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

JOSÉ ELITON ALBUQUERQUE FILHO

Joint-task learning to improve super-resolution of aerial images

Dissertação apresentada como requisito
parcial para a obtenção do grau
de Mestre em Ciência da Computação

Advisor: Prof. Dr. Claudio Rosito Jung

Porto Alegre
August 2022

CIP — CATALOGING-IN-PUBLICATION

Albuquerque Filho, José Eliton

Joint-task learning to improve super-resolution of aerial images / José Eliton Albuquerque Filho. – 2022.

83 f.

Orientador: Claudio Rosito Jung.

Dissertação (Mestrado) – Universidade Federal do Rio Grande do Sul, Instituto de Informática, Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2022.

1. aprendizado profundo. 2. super resolução. 3. segmentação semântica. 4. tarefas conjuntas. 5. imagens aéreas. 6. qualidade perceptual. I. Jung, Claudio Rosito. II. Título.

“Don’t be pushed around by the fears in your mind.
Be led by the dreams in your heart.”
— Roy T. Bennett

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Claudio Rosito Jung, for your guidance, patience and support throughout this project. I am extremely grateful that you took me on as a student and provided your extensive knowledge and support throughout these years. Thanks for my family and my wife for always being there for me, giving me unfailing support and continuous encouragement whenever possible. I would also like to thank my friends, specially the ones from the First Geoinformation Center, for the insightful conversations and steady assistance.

Aprendizagem de tarefas conjuntas para melhorar a super-resolução de imagens aéreas

RESUMO

Redes de aprendizado profundo tornaram-se uma abordagem muito popular para resolver vários problemas de visão computacional. Entre eles, a super resolução (SR) é uma tarefa particularmente desafiadora, devido à sua natureza mal-posta, uma vez que uma imagem super resolvida pode ser originada de várias imagens de baixa resolução (LR), e a dificuldade em sintetizar informações coerentes em maior resolução, possivelmente levando a artefatos visuais ou texturas inconsistentes. Isso é facilmente verificado no contexto de sensoriamento remoto, onde as técnicas de restauração de imagens enfrentam dificuldades na replicação de superfícies terrestres do mundo real, tendo no entanto um grande potencial para gerar dados de alta resolução (HR) a partir de imagens LR. Embora existam vários métodos SR na literatura, poucos deles focam na qualidade perceptual das imagens SR, falhando em recuperar informações detalhadas inerentes às imagens aéreas. Uma das principais razões para isso é a dificuldade em definir uma imagem "boa"na perspectiva da máquina, fato não alcançável para métricas comuns de pixel como PSNR e SSIM. Neste contexto, este trabalho propõe um procedimento de treinamento conjunto de ponta a ponta para gerar imagens SR perceptualmente melhores: usando um módulo SR baseado em Redes Generativas Adversariais (GAN) e um módulo de segmentação semântica, é possível induzir o gerador a produzir estruturas e informações texturais mais coerentes usando uma função objetiva de segmentação capaz de capturar detalhes de textura em dados sintetizados, fato corroborado por resultados experimentais.

Palavras-chave: aprendizado profundo, super resolução, segmentação semântica, tarefas conjuntas, imagem aérea, qualidade perceptual.

ABSTRACT

Deep learning networks have become a very popular approach for solving multiple computer vision problems. Amongst them, super resolution (SR) is a particularly challenging task because of its ill-posed nature, since one super resolved image could be originated from multiple low resolution (LR) counterparts, and the difficulty in synthesizing coherent information at increased resolution, possibly leading to visual artifacts or inconsistent textures. This is readily verified in the context of remote sensing, where image restoration techniques face difficulties in replicating real-world land surfaces, having though a great potential for generating high-resolution (HR) data from LR images. While there are multiple SR methods in the literature, few of them focus on the perceptual quality of SR images, failing to recover detailed information inherent in aerial imagery. One of the main reasons for that is the difficulty in defining a "good-looking" image in the perspective of the machine, a fact not achievable for common pixel-wise metrics like PSNR and SSIM. In this context, this work proposes an end-to-end joint training procedure to generate better perpetually-wise SR images: by using a SR module based on Generative Adversarial Network (GAN) and a semantic segmentation module, it is possible to induce the generator network to produce more coherent structures and textural information by using a segmentation loss capable of capturing texture details on synthesized data, a fact corroborated by experimental results.

Keywords: Deep learning. super resolution. semantic segmentation. joint tasks. aerial imagery. perceptual quality.

