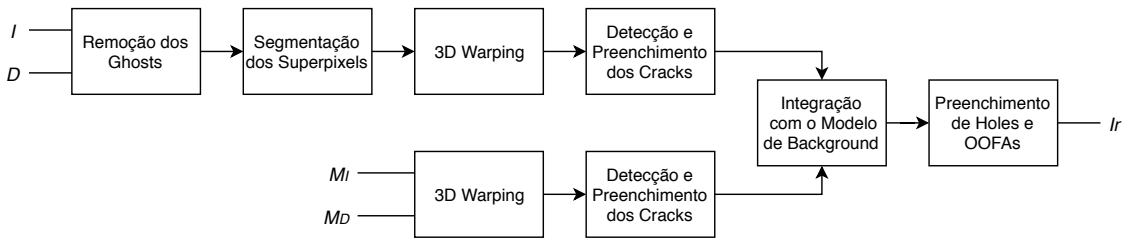
onde $M_Q(\boldsymbol{p})$ indica a quantidade de quadros utilizados para a definição da intensidade de um dado *pixel* $\boldsymbol{p}$. Ou seja, define-se $M_Q(\boldsymbol{p}) = \lfloor \alpha_{M_I} \cdot T \rfloor$ inicialmente, e quando o *pixel* é atualizado, incrementa-se seu valor. Em caso de substituição de $\boldsymbol{p}$, define-se $M_Q(\boldsymbol{p}) = Q_q$.

5.4 Integração do Modelo de *Background* com a Abordagem DHS

O modelo de *background* proposto pode ser combinado com qualquer abordagem DIBR, uma vez que este contém as mesmas informações requeridas para a geração de vistas sintéticas. Genericamente, para utilizar informação do modelo de *background*, basta projetá-lo para o ponto de vista virtual com 3D *image warping* e copiar adequadamente a sua informação para o interior das *disocclusions* na vista sintética. Entretanto, observa-se que tanto a forma como a informação é copiada quanto o estágio do *pipeline* da abordagem DIBR em que isso é realizado podem influenciar no resultado final produzido. Com o objetivo de validar o método proposto e fornecer um exemplo real de aplicação, desenvolveu-se um processo de integração do modelo de *background* com a abordagem DHS (descrita no Capítulo 4). Esta abordagem foi escolhida em razão de sua peculiaridade, relativa ao uso de um mapa de *superpixels* auxiliar, e por apresentar melhores resultados experimentais (apresentados no próximo capítulo) quando comparada com a técnica ATA (descrita no Capítulo 3).

A Figura 5.3 apresenta um diagrama de blocos que detalha o processo de integração do modelo de *background* com a abordagem DIBR. Na sequência de passos exibida no *pipeline*, pode-se observar que até a etapa do 3D *warping* nada foi alterado no fluxo original do DHS. Nesta etapa, inicia-se o processamento do modelo de *background*, onde este é projetado para o ponto de vista virtual juntamente com a imagem de referência.

Figura 5.3: Diagrama de blocos que detalha o processo de integração do modelo de *background* proposto com a abordagem DHS.



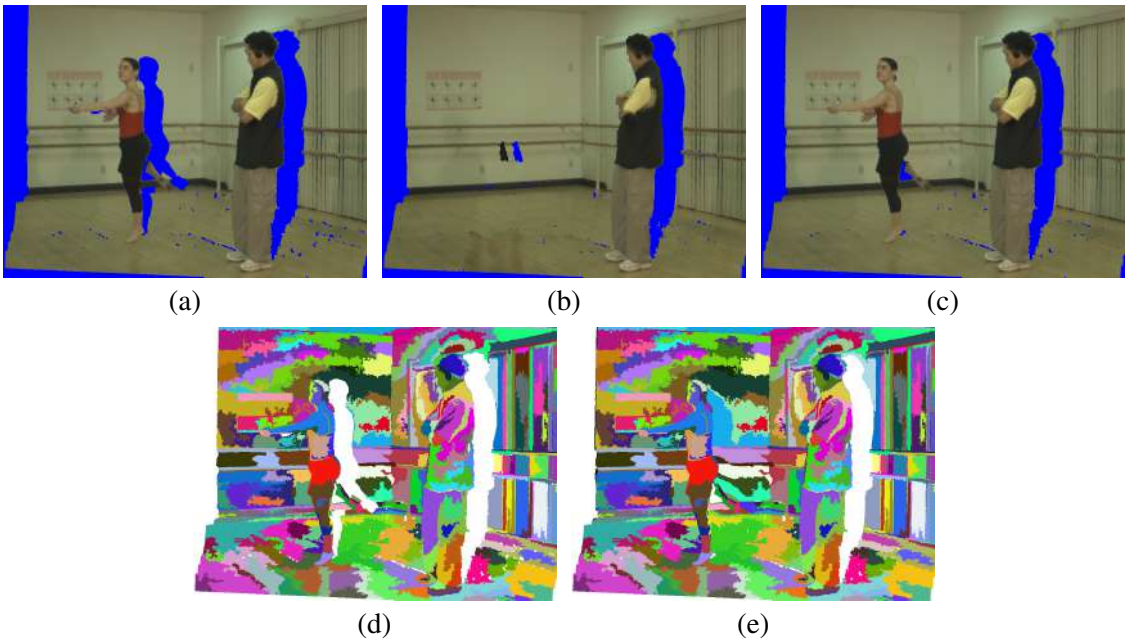
Fonte: O autor.

No caso do modelo, projeta-se M_I com base em M_D e em parâmetros de câmera (iguais aos definidos para o ponto de vista de referência), produzindo M_{Iw}. Ao mesmo tempo, projeta-se M_D, gerando M_{D_w}. Em seguida, são removidos os *cracks* translúcidos e preenchidas ambas as formas do artefato com a solução empregada no DHS, tanto em I_w como em M_{I_w} (e nos respectivos mapas). Então, inicia-se o processo de integração do modelo de *background* projetado com a vista sintética.

No processo de integração proposto para a DHS, realiza-se o preenchimento parcial ou total de cada *hole* Ω de I_w individualmente, copiando informação diretamente do modelo de *background* projetado. Para isso, inicialmente, são identificados os *holes* da classe BG (que correspondem às *disocclusions*), utilizando o processo de classificação descrito na Subseção 4.5.1. Então, com base no conteúdo individual da borda $\delta\Omega$ de cada Ω, estima-se um limiar T_Ω, que permite controlar o processo de cópia, para evitar que conteúdo de *foreground* seja reproduzido. O valor deste limiar é determinado dinamicamente, por meio do cálculo da média truncada, seguindo o processo definido na Seção 3.3. Este, permite separar o conteúdo da vizinhança de uma *disocclusion* em duas partes, *background* e *foreground*. Então, com base nisso, se um dado *pixel* $\boldsymbol{p} \in \Omega$ estiver vazio em I_w e contiver informação em M_{I_w}, seu conteúdo deverá ser copiado tanto para a imagem colorida quanto para o mapa de disparidades projetado, desde que $M_{D_w}(\boldsymbol{p}) \leq T_\Omega$. Na Figura 5.4(c), exibe-se o resultado do preenchimento da vista sintética exibida em (a), com o conteúdo do modelo de *background* exposto em (b).

Abordagens DIBR podem possuir determinadas particularidades, como no caso da DHS que emprega *superpixels* como estrutura auxiliar do algoritmo de preenchimento de *disocclusions* e OOFAs. Neste caso, além de copiar informação de cor e disparidade, faz-se necessário atribuir adequadamente os rótulos de *superpixel* para a região preenchida. Uma solução direta para este problema seria manter um mapa com a informação temporal dos rótulos associados ao modelo. No entanto, o algoritmo de estimativa do *superpixels* não produz segmentos temporalmente coerentes, e os mapas computados individualmente para os instantes t e $t+1$ podem ser completamente diferentes. Para produzir conteúdo para estas regiões, foi desenvolvido um processo que replica, gradativamente, os rótulos da borda de *background* de cada *disocclusion* individualmente. Para identificar os rótulos que pertencem ao *background*, utiliza-se o limiar T_Ω. Neste caso, se mais de 70% dos *pixels* de um dado rótulo em $\delta\Omega$ possuir valor de disparidade menor que T_Ω, o rótulo é classificado como apto a ser replicado, por pertencer ao *background*. Os rótulos que não obedecem esta restrição são desconsiderados. Então, no processo desenvolvido, para

Figura 5.4: Resultado produzido pela integração do modelo de *background* com a abordagem DHS. Em (c), exibe-se o resultado do preenchimento da vista sintética (a) com o conteúdo do modelo de *background* (b). Em (d), apresenta-se o mapa de *superpixels* antes da integração e, em (e), após este processo.



(a) (b) (c)

(d) (e)

Fonte: O autor, com imagens adaptadas do *dataset Ballet* de (ZITNICK et al., 2004).

cada *pixel* copiado, seleciona-se o vizinho – considerando 4-vizinhança – que minimiza a distância SSD de cor e possui rótulo classificado como pertencente ao *background*. Caso existam vizinhos válidos somente com rótulo de *foreground* ou ainda sem indicação, o *pixel* permanece sem indicação de *superpixel*. Este processo é repetido iterativamente até que nenhum *pixel* receba um novo rótulo. Os *pixels* isolados ou somente com vizinhos no *foreground* que permanecem sem atribuição de *superpixel* são apagados ao final. Um exemplo do resultado produzido por este processo pode ser visualizado nas Figuras 5.4(e), que corresponde a (d) reconstruída.

Como pode ser visto no diagrama da Figura 5.3, após a integração da imagem sintética com o modelo de *background*, realiza-se o preenchimento dos *holes* e OOFAs normalmente. Ao fim, tem-se uma imagem sintética resultante I_r, sem artefatos e sem *holes*, onde as *disocclusions* foram preenchidas parcial ou totalmente com conteúdo do modelo de *background*.

5.5 Conclusões do Capítulo

Neste capítulo, descreveu-se um novo método para a modelagem de *background* para sequências de vídeo gerados com câmera estática. Inicialmente, foi detalhada a primeira fase do método proposto, que compreende a etapa de construção do modelo de *background*. Neste processo, estima-se o conteúdo do modelo inicial, com base na informação de disparidade e cor de cada *pixel* individualmente, utilizando uma quantidade predeterminada de quadros, e empregando verificações que não permitem a inclusão de possíveis *outliers* e artefatos. Em seguida, foi apresentada a segunda fase, que compreende o incremento do modelo de *background*. Neste etapa, avalia-se o conteúdo de novos quadros de vídeo, com o objetivo de substituir informação de *foreground* e atualizar o modelo, para torná-lo mais completo e confiável. Por fim, descreveu-se detalhadamente o processo de integração do modelo de *background* com a abordagem DHS. Este processo deixa claro como a cópia de informação deve ser realizada, para que elementos de *foreground* não sejam inseridos na vista sintética. Além disso, permite compreender o que pode ser feito quando peculiaridades como a atribuição de rótulos do mapa de *superpixels* nas regiões preenchidas se faz necessária.

Atualmente, o método proposto não prevê soluções para a estimativa de modelos de *background* quando existe movimento de câmera na captação do vídeo. Entretanto, para que a solução proposta contemple este caso, pode-se realizar a compensação de movimento entre a posição do modelo de *background* no instante $t-1$ e do novo quadro no t, para que as coordenadas das imagens coincidam. Para isto, pode-se utilizar uma solução existente, como as apresentadas em (LIE; HSIEH; LIN, 2018; LUO et al., 2019), e então o método proposto poderá gerar o modelo de *background* para vídeos captados dinamicamente, sem outras alterações. Entretanto, observa-se que alguns problemas relacionados ao movimento de câmera – como vibração excessiva – podem levar estas soluções a erros. Se algum destes problemas for identificado nos testes, uma nova solução mais robusta deverá ser desenvolvida em trabalhos futuros.

6 RESULTADOS EXPERIMENTAIS

Neste capítulo, apresentam-se os resultados experimentais obtidos com as abordagens propostas, além de um comparativo com diversos trabalhos competitivos – recentemente publicados na literatura – com o objetivo de comprovar a efetividade das soluções desenvolvidas. Na Seção 6.1, descrevem-se os *datasets* utilizados no comparativo, juntamente com as métricas empregadas na avaliação quantitativa. Em seguida, na Seção 6.2, exibe-se uma avaliação qualitativa e quantitativa das abordagens, tanto no contexto de vídeos como de fotografias, utilizando mapas de disparidade *ground truth*. Na seção seguinte, para produzir uma análise realista, mede-se o impacto causado pelo uso de diferentes mapas de disparidade, produzidos por algoritmos recentes de casamento estéreo, em métodos DIBR de síntese de vistas. Todos os casos de teste apresentados neste capítulo foram executados com os parâmetros definidos na Tabela 6.1.

6.1 *Datasets* e Métricas de Avaliação

Para validar as abordagens propostas, foram realizados inúmeros testes com *datasets* disponíveis na internet, tanto no contexto de vídeos como de fotografias, utilizando mapas de disparidade *ground truth* e reais (estimados com algoritmos de casamento estéreo). Os resultados produzidos foram quantificados com duas métricas amplamente utilizadas no contexto de síntese de vistas e, adicionalmente, com uma destinada a avaliação espacial e temporal de vídeos e uma específica para DIBR nos testes com mapas de disparidade reais.

Para a avaliação no contexto de fotografia, foram empregados os *datasets* de Middlebury (HIRSCHMULLER; SCHARSTEIN, 2007). Para cada *dataset*, são fornecidas imagens de 7 pontos de vista distintos e dois mapas de disparidade *ground truth*, associados aos pontos de vista 1 e 5. As imagens são disponibilizadas em três tamanhos distintos. Nos testes com mapas de disparidade *ground truth*, descritos na Seção 6.2, foram utilizadas as imagens *full-size*, que possuem entre 1240 e 1396 *pixels* de largura, por 1110 de altura. Já na análise apresentada na Seção 6.3, foram utilizadas imagens *half-size*, com tamanho entre 620 e 698 *pixels* de largura, por 370 de altura, devido a falhas produzidas pelos algoritmos de casamento estéreo disponibilizados por alguns autores, não adequados às dimensões de *full-size*.

Para validar as abordagens em vídeos, foram usados os *datasets* Ballet e Break-

Tabela 6.1: Sumário com os valores definidos nos parâmetros das abordagens propostas, em todos os testes realizados para a produção dos resultados experimentais.

Notação	Valor	Descrição
$Hl_{\#}$	$\#$	Elemento estruturante binário em formato de linha, onde $\#$ representa o tamanho
$Hd_{\#}$	$\#$	Elemento estruturante binário em formato de diamante, onde $\#$ representa o tamanho
λ	5	Limiar que determina a variação máxima de disparidade permitida dentro de um objeto, considerando dois *pixels* vizinhos
α	10%	Percentual de *pixels* removidos do computo do *trimmed mean* definido no extrator *foreground–background*
γ	11	Limiar que determina a variação mínima para um dado *pixel* ser classificado como pertencente ao *foreground*
ρ	255	Fator de normalização definido originalmente no termo de dados do algoritmo de *inpainting* de (CRIMINISI; PEREZ; TOYAMA, 2004)
N	69	Tamanho definido para a janela de busca na imagem de referência
β	35	Limiar que determina o erro máximo aceitável no comparativo entre dois *patches*
k	350	Número de *superpixels* definido no processo de segmentação
σ^2_{MAX}	1521	Variância máxima dada pela quantidade de rótulos em um *patch* de 9×9
τ	10	Parâmetro que controla o decaimento da função exponencial que estima a correspondência de rótulos entre dois *patches*
T	5	Número de quadros utilizados para gerar o modelo de *background* inicial
α_{M_D}	10%	Percentual da amostra que deve ser removida para o cômputo do valor mínimo de disparidade associado a um *pixel* do modelo de *background*
ν	5	Margem de flutuação aceitável para a disparidade de um dado *pixel* durante a sucessão de quadros em vídeos
α_{M_I}	$0,3$	Valor utilizado para determinar a fração da amostra que será empregada na estimativa de cor para um dado *pixel* do modelo de *background* inicial
σ_{M_D}	1	Limiar que determina a máxima variação de disparidade para um dado *pixel* no processo de incremento do modelo de *background*
Q_q	1	Quantidade mínima de repetições consecutivas necessárias para a substituição do conteúdo de um dado *pixel* no modelo de *background*

Fonte: O Autor.

dancers de (ZITNICK et al., 2004), PoznanHall2 e PoznanStreet de (SCHWARZ; MARPE; WIEGAND, 2010) e Dancer de (RUSANOVSKYY; AFLAKI; HANNUKSELA, 2011). Em todos os testes, realizou-se a avaliação dos 100 primeiros quadros de cada vídeo, se-

guindo a definição dada em (LUO; ZHU, 2017). Os *datasets* Ballet e Breakdancers foram capturados sob 8 pontos de vista distintos, para os quais são fornecidos mapas de disparidade *ground truth*. Em ambos, as imagens têm tamanho de 1024×728 *pixels*. Para os *datasets* de (SCHWARZ; MARPE; WIEGAND, 2010) e (RUSANOVSKYY; AFLAKI; HANNUKSELA, 2011) são fornecidos vídeos para 9 pontos de vista, com resolução de 1920×1088 *pixels*, e informação de disparidade *ground truth* associada. Observa-se que no caso de PoznanHall2 e Dancer, o *background* não é estático, pois existe movimento de câmera. As fotografias e vídeos foram gerados em cenários reais, exceto no caso do *dataset* Dancer, que foi produzido sinteticamente por técnicas de computação gráfica.