LIST OF ABBREVIATIONS AND ACRONYMS

ACC	Accuracy
ANN	Aritificial Neural Network
CNN	Convolutional Neural Network
BCE	Binary Cross Entropy
CV	Computer Vision
DL	Deep Learning
GAN	Generative Adversarial Network
HR	High Resolution
IOU	Intersection Over Union
LR	Low Resolution
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
MACC	Mean Accuracy
MIOU	Mean Intersection Over Union
ML	Machine Learning
MSE	Mean Squared Error
NIQE	Natural Image Quality Evaluator
PI	Perceptual Index
PIRM	Perceptual Image Super Resolution Challenge
PSNR	Peak Signal to Noise Ratio
R-CNN	Region-based Convolutional Neural Network
SISR	Single Image Super Resolution
SOTA	State Of The Art
SSIM	Structural Similarity
SR	Super Resolution
UAV	Unmanned Aerial Vehicle

LIST OF SYMBOLS

L_x	Loss relative to x
$\alpha\beta\gamma\delta$	Loss contribution factors
ζ	Degradation function
ϕ_i	Feature maps from layer i
$\mathbb{E}$	Expectancy operator
$\sum_{i=0}^{N} x_i$	Sum of x_i from $i = 0$ to N
G	Generator
D	Discriminator
p_{ij}	Quantity of pixels from class i classified on j
G_θ	Parameters of network G
$\min_{G_\theta}$	Optimized parameters θ over G
$\|x\|$	Module of x
$\|\|x\|\|_n$	n-th norm of x

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

1.1 Motivation

Image super-resolution (SR) is an image restoration process that aims to recover high-resolution (HR) images from low-resolution (LR) samples as accurately as possible. It is a challenging problem in the computer vision (CV) field because of the difficulty in mapping the LR to the HR space, especially because of the ill-posed nature of this problem, since one low-resolution input could represent multiple high-resolution counterparts (YANG; HUANG, 2017).

Multiple classic algorithms tackle the SR problem (FREEMAN; JONES; PASZTOR, 2002; GLASNER; BAGON; IRANI, 2009; FARSIU et al., 2004), but in recent years we verified a hasty growth of machine learning (ML) methods that accomplish immense achievements when compared to other methodologies that do not employ ML techniques. This is mostly due to the development of deep neural networks, which are mathematical structures that simulate the human brain and can extract multiple features from data.

Deep Learning (DL), a branch of machine learning that learns the hierarchical representation of data, displays superior handling of unstructured data that translates into a complex yet efficient robust algorithm modeling (SCHMIDHUBER, 2015; ROHITH; KUMAR, 2020). The capacity of extracting high and low-level abstractions allows deep learning methods to be extended in a range of fields, such as medicine (RONNEBERGER; FISCHER; BROX, 2015; LI et al., 2018) and remote sensing (ZHANG; LIU; WANG, 2018; FANG et al., 2018). The latter domain, which is remarkably complex because of intrinsic properties of spatial data, will be discussed in this work.

The wide usage of satellite imagery in multiple fields, such as agriculture, city planning, military applications, and environmental monitoring suggests a great demand for satellite products, which oftentimes have low resolution due to limitations of the imaging equipment or communication bandwidth. In particular, generating high-resolution data – either in temporal or spatial domains – can be helpful in several applications. Modern remote sensors on satellites can provide very good spatial resolution (< 1m). However, their launch costs are astronomically high: the WorldView-4 satellite shown in Fig.1.1, for example, is a commercial earth observation satellite launched in November 2016, which had an estimated cost of $ 835 million dollars (SMITH, 2012). This redeems images that

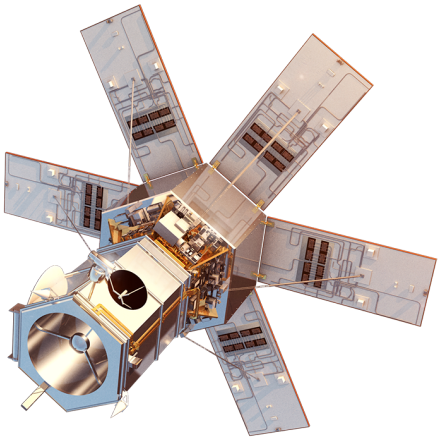


Figure 1.1 – Satellite World-View 4. Photo: Digital Globe.

are usually under very restrictive licenses or are financially prohibitive.

Unmanned aerial vehicle (UAV) based solutions, mostly capable of delivering high-resolution (HR) imagery at a low temporal range, could be an “affordable” alternative to satellite products, but can only cover relatively small areas due to limited UAV autonomy. Therefore, low-resolution (LR) satellite imagery is still used in applications where higher resolution data is more helpful (DAI et al., 2016). In this context, enhancement of LR data is a useful way of achieving better quality in imagery where visual quality is essential (DAI et al., 2016; THORNTON; ATKINSON; HOLLAND, 2006; XU; LIN; MENG, 2017; REETH et al., 2012).