Em todas as avaliações quantitativas foram empregadas as métricas PSNR (*Peak Signal-to-Noise Ratio*)[1] e SSIM (*Structural Similarity Index*)[2] (WANG et al., 2004), amplamente utilizadas para quantificar a similaridade entre duas imagens em síntese de vistas, compressão e outras aplicações. O PSNR tem como objetivo medir a razão entre o máximo valor de intensidade que um *pixel* pode assumir e o erro médio computado entre duas imagens. O cálculo da métrica PSNR computada entre uma imagem *ground truth* I_{GT} (referência) e a resultante do processo de síntese I_r (por exemplo) se dá pela seguinte equação:

$$PSNR(I_{GT}, I_r) = 10 \cdot \log_{10} \Big(\frac{{R_I}^2}{MSE(I_{GT}, I_r)} \Big), \tag{6.1}$$

onde R_I é definido de acordo com o máximo valor que um *pixel* pode assumir, estipulado de acordo com o tipo de dados utilizado para a representação da imagem. Por exemplo, para uma imagem representada por inteiros de 8 bits, deve-se definir $R_I = 255$. Já o MSE (*Mean Square Error*), corresponde a média dos erros quadráticos, computado com base no comparativo *pixel* a *pixel* entre I_{GT} e I_r. Neste caso, um alto valor de PSNR representa um baixo erro médio em relação as duas imagens comparadas, o que indica uma maior qualidade. Esta métrica permite mensurar diretamente a similaridade entre duas imagens. Contudo, desconsidera o quanto isto afeta a percepção do sentido visual humano (SUN; LIU; YANG, 2012). Por este motivo, como complemento, foi empregada a métrica SSIM, que foi projetada para avaliar principalmente elementos que afetam a percepção visual. Para quantificar isso, a métrica se baseia na análise da estrutura, iluminação e do contraste das imagens, como pode ser visto na equação abaixo:

$$SSIM(I_{GT}, I_r) = lu(I_{GT}, I_r)^a \cdot co(I_{GT}, I_r)^b \cdot st(I_{GT}, I_r)^c, \tag{6.2}$$

[1]Página com detalhes da implementação usada: <https://www.mathworks.com/help/images/ref/psnr.html>.

[2]Página para download do código fonte: <https://www.ece.uwaterloo.ca/~z70wang/research/ssim/>.

onde as funções se referem a comparação de luminância lu, contraste co e estrutura st. Na equação, $a > 0$, $b > 0$ e $c > 0$ são parâmetros utilizados para o ajuste relativo da importância dos três componentes, por padrão definidos com valor 1. Assim como para o PSNR, um alto valor de SSIM – que varia entre 0 e 1 – indica maior similaridade.

Para a avaliação em cenário real, com mapas de disparidade gerados por algoritmos de casamento estéreo (apresentada na Seção 6.3), foi utilizada a métrica *Morphological-Wavelet* PSNR (MW-PSNR)[3] (SANDIĆ-STANKOVIĆ; KUKOLJ; CALLET, 2016), que objetiva mensurar a qualidade de vistas sintéticas produzidas especificamente por abordagens DIBR. Esta é calculada por meio da Equação 6.1, mas com a substituição do MSE pelo MW-MSE (*Multi-scale Wavelet Mean Squared Error*). Neste caso, o MW-MSE corresponde ao MSE médio de todas as subbandas das imagens decompostas em *wavelets* morfológicas.

Métricas para a estimativa de qualidade de vídeos não podem se basear somente na análise espacial das imagens, devendo também avaliar o domínio temporal. Por este motivo, para realizar uma análise quantitativa mais assertiva das abordagens propostas, aplicou-se a métrica STRRED (*Spatio-Temporal-Reduced Reference Entropic Differences*)[4] (SOUNDARARAJAN; BOVIK, 2013) nos vídeos produzidos. O cômputo de STRRED se dá pelo produto de dois índices (SRRED e TRRED), que se correlacionam com os julgamentos humanos de qualidade. Neste caso, SRRED se baseia no cálculo de diferenças entrópicas espaciais, enquanto TRRED estima apenas as temporais (considerando quadros consecutivos). Todas as métricas foram computadas com os valores de parâmetro padrão, nas imagens convertidas para escala de cinza.

6.2 Avaliação com Mapas de Disparidade *Ground Truth*

Para validar as abordagens propostas, realizou-se um comparativo quantitativo e qualitativo empregando *datasets* de fotografia e vídeo. Na avaliação, foram utilizados diferentes casos de teste, definidos de acordo com os trabalhos de (LUO; ZHU, 2017; LUO et al., 2016; LUO et al., 2019). Estes artigos também serviram como guia para a seleção dos trabalhos usados no comparativo, e foram elaborados com base em outras avaliações apresentadas na literatura (como de (AHN; KIM, 2013)), se consolidando hoje

[3]Página para download do código fonte: <https://sites.google.com/site/draganasandicstankovic/code/mw-psnr>.

[4]Página para download do código fonte: <http://live.ece.utexas.edu/research/quality/strred.rar>.

como uma representação do estado-da-arte. Abaixo detalhamos esses trabalhos.

CRI – algoritmo de *inpainting* de (CRIMINISI; PEREZ; TOYAMA, 2004), utilizado como base para a elaboração das abordagens propostas;

VSRS – MPEG *View Synthesis Reference Software* na versão 3.5 (TANIMOTO et al., 2008; MORI et al., 2008);

DAR – abordagem proposta por Daribo and Saito (2011);

AHN – abordagem proposta por Ahn and Kim (2013);

DHHF – *Depth Adaptive* HHF (SOLH; ALREGIB, 2012b);

SHF – *pipeline* proposto por Oliveira et al. (2015);

FRHF – método FRHF de (LUO; ZHU, 2017);

FRBGE – método FRHF com BGE de (LUO; ZHU, 2017);

YAO – abordagem baseada em uma imagem de *background* estável temporalmente, proposta por Yao et al. (2014);

NEWSON – extensão do algoritmo PatchMatch de Barnes et al. (2009) para o caso espaço-temporal, proposto por Newson et al. (2014).

LUO – método baseado na estimativa de um modelo de *background* que emprega informação temporal, proposto por Luo et al. (2016);

PNT3 – versão do *framework* desenvolvido por Luo et al. (2019) que utiliza extração de *foreground*, um GMM modificado e o algoritmo PatchMatch (BARNES et al., 2009);

PNT5 – a mesma abordagem adotada em PNT3, mas substituindo o PatchMatch pelo método de Huang et al. (2014).

Na listagem, antes da referência de cada abordagem, exibe-se o nome utilizado para indicá-la nos comparativos. Os resultados quantitativos exibidos nesta seção foram extraídos dos trabalhos de (LUO; ZHU, 2017; LUO et al., 2016; LUO et al., 2019), exceto para o SHF, que foi implementado e avaliado juntamente com as abordagens propostas, seguindo as definições dadas no artigo. A avaliação qualitativa foi confeccionada a partir de imagens – com compressão – cedidas por (LUO; ZHU, 2017) e com os resultados gerados pelas implementações.

Neste trabalho, para avaliar as diferentes abordagens no contexto de fotografia, foram utilizados 11 *datasets* de Middlebury (HIRSCHMULLER; SCHARSTEIN, 2007). Nos testes, para gerar as vistas sintéticas, a imagem do ponto de vista de referência 1 foi projetada para o virtual 2. Então, foram computadas as métricas entre o resultado produzido e a imagem real do ponto de vista 2 (*ground truth*), para medir a efetividade das abordagens. Os *datasets* de Middlebury não possuem mapa de disparidades para a vista 2, por isso, o trabalho de (DARIBO; SAITO, 2011) não pôde ser relacionado nestes testes.

Para produzir uma avaliação consistente no contexto de vídeos, foram executados 14 casos de teste, com configurações distintas. Para Ballet (BA) e Breakdancers (BR) de (ZITNICK et al., 2004), a vista 4 foi projetada para 1 e 3, e a vista 5, para 2 e 4. Nos *datasets* de (SCHWARZ; MARPE; WIEGAND, 2010), a vista 7 de PoznanHall2 (PH) foi usada como referência para produzir as vistas 5 e 6 e, em PoznanStreet (PS), foram geradas as vistas 4 e 5 a partir da 3. Por fim, em Dancer (DA) de (RUSANOVSKYY; AFLAKI; HANNUKSELA, 2011), projetou-se a vista 1 para 5 e a 5 para 9.

6.2.1 Avaliação em Fotografias

Nas Tabelas 6.2 e 6.3, exibe-se uma análise quantitativa das fotografias sintéticas produzidas pelas abordagens em termos de PSNR e SSIM, respectivamente, com os melhores resultados destacados em negrito. Como pode ser observado, as abordagens propostas, descritas nos Capítulos 3 (ATA) e 4 (DHS), apresentam os melhores resul-

Tabela 6.2: Comparativo entre as diferentes abordagens, na avaliação quantitativa da métrica PSNR, no contexto de fotografia. Apresenta-se em cada linha os resultados para cada *dataset*, para as abordagens dispostas nas colunas. Os melhores resultados encontram-se destacados em negrito.

Dataset	CRI	VSRS	AHN	DHHF	SHF	FRHF	ATA	DHS
Aloe	28,10	27,97	26,32	28,24	28,55	**29,91**	29,35	29,76
Art	27,08	25,83	26,42	27,52	28,01	27,85	30,02	**30,98**
Bowling1	30,15	31,11	28,66	26,00	30,77	31,97	32,23	**33,02**
Bowling2	27,02	27,85	25,27	25,75	26,78	28,44	30,48	**30,58**
Lampshade1	29,93	30,12	30,33	28,50	31,97	34,82	35,35	**36,27**
Lampshade2	30,22	30,95	32,28	30,11	32,44	34,66	36,02	**36,97**
Laundry	29,69	28,30	28,53	28,76	30,46	28,85	29,99	**30,65**
Midd2	29,40	29,84	25,83	30,11	29,06	30,58	29,81	**31,98**
Reindeer	31,75	31,50	30,50	29,30	29,24	30,39	33,30	**34,27**
Flowerpots	21,64	24,21	19,48	20,64	20,84	26,59	**30,29**	29,46
Monopoly	30,35	30,18	29,47	31,06	30,58	32,21	32,37	**33,52**
Média	28,67	28,90	27,55	27,82	28,97	30,57	31,74	**32,50**
Desvio Padrão	2,74	3,47	2,32	2,90	3,18	2,66	2,31	2,55

Fonte: O Autor.

Tabela 6.3: Comparativo entre as diferentes abordagens, na avaliação quantitativa da métrica SSIM ($\times 10^{-1}$), no contexto de fotografia. Apresenta-se em cada linha os resultados para cada *dataset*, para as abordagens dispostas nas colunas. Os melhores resultados encontram-se destacados em negrito.

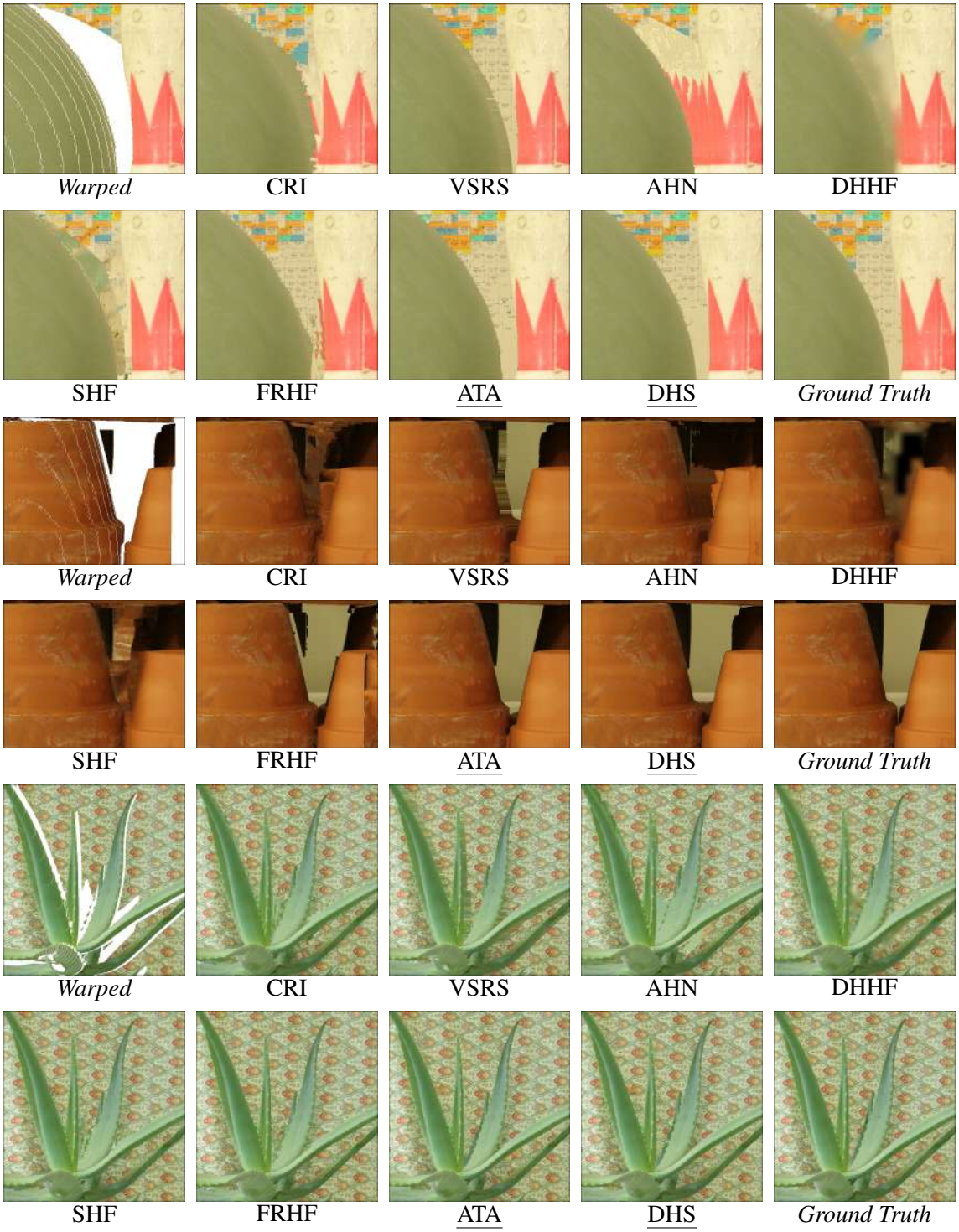
Dataset	CRI	VSRS	AHN	DHHF	SHF	FRHF	ATA	DHS
Aloe	8,988	9,121	8,929	9,053	9,098	**9,172**	9,153	9,169
Art	9,041	9,119	9,016	9,174	9,157	9,090	9,279	**9,313**
Bowling1	9,398	9,488	9,336	9,450	9,448	9,499	9,526	**9,559**
Bowling2	9,130	9,249	9,029	9,283	9,183	9,334	9,401	**9,409**
Lampshade1	9,456	9,551	9,463	9,544	9,517	9,590	9,656	**9,705**
Lampshade2	9,503	9,573	9,505	9,579	9,538	9,580	9,667	**9,695**
Laundry	9,317	9,384	9,262	9,327	9,382	9,256	9,391	**9,421**
Midd2	9,281	9,408	9,087	9,411	9,382	9,412	9,434	**9,492**
Reindeer	9,376	**9,499**	9,285	9,338	9,316	9,382	9,445	**9,499**
Flowerpots	9,072	9,306	8,826	9,415	9,043	9,389	9,609	**9,618**
Monopoly	9,419	9,507	9,432	9,492	9,473	9,515	9,550	**9,586**
Média	9,271	9,382	9,197	9,370	9,322	9,384	9,464	**9,497**
Desvio Padrão	0,182	0,231	0,163	0,158	0,175	0,162	0,159	0,162

Fonte: O Autor.

tados médios em ambas métricas. Com relação ao PSNR, considerando os 11 *datasets* empregados no comparativo, DHS obteve melhores resultados em 9, com ganho médio de $0,76db$ sobre o ATA e $1,93db$ sobre o FRHF. Já no caso do SSIM, os resultados se destacam ainda mais, pois o DHS só não obteve o melhor resultado para o *dataset* Aloe. Nesta métrica, a abordagem DHS apresentou um ganho médio de $0,0033$ e $0,0113$ sobre o ATA e o FRHF, respectivamente.

Um comparativo qualitativo entre os resultados obtidos pelas diferentes abordagens nos *datasets* de Middlebury é apresentado na Figura 6.1. Para auxiliar na inspeção visual, exibe-se a imagem projetada (*warped*) com os *holes* e alguns artefatos destacados em branco e a imagem real correspondente ao ponto de vista virtual (*ground truth*) para cada fotografia. Nos resultados para o *dataset* Bowling1, nas duas primeiras linhas, pode-se observar que ATA e DHS reconstruíram coerentemente a textura do *background*, sem a introdução de informação dos objetos no *foreground* (a bola verde e o pino de boliche), diferentemente de todas as outras abordagens. Nas próximas duas linhas, relativas a Flowerpots, *warped* possui pouco conteúdo relativo ao *background* para orientar o preenchimento dos *holes* (como pode ser observado no comparativo com o *ground truth*) e, mesmo assim, as abordagens propostas conseguiram reconstruir estas regiões adequadamente, o que comprova a eficácia do uso da imagem de referência como fonte de informação para a busca por *patches*. Neste mesmo caso de teste, destaca-se negativamente a reconstrução produzida por DHHF e VSRS, que apresentam um resultado similar a um borrão nas regiões reconstruídas pelos algoritmos de *inpainting*, mostrando sua incapacidade de reconstruir textura. Abaixo, nas duas últimas linhas, apresenta-se o resultado obtido para o *dataset* Aloe, para o qual a abordagem FRHF possui os melhores resulta-

Figura 6.1: Comparativo visual entre fotografias sintéticas produzidas com as abordagens avaliadas. Nas duas primeiras linhas são exibidos os resultados para o *dataset* Bowling1 e, em seguida, para Flowerpots e Monopoly, respectivamente. Na primeira imagem, apresenta-se o resultado da projeção para o ponto de vista virtual (*warped*) – com *holes* em branco e artefatos visíveis – e, na última, a imagem real (*ground truth*).