Enhancing low-resolution satellite data is a particularly challenging problem for multiple reasons (SHERMEYER; ETTEN, 2019; ETTEN, 2019; ETTEN, 2019): objects invariant to rotation and orientation, small spatial extents of some objects and their clusters, training example frequency, and massive raw data. Such problems reduce the image restoration capability by encumbering the replication of high-frequency data that are very distinguishable to the human eye.

Despite the good results achieved by DL algorithms, the replication of specific textures or the generation of undesired artifacts are still challenges (ZHAO et al., 2019). One cause is the difference between human and machine perception: metrics for measuring the reconstruction quality are based on machine perception; thus improvements in SR images are not equally perceived between humans and machines. This is noticeably verified in Figure 1.2, where the image on the right displays the highest Peak Signal-to-Noise Ratio (PSNR), which is a well-known quality assessment metric but is clearly worse (visually) than the image in the center. There are multiple propositions of score functions (GOODFELLOW et al., 2014; WANG et al., 2018c; RABBI et al., 2020) that aim to

(a) Original HR image (b) PSNR:29.9049 (c) PSNR:29.9197

Figure 1.2 – PSNR comparison of (a) an original HR image, (b) a slightly modified high-resolution image and (c) a filtered high-resolution image using bicubic interpolation. Although having similar PSRN values, their perceptual quality are very different

“humanize" comparisons between the original and reconstructed images, but defining an “ideal" perception score index similar to a human-based opinion score is still a research challenge.

Super-resolution methodologies that focus on generating perceptually better images mostly focus on tailoring objective functions that influence optimization of the ML model (VASU; MADAM; RAJAGOPALAN, 2018). One of the most famous examples is the perceptual loss proposed by Johnson, Alahi e Fei-Fei (2016), which calculates the differences of intermediates features of a VGG-19 model (SIMONYAN; ZISSERMAN, 2014) when using the ground truth and reconstructed images as inputs. Tailoring such functions and combining them with natural pixel-to-pixel objective functions, such as mean squared error (MSE), have proven to create state-of-the-art (SOTA) strategies in the super-resolution domain (VASU; MADAM; RAJAGOPALAN, 2018).

In this context, it is noticeable that a task-oriented network could act as an objective function, providing semantic guidance to the SR module. Such semantics inputs would be able to capture further details about local texture, thus inspiring the SR network conditioning to produce sharper images with realistic textures. In this work, we use an additional loss function related to semantic segmentation as a task-oriented objective function. Our hypothesis is that segmentation results of a super-resolved image can condition the SR module to generate class-aware textures, thus improving the perceptual scores of reconstructed data.

1.2 Main goals

The generation of high-resolution imagery from low-resolution sensors has a direct impact on multiple applications where the usage of HR data is essential. But most of the existing SR methods are still far away from reconstructing realistic textures, since failure in reconstructing textures, especially the information-rich regions, leads to blurry, overly smooth and unnatural appearance of synthesized images (VASU; MADAM; RAJAGOPALAN, 2018). The main goal of this thesis is to develop a deep SR approach that can effectively produce realistic images, particularly in textured regions. For that purpose, the following specific goals were defined:

- introduction of task-tailored functions based on semantic segmentation for training a super-resolution approach;
- study and evaluation of different baseline approaches for super-resolution and semantic segmentation;
- comparison of the proposed strategy with state-of-the-art (SOTA) super-resolution approaches using perceptual quality assessment metrics.

1.3 Contributions

This work proposes a super-resolution methodology capable of generating high-detailed images with better perceptual quality indices than existing approaches. The main contribution of this work is the introduction of a task-driven joint learning strategy that uses a segmentation module as a component of the loss function, backed by a Generative Adversarial Network (GAN) module responsible for the super-resolution itself. The proposed method yielded better perceptual metrics for two baseline GAN-based modules: ESRGAN (WANG et al., 2018c), a classic model for super-resolving images using Generative Adversarial Nets, and the SAGAN (ZHANG et al., 2019), a pioneer work to propose an ensemble of GAN and attention networks for the SR task. Regarding the semantic segmentation modules, we also explored two methods: the widely known UNet (RONNEBERGER; FISCHER; BROX, 2015) and the state-of-the-art HRNet (WANG; CHEN; HOI, 2020). Even though we specifically observed great improvements for only four combinations of SR-Segmentation modules, this technique is generic and applicable to any combination of super resolution and segmentation networks.