Fonte: O Autor, com imagens adaptadas de (HIRSCHMULLER; SCHARSTEIN, 2007).

dos quantitativos. O pequeno ganho obtido sobre as abordagens propostas não se reflete claramente em um comparativo visual pois, diferentemente do ATA e do DHS, o FRHF introduz conteúdo de *foreground* na região reconstruída.

6.2.2 Avaliação em Vídeos sem Informação Temporal

Os resultados quantitativos para os testes com vídeo – na média dos 100 primeiros quadros – são exibidos nas Tabelas 6.4 e 6.5, relativas as métricas PSNR e SSIM, respectivamente. A coluna *dataset* denota o nome do vídeo e, adicionalmente, informação de projeção. Por exemplo, o nome "BA41" se refere ao vídeo Ballet, com a projeção da vista de referência 4 para a virtual 1. Em ambas tabelas, os melhores resultados estão destacados em negrito. Neste ponto, se destaca a abordagem DHS, que apresenta a maior média de resultados considerando todos os casos de teste em ambas as métricas. No comparativo com as outras duas abordagens mais bem ranqueadas, ATA e FRBGE, DHS apresenta uma vantagem de $0,2db$ e $0,31db$ no PSNR, e no SSIM $0,0002$ e $0,0024$, respectivamente. Considerando os testes individuais, dentre os 14 vídeos sintéticos produzidos, DHS apresentou melhores resultados em 9 para o PSNR e 7 no SSIM. Observa-se que nos casos de teste onde as abordagens propostas não apresentam os melhores resultados quantitativos, estas aproximam-se muito do valor obtido pela concorrente mais bem classificada, justificando os bons resultados médios.

A Figura 6.2 exibe o primeiro quadro dos testes realizados com os vídeos de (ZIT-

Tabela 6.4: Comparativo entre as diferentes abordagens, na avaliação quantitativa da métrica PSNR, no contexto de vídeos. Apresenta-se em cada linha a média para os primeiros 100 quadros de cada caso de teste e o desvio padrão (DP) correspondente, para as abordagens dispostas nas colunas. Os melhores resultados encontram-se destacados em negrito.

Dataset	CRI	VSRS	DAR	AHN	DHHF	SHF		FRHF	FRBGE	ATA		DHS	
	PSNR	PSNR	PSNR	PSNR	PSNR	PSNR	DP	PSNR	PSNR	PSNR	DP	PSNR	DP
BA41	22,76	22,23	22,56	23,27	14,98	21,34	0,57	23,54	24,08	23,77	0,33	**24,13**	0,22
BA43	25,08	25,93	27,63	28,15	20,60	27,76	0,35	28,72	28,77	29,47	0,21	**29,54**	0,14
BA52	24,38	23,89	23,97	24,29	16,42	23,03	0,55	25,10	**25,89**	24,70	0,42	25,29	0,18
BA54	26,56	27,60	29,60	30,54	24,31	30,17	0,50	31,93	**32,06**	31,72	0,35	31,68	0,33
BR41	25,87	27,03	26,92	26,91	26,14	26,64	0,83	27,07	27,15	27,50	0,75	**27,73**	0,67
BR43	29,74	29,61	30,20	30,40	29,01	30,27	0,70	30,41	30,52	30,97	0,58	**31,07**	0,54
BR52	26,23	26,40	27,55	27,32	26,44	27,14	0,96	27,66	27,69	28,18	0,83	**28,41**	0,81
BR54	30,24	30,25	30,86	30,27	29,38	30,73	0,73	31,14	31,20	31,53	0,60	**31,71**	0,67
PH57	29,97	29,91	29,83	29,94	26,88	30,34	1,26	29,98	30,03	30,19	1,46	**30,36**	1,40
PH67	33,03	32,90	33,02	32,93	32,35	**33,59**	1,27	33,22	33,24	33,32	1,27	33,45	1,28
PS34	30,80	**31,75**	31,24	30,73	30,53	31,49	0,19	31,63	31,64	31,63	0,22	31,56	0,22
PS35	28,19	28,48	28,11	27,91	27,59	28,15	0,21	28,19	**28,51**	28,23	0,21	28,26	0,21
DA15	26,61	26,35	27,44	27,62	26,74	28,25	1,94	28,07	28,10	28,87	2,05	**29,07**	1,93
DA59	26,54	26,16	27,40	27,17	26,69	28,75	1,45	27,96	27,99	28,24	1,38	**28,90**	1,56
Média	27,57	27,75	28,31	28,39	25,58	28,40	0,82	28,90	29,06	29,17	0,76	**29,37**	0,73

Fonte: O Autor.

Tabela 6.5: Comparativo entre as diferentes abordagens, na avaliação quantitativa da métrica SSIM ($\times 10^{-1}$), no contexto de vídeos. Apresenta-se em cada linha a média para os primeiros 100 quadros de cada caso de teste e o desvio padrão (DP) correspondente, para as abordagens dispostas nas colunas. Os melhores resultados encontram-se destacados em negrito.

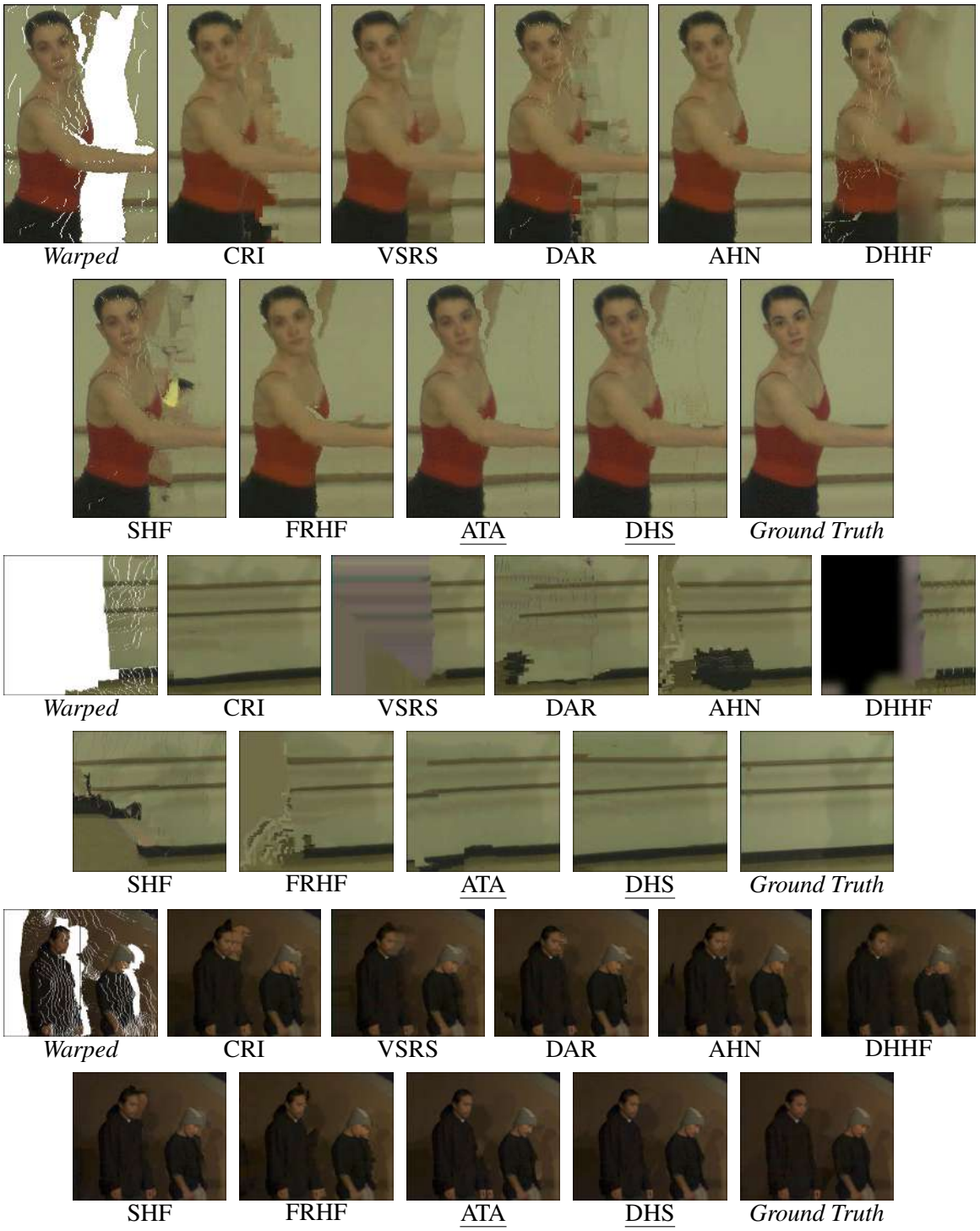
Dataset	CRI	VSRS	DAR	AHN	DHHF	SHF		FRHF	FRBGE	ATA		DHS	
	SSIM	SSIM	SSIM	SSIM	SSIM	SSIM	DP	SSIM	SSIM	SSIM	DP	SSIM	DP
BA41	7,391	7,611	7,130	7,478	6,355	6,670	0,221	7,551	7,661	7,674	0,034	**7,689**	0,036
BA43	8,388	8,514	8,351	8,465	7,917	8,393	0,028	8,501	8,517	**8,623**	0,013	8,607	0,011
BA52	7,422	7,654	7,200	7,385	6,429	6,947	0,083	7,546	**7,668**	7,523	0,045	7,551	0,038
BA54	8,468	8,584	8,448	8,545	8,181	8,524	0,025	8,634	8,646	**8,692**	0,018	8,667	0,020
BR41	7,639	7,814	7,639	7,737	7,537	7,433	0,186	7,750	7,731	7,825	0,054	**7,833**	0,052
BR43	8,197	8,126	8,151	8,223	8,006	8,169	0,052	8,227	8,240	8,250	0,032	**8,255**	0,029
BR52	7,660	7,606	7,675	7,737	7,480	7,571	0,088	7,762	7,768	7,864	0,054	**7,877**	0,051
BR54	8,217	8,133	8,177	8,225	7,995	8,199	0,050	8,234	8,243	8,292	0,035	**8,297**	0,033
PH57	8,506	8,632	8,511	8,511	8,559	8,717	0,114	8,669	8,671	8,720	0,118	**8,726**	0,117
PH67	8,743	8,871	8,754	8,741	8,908	**8,964**	0,073	8,916	8,917	8,962	0,072	**8,964**	0,072
PS34	8,579	8,807	8,671	8,571	8,829	8,845	0,009	8,842	8,839	**8,855**	0,009	8,853	0,009
PS35	8,148	8,278	8,174	8,128	8,206	8,188	0,026	8,244	**8,285**	8,251	0,023	8,266	0,023
DA15	9,220	9,425	9,160	9,253	9,411	9,477	0,042	9,444	**9,559**	9,487	0,048	9,478	0,047
DA59	9,214	9,412	9,230	9,244	9,415	**9,470**	0,024	9,436	9,437	9,467	0,034	9,466	0,028
Média	8,271	8,391	8,234	8,303	8,088	8,255	0,073	8,411	8,442	8,464	0,042	**8,466**	0,041

Fonte: O Autor.

NICK et al., 2004), em um comparativo entre as abordagens analisadas. Nas duas primeiras linhas, apresenta-se o resultado de BA43, com uma grande *disocclusion* a ser estimada ao lado do corpo da bailarina. Neste teste, fica evidente que abordagens que empregam informação de disparidade como AHN, FRHF, ATA e DHS, apresentam resultados mais coerentes do que algoritmos de *inpainting* convencionais, como CRI, VSRS e DHHF. Neste mesmo exemplo, destaca-se a presença de diversas ocorrências de *cracks* translúcidos não tratados (no corpo da bailarina) nos resultados de DAR, DHHF e SHF. Abaixo, apresenta-se o resultado produzido em parte de uma OOFA do caso de teste BA41, onde se destacam as reconstruções produzidas pelas abordagens propostas, em especial DHS. Como discutido anteriormente, estas regiões possuem características distintas das *disocclusions* e, utilizar a mesma abordagem em todos *holes*, faz com que artefatos sejam gerados, o que fica evidente em DAR, AHN, DHHF, SHF e FRHF. Nas últimas duas linhas apresenta-se o resultado para BR52, com inúmeras ocorrências de *cracks*, *disocclusions* e OOFA (na extremidade esquerda). Neste teste, em específico, exalta-se o resultado produzido pela abordagem DHS, que foi a única a não produzir prolongamentos no corpo dos atores, gerando uma imagem sintética sem artefatos visíveis.

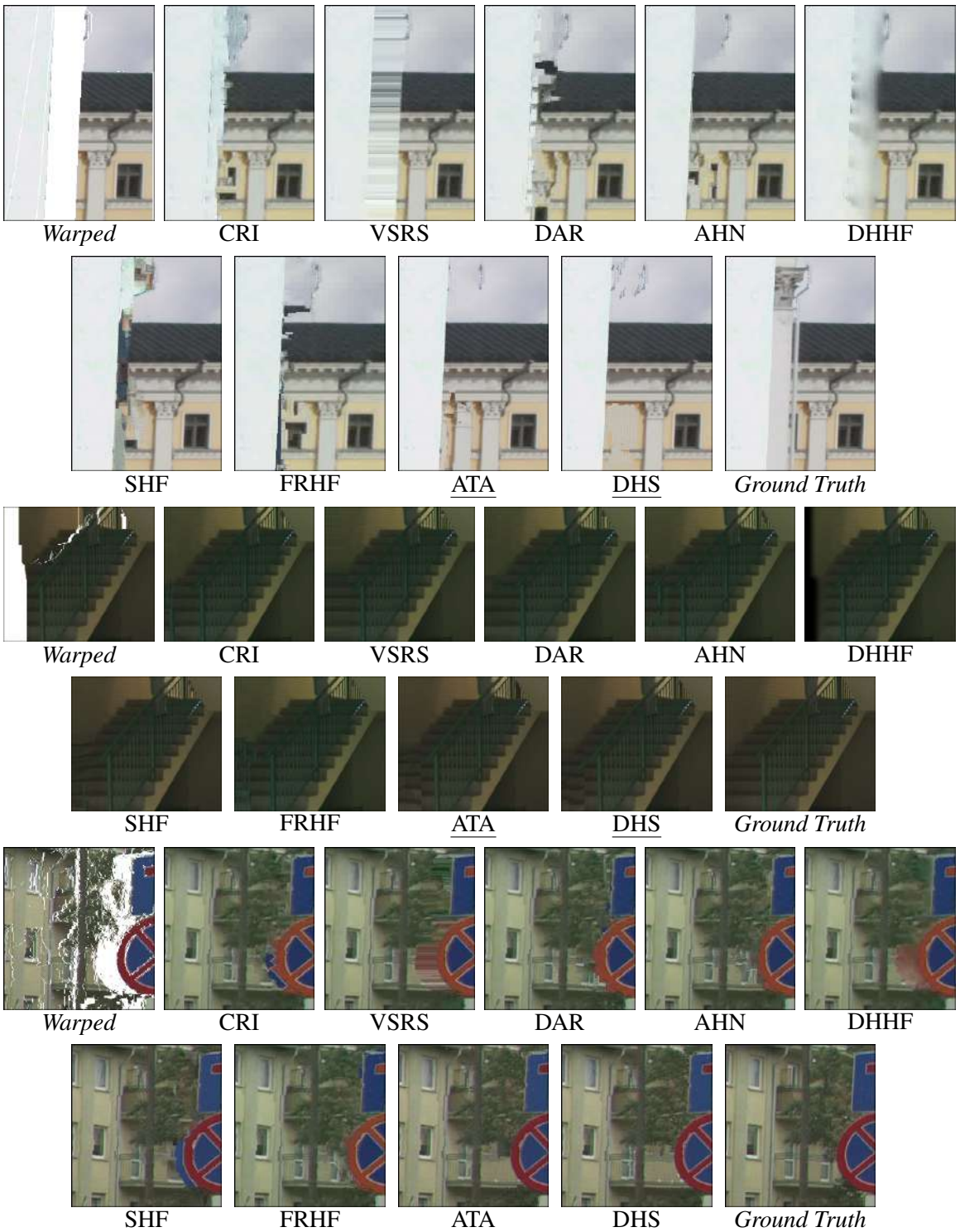
Na Figura 6.3, são exibidos os resultados para o primeiro quadro dos vídeos dos *datasets* de (SCHWARZ; MARPE; WIEGAND, 2010). Inicialmente, apresenta-se o resultado para Dancer, no teste DA15. Nesse caso, exibe-se em *warped* uma *disocclusion* em branco, que deve ser reconstruída com uma parte do prédio que sequer fica visível na vista de referência. Uma reconstrução exata (ou até mesmo aproximada), nestes casos,

Figura 6.2: Comparativo visual entre o resultado produzido pelas abordagens avaliadas, para o primeiro quadro de diferentes vídeos. Nas duas primeiras linhas são exibidos os resultados para o caso de teste BA43, em seguida BA41 e BR52 por último. Na primeira imagem, apresenta-se o resultado da projeção para o ponto de vista virtual (*warped*) – com *holes* em branco e artefatos visíveis – e, na última, a imagem real (*ground truth*).



Fonte: O Autor, com imagens adaptadas de (ZITNICK et al., 2004).

Figura 6.3: Comparativo visual entre o resultado produzido pelas abordagens avaliadas, para o primeiro quadro de diferentes vídeos. Nas duas primeiras linhas são exibidos os resultados para o caso de teste DA59 e, em seguida, para PH57 e PS35. Na primeira imagem, apresenta-se o resultado da projeção para o ponto de vista virtual (*warped*) – com *holes* em branco e artefatos visíveis – e, na última, a imagem real (*ground truth*).



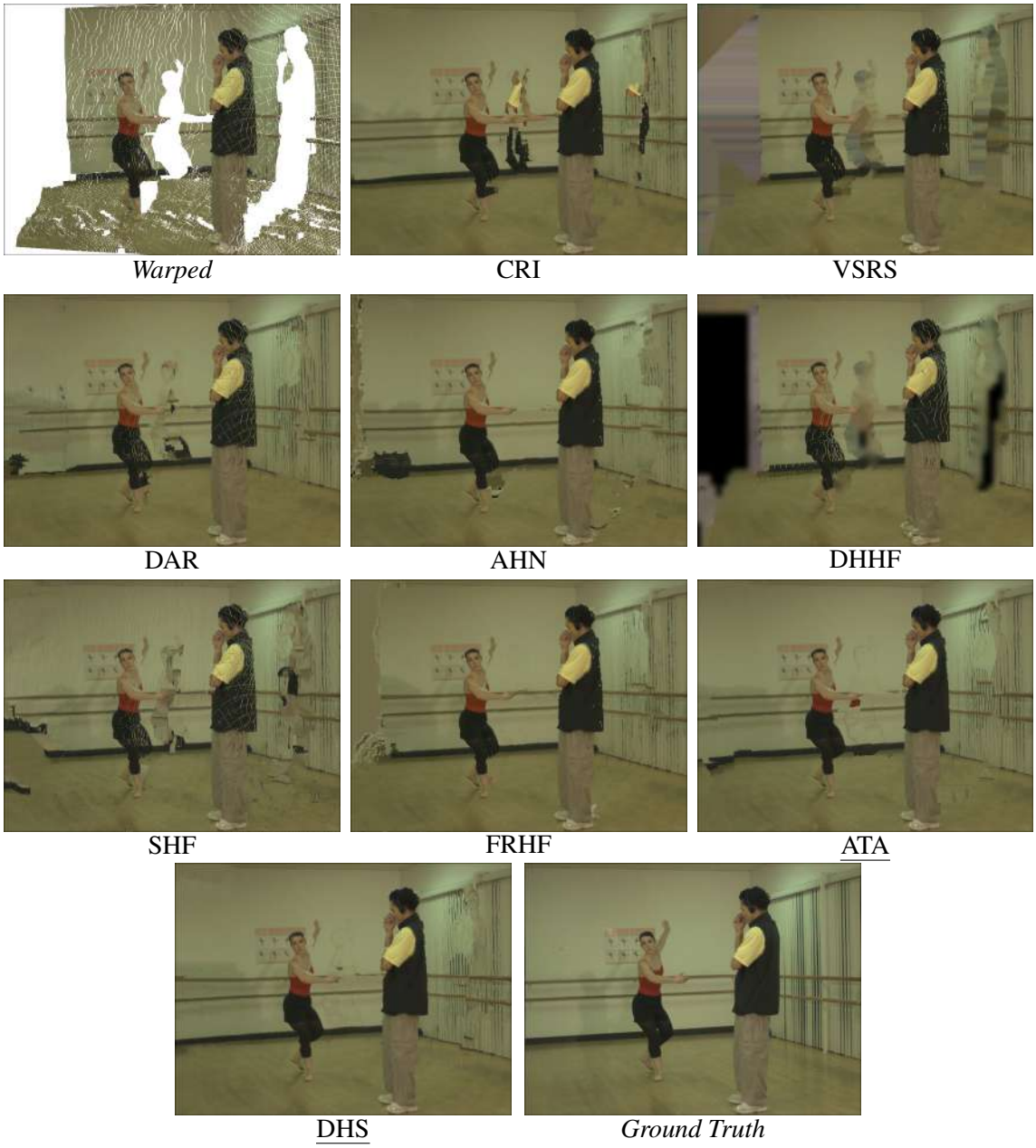
Fonte: O Autor, com imagens adaptadas de (SCHWARZ; MARPE; WIEGAND, 2010; RUSANOVSKYY; AFLAKI; HANNUKSELA, 2011).

torna-se praticamente impossível, o que fica evidente nos resultados. Portanto, o melhor que se pode produzir é uma aproximação do que se presume ser exibido nesta região, ou seja, um prolongamento do céu e do prédio. Posto isso, destaca-se o resultado produzido por ambas abordagens propostas, que reconstroem de forma aproximada ambas estruturas do *background*. Neste mesmo teste, destaca-se negativamente os resultados produzidos pelas demais abordagens, pois sequer remetem a alguma estrutura que poderia compor a região. No teste abaixo, PH57, dá-se enfase para a reconstrução da OOFA à esquerda, correspondente a uma parte da escada. Como pode ser visto, nenhum dos resultados produzidos foi totalmente adequado, e até mesmo extensões do corrimão em verde foram produzidas. Mesmo assim, DHS e DAR produzem o resultado que mais se aproxima do *ground truth*. Por fim, o caso de teste PS35 exibe algumas *disocclusions* que precisam ser reconstruídas, em uma região que contém diversos elementos. Neste caso, destaca-se novamente o resultado produzido por ATA e DHS, que realizaram uma reconstrução coerente, sem incluir sombras e artefatos oriundos da placa no *foreground*, como ocorre com as outras abordagens.

Na Figura 6.4, apresenta-se um comparativo de resultado para o caso de teste BA41, o qual emprega um *baseline* pouco usual, devido ao seu tamanho. Neste teste, tem-se como objetivo analisar o desempenho das abordagens DIBR, quando *holes* são revelados em maiores proporções na vista sintética. No comparativo, exibe-se o primeiro quadro do vídeo, para o qual se faz necessário reconstruir $34,22\%$ do conteúdo da imagem, como destacado em branco na figura projetada (*warped*). Nos resultados, pode-se notar que CRI, em específico, produz diversas extensões do corpo dos atores. Já VSRS e DHHF, produzem resultados incoerentes com o *ground truth*, similares a um borrão. Em DAR, DHHF e SHF, fica evidente a necessidade da existência de uma técnica para a remoção de *cracks* translúcidos. Neste sentido, apesar de Luo and Zhu (2017) não proporem mecanismos para a remoção deste artefato no FRHF, não são apresentadas ocorrências em seus resultados, o que evidência o uso de algum pré-processamento de suas imagens. Os resultados de AHN, assim como de DAR, SHF e FRHF, deixam clara a necessidade do uso de uma abordagem específica para o preenchimento das OOFAs, considerando os resultados produzidos por seus algoritmos nestas regiões. Com uma simples inspeção visual, pode-se notar que os resultados apresentados em ATA e DHS se assemelham mais com o *ground truth*. Isto demonstra que apesar de não ser perfeita, a reconstrução produzida por ambas se mostra mais adequada que a dos trabalhos comparados, independentemente da distância determinada no *baseline*, o que justifica os resultados quantitativos apresen-

tados.

Figura 6.4: Comparativo visual do resultado produzido pelas abordagens avaliadas, utilizando *baseline* grande, no caso de teste BA41. Na primeira imagem, apresenta-se o resultado da projeção para o ponto de vista virtual (*warped*) – com *holes* em branco e artefatos visíveis – e, na última, a imagem real (*ground truth*).



Fonte: O Autor, com imagens adaptadas de (ZITNICK et al., 2004).

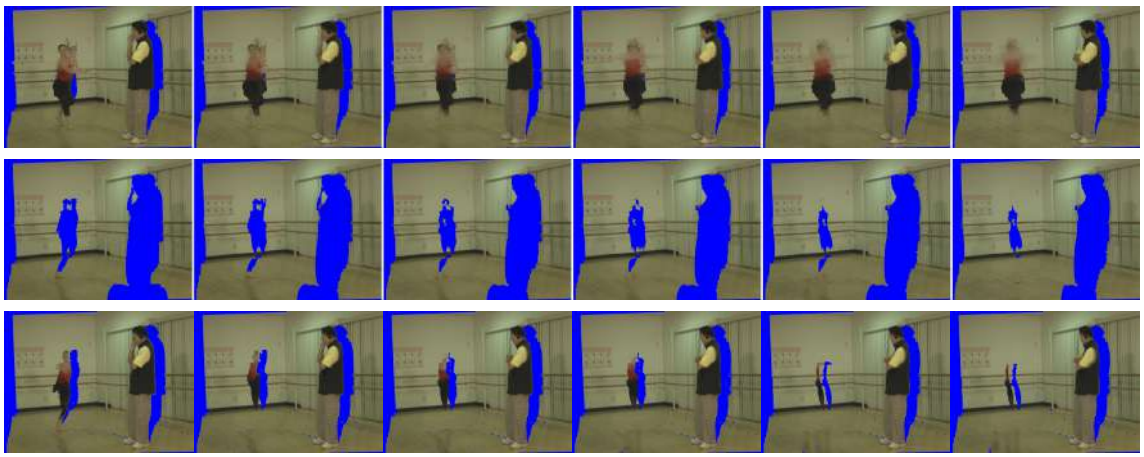
6.2.3 Avaliação em Vídeos com Informação Temporal

Como discutido anteriormente, métodos que exploram informação temporal podem estabelecer diferentes estratégias para utilizar o conteúdo de quadros passado e/ou futuros para o preenchimento de *holes* e manutenção da consistência temporal em vídeos. Dentre as estratégias adotadas, a mais difundida se baseia na formulação de um modelo (ou imagem) base, que acumula conteúdo do *background* ao longo dos quadros. Neste caso, o modelo é incluído no *pipeline* DIBR, como parte do processo de preenchimento das *disocclusions* normalmente.

Modelos de *background* podem conter informação de *foreground* ou não, dependendo da estratégia adotada. No método proposto por exemplo, permite-se que algumas partes do *foreground* permaneçam no modelo até que informação de *background* seja exibida no mesmo local. Na Figura 6.5, apresenta-se um comparativo entre os modelos de *background* produzidos com o método proposto e com duas variações de um GMM modificado, que fazem parte do *framework* de (LUO et al., 2019). Os modelos exibidos na primeira linha contêm elementos de *foreground*, assim como os gerados com a abordagem proposta (exibidos na última linha). Nos resultados, pode-se perceber que os modelos produzidos com o GMM contêm muito mais conteúdo de *foreground* do que os gerados com o método proposto, principalmente no corpo da atriz. Na segunda linha da figura, exibem-se os modelos produzidos com o mesmo GMM, mas com a remoção de objetos do *foreground*. Como pode ser visto nas imagens, a remoção não foi tão bem sucedida, pois parte do pé da atriz foi mantida nos modelos. Quando problemas como este ocorrem, abordagens DIBR tendem a ser induzidas ao erro (seja na etapa de cópia de informação ou no processo de *inpainting*), porque consideram que o modelo é composto somente por conteúdo do *background*. Por outro lado, isto não ocorre com métodos que mantém os elementos de *foreground* no modelo, pois estes sempre precisam validar o conteúdo antes da cópia, para evitar erros.

Para validar o método proposto, foi desenvolvido um processo que tem como objetivo integrar modelos de *background* com a abordagem DHS (descrito na Seção 5.4). Esta integração gera uma nova abordagem DIBR, que emprega informação temporal para preencher parcial ou totalmente *disocclusions*, referenciada como DHS+B no restante do texto. Para avaliar os resultados produzidos por DHS+B, foram repetidos os testes em vídeos, mas desconsiderando os *datasets* que foram capturados com o uso de movimentação de câmera, porque a abordagem proposta não possui mecanismos de compensação de

Figura 6.5: Comparativo visual entre diferentes abordagens destinadas a construção de modelos de *background*, em teste no *dataset Ballet* com a projeção da vista 5 para a 4, nos quadros 10, 20, 30, 40, 50 e 60, exibidos em ordem nas colunas. Na primeira e segunda linhas são apresentados os resultados para o GMM de Luo et al. (2019) sem e com a remoção de *foreground*, respectivamente. Na última linha, são exibidos os resultados obtidos pela abordagem proposta.



Fonte: O autor e Luo et al. (2019).

movimento, como exposto anteriormente. Com os resultados, formulou-se a Tabela 6.6, que exibe um comparativo entre DHS+B e diferentes abordagens que empregam informação temporal no processo de síntese, utilizando a métrica PSNR. Como pode ser visto, a abordagem proposta obteve o maior valor médio, apresentando melhores resultados em 5 dos 8 casos de tese, superando trabalhos muito recentes (como PNT3 e PNT5). Neste comparativo, o ganho médio sobre a segunda melhor abordagem foi de $0,17db$.

Na Tabela 6.7, apresentam-se os resultados obtidos pelas diferentes abordagens que utilizam informação temporal para a métrica SSIM. Nos resultados, pode-se observar

Tabela 6.6: Comparativo entre as diferentes abordagens, na avaliação quantitativa da métrica PSNR, no contexto de vídeos com informação temporal. Apresenta-se em cada linha a média para os primeiros 100 quadros de cada caso de teste e o desvio padrão (DP) correspondente, para as abordagens dispostas nas colunas. Os melhores resultados encontram-se destacados em negrito.

Dataset	YAO	NEWSON	LUO	PNT3	PNT5	DHS+B	
	PSNR	PSNR	PSNR	PSNR	PSNR	PSNR	DP
BA41	23,05	22,31	24,15	**24,33**	23,75	24,22	0,15
BA43	25,61	24,70	28,86	29,31	29,33	**29,61**	0,14
BA52	24,81	24,08	**25,76**	25,33	25,14	25,47	0,14
BA54	27,53	26,09	32,00	**32,06**	31,77	32,03	0,22
BR41	27,09	27,97	27,30	27,59	27,29	**27,99**	0,57
BR43	30,53	30,62	30,60	30,75	30,77	**31,00**	0,55
BR52	27,72	28,81	27,85	28,55	27,73	**28,82**	0,66
BR54	31,50	31,55	**31,73**	31,62	31,37	**31,73**	0,57
Média	27,23	27,02	28,53	28,69	28,39	**28,86**	0,38

Fonte: O Autor.

Tabela 6.7: Comparativo entre as diferentes abordagens, na avaliação quantitativa da métrica SSIM ($\times 10^{-1}$), no contexto de vídeos com informação temporal. Apresenta-se em cada linha a média para os primeiros 100 quadros de cada caso de teste e o desvio padrão (DP) correspondente, para as abordagens dispostas nas colunas. Os melhores resultados encontram-se destacados em negrito.

Dataset	YAO	NEWSON	LUO	PNT3	PNT5	DHS+B	
	SSIM	SSIM	SSIM	SSIM	SSIM	SSIM	DP
BA41	7,496	7,618	7,759	7,786	**7,795**	7,690	0,031
BA43	8,429	8,440	8,570	8,589	8,599	**8,623**	0,014
BA52	7,563	7,649	7,768	7,756	**7,826**	7,587	0,027
BA54	8,524	8,485	8,665	8,657	8,662	**8,705**	0,018
BR41	7,694	**7,879**	7,763	7,875	7,754	7,858	0,041
BR43	8,227	8,235	8,248	8,248	8,236	**8,253**	0,028
BR52	7,712	7,895	7,776	7,885	7,740	**7,904**	0,040
BR54	8,255	8,259	8,275	8,214	8,246	**8,295**	0,032
Média	7,988	8,058	8,103	**8,126**	8,107	8,114	0,029

Fonte: O Autor.

que em 5 dos 8 casos de teste a abordagem proposta obteve o maior SSIM médio. Já na média geral, o maior valor foi obtido pela abordagem PNT3, que superou o DHS+B por $0,001$. Considerando os testes individuais, PNT3 produziu resultados mais relevantes para BA52 e BA41, o que aumentou a média da abordagem, fazendo com que esta superasse DHS+B. Contudo, considerando os testes individuais, tanto em razão do SSIM quanto do PSNR a abordagem proposta apresentou melhores resultados que PNT3 e todos os demais trabalhos. Observa-se que um comparativo visual entre o resultado das diferentes abordagens não pode ser disponibilizado, uma vez que as imagens geradas com os métodos comparados não foram disponibilizadas pelos autores.

Para complementar a avaliação do DHS+B, produziu-se um comparativo quantitativo e qualitativo entre os resultados da abordagem e os produzidos por DHS, SHF e ATA. Considerando que DHS+B corresponde a uma extensão do DHS que emprega informação temporal, decidiu-se analisar os resultados produzidos por ambas (e pelas outras abordagens) com uma métrica própria para quantificar a qualidade de vídeos não somente no domínio espacial (como ocorre com PSNR e SSIM) mas também temporal. Para tanto, optou-se pela métrica STRRED, com base na qual foi formulada a Tabela 6.8, que exibe os valores computados para cada uma das abordagens nos casos de tese. Como esperado, o DHS+B apresenta os melhores resultados no comparativo, superando as demais abordagens em 9 dos 10 comparativos realizados. Neste teste, a extensão proposta para o DHS não superou a abordagem original apenas no caso de teste BA52, mas mesmo assim ainda apresentou melhores resultados que SHF e ATA. Devido ao fato de o DHS+B não possuir mecanismos para compensação de movimento, não foi possível produzir resultados para 6 casos de teste exibidos na tabela. Para estes testes, DHS se mostrou melhor que as demais

Tabela 6.8: Comparativo entre as diferentes abordagens, na avaliação quantitativa da métrica STRRED, no contexto de vídeos. Apresenta-se em cada linha a média considerando os primeiros 100 quadros de cada caso de teste e o desvio padrão (DP) correspondente, para cada uma das abordagens dispostas nas colunas. Os melhores resultados encontram-se destacados em negrito.