1.4 Text outline

The document is organized as follows. Chapter 2 describes an overview of Single Image Super Resolution, the main subject of this thesis, by summarizing the problem formulation and providing recent improvements on the super resolution topic. It also provides additional information about other topics related to this thesis, such as attention networks, semantic segmentation, joint learning and perceptual quality. Chapter 3 describes the proposed method whilst describing the tools employed in the experiments, such as data sets, training procedure, and evaluation policies. Chapter 4 provides the experimental results of multiple runs aggregating different data sets, SR and segmentation networks by showing reconstruction and segmentation metrics of super-resolved images. This chapter also discusses the generalization capacity of the proposed method over different data sources, and finally assembles a comparison between multiple state-of-the-art SR procedures. Chapter 5 recapitulates the research objectives, results and contributions, while providing future work expansions over the original proposition. Finally, Appendices A, B and C contains image outputs of experiments described on Chapter 4.

2 BACKGROUND AND RELATED WORK

This chapter presents the background and related work about Single Image Super Resolution (SISR), semantic segmentation, proposals of joint-learning techniques and an overview of perceptual quality on image restoration processes.

2.1 Background

This section describes major concepts used in this work, with details about the mathematical background behind major subjects, such as Image Super Resolution, Semantic Segmentation and Perceptual Quality.

2.1.1 Single Image Super Resolution

Image Super Resolution is the process of generating high-resolution images from lower resolution inputs. The SR task has been studied for decades (IRANI; PELEG, 1991). Sampling methods such as bicubic and Lanczos (DUCHON, 1979) interpolations are some of the first methods to super-resolve images, producing quite often blurry results with aliasing artifacts.

SR methods can be divided in two families: the classical multi-image super-resolution, where sets of unaligned low-resolution pictures of the same scene impose linear constraints for building the high-resolution space, and Single Image Super Resolution (SISR), where the method learns correspondences between low and high-resolution patches of image pairs. Due to numeric limitations to generate images with great scale factors in the first approach, SISR became a default approach to reconstruct images from lower resolution inputs(GLASNER; BAGON; IRANI, 2009).

Popular SISR approaches use exemplar-based learning to exploit differences between multiple scale representations of the same image (YANG; HUANG; YANG, 2010) or prior knowledge under the form of large external databases or dictionary-based methods (TIMOFTE; ROTHE; GOOL, 2016; YANG et al., 2012). External priors extracted from large collections of image pairs are, however, very expensive and only produce marginal gains at the SR task (LIANG et al., 2021).

In recent years, the evolution of neural network approaches has shown superiority

over super-resolution tasks from other domains. This is due to the active exploration of deep learning (DL) techniques, mainly supported by the development of efficient computing hardware and sophisticated algorithms. The strong capacity of DL methods made them achieve state-of-the-art performance in multiple SR benchmarks (WANG; CHEN; HOI, 2020).

Current deep learning methods use feed-forward networks to learn a mapping function G between a pair of high-resolution I_{HR} and low-resolution I_{LR} images, defined by

$$I_{HR} \approx G(I_{LR}; \theta), \tag{2.1}$$

where θ are the parameters of G. The LR image is usually generated through a degradation process that is unknown and can be affected by multiple factors, such as compression artifacts, anisotropic degradations, sensor noise and speckle noise (WANG; CHEN; HOI, 2020). It is common to apply a unique downsampling operation to obtain a low-resolution counterpart:

$$I_{LR} = Deg(I_{HR}; \zeta; s), \tag{2.2}$$

where Deg and ζ are the degradation function and its parameters, respectively, and s is the scaling factor.

In order to optimize G_θ, it is necessary to define a cost function L_θ to compute the reconstruction quality of super-resolved images I_{SR}. Such optimization could be viewed as

$$\min_\theta \sum_n L_\theta(I^n_{SR}, I^n_{HR}) \tag{2.3}$$

over the n training pairs. Choosing the Mean Squared Error (MSE), a widely used metric for quantitatively evaluating the image restoration quality, as a loss function, Equation (2.3) can be rewritten as

$$\min_\theta \sum_n ||G_\theta(Deg(I^n_{HR})) - I^n_{HR}||^2. \tag{2.4}$$

As most methodologies aim to improve distortion measures, such as Peak Signal-to-Noise Ratio (PSNR), several SR approaches explore only the MSE loss function (note that PSNR and MSE are closely related). It brings, however, blur and over-smoothed textures that are introduced in the regression-to-the-mean problem, usually caused by conventional MSE-oriented loss functions (WANG et al., 2018b).

In this work, we will focus on a single-sensor spatial super resolution of aerial im-

agery. This description is necessary to clear out that other types of SR, such as mono/multi-sensor temporal SR (which is often studied in the remote sensing context), won't be adopted in this work.

2.1.2 Generative Adversarial Networks

A major progress in the generation of realistic images was made by Generative Adversarial Networks (GAN). Proposed by Goodfellow et al. (2014), Generative Adversarial Nets consist of two adversarial models: a generative model G_{gan} responsible for capturing the data distribution and a discriminative model D_{gan} that estimates the probability of sample being a (original) training sample or a (fake) data produced by G_{gan}. In other words, G_{gan} builds a mapping function from between the input distribution and the data space (in our case, the I_{HR} space) and D_{gan} outputs a single scalar representing the probability of G_{gan} producing original or fake data.

Both generator and discriminator are optimized simultaneously by minimizing an objective function

$$L(G_{gan}, D_{gan}) = \mathbb{E}[\log\left(D_{gan}(I_{HR})\right)] + \mathbb{E}[\log\left(1 - D_{gan}(G_{gan}(I_{LR}))\right)], \tag{2.5}$$

where $\mathbb{E}$ is the expectancy operator over the training samples. Notice that the "simultaneous" optimization of both G_{gan} and D_{gan} create a min-max game, where both generator (in order to fool the discriminator) and D_{gan} (in order to distinguish real and fake data) are continuously optimized. When the optimization via backpropagation occurs, the generator is trained to minimize

$$L_g(G_{gan}, D_{gan}) = \log(1 - D_{gan}(G_{gan}(I_{LR})), \tag{2.6}$$

while the discriminator aims to minimize

$$L_d(G_{gan}, D_{gan}) = \log(D_{gan}(I_{HR})) + \log(1 - D_{gan}(G_{gan}(I_{LR})). \tag{2.7}$$

The GAN loss described by Goodfellow et al. (2014) is, in practice, hard to optimize, since Equation 2.5 may not provide sufficient gradient for G_{gan} to learn well. According to the authors, early in learning, when G_{gab} is not capable of replication the original data distribution, D_{gan} reject samples with high confidence because of the

clear difference from the training data. In this scenario, the second part of Equation 2.5, $\log\left(1 - D_{gan}(G_{gan}(I_{LR}))\right)$ provides very small gradients, difficulting the training procedure. Rather than using the latter part of the equation, we train G_{gan} to maximize $\log\left(D_{gan}(G_{gan}(I_{LR}))\right)$ because it provides much stronger gradients early in learning. Therefore, the Equation 2.6 is, in this study, modified to

$$L_g(G_{gan}, D_{gan}) = -\log(D_{gan}(G_{gan}(I_{LR})). \tag{2.8}$$

2.2 Semantic Segmentation

Semantic segmentation is the task of classification and / or clustering correlated parts of images on a region or pixel level. This is a core (and challenging) computer vision problem since it requires a full understanding of a scene to correctly infer it: aspects such as color, texture, luminosity and perspective variation could cause algorithms to miss-segment even the most "trivial" regions. Multiple applications nourish from describing an entire scene on a pixel level such as medical diagnosis (OUAHABI; TALEB-AHMED, 2021), autonomous driving (CORDTS et al., 2016) and remote sensing (YUAN; SHI; GU, 2021).

The majority of semantic segmentation tasks assigns one single label to each image pixel (LATEEF; RUICHEK, 2019). Its formulation can be simply stated by finding a way to assign every pixel from the image I to the label space $\mathcal{L} = \{l_1, l_2, ..., l_{N_c}\}$ with N_c classes, being sometimes assigned to $N_c + 1$ classes when treating l_0 as a background or void class. The labeled image I_M has, therefore, the same shape of I and has pixel values from $\mathcal{L}$.

Traditional algorithms are heavily based on Markov Random Fields (GEMAN; GEMAN, 1984) (indirect probabilistic graph model) that generate a hierarchical approach of clustering an image by assigning random variables for every pixel in the image. The application of Markov properties on indirect graphs created, since the 80s, an optimized approach to finding similarities between pixels in the feature space, but was unable to capture global impressions of a scene. From clustering algorithms using information from contour and edges, Ren e Malik (2003) is famous for proposing a superpixel-based approach to cluster segments based on intra- and inter-region similarity by classical descriptors, such as contour, texture and brightness. It brought lower computational complexity since it employs the Normalized Cuts (SHI; MALIK, 2000) algorithm, an optimized graph

partition algorithm solved as a generalized eigenvalue problem.

Despite the popularity of such methods, advances in deep neural networks revolutionized the semantic segmentation task. According to Garcia-Garcia et al. (2017), deep learning architectures displayed such improvements in terms of accuracy and sometimes even in efficiency that they easily surpassed non-DL approaches by far, mostly due its the great capacity of recognizing either low-level features that describe local properties and high-level structures that capture global object information.