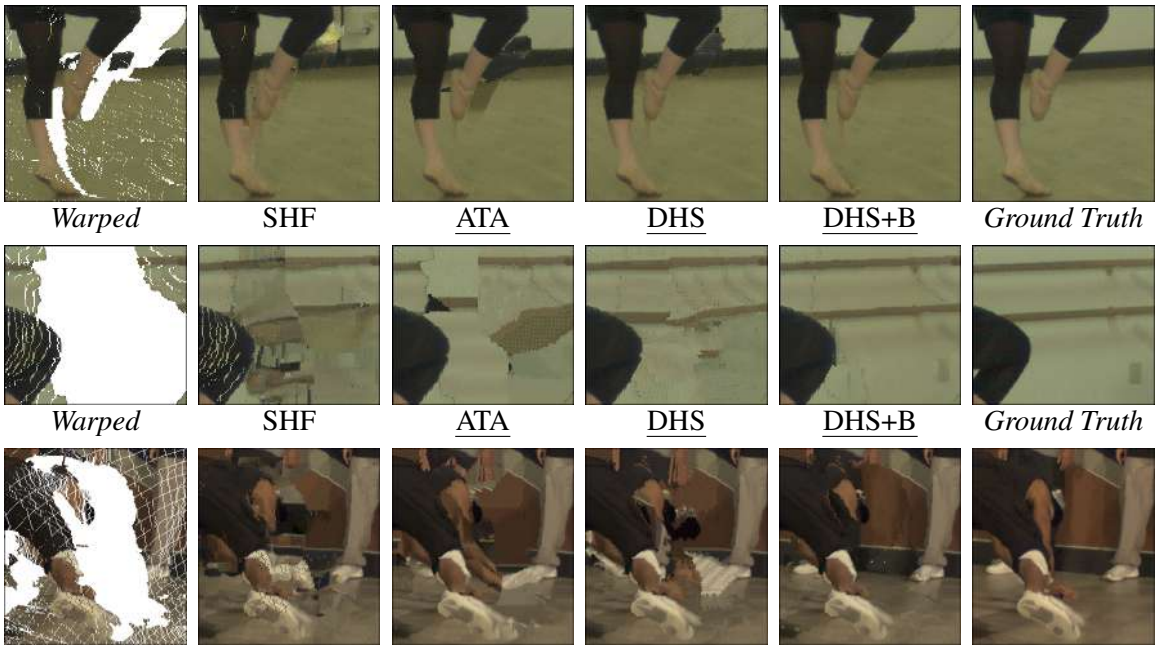
Dataset	SHF		ATA		DHS		DHS+B	
	STRRED	DP	STRRED	DP	STRRED	DP	STRRED	DP
BA41	2199,16	574,88	720,07	187,07	501,63	124,39	**462,61**	92,62
BA43	297,61	84,59	95,08	32,79	87,85	29,74	**68,67**	25,19
BA52	1598,46	415,43	783,61	218,24	**491,22**	113,94	531,08	142,30
BA54	247,02	89,83	78,84	29,97	73,79	25,68	**54,88**	19,74
BR41	486,43	189,46	341,28	137,68	299,74	122,19	**222,49**	76,65
BR43	109,73	50,31	75,20	30,69	68,15	26,46	**66,23**	23,45
BR52	475,91	201,45	306,90	124,99	270,54	111,49	**204,09**	65,15
BR54	121,18	42,89	82,26	29,76	72,82	28,19	**68,37**	22,55
PH57	**139,41**	40,17	158,69	49,73	159,66	50,89	–	–
PH67	**50,08**	20,02	57,08	21,09	61,88	25,99	–	–
PS34	**20,83**	5,36	21,26	5,97	21,40	5,35	–	–
PS35	73,69	17,30	70,96	14,52	**66,79**	15,44	–	–
DA15	61,87	43,15	50,32	43,65	**42,14**	31,74	–	–
DA59	68,85	34,44	58,16	37,07	**46,06**	23,77	–	–

Fonte: O Autor.

abordagens em PS35, DA15 e DA59, e SHF em PH57, PH67 e PS34.

Para ilustrar a diferença entre o resultado produzido pela abordagem DHS+B e as demais, foi elaborado um comparativo visual, exibido na Figura 6.6. Nas imagens, pode-se perceber que os resultados produzidos por DHS+B tendem a ser mais similares ao *ground truth*. Isto se dá principalmente pelo fato da abordagem diminuir a possibilidade de que artefatos possam ser produzidos pelo algoritmo de preenchimento, por reconstruir grande boa parte das *disocclusions* com cópia de informação real proveniente do modelo de *background*. As imagens produzidas para BA52 tornam evidente a vantagem trazida por esta extensão, pois ao ser criado um *hole* na parede, onde um espelho de tomada é exibido no *ground truth*, o DHS+B foi a única abordagem que conseguiu reproduzir o item. Já as outras abordagens, mesmo com um algoritmo de *inpainting* muito robusto, não poderiam sequer recriar este item, uma vez que ele não está visível em parte alguma da imagem de referência. Entretanto, mesmo com a inclusão de informação temporal, erros de reconstrução ainda podem ser visualizados nos resultados, como no caso de teste BA52, próximo do joelho da atriz. Isto justifica em parte o resultado quantitativo obtido para este mesmo caso de teste. Outro problema que justifica o resultado ruim está associado à falta de padrão e aos defeitos encontrados nos *datasets* de (ZITNICK et al., 2004), que prejudicam o processo de reatribuição dos rótulos de *superpixel* para as regiões copiadas. Neste caso, os critérios de estimativa de similaridade adotados pelo processo de redistribuição dos rótulos são deturpados, considerando que a mesma região pode ter valores de cor e disparidade diferentes em quadros consecutivos.

Figura 6.6: Comparativo visual entre o resultado produzido pelas abordagens avaliadas no contexto temporal. Na primeira linha, exibe-se o resultado para o caso de teste BA43 (no quadro 57), na seguinte para BA52 (quadro 90) e na última para BR52 (quadro 38). Na primeira imagem, apresenta-se o resultado da projeção para o ponto de vista virtual (*warped*) – com *holes* em branco e artefatos visíveis – e, na última, a imagem real (*ground truth*).



Fonte: O Autor, com imagens adaptadas de (ZITNICK et al., 2004).

6.3 Avaliação com Mapas de Disparidade Produzidos por Algoritmos de casamento estéreo

Nesta seção, apresenta-se um estudo comparativo que tem como objetivo avaliar a qualidade das vistas sintetizadas por diferentes abordagens DIBR quando alimentadas com mapas de disparidade realísticos, gerados com algoritmos de casamento estéreo. Além disso, analisa-se a correlação entre métricas de avaliação usadas para síntese de vistas e casamento estéreo. Para produzir resultados consistentes, foram selecionados cinco algoritmos de casamento estéreo (YIN et al., 2017; MOZEROV; WEIJER, 2015; TANIAI et al., 2018; ZHANG et al., 2015; MOZEROV; WEIJER, 2019), ranqueados de acordo com quatro métricas comumente usadas (HIRSCHMULLER; SCHARSTEIN, 2007). No contexto de DIBR, comparou-se ATA e DHS com outros três trabalhos (AHN; KIM, 2013; SOLH; ALREGIB, 2012b; OLIVEIRA et al., 2015). Parte dos resultados produzidos pela análise descrita nesta seção geraram o artigo "*On the Performance of DIBR Methods When Using Depth Maps From State-of-the-Art Stereo Matching Algo-*

rithms" que foi apresentado na conferência *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) em maio de 2019. Na literatura foram publicados dois trabalhos com análises similares (LU; YANG; LAFRUIT, 2009; FUHR et al., 2013), focados em interpolação de vistas, os quais foram detalhados na Subseção 2.2.3. Algumas das conclusões obtidas por estes autores corroboram com o que ocorre no cenário de extrapolação de vistas com DIBR, como será discutido mais a frente juntamente com os resultados produzidos por esta análise.

6.3.1 Seleção dos Algoritmos de casamento estéreo

Há um número crescente de algoritmos de casamento estéreo, e selecionar aquele que tem melhor desempenho para uma dada tarefa não é trivial (FUHR et al., 2013). Além disso, diferentes métricas de erro quantitativo capturam diferentes tipos de erros. Neste trabalho, foram selecionadas quatro métricas para casamento estéreo usadas no conhecido benchmark Middlebury (HIRSCHMULLER; SCHARSTEIN, 2007), `bad 2.0`, `avgerr`, `rms` e `a95`. Elas formam um subconjunto representativo de todas as métricas em Middlebury, favorecendo diferentes tipos de erros na correspondência.

Erros com base na porcentagem de *pixels* "ruins" classificam os *pixels* em bons ou ruins com base em um limite de aceitação para as diferenças de disparidade. *Pixels* ruins têm a mesma penalidade, independentemente do quão longe eles estão do valor real. Esse é o caso da métrica `bad 2.0`, para a qual se define o limite como $2,0$. Por outro lado, somas/médias de erros absolutos ou quadrados penalizam cada vez mais as estimativas longe do *ground truth*, de modo que poucos *pixels* ruins podem corromper o resultado. A métrica `avgerr` captura o erro absoluto médio, `a95` o percentil 95 da distribuição do erro e `rms` o erro do RMS (*root mean square*), todos em *pixels*. Foram classificados os algoritmos listados no índice de referência[5] em ordem crescente, na qual a pontuação do ranking se dá pela soma das respectivas posições de acordo com `bad 2.0`, `avgerr`, `rms` e `a95` para o conjunto denso de teste e desconsiderando máscaras de remoção de oclusão. Quanto menor a pontuação da métrica combinada, melhor. Para gerar a avaliação, foram selecionados os cinco algoritmos de melhor desempenho e com código-fonte disponível. De acordo com a métrica combinada, os trabalhos obtiveram a seguinte classificação:

[5] Algoritmos classificados utilizando <http://vision.middlebury.edu/stereo/eval3> em 29/10/2018.

3º – O método de otimização global baseado em *Markov Random Fields* (MRF) de Taniai et al. (2018);

6º – A extensão da técnica padrão de *Belief Propagation* (BP) proposta em (KOLMOGOROV, 2006) para modelos totalmente conectados *conditional random field* (CRF) com afinidade por distância geodésica de Mozerov and Weijer (2019);

21º – O algoritmo de casamento estéreo global que se baseia na triangulação 2D da vista de referência para estimativa de disparidade de Zhang et al. (2015);

39º – A abordagem de correspondência de custos baseada em *dictionary learning* para comparativo de *patches* de imagens de Yin et al. (2017);

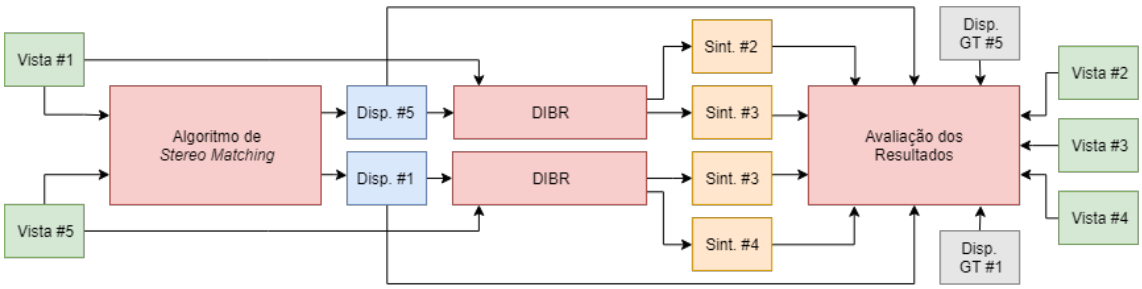
42º – O método de Mozerov and Weijer (2015), que explora tanto o potencial de filtragem de custos quanto a minimização de energia.

6.3.2 Metodologia Adotada na Avaliação

Para produzir os resultados, avalia-se o desempenho de todos os métodos DIBR selecionados usando a disparidade estimada por todos os algoritmos de casamento estéreo considerados (juntamente com a disparidade *ground truth*). Os parâmetros dos métodos foram definidos de acordo com os respectivos artigos ou, se ausentes, de acordo com os códigos fonte disponibilizados. São necessárias pelo menos três vistas para cada teste, duas utilizadas na estimativa das disparidades e uma terceira para avaliar a vista sintética.

A Figura 6.7 descreve o *pipeline* empregado na avaliação experimental. Para alimentar os algoritmos de casamento estéreo, são utilizadas as vistas 1 e 5 (as quais possuem *ground truth*), produzindo dois mapas de disparidades, um para cada vista de referência. Juntamente com a imagem colorida, enviam-se os mapas de disparidade estimados para cada abordagem DIBR. Então, com base na informação (V+D) de 1 são produzidas as vistas sintéticas 2 e 3, e em 5, são geradas 4 e 3, formando ao todo imagens para quatro pontos de vista virtuais de cada conjunto de imagens. Com isso, torna-se possível avaliar a qualidade de ambos os mapas de disparidade e vistas sintetizadas em relação aos respectivos *ground truths*.

Figura 6.7: *Pipeline* empregado na avaliação experimental das abordagens DIBR alimentadas por mapas de disparidade produzidos por diferentes algoritmos de casamento estéreo. Na ilustração, "Vista" refere-se a imagem de referência, "Disp." a mapas de disparidade gerados com algoritmos de casamento estéreo, "Sint." a vistas sintéticas produzidas com abordagens DIBR e, por fim, "Disp. GT" corresponde aos mapas de disparidade *ground truth*.



Fonte: O Autor.

6.3.3 Resultados e Discussão

Com base no *pipeline* definido na Subseção 6.3.2, pode-se comparar os resultados variando as técnicas (i) casamento estéreo e (ii) DIBR, avaliando-as por meio de métricas como (iii) `bad 2.0`, `avgerr`, `rms` e `a95`, e (iv) PSNR, SSIM e MW-PSNR. Observa-se a relação entre essas variáveis formam um hiper-cubo quadridimensional. Além disso, há duas dimensões adicionais: os conjuntos de imagens de (HIRSCHMULLER; SCHARSTEIN, 2007) e os dois mapas de disparidade e as quatro vistas sintetizadas para cada conjunto de imagens.

A Tabela 6.9 apresenta resultados médios de PSNR, SSIM e MW-PSNR (para as vistas de $21 \times 4 = 84$ *datasets*) para todas as abordagens DIBR e algoritmos de casamento estéreo, além do mapa de disparidade *ground truth*. Os melhores resultados para cada métrica em cada algoritmo de casamento estéreo estão destacados em negrito e o maior valor geral em itálico. Como pode ser observado, o uso de mapas de disparidade *ground truth* apresenta melhores resultados com a abordagem DHS. Entretanto, este caso representa uma exceção, e isto pode ser comprovado com uma análise mais atenta dos resultados quantitativos obtidos por todas as outras abordagens, onde os algoritmos de casamento estéreo superaram o *ground truth* em todas as métricas. Os resultados produzidos por DHS juntamente com os métodos de casamento estéreo (YIN et al., 2017; MOZEROV; WEIJER, 2019; TANIAI et al., 2018) são muito próximos do obtido com o *ground truth*, o que reforça a qualidade das vistas produzidas com mapas reais.

Ao contrário do que ocorre com os algoritmos de casamento estéreo seleciona-

Tabela 6.9: Média dos resultados no cenário real nas métricas PSNR, SSIM e MW-PSNR, respectivamente. As colunas apresentam os algoritmos de casamento estéreo, o mapa de disparidade *ground truth* (GT) e a média, enquanto as linhas triplas delimitam as abordagens DIBR.

	TSGO	DDL	MeshStereo	OVOD	LocalExp	GT	Média
	24,334	31,960	16,883	31,143	30,918	29,273	27,333
HHF	**0,7276**	**0,9532**	0,5331	**0,9527**	**0,9536**	0,9472	0,8324
	26,258	32,926	20,207	32,593	32,025	31,475	28,242
	24,548	31,335	18,254	31,024	30,128	28,713	27,418
AHN	0,7174	0,9393	0,5258	0,9427	0,9402	0,9290	0,8446
	26,225	30,892	20,363	31,419	30,478	30,075	29,247
	24,685	32,196	18,230	31,793	31,811	30,066	28,130
SHF	0,7229	0,9450	0,5314	0,9496	0,9499	0,9422	0,8402
	25,881	28,854	20,080	30,050	28,951	29,649	27,244
	24,859	**32,626**	18,292	32,052	32,041	31,721	28,599
ATA	0,7236	0,9512	0,5368	0,9517	0,9531	0,9522	0,8448
	26,435	32,848	20,432	32,717	32,330	32,609	29,562
	24,825	32,577	**18,307**	**32,220**	**32,176**	***32,879***	**28,831**
DHS	0,7235	0,9503	**0,5375**	0,9519	0,9528	***0,9559***	**0,8453**
	26,506	**33,352**	**20,462**	**33,117**	**32,885**	***33,499***	**29,970**

Fonte: O Autor.

dos, os mapas *ground truth* não têm valores de disparidade em regiões classificadas como desconhecidas (HIRSCHMULLER; SCHARSTEIN, 2007), ou seja, regiões sem correspondências nas vistas esquerda e direita. Portanto, na prática, os métodos DIBR têm muito mais *pixels* para estimar ao usar o mapa de disparidades *ground truth*, o que pode justificar em parte este resultado.

Considerando as médias dos resultados produzidos por todas as abordagens DIBR, DHS e ATA destacam-se em primeiro e segundo lugares, na classificação de resultados das três métricas, indicando que independentemente do mapa de disparidades fornecido como entrada, estas produzem resultados quantitativos satisfatórios. Isto indica que ambas não possuem relação de dependência com mapas de disparidade *ground truth*, ou funcionam bem apenas quando empregados junto a algum algoritmo de casamento estéreo específico.

A ordem de classificação relativa entre os algoritmos casamento estéreo considerados de acordo com `bad 2.0`, `avgerr`, `rms` e a pontuação combinada de todas as métricas analisadas é a mesma apresentada na Subseção 6.3.1. A ordem de classificação baseada na métrica `a95` troca o primeiro e o segundo algoritmos de melhor desempenho, e também os dois últimos. Entretanto, se forem usadas as métricas de síntese de vistas para classificação, são definidas ordens diferentes, como pode ser visto na Tabela 6.9. Especificamente, o método em (ZHANG et al., 2015) apresentou um desempenho pior,

provavelmente porque o código disponibilizado é uma aproximação do método real publicado.