Optimization of semantic segmentation models uses specific loss functions rather different from the ones used in super resolution tasks. As the choice of objective function is essential to instigate the learning process of the algorithm, multiple propositions, sometimes forged to adhere to specific domains, were proposed: common ones are the Categorical Cross Entropy and its weighed or balanced variations, the Focal Loss (LIN et al., 2017), seen as an adaptation of the BCE that works well for highly imbalanced class scenario, and the Dice Loss (MILLETARI; NAVAB; AHMADI, 2016), an adaptation of the dice coefficient (of F-score). For this work, a multi-class cross entropy is adopted as a loss function, described by

$$L_{seg} = \sum_{i=0}^{N_c} -t_i \log q_i, \tag{2.9}$$

where N_c represents the number of classes, t_i is equal to 1 if the analyzed ground-truth pixel is from class i or 0 otherwise, and q_i is the normalized probability of such pixel being classified as class i. This equation produces, therefore, low values of L_{seg} when pixels are correctly predicted with high confidence, since the logarithmic factor will tend to zero, or very high values for low-confidence correct classification.

From the many proposed criteria to evaluate the performance of segmentation models, popular functions are the pixel accuracy and Intersection over Union (IoU) and its mean variations for multi-class problems. Pixel accuracy is defined as the ratio of correctly classified pixels divided by the total number of pixels in a image:

$$Acc = \frac{\sum_{i=0}^{N} p_{ii}}{\sum_{i=0}^{N} \sum_{j=0}^{N} p_{ij}}, \tag{2.10}$$

where p_{ij} denotes the quantity of pixels from the class i classified as j from N analyzed classes and the background. In other words, p_{ii} represents the number of true positives, while p_{ij} and p_{ji} (for $i \neq j$) can be interpreted as false positives and false negatives, respectively. The pixel accuracy can also be calculated per-class by modifying Equation

2.10: the numerator should represent the sum of correctly predicted pixels from class i and the denominator need to calculate the overall quantity of pixels of the $i-th$ class. Its mean multi-class variant has a similar formulation, being averaged between the quantity of classes described in the problem:

$$mAcc = \frac{1}{N+1}\sum_{i=0}^{N}\frac{p_{ii}}{\sum_{j=0}^{N}p_{ij}}. \tag{2.11}$$

The Intersection over Union (IoU) or the Jaccard Index computes the ratio between intersection and union of two sets (ground truth and predicted segmentation), which can be reinterpreted as the ratio of true positives (intersection) over the sum of true positives, false positives and false negatives (union). It ranges between 0 and 1, and for the ith class, the IoU is formulated by

$$IoU = \frac{p_{ii}}{\sum_{j=0}^{N}p_{ij} + \sum_{j=0}^{N}p_{ji} - p_{ii}}. \tag{2.12}$$

Mean-IoU (mIoU) is defined as the average IoU over all N_c classes, according to Equation 2.13. It stands out as the most used segmentation metric for challenges and researchers due to its representativeness and simplicity (GARCIA-GARCIA et al., 2017).

$$mIoU = \frac{1}{N+1}\sum_{i=0}^{N}\frac{p_{ii}}{\sum_{j=0}^{N}p_{ij} + \sum_{j=0}^{N}p_{ji} - p_{ii}}. \tag{2.13}$$

Other common reported metrics to verify segmentation performance are mostly based in combinations true positive, false positive and false negative values, like the Recall, F1 and Dice scores, but their similarity with the already well-known IoU made them a bit less popular.

2.3 Joint Learning

There is a recent focus on creating image processing pipelines that gather operations from multiple computer vision tasks, such as super-resolution, semantic segmentation, object detection and instance segmentation. These tasks were traditionally tackled in isolation by using a tailored neural network to optimize each problem. However, the multi-tasking capabilities of the human brain motivated researchers to develop multi-task learning, aiming to optimize one or many tasks by using a pool of concurrent task-oriented

models.

This is very noticeable in the super-resolution task, which is often used as a tool to enhance tasks such as object detection (HARIS; SHAKHNAROVICH; UKITA, 2018b) and image segmentation (PEREIRA; SANTOS, 2020). Such enhancements are noticeable in specific applications, such as remote sensing: super-resolving images is proven to be a powerful proxy to enhance the objection detection task, backed up by the work of Shermeyer e Etten (2019) that reports benefits of over 30% in mean Average Precision (mAP) when super-resolving native 30cm imagery to 15cm.

For a generic multi-task problem, with specific loss functions L_i and task-specific weights w_i, the optimization goal L is typically given by

$$L = \sum_i w_i Li, \tag{2.14}$$

which is often minimized by using stochastic gradient descent. Whereas the collective training of multiple models can induce difficulties in optimizing multiple models, the weight balance is deemed essential to hinder task gradients conflicts or harmonize gradient magnitudes. Multiple authors propose task balancing approaches by adapting the task weights w_i (SENER; KOLTUN, 2018), the task-specific gradients (CHEN et al., 2018) or prioritizing tasks dynamically (GUO et al., 2018).