Foi investigada ainda, a relação entre as métricas casamento estéreo e DIBR. A Tabela 6.10 apresenta a mais forte correlação de Spearman (SPEARMAN, 1904) e as técnicas correspondentes de casamento estéreo e DIBR, para todas as combinações de métricas. Pode-se notar que os resultados sugerem que as métricas `bad 2.0` e MW-PSNR têm um relacionamento negativo bastante forte. Esta análise também indica que não se espera ter valores de SSIM e PSNR necessariamente maiores para vistas sintetizadas quando são escolhidos algoritmos de casamento estéreo que minimizam as métricas de erro `bad 2.0`, `avgerr`, `rms` e `a95`. As descobertas estão de acordo com as mostradas em (FUHR et al., 2013) para o cenário de interpolação de vista.

Os tipos de artefatos DIBR relatados na literatura estão relacionados aos mapas de disparidade *ground truth*. Portanto, os mapas de disparidade baseados em casamento estéreo podem diferir, levando as abordagens de DIBR a apresentar resultados contrastantes. Entretanto, com uma inspeção visual, pode-se notar que os *cracks*, *ghosts* e OOFAs apresentaram os mesmos padrões relatados na literatura (e descritos na Seção 2.1), embora os dois primeiros apareçam mais intensamente. No entanto, regiões de *disocclusion* estão contaminadas com camadas de disparidade sobre-segmentadas devido à estimativa de disparidade imprecisa, produzindo um efeito de distorção, especialmente nos objetos de *foreground* e suas bordas. Este efeito, tende a prejudicar pouco abordagens como HHF, que se baseiam na estimativa de sucessivas médias. Diferentemente, para técnicas baseadas em *patches* AHN, SHF, ATA e DHS, este problema pode produzir muitos artefatos incoerentes. Essas técnicas executam o processo de reconstrução usando um *patch* como um modelo (selecionado a partir da borda do *hole* alvo), para a busca e cópia de conteúdo de informação válida da vista sintética. As abordagens baseadas em *patches* classificam as bordas de *disocclusions* com base no mapa de disparidades projetado para definir modelos compostos por informação de *background*. No entanto, a classificação em si pode

Tabela 6.10: Análise da correlação para métricas de casamento estéreo e síntese de vistas. Os nomes dentro das células indicam os métodos para os quais a correlação máxima foi alcançada.

	`bad 2.0`	`avgerr`	`rms`	`a95`
PSNR	$-0,39$ [OVOD,SHF]	$-0,37$ [OVOD,SHF]	$-0,34$ [MS,SHF]	$-0,40$ [OVOD,SHF]
SSIM	$-0,40$ [MS,SHF]	$-0,47$ [MS,AHN]	$-0,44$ [MS,AHN]	$-0,33$ [MS,AHN]
MW-PSNR	$-0,80$ [LE,SHF]	$-0,74$ [LE,SHF]	$-0,68$ [LE,SHF]	$-0,65$ [LE,SHF]

Fonte: O Autor.

não ser satisfatória, pois a disparidade não apresenta consistência em relação a textura, como ilustrado nas Figuras 6.8 e 6.9.

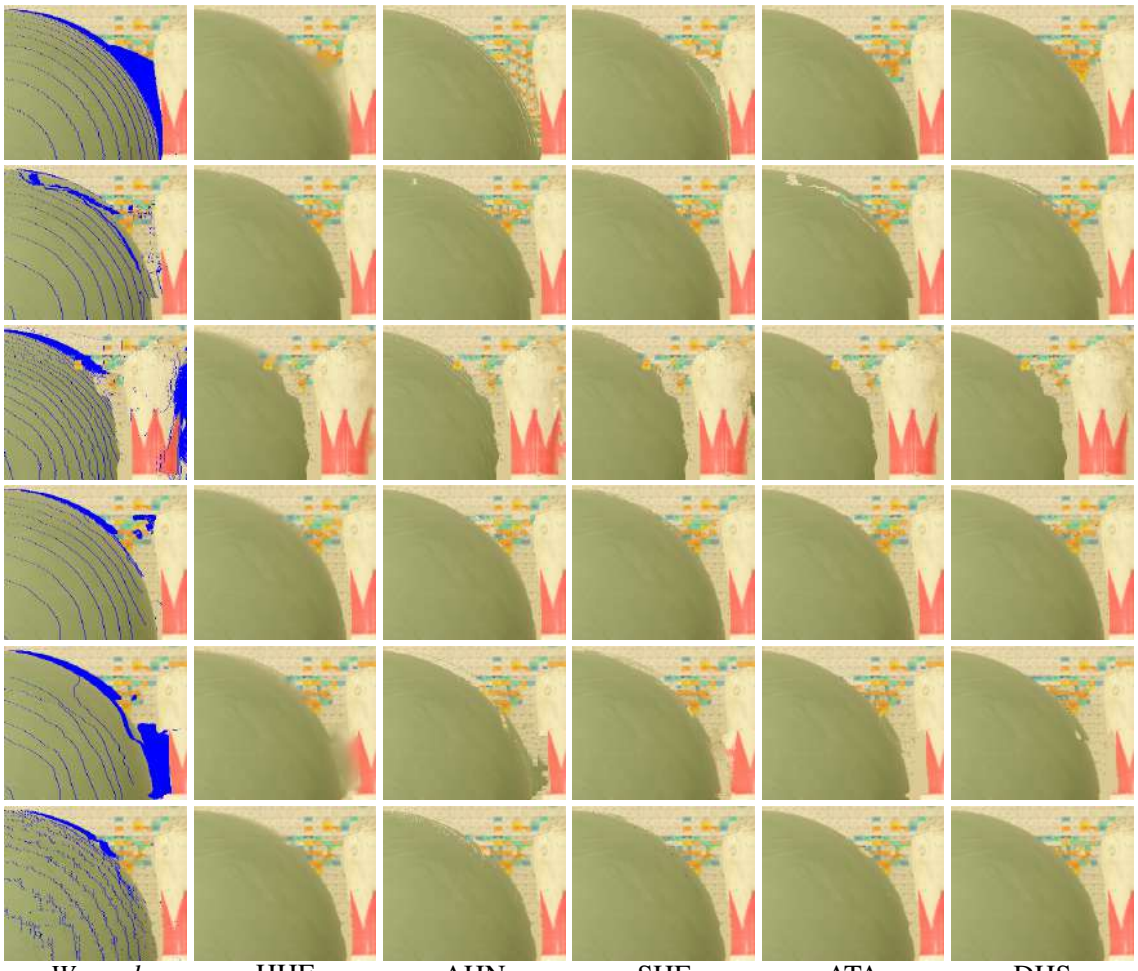
A Figura 6.8 apresenta os resultados das diferentes abordagens, quando alimentadas pelos diferentes algoritmos de casamento estéreo, para o *dataset* Bowling1. Como pode ser visto, tanto ATA quanto DHS apresentaram resultados satisfatórios, independentemente da imagem projetada dada como entrada, preenchendo a *disocclusion* – em azul – com conteúdo de *background*, sem produzir artefatos. Diferentemente, abordagens como AHN e SHF, produziram novos artefatos, influenciados por erros do mapa de disparidades, mostrando-se mais suscetíveis a inconformidades produzidas por algoritmos de casamento estéreo. Na Figura 6.9, são apresentados resultados para o *dataset* Baby1. Novamente, as abordagens propostas propagaram na *disocclusion* informação adequada, que neste caso representa um grande desafio, devido a complexidade da textura do mapa exibida ao fundo. Por outro lado, a abordagem HHF, em todas as reconstruções apresentou um resultado similar a um borrão, com maior destaque para o resultado gerado com o mapa criado com o algoritmo MeshStereo (ZHANG et al., 2015). Este resultado qualitativo, acaba contrapondo a média de SSIM apresentada pela abordagem na Tabela 6.9, que supera ATA e DHS em algumas situações. Mais resultados estão disponíveis em: <http://www.inf.ufrgs.br/~mwalter/dibrxsm/>.

6.4 Conclusões do Capítulo

Neste capítulo, foram apresentados os resultados experimentais relativos a esta tese. Primeiramente, exibiu-se uma avaliação das abordagens propostas utilizando mapas de disparidade *ground truth*, tanto no contexto de vídeos como de fotografias. Após, apresentou-se uma análise em cenário realista, onde DHS, ATA e outras três abordagens foram testadas com mapas de disparidades gerados por cinco algoritmos de casamento estéreo.

Para avaliar as abordagens propostas utilizando mapas de disparidade *ground truth*, foram realizados diversos testes no contexto de fotografia, utilizando os *datasets* de Middlebury. Com o resultado dos testes, constatou-se que as abordagens propostas produzem os melhores resultados qualitativos, reconstruindo coerentemente até mesmo regiões que possuem textura complexa. Na análise quantitativa, ATA e DHS obtiveram vantagens médias de no mínimo $1,17db$ e $1,93db$ no PSNR e no SSIM $0,008$ e $0,0113$, sobre as demais abordagens. O mesmo se repete na avaliação com vídeos, onde as reconstru-

Figura 6.8: Resultados para as abordagem DIBR usando os mapas de disparidade produzidos pelos algoritmos de casamento estéreo para o *dataset* Bowling1, projetando a vista real 1 para a virtual 3. A primeira coluna apresenta a vista projetada com base no mapa de disparidade *ground truth*, e nos algoritmos de casamento estéreo (TANIAI et al., 2018), (ZHANG et al., 2015), (MOZEROV; WEIJER, 2019), (MOZEROV; WEIJER, 2015) e (YIN et al., 2017), respectivamente para cada linha. As outras colunas apresentam vistas sintetizadas por cada uma das abordagens DIBR, como indicado, para cada algoritmo de casamento estéreo em cada linha.



Fonte: O Autor, com imagens adaptadas de (HIRSCHMULLER; SCHARSTEIN, 2007).

ções exibidas demonstram a capacidade de preenchimento das abordagens propostas, até mesmo em regiões como OOFAs.Na análise quantitativa de vídeos sem informação temporal, a vantagem apresentada pelas abordagens ATA e DHS, respectivamente, se reduz um pouco, apresentando a diferença de $0,11db$ e $0,31db$ no PSNR, e no SSIM $0,0022$ e $0,0024$, para a competitiva técnica FRBGE. Também no contexto de vídeos, foram avaliados os resultados produzidos pela abordagem DHS+B, que foi comparada com trabalhos que empregam informação temporal no processo de síntese. Na métrica PSNR, DHS+B apresentou uma superioridade média de $0,17db$ sobre a segunda melhor abordagem. Já

Figura 6.9: Resultados para as abordagem DIBR usando os mapas de disparidade produzidos pelos algoritmos de casamento estéreo para o *dataset* Baby1, projetando a vista real 1 para a virtual 3. A primeira coluna apresenta a vista projetada com base no mapa de disparidade *ground truth*, e nos algoritmos de casamento estéreo (TANIAI et al., 2018), (ZHANG et al., 2015), (MOZEROV; WEIJER, 2019), (MOZEROV; WEIJER, 2015) e (YIN et al., 2017), respectivamente para cada linha. As outras colunas apresentam vistas sintetizadas por cada uma das abordagens DIBR, como indicado, para cada algoritmo de casamento estéreo em cada linha.

Fonte: O Autor, com imagens adaptadas de (HIRSCHMULLER; SCHARSTEIN, 2007).

no SSIM, a abordagem PNT3 apresentou uma pequena vantagem nos resultados médios sobre a DHS+B, de $0,001$. Entretanto, apesar desta vantagem na média, a abordagem proposta obteve melhores resultados em 5 dos 8 comparativos realizados. Adicionalmente, fez-se uma análise do DHS+B em relação ao próprio DHS original e outras abordagens, utilizando uma métrica que avalia vídeos nos dominios espacial e temporal (a STRRED), onde a extensão com o modelo de *background* melhores resultados em 9 dos 10 testes realizados.

Na avaliação em cenários reais, as abordagens propostas também produziram os melhores resultados médios nas métricas PSNR, SSIM, e MW-PSNR. Considerando tanto mapas de disparidades *ground truth* quanto produzidos com casamento estéreo, os melhores resultados foram apresentados pela abordagem proposta DHS, seguida da ATA.

O mesmo se comprovou na análise qualitativa, que demonstrou a capacidade de ambas em produzir reconstruções coerentes mesmo diante de erros produzidos pelos mapas de disparidade reais. Além disso, provou-se experimentalmente que: (i) abordagens DIBR geralmente produzem melhores resultados com mapas de disparidade estimados com casamento estéreo do que com os fornecidos como *ground truth*. Mesmo no caso onde isto não ocorre (como no teste com DHS), o resultado produzido com mapas estimados aproxima-se muito do gerado com *ground truth*; (ii) abordagens DIBR são classificadas de forma diferente quando alimentadas por mapas de disparidade gerados com algoritmos casamento estéreo ou disparidades *ground truth*, considerando as métricas PSNR e SSIM; (iii) algoritmos que minimizam as medidas de erro para casamento estéreo não resultam necessariamente em melhores vistas sintéticas de acordo com o SSIM e o PSNR; (iv) o MW-PSNR tem uma forte correlação negativa com as métricas de casamento estéreo e pode ser mais útil para avaliar métodos de DIBR do que PSNR e SSIM; (v) e mapas de disparidade baseados em casamento estéreo contêm erros que "enganam" algumas das técnicas de DIBR, indicando que elas podem não estar preparadas para aplicação em cenário real.

7 CONSIDERAÇÕES FINAIS

7.1 Conclusões

Nesta tese, foram abordados os diferentes problemas relacionados à síntese de vistas com DIBR. Como definido, os principais desafios referentes a viabilidade técnica do modelo estão relacionados à detecção e tratamento de *cracks* (vazios e translúcidos) e *ghosts*, e principalmente ao preenchimento de *holes*. Para solucionar estes problemas, são necessárias abordagens especializadas, adaptadas às características específicas de cada artefato, e ao contexto das regiões que precisam ser reconstruídas. Em relação a fotografias, soluções no domínio espacial apenas são suficientes, mas para vídeos, faz-se necessário empregar mecanismos que auxiliem na coerência temporal durante a sucessão de quadros.

Com base na análise do processo de síntese de vistas com fotografias, foi desenvolvida a abordagem ATA. No *pipeline* proposto, o primeiro passo consiste na projeção da imagem de referência para o ponto de vista virtual. Após, com uma técnica baseada em filtros morfológicos, são detectados os *cracks* vazios e translúcidos em um único passo, os quais são preenchidos por um algoritmo próprio para este tipo de área, o HHF. No passo seguinte, candidatos a *ghost* são identificados na região de *background* das *disocclusions* (segmentada com o extrator *foreground-background* desenvolvido), e pontos classificados como ocorrências do artefato são projetados para sua posição correspondente no *foreground*, definida por um valor de disparidade estimado. Por fim, *disocclusions* e OOFAs são preenchidas com adaptações apropriadas do algoritmo de *inpainting* de (CRIMINISI; PEREZ; TOYAMA, 2004), com um esquema de busca que utiliza *patches* de tamanho adaptativo, extraídos da imagem de referência, e que explora o conceito de localidade espacial. No caso específico das *disocclusion*, as regiões de *background* são segmentadas por meio do extrator proposto, para evitar que a busca seja realizada em locais inadequados.

Considerando os avanços produzidos com o desenvolvimento de ATA, junto à uma pesquisa por estruturas que pudessem tornar o processo de síntese mais robusto, e que explorasse melhor o conceito de vizinhança na busca por conteúdo para preenchimento, elaborou-se a abordagem DHS. Nesta, os *ghosts* são removidos antes mesmo do processo de projeção, por meio de uma técnica de pós-processamento do mapa de disparidades, que evita que o artefato seja formado. Esta pode ser empregada no refinamento de mapas de disparidade, não somente no contexto deste trabalho, mas como um passo complementar

para algoritmos de casamento estéreo. Para o preenchimento dos *holes*, foi proposto um novo método, que se baseia no uso de *superpixels* hierárquicos para a compreensão de contexto e estrutura da imagem. Este método trata cada *hole* individualmente, e analisa os elementos que compõem sua vizinhança por meio dos rótulos de *superpixel*, para definir a melhor estratégia de preenchimento. Então, reconstrói-se a região vazia seguindo uma ordem de preenchimento adequada às características de sua vizinhança, realizando a cópia iterativa de *patches*. Estes são selecionados na imagem de referência, obedecendo um processo de avaliação que considera similaridade por cor, rótulo e, no caso das OOFAs, distribuição de disparidade na parte a ser copiada.

Além das abordagens ATA e DHS, desenvolveu-se um método para a geração de modelos de *background* para vídeos captados com câmera estacionária. Este método pode ser utilizado em conjunto com abordagens DIBR para o preenchimento parcial ou total de *holes*, ou por outras aplicações que necessitem da separação de elementos de *foreground* e *background*. O método proposto se divide em duas fases, construção e incremento do modelo de *background*. Na primeira fase, formula-se um modelo inicial utilizando o conteúdo dos primeiros T quadros do vídeo, com base na informação de disparidade e cor associada a cada *pixel*, em um processo que evita artefatos e *outliers*, e define a imagem colorida com base no método de (JUNG, 2009). Após, na segunda fase, inicia-se o processo de incremento do modelo, o qual tem como objetivo identificar regiões de *background* que são reveladas durante a sucessão de quadros. Para isso, analisam-se a frequência de ocorrência e as mudanças de disparidade de cada *pixel*, para substituir de maneira confiável o conteúdo de *foreground* restante por informação de *background*. Como complemento, para validar o método proposto, foi desenvolvido um processo de integração do modelo de *background* com DHS, que permite compor uma versão temporal da abordagem, denominada DHS+B.