Regarding applications involving spatial imagery, image super-resolution is mostly used as a proxy to improve other image recognition tasks (MOSTOFA et al., 2020; PANG et al., 2019; PEREIRA; SANTOS, 2020). This work proposes the opposite: a methodology where image SR is improved by another task, namely semantic segmentation. The segmentation maps could leverage important spatial information that would be used in the optimization of the super-resolution module, serving as a perceptual evaluator of super-resolved images by analyzing the texture reconstruction quality face another neural network.

2.4 Perceptual Quality

Perceptual Quality aims to propose a human-based perception of image quality. It aims to describe a "viewer experience", which is a difficult task because of the limited understanding of the Human Visual System. Whereas a subjective assessment of visual quality is the best indicator of image perceptual quality, they are time-consuming, cum-

bersome, and impractical (MOORTHY; BOVIK, 2011). Therefore, objective assessment of image quality became standard in the big data era, where machine learning procedures are notably used (FANG et al., 2020).

Despite advances in reconstruction quality and speed, many neural networks fail to reconstruct realistic and visually appealing images. This is mostly due to the optimization based on pixel-based distortion scores, such as MSE, which encourages the network to find an average of multiple plausible solutions, leading to blurry, over smooth and unnatural aspect in the output, especially in information-rich regions (VASU; MADAM; RAJAGOPALAN, 2018; WANG et al., 2018b).

Pixel-wise losses do not capture well the perceptual differences between images, which can be easily verified by offsetting identical images by only one pixel: the high image correlation prior to our eyes does not translate into low per-pixel losses. Hence, we notice an up-rise in studies that bring new ways of improving the perceived quality of reconstructed images (VASU; MADAM; RAJAGOPALAN, 2018; SAJJADI; SCHOLKOPF; HIRSCH, 2017), such as the development of tailored objective functions that could cope with the human perception.

Perceptual loss (JOHNSON; ALAHI; FEI-FEI, 2016) was introduced as an alternative to conventional per-pixel losses between SR and HR images. It uses high-level feature representations of a pre-trained neural network to perceive the difference between images, thus being able to distinguish semantic differences that are not captured by per-pixel losses. The perceptual loss function L_{per} is defined by the squared euclidean distance of intermediate features ϕ_j from a pre-trained VGG network (SIMONYAN; ZISSERMAN, 2014) trained on the ImageNet (RUSSAKOVSKY et al., 2014) dataset:

$$L_{per} = ||\phi_j(I_{SR}) - \phi_j(I_{HR})||_2^2, \tag{2.15}$$

where L_{per} is averaged by the dimensions (width, height, and depth) of the feature map ϕ_j. Perceptual losses are responsible for breakthrough enhancements in image reconstruction quality, since intermediate feature maps of the VGG network predicted very well the texture disparities in low and high feature depths between reconstructed and original images. This enabled the generation of higher quality images with more coherent textures, as is the case of EnhanceNet (SAJJADI; SCHOLKOPF; HIRSCH, 2017), which employs the perceptual loss in the feature space to overcome the high smoothness of conventional SR methods, as shown in Figure 2.1

Besides the perceptual loss, GANs are well known to generate realistic images in

Figure 2.1 – Comparison between SOTA Super-Resolution techniques to improve PSNR (left) and more plausible results produced by EnhanceNet (right) at 4x scale. Source: (SAJJADI; SCHOLKOPF; HIRSCH, 2017)

SISR. Using the so-called adversarial loss to optimize a model results in restoration of fine details and common patterns, notably verified in the studies of Sajjadi, Scholkopf e Hirsch (2017) and Ledig et al. (2016). These works employ, in fact, a combination of multiple loss functions that searches an equilibrium between distortion and perceptual quality.

From the literature that evaluates such equilibrium, the work of Blau e Michaeli (2018) is remarkably important because it formulates the perception-distortion tradeoff and proves that there is a region in the perception-distortion plane (displayed in Figure 2.2) that cannot be attained regardless of the chosen algorithm. Therefore, an optimal model, which would be close to the boundary of the unattainable perception versus distortion region, can only improve either perceptual quality or distortion at the expense of the other. This explains the difficulty of creating networks that perform very well in both perceptive and pixel-wise metrics.

2.5 Restoration metrics

For applications where images are viewed by human beings, user studies are the the ultimate way to evaluate the performance of a model. For example, multiple persons can be requested to rank synthetic images and the original high-quality image according to their visual quality, and a comparison between random image pairs from different methods can be done. In practice, however, subjective evaluation is usually inconvenient, due to

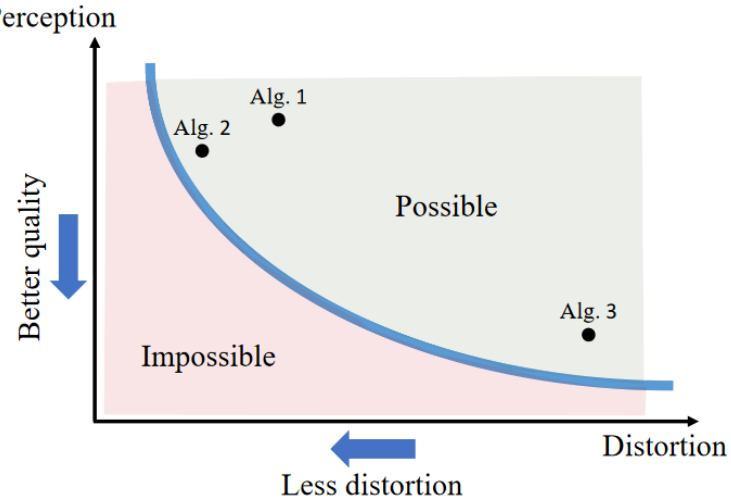


Figure 2.2 – The perception-distortion tradeoff region. Source: (BLAU; MICHAELI, 2018)

cost and/or time expended. Objective image quality assessment metrics is the process of developing quantitative metrics that can automatically perceive image quality (WANG et al., 2004). Such metrics can be classified as full-reference, meaning that a complete reference image is available, no-reference, which blindly evaluates pictures, or reduced-reference, where only parts of the reference image are available, such as sets of extracted features.

According to Wang et al. (2004), the most widely-used full-reference quality metrics are the Mean Squared Error, also referred to as L_2 loss, and the Peak Signal-to-Noise Ratio (PSNR), because they are simple to calculate and have clear optimization purposes. The PSNR, usually expressed in decibels, calculates the ratio between the maximum possible power of a signal and the noise present on it. It is given by

$$PSNR_I = 10 * \log_{10} \left(\frac{MAX_I^2}{MSE} \right), \tag{2.16}$$

where MAX_I represents the maximum possible pixel value of the image, being equal to 255 in 8-bit images. We can now easily see why many SR tasks are optimized by using solely the L_2 loss: models that minimize the mean squared error also maximize PSNR.

Wang et al. (2004) proposed a full-reference quality index by extracting structural information from a scene. Their method quantified image degradation as perceived changes in scene structures by analyzing discrepancies in luminance and contrast measurements. Their proposal, called Structural Similarity (SSIM), was quickly adopted as a reconstruction metric because it did not attempt to predict image quality by accumulating errors associated with simple patterns. The SSIM index between images I_A and I_B is

calculated by

$$SSIM(I_A, I_B) = \frac{(2\mu_{I_A}\mu_{I_B} + c_1)(2\sigma_{I_A I_B} + c_2)}{(\mu_{I_A}^2 + \mu_{I_B}^2 + c_1)(\sigma_{I_A}^2 + \sigma_{I_B}^2 + c_2)}, \tag{2.17}$$

where μ_{I_A} and $\sigma_{I_A}^2$ represent the mean and the variance of I_A, $\sigma_{I_A I_B}$ is the covariance between I_A and I_B and c_1 and c_2 are constants that depends on the dynamic range of the image (typically $2^{nbits} - 1$, $nbits$ being the number of bits that define a pixel).

Traditional metrics like PSNR and SSIM, which rely on low-level differences between pixels, fail to measure the reconstruction quality in a perceived visual manner (JOHNSON; ALAHI; FEI-FEI, 2016). Therefore, multiple studies aimed to propose a fitting similarity descriptor based on human judgment of quality (ZHANG et al., 2018a; JOHNSON; ALAHI; FEI-FEI, 2016). This is a unique challenge due to the high dimensionality of visual patterns and the subjective notion of similarity face the human perception. In this context, Zhang et al. (2018a) proposed the Learned Perceptual Image Patch Similarity (LPIPS), a framework that evaluates distances in deep feature spaces. The authors refine the idea of using feature embeddings of trained networks as elements to calculate a "perceptual distance" between inputs, as first seen in Johnson, Alahi e Fei-Fei (2016), by adding normalization and calibration procedures when computing feature distances. They also use a small network to predict a perceptual judgment between a pair of images.

No-reference image quality assessment can define the visual quality of a reconstructed image without the need of a ground-truth counterpart