As abordagens propostas foram avaliadas exaustivamente no contexto de vídeos e fotografias, utilizando mapas de disparidade *ground truth*. A avaliação quantitativa demonstra que ATA e DHS apresentam os melhores resultados no comparativo com outras 8 abordagens relevantes descritas na literatura que não utilizam informação temporal. No caso específico do DHS, no comparativo com as competitivas abordagens FHRF e FRBGE de (LUO; ZHU, 2017), por exemplo, são obtidas vantagens médias de $1,93$ e $0,31db$ no PSNR e, no SSIM $0,0113$ e $0,0024$, considerando os *datasets* de fotografia e vídeos, respectivamente. A análise qualitativa dos resultados reforça a capacidade de remoção de artefatos e o potencial de reconstrução de ambas abordagens, exibindo imagens

sintéticas limpas e com aparência realista, quase na totalidade dos casos de testes, exceto para *baselines* muito grandes, como no caso de BA41 (Figura 6.4). Contudo, mesmo nestes testes, os resultado produzidos por ATA e DHS são mais similares ao *ground truth* do que os gerados pelas outras abordagens comparadas. Um fator de destaque que pode ser observado nos resultados qualitativos principalmente, remete ao uso da imagem de referência para a busca de informação no processo de preenchimento, que possibilita uma reconstrução adequada até mesmo quando existe pouco conteúdo como referência na vista projetada. Em relação ao uso de informação estrutural ou de contexto, destacam-se as reconstruções executadas pelo algoritmo de preenchimento desenvolvido para o DHS. Este permitiu demonstrar por meio do uso de *superpixels* hierárquicos, que este tipo de representação estrutural faz com que os *holes* possam ser reconstruídos com maior precisão, em quaisquer situações, como comprovam os resultados experimentais.

A versão temporal DHS+B foi avaliada no comparativo com outras 5 abordagens que exploram este recurso de diferentes formas. Nesta avaliação, DHS+B apresentou melhores resultados em 5 casos de teste para as métricas PSNR e SSIM, considerando um total de 8. Nos resultados médios, a abordagem temporal apresentou um ganho médio de $0,17db$ sobre a segunda melhor abordagem para o PSNR, e para o SSIM uma perda média de $0,001$ com relação somente ao PNT3. Para avaliar a melhora produzida por DHS+B em relação ao DHS, foram computados os resultados apresentados pelas abordagens na métrica STRRED (própria para a análise espacial e temporal de vídeo), onde a extensão apresentou melhores resultados em 9 de 10 casos de teste. Neste comparativo também foram incluídas as abordagens SHF e ATA, onde DHS+B apresentou o melhor resultado em todos os testes.

Como complemento, foi realizada uma avaliação em cenário real das abordagens propostas, em um comparativo com HHF, AHN e SHF, empregando mapas de disparidade produzidos por cinco algoritmos de casamento estéreo diferentes. Nos testes, foram empregadas as métricas de síntese de vista PSNR, SSIM e MW-PSNR para medir a qualidade das vistas sintéticas produzidas. Esta análise comprovou a efetividade das abordagens propostas, uma vez que estas apresentaram os melhores resultados quantitativos e qualitativos, considerando todos os testes, com os diferentes mapas de disparidade. Por meio destes testes foi possível confirmar que as abordagens desenvolvidas são capazes de remover completamente os artefatos, até mesmo nestas situações. Como contribuição adicional, a avaliação em cenário real produziu uma análise do impacto dos mapas reais na produção de vistas sintéticas com DIBR. Esta análise permitiu que conclusões

importantes fossem obtidas como, por exemplo: *(i)* mapas de disparidade reais podem apresentar, em alguns casos, resultados até melhores que os produzidos com *ground truth* na síntese de vistas; *(ii)* algoritmos bem classificados nas métricas de casamento estéreo não necessariamente geram melhores imagens sintéticas.

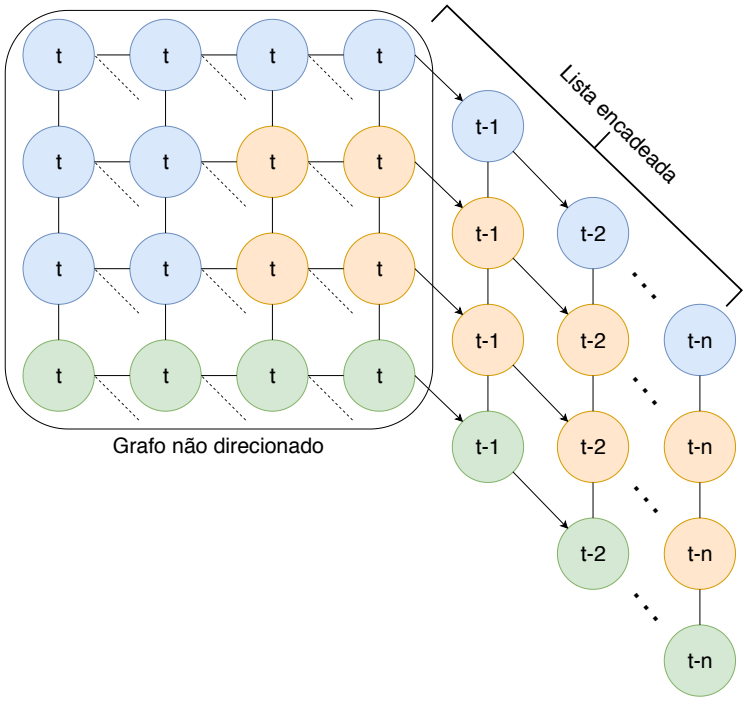
7.2 Trabalhos Futuros

Considerando que o método DHS+B ainda pode ser aprimorado, por exemplo, com a inclusão de uma estratégia que habilite seu uso em vídeos captados com movimento de câmera e com a remoção de possíveis inconsistências ainda presentes na abordagem, tem-se como um objetivo breve investigar estas possibilidades. Além disso, uma alternativa a ser estudada consiste na exploração de algoritmos de segmentação para a remoção de objetos de *foreground*, que possibilitem a produção de um modelo de *background* limpo, ainda mais preciso do que o produzido por Luo et al. (2019). Neste caso, o próprio método criado como parte do DHS para a geração do mapa de regiões candidatas a *ghost* pode ser considerado como ponto de partida, uma vez que este traz uma boa segmentação dos elementos de *foreground*, que corresponde as regiões de transição entre objetos na cena.

Os *superpixels* hierárquicos representam a estrutura da imagem de um modo confiável e, portanto, podem servir como base para a análise de coesão temporal no contexto de vídeos. Neste caso, deve-se analisar as mudanças produzidas nos *superpixels* durante a sucessão de quadros. Por exemplo, mudanças podem indicar que artefatos foram produzidos por erros no mapa de disparidades. Além disso, o registro de informação de *superpixels* de quadros passados pode ser útil para reconstruir parcialmente *holes* em quadros futuros. Neste contexto, como uma alternativa para o controle do processo de síntese de vídeos, pode-se explorar a estrutura de dados base – grafo – do algoritmo SH, empregado na estimativa dos *superpixels*, para controle de informação temporal. Em (YANG, 2015), utiliza-se um grafo como estrutura base para a geração de mapas de disparidade por meio de casamento estéreo, com uma extensão para exploração de informação temporal. Esta extensão baseia-se no uso de uma lista encadeada associada a cada posição de pixel da imagem, que permite produzir mapas de disparidade coerentes temporalmente para aplicação em vídeos.

Na Figura 7.1, apresenta-se uma adaptação desta proposta para o contexto de síntese de vista. Como pode ser visto, um grafo não direcionado representa a segmentação

Figura 7.1: Ilustração com a segmentação de *superpixels* em um grafo não direcionado, no instante t de um vídeo. Neste caso, para preservar informação temporal, associa-se a cada *pixel* (com um vértice correspondente no grafo) uma lista encadeada com conteúdo de quadros passados (instantes $t-1$, $t-2$, ..., $t-n$).



Fonte: O Autor.

em *superpixels* do quadro em um dado instante t, e uma lista encadeada mantém o histórico relativo à informação de cada posição de *pixel* ao longo do tempo (instantes $t-1$, $t-2$, ..., $t-n$). Esta estrutura pode ser utilizada, por exemplo, como fonte de informação para o preenchimento parcial de *holes*, com pesquisas não somente no domínio espacial, mas também temporal. Do mesmo modo, pode-se manter a coerência durante a sucessão de quadros, para evitar que *outliers* produzam efeitos como *flickering* em vídeos, por meio de mecanismos de validação.

Ainda, como outra alternativa, acredita-se que em um futuro breve será possível a produção de uma segmentação semântica de imagens com um grau de precisão bastante alto, o que fica evidenciado pelos bons resultados apresentados por trabalhos recentes como (LIU et al., 2019; FU et al., 2019). Quando isso for viabilizado efetivamente, será possível identificar quaisquer objetos nas cenas, e esta informação poderá auxiliar em diferentes partes do *pipeline* DIBR. Algoritmos de *inpainting* por exemplo, poderão utilizar informação de objetos da mesma classe semântica exibidos em outras imagens para a reconstrução dos *holes*. Da mesma forma, mecanismos para a manutenção da coesão

temporal em vídeos poderão rastrear os objetos ao longo dos quadros, não permitindo que inconsistências sejam incluídas nas vistas sintéticas.

7.3 Artigos Publicados e Futuras Submissões

Abaixo, encontram-se listados os artigos publicados durante o desenvolvimento desta tese.

1. **Oliveira, A. Q.**, Walter, M. and Jung, C. R.; An Artifact-type Aware DIBR Method for View Synthesis; *IEEE Signal Processing Letters*, v. 25, n. 11, p. 1705–1709, 2018.

2. **Oliveira, A. Q.**, Silveira, T. L. T., Walter, M. and Jung, C. R.; On the Performance of DIBR Methods when Using Depth Maps from State-of-the-Art Stereo Matching Algorithms; In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

Além dos trabalhos já publicados, um novo artigo será submetido para um periódico internacional da área, relativo a abordagem DHS, apresentada no Capítulo 4, que já possui resultados qualitativos e quantitativos melhores que abordagens recentes, publicadas em veículos reconhecidos, como (LUO; ZHU, 2017) e (OLIVEIRA; WALTER; JUNG, 2018). Neste mesmo artigo, também será considerada a inclusão da extensão da abordagem para o domínio temporal, que emprega o método desenvolvido para construção de modelos de *background* (descrito no Capítulo 5), com possíveis aprimoramentos.

REFERÊNCIAS

ACHANTA, R.; SHAJI, A.; SMITH, K.; LUCCHI, A.; FUA, P.; SÜSSTRUNK; SABINE. **Slic superpixels**. [S.l.], 2010. 155–162 p.

ACHANTA, R.; SHAJI, A.; SMITH, K.; LUCCHI, A.; FUA, P.; SÜSSTRUNK, S. et al. Slic superpixels compared to state-of-the-art superpixel methods. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE Press, Piscataway, NJ, USA, v. 34, n. 11, p. 2274–2282, 2012.

ACHANTA, R.; SüSSTRUNK, S. Superpixels and polygons using simple non-iterative clustering. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** Piscataway, NJ, USA: IEEE Press, 2017. (CVPR'17), p. 4895–4904.

AHN, I.; KIM, C. A novel depth-based virtual view synthesis method for free viewpoint video. **IEEE Transactions on Broadcasting**, IEEE Press, Piscataway, NJ, USA, v. 59, n. 4, p. 614–626, 2013.

AHN, J.; KWAK, S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** Piscataway, NJ, USA: IEEE Press, 2018. (CVPR'18).

ASHIKHMIN, M. Synthesizing natural textures. In: SYMPOSIUM ON INTERACTIVE 3D GRAPHICS. **Proceedings...** New York, NY, USA: ACM, 2001. p. 217–226.

BARNARD, S. T.; FISCHLER, M. A. Computational stereo. **ACM Computing Surveys**, ACM, New York, NY, USA, v. 14, n. 4, p. 553–572, dez. 1982.

BARNES, C.; SHECHTMAN, E.; FINKELSTEIN, A.; GOLDMAN, D. B. PatchMatch: A randomized correspondence algorithm for structural image editing. **ACM Transactions on Graphics (Proc. SIGGRAPH)**, v. 28, n. 3, ago 2009.

BAY, H.; TUYTELAARS, T.; GOOL, L. V. Surf: Speeded up robust features. In: EUROPEAN CONFERENCE ON COMPUTER VISION. **Proceedings...** Berlin, Heidelberg: Springer-Verlag, 2006. (ECCV'06), p. 404–417.

BERGH, M. Van den; BOIX, X.; ROIG, G.; CAPITANI, B. de; GOOL, L. V. Seeds: Superpixels extracted via energy-driven sampling. In: **Computer Vision – ECCV 2012**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 13–26.

BINDEMANN, M. Scene and screen center bias early eye movements in scene viewing. **Vision Research**, v. 50, n. 23, p. 2577 – 2587, 2010. Vision Research Reviews.

BOUWMANS, T. Traditional and recent approaches in background modeling for foreground detection: An overview. **Computer Science Review**, v. 11-12, p. 31–66, 2014.

BOUWMANS, T.; JAVED, S.; SULTANA, M.; JUNG, S. K. Deep neural network concepts for background subtraction:a systematic review and comparative evaluation. **Neural Networks**, v. 117, p. 8–66, 2019.

CANNY, J. A computational approach to edge detection. **IEEE Transactions on pattern analysis and machine intelligence**, IEEE Press, Piscataway, NJ, USA, n. 6, p. 679–698, 1986.

CHEN, L.; PAPANDREOU, G.; KOKKINOS, I.; MURPHY, K.; YUILLE, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE Press, Piscataway, NJ, USA, v. 40, n. 4, p. 834–848, abr. 2018.

CHO, J.-H.; SONG, W.; CHOI, H.; KIM, T. Hole filling method for depth image based rendering based on boundary decision. **IEEE Signal Processing Letters**, IEEE Press, Piscataway, NJ, USA, v. 24, n. 3, p. 329–333, 2017.

CRIMINISI, A.; PEREZ, P.; TOYAMA, K. Region filling and object removal by exemplar-based image inpainting. **IEEE Transactions on Image Processing**, IEEE Press, Piscataway, NJ, USA, v. 13, n. 9, p. 1200–1212, 2004.

DAI, J.; NGUYEN, T. View synthesis with hierarchical clustering based occlusion filling. In: IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING. **Proceedings...** Piscataway, NJ, USA: IEEE Press, 2017. (ICIP'17), p. 1387–1391.

DARIBO, I.; SAITO, H. A novel inpainting-based layered depth video for 3dtv. **IEEE Transactions on Broadcasting**, IEEE Press, Piscataway, NJ, USA, v. 57, n. 2, p. 533–541, 2011.

DOLLAR, P.; ZITNICK, C. L. Structured forests for fast edge detection. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION. **Proceedings...** Piscataway, NJ, USA: IEEE Press, 2013. (ICCV'13).

DU, S.-P.; DIDYK, P.; DURAND, F.; HU, S.-M.; MATUSIK, W. Improving visual quality of view transitions in automultiscopic displays. **ACM Transactions on Graphics**, ACM, New York, NY, USA, v. 33, n. 6, p. 192:1–192:9, nov 2014.

EFROS, A. A.; LEUNG, T. K. Texture synthesis by non-parametric sampling. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION. **Proceedings...** Piscataway, NJ, USA: IEEE Press, 1999. (ICCV'99), p. 1033–1038.

EIGEN, D.; FERGUS, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION. **Proceedings...** Piscataway, NJ, USA: IEEE Press, 2015. (ICCV'15), p. 2650–2658.

FEHN, C. Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3d-tv. In: STEREOSCOPIC DISPLAYS AND VIRTUAL REALITY SYSTEMS. **Proceedings...** [S.l.], 2004. p. 93–104.

FICKEL, G. P. **Video view interpolation using temporally adaptive 3D meshes**. Thesis (PhD) — Universidade Federal do Rio Grande do Sul, Instituto de Informática, 2015.

FISCHLER, M. A.; BOLLES, R. C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. **Communications of the ACM**, ACM, New York, NY, USA, v. 24, n. 6, p. 381–395, jun. 1981.

FU, J.; LIU, J.; WANG, Y.; ZHOU, J.; WANG, C.; LU, H. Stacked deconvolutional network for semantic segmentation. **IEEE Transactions on Image Processing**, p. 1–1, 2019.

FUHR, G.; FICKEL, G. P.; DAL'AQUA, L. P.; JUNG, C. R.; MALZBENDER, T.; SAMADANI, R. An evaluation of stereo matching methods for view interpolation. In: IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING. **Proceedings...** Piscataway, NJ, USA: IEEE Press, 2013. (ICIP'13), p. 403–407.

GAUTIER, J.; MEUR, O. L.; GUILLEMOT, C. Depth-based image completion for view synthesis. In: 3DTV CONFERENCE: THE TRUE VISION - CAPTURE, TRANSMISSION AND DISPLAY OF 3D VIDEO. **Proceedings...** Piscataway, NJ, USA: IEEE Press, 2011. (3DTV-CON'11), p. 1–4.

GRADY, L. Random walks for image segmentation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE Press, Piscataway, NJ, USA, v. 28, n. 11, p. 1768–1783, nov. 2006.

GUO, X.; LI, H.; YI, S.; REN, J.; WANG, X. Learning monocular depth by distilling cross-domain stereo networks. In: EUROPEAN CONFERENCE ON COMPUTER VISION. **Proceedings...** New York, USA: Springer-Verlag, 2018. (ECCV'18).

HARIHARAN, B.; ARBELáEZ, P.; BOURDEV, L.; MAJI, S.; MALIK, J. Semantic contours from inverse detectors. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION. **Proceedings...** Piscataway, NJ, USA: IEEE Press, 2011. (ICCV'11), p. 991–998.

HIRSCHMULLER, H.; SCHARSTEIN, D. Evaluation of cost functions for stereo matching. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** Piscataway, NJ, USA: IEEE Press, 2007. (CVPR'07), p. 1–8.

HUANG, J. B.; KANG, S. B.; AHUJA, N.; KOPF, J. Image completion using planar structure guidance. **ACM Transactions on Graphics**, Association for Computing Machinery (ACM), v. 33, n. 4, jan 2014.

JAMPANI, V.; SUN, D.; LIU, M.-Y.; YANG, M.-H.; KAUTZ, J. Superpixel sampling networks. In: EUROPEAN CONFERENCE ON COMPUTER VISION. **Proceedings...** New York, USA: Springer-Verlag, 2018. (ECCV'18).

JUNG, C. R. Efficient background subtraction and shadow removal for monochromatic video sequences. **IEEE Transactions on Multimedia**, v. 11, n. 3, p. 571–577, April 2009.

KAWAI, N.; SATO, T.; YOKOYA, N. Image inpainting considering brightness change and spatial locality of textures and its evaluation. In: **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. [S.l.: s.n.], 2009. v. 5414 LNCS, p. 271–282.

KELLNHOFER, P.; DIDYK, P.; WANG, S.-P.; SITTHI-AMORN, P.; FREEMAN, W.; DURAND, F.; MATUSIK, W. 3dtv at home: Eulerian-lagrangian stereo-to-multiview conversion. **ACM Transactions on Graphics**, ACM, New York, NY, USA, v. 36, n. 4, p. 146:1–146:13, jul. 2017.

KOLMOGOROV, V. Convergent tree-reweighted message passing for energy minimization. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE Press, Piscataway, NJ, USA, v. 28, n. 10, p. 1568–1583, out. 2006.

KOMODAKIS, N.; TZIRITAS, G. Image completion using efficient belief propagation via priority scheduling and dynamic pruning. **IEEE Transactions on Image Processing**, IEEE Press, Piscataway, NJ, USA, v. 16, n. 11, p. 2649–2661, nov 2007.

KöPPEL, M.; MüLLER, K.; WIEGAND, T. Filling disocclusions in extrapolated virtual views using hybrid texture synthesis. **IEEE Transactions on Broadcasting**, IEEE Press, Piscataway, NJ, USA, n. 99, p. 1–13, 2016.

LIE, W.; HSIEH, C.; LIN, G. Key-frame-based background sprite generation for hole filling in depth image-based rendering. **IEEE Transactions on Multimedia**, IEEE Press, Piscataway, NJ, USA, v. 20, n. 5, p. 1075–1087, maio 2018.

LIU, C.; CHEN, L.-C.; SCHROFF, F.; ADAM, H.; HUA, W.; YUILLE, A. L.; FEI-FEI, L. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In: **IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2019.

LIU, F.; SHEN, C.; LIN, G.; REID, I. D. Learning depth from single monocular images using deep convolutional neural fields. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE Press, Piscataway, NJ, USA, v. 38, n. 10, p. 2024–2039, 2016.

LONG, J.; SHELHAMER, E.; DARRELL, T. Fully convolutional networks for semantic segmentation. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** Piscataway, NJ, USA: IEEE Press, 2015. (CVPR'15).

LU, J.; YANG, Q.; LAFRUIT, G. Interpolation error as a quality metric for stereo: Robust, or not? In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING. **Proceedings...** Piscataway, NJ, USA: IEEE Press, 2009. (ICASSP'09), p. 977–980.

LUO, G.; ZHU, Y. Foreground removal approach for hole filling in 3d video and fvv synthesis. **IEEE Transactions on Circuits and Systems for Video Technology**, IEEE Press, Piscataway, NJ, USA, v. 27, n. 10, p. 2118–2131, out. 2017.

LUO, G.; ZHU, Y.; LI, Z.; ZHANG, L. A hole filling approach based on background reconstruction for view synthesis in 3d video. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** Piscataway, NJ, USA: IEEE Press, 2016. (CVPR'16), p. 1781–1789.

LUO, G.; ZHU, Y.; WENG, Z.; LI, Z. A disocclusion inpainting framework for depth-based view synthesis. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, p. 1–1, 2019.

MARK, W. R.; MCMILLAN, L.; BISHOP, G. Post-rendering 3d warping. In: SYMPOSIUM ON INTERACTIVE 3D GRAPHICS. **Proceedings...** [S.l.], 1997. p. 7–16.

MCMILLAN, L. **An Image-Based Approach to Three-Dimensional Computer Graphics**. Thesis (PhD) — University of North Carolina, Chapel Hill, NC, USA, 1997.

MOHAMED, S. S.; TAHIR, N. M.; ADNAN, R. Background modelling and background subtraction performance for object detection. In: **International Colloquium on Signal Processing its Applications**. [S.l.: s.n.], 2010. p. 1–6.

MORI, Y.; FUKUSHIMA, N.; FUJII, T.; TANIMOTO, M. View generation with 3d warping using depth information for ftv. In: 3DTV CONFERENCE: THE TRUE VISION - CAPTURE, TRANSMISSION AND DISPLAY OF 3D VIDEO. **Proceedings...** Piscataway, NJ, USA: IEEE Press, 2008. (3DTV-CON'08), p. 229–232.

MOZEROV, M. G.; WEIJER, J. van de. Accurate stereo matching by two-step energy minimization. **IEEE Transactions on Image Processing**, IEEE Press, Piscataway, NJ, USA, v. 24, n. 3, p. 1153–1163, 2015.

MOZEROV, M. G.; WEIJER, J. van de. One-view occlusion detection for stereo matching with a fully connected crf model. **IEEE Transactions on Image Processing**, IEEE Press, Piscataway, NJ, USA, v. 28, n. 6, p. 2936–2947, jun. 2019.

MUDDALA, S. M. **Free View Rendering for 3D Video: Edge-Aided Rendering and Depth-Based Image Inpainting**. Thesis (PhD) — Mid Sweden University, Department of Information and Communication Systems, jun. 2015.

MUDDALA, S. M.; OLSSON, R.; SJÖSTRÖM, M. Disocclusion handling using depth-based inpainting. In: INTERNATIONAL CONFERENCES ON ADVANCES IN MULTIMEDIA. **Proceedings...** [S.l.]: International Academy, Research and Industry Association (IARIA), 2013. p. 136–141.

MUDDALA, S. M.; SJÖSTRÖM, M.; OLSSON, R. Virtual view synthesis using layered depth image generation and depth-based inpainting for filling disocclusions and translucent disocclusions. **Journal of Visual Communication and Image Representation**, Elsevier, v. 38, p. 351–366, 2016.

NDJIKI-NYA, P.; KOPPEL, M.; DOSHKOV, D.; LAKSHMAN, H.; MERKLE, P.; MULLER, K.; WIEGAND, T. Depth image-based rendering with advanced texture synthesis for 3-d video. **IEEE Transactions on Multimedia**, IEEE Press, Piscataway, NJ, USA, v. 13, n. 3, p. 453–465, jun. 2011.

NEWSON, A.; ALMANSA, A.; FRADET, M.; GOUSSEAU, Y.; PÉREZ, P. Video inpainting of complex scenes. **SIAM Journal on Imaging Sciences**, SIAM, v. 7, n. 4, p. 1993–2019, 2014.

OH, K.-J.; YEA, S.; HO, Y.-S. Hole filling method using depth based in-painting for view synthesis in free viewpoint television and 3-d video. In: PICTURE CODING SYMPOSIUM. **Proceedings...** [S.l.], 2009. (PCS'09), p. 1–4.

OLIVEIRA, A.; FICKEL, G.; WALTER, M.; JUNG, C. Selective hole-filling for depth-image based rendering. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING. **Proceedings...** Piscataway, NJ, USA: IEEE Press, 2015. (ICASSP'15), p. 1186–1190.

OLIVEIRA, A. Q. **Síntese de vistas com depht-image-based rendering (DIBR)**. Dissertation (Master) — Universidade Federal do Rio Grande do Sul (UFRGS), Instituto de Informática (INF), abr. 2016.

OLIVEIRA, A. Q. de; WALTER, M.; JUNG, C. R. An artifact-type aware dibr method for view synthesis. **IEEE Signal Processing Letters**, IEEE Press, Piscataway, NJ, USA, v. 25, n. 11, p. 1705–1709, nov. 2018.

OLIVEIRA, M. M.; BOWEN, B.; MCKENNA, R.; CHANG, Y. sung. Fast digital image inpainting. In: INTERNATIONAL CONFERENCE ON VISUALIZATION, IMAGING AND IMAGE PROCESSING. **Proceedings...** [S.l.]: ACTA Press, 2001. p. 261–266.

PETERSEN, K. B.; PEDERSEN, M. S. et al. The matrix cookbook. **Technical University of Denmark**, v. 7, n. 15, p. 510, 2008.

PURICA, A.; MORA, E.; PESQUET-POPESCU, B.; CAGNAZZO, M.; IONESCU, B. Improved view synthesis by motion warping and temporal hole filling. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING. **Proceedings...** Brisbane, Australia, 2015. (ICASSP'15, v. 1), p. 1191–1195.

RAHAMAN, D. M. M.; PAUL, M. Virtual view synthesis for free viewpoint video and multiview video compression using gaussian mixture modelling. **IEEE Transactions on Image Processing**, IEEE Press, Piscataway, NJ, USA, v. 27, n. 3, p. 1190–1201, mar. 2018.

RUSANOVSKYY, D.; AFLAKI, P.; HANNUKSELA, M. **Undo Dancer 3DV sequence for purposes of 3DV standardization**. Geneva, Switzerland, Tech. Rep. MPEG/M20028, 2011. v. 2011.

SANDIĆ-STANKOVIĆ, D.; KUKOLJ, D.; CALLET, P. L. Multi-scale synthesized view assessment based on morphological pyramids. **Journal of Electrical Engineering**, v. 67, n. 1, p. 3–11, 2016.

SCHARSTEIN, D.; SZELISKI, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. **International Journal of Computer Vision**, v. 47, n. 1, p. 7–42, abr. 2002.

SCHMEING, M.; JIANG, X. Depth image based rendering. In: **Pattern Recognition, Machine Intelligence and Biometrics**. [S.l.]: Springer, 2011. p. 279–310.

SCHMEING, M.; JIANG, X. Faithful disocclusion filling in depth image based rendering using superpixel-based inpainting. **IEEE Transactions on Multimedia**, IEEE Press, Piscataway, NJ, USA, v. 17, n. 12, p. 2160–2173, dez. 2015.

SCHWARZ, H.; MARPE, D.; WIEGAND, T. **Description of exploration experiments in 3D video coding**. Dresden, Germany, Tech. Rep. MPEG2010/N11274, 2010. v. 2010.

SCHWARZ, S.; OLSSON, R.; SJöSTRöM, M. Depth sensing for 3dtv: A survey. **IEEE MultiMedia**, IEEE Press, Piscataway, NJ, USA, v. 20, n. 4, p. 10–17, out. 2013.

SHIBAHARA, T.; AOKI, T.; NAKAJIMA, H.; KOBAYASHI, K. A sub-pixel stereo correspondence technique based on 1d phase-only correlation. In: **IEEE International Conference on Image Processing**. [S.l.: s.n.], 2007. v. 5, p. 221–224.

SOLH, M.; ALREGIB, G. Depth adaptive hierarchical hole filling for dibr-based 3d videos. In: IS&T/SPIE ELECTRONIC IMAGING. **Proceedings...** [S.l.], 2012. p. 829004–829004.

SOLH, M.; ALREGIB, G. Hierarchical hole-filling for depth-based view synthesis in ftv and 3d video. **IEEE Journal of Selected Topics in Signal Processing**, IEEE Press, Piscataway, NJ, USA, v. 6, n. 5, p. 495–504, 2012.

SOUNDARARAJAN, R.; BOVIK, A. C. Video quality assessment by reduced reference spatio-temporal entropic differencing. **IEEE Transactions on Circuits and Systems for Video Technology**, v. 23, n. 4, p. 684–694, abr 2013.

SPEARMAN, C. The proof and measurement of association between two things. **The American Journal of Psychology**, v. 15, n. 1, p. 72–101, 1904.

STUTZ, D. Superpixel segmentation: An evaluation. In: GALL, J.; GEHLER, P.; LEIBE, B. (Ed.). **Pattern Recognition**. Cham: Springer International Publishing, 2015. p. 555–562.

STUTZ, D.; HERMANS, A.; LEIBE, B. Superpixels: An evaluation of the state-of-the-art. **Computer Vision and Image Understanding**, v. 166, p. 1–27, 2018.

SUN, C.; LIU, X.; YANG, W. An efficient quality metric for dibr-based 3d video. In: **IEEE International Conference on High Performance Computing and Communication**. [S.l.: s.n.], 2012. p. 1391–1394.

SZELISKI, R. **Computer vision: algorithms and applications**. [S.l.]: Springer Science & Business Media, 2010.

TANIAI, T.; MATSUSHITA, Y.; SATO, Y.; NAEMURA, T. Continuous 3d label stereo matching using local expansion moves. **IEEE transactions on pattern analysis and machine intelligence**, IEEE Press, Piscataway, NJ, USA, v. 40, n. 11, p. 2725–2739, 2018.

TANIMOTO, M.; FUJII, T.; SUZUKI, K.; FUKUSHIMA, N.; MORI, Y. **Reference softwares for depth estimation and view synthesis**. Archamps, France, Technical Report M15377, 2008. v. 2008.

TELEA, A. An Image Inpainting Technique Based on the Fast Marching Method. **Journal of Graphics Tools**, v. 9, n. 1, p. 23–34, 2004.

TIAN, S.; ZHANG, L.; MORIN, L.; DéFORGES, O. A benchmark of dibr synthesized view quality assessment metrics on a new database for immersive media applications. **IEEE Transactions on Multimedia**, IEEE Press, Piscataway, NJ, USA, v. 21, n. 5, p. 1235–1247, maio 2019.

TONIETTO, L.; WALTER, M.; JUNG, C. R. Patch-based texture synthesis using wavelets. In: BRAZILIAN SYMPOSIUM ON COMPUTER GRAPHICS AND IMAGE PROCESSING. **Proceedings...** [S.l.], 2005. (SIBGRAPI'05), p. 383–389.

VISWANATH, A.; BEHERA, R. K.; SENTHAMILARASU, V.; KUTTY, K. Background modelling from a moving camera. **Procedia Computer Science**, v. 58, p. 289–296, 2015. Second International Symposium on Computer Vision and the Internet (VisionNet'15).

WANG, C.; BUENAPOSADA, J. M.; ZHU, R.; LUCEY, S. Learning depth from monocular videos using direct methods. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** [S.l.], 2018. (CVPR'18), p. 2022–2030.

WANG, Z.; BOVIK, A. C.; SHEIKH, H. R.; SIMONCELLI, E. P. Image quality assessment: from error visibility to structural similarity. **IEEE Transactions on Image Processing**, IEEE Press, Piscataway, NJ, USA, v. 13, n. 4, p. 600–612, 2004.

WEI, X.; YANG, Q.; GONG, Y.; AHUJA, N.; YANG, M. Superpixel hierarchy. **IEEE Transactions on Image Processing**, IEEE Press, Piscataway, NJ, USA, v. 27, n. 10, p. 4838–4849, out. 2018.

WEST, D. B. **Introduction to graph theory**. [S.l.]: Prentice hall Upper Saddle River, NJ, 1996.

XU, L.; YAN, Q.; XIA, Y.; JIA, J. Structure extraction from texture via relative total variation. **ACM Transactions on Graphics**, ACM, New York, NY, USA, v. 31, n. 6, p. 139:1–139:10, nov. 2012.

XU, Y.; DONG, J.; ZHANG, B.; XU, D. Background modeling methods in video analysis: A review and comparative evaluation. **CAAI Transactions on Intelligence Technology**, v. 1, n. 1, p. 43–60, 2016.

YANG, Q. Stereo matching using tree filtering. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE Press, Piscataway, NJ, USA, v. 37, n. 4, p. 834–846, abr. 2015.

YANG, X.; LIU, J.; SUN, J.; LI, X.; LIU, W.; GAO, Y. Dibr based view synthesis for free-viewpoint television. In: 3DTV CONFERENCE: THE TRUE VISION - CAPTURE, TRANSMISSION AND DISPLAY OF 3D VIDEO. **Proceedings...** Piscataway, NJ, USA: IEEE Press, 2011. (3DTV-CON'11, 1), p. 1–4.

YAO, C.; TILLO, T.; ZHAO, Y.; XIAO, J.; BAI, H.; LIN, C. Depth map driven hole filling algorithm exploiting temporal correlation information. **IEEE Transactions on Broadcasting**, v. 60, n. 2, p. 394–404, June 2014.

YIN, J.; ZHU, H.; YUAN, D.; XUE, T. Sparse representation over discriminative dictionary for stereo matching. **Pattern Recognition**, v. 71, p. 278–289, 2017.

ZHANG, C.; LI, Z.; CHENG, Y.; CAI, R.; CHAO, H.; RUI, Y. Meshstereo: A global stereo model with mesh alignment regularization for view interpolation. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION. **Proceedings...** Piscataway, NJ, USA: IEEE Press, 2015. (ICCV'15), p. 2057–2065.

ZHANG, L.; VAZQUEZ, C.; KNORR, S. 3d-tv content creation: automatic 2d-to-3d video conversion. **IEEE Transactions on Broadcasting**, IEEE Press, Piscataway, NJ, USA, v. 57, n. 2, p. 372–383, 2011.

ZINGER, S.; DO, L.; WITH, P. de. Free-viewpoint depth image based rendering. **Journal of Visual Communication and Image Representation**, Elsevier Inc., v. 21, n. 5-6, p. 533–541, jul. 2010.

ZITNICK, C. L.; KANG, S. B.; UYTTENDAELE, M.; WINDER, S.; SZELISKI, R. High-quality video view interpolation using a layered representation. **ACM Transactions on Graphics**, ACM, New York, NY, USA, v. 23, n. 3, p. 600–608, ago. 2004.

ZOLLHÖFER, M.; STOTKO, P.; GÖRLITZ, A.; THEOBALT, C.; NIESSNER, M.; KLEIN, R.; KOLB, A. State of the art on 3d reconstruction with rgb-d cameras. In: COMPUTER GRAPHICS FORUM. **Proceedings...** [S.l.]: Wiley Online Library, 2018. v. 37, n. 2, p. 625–652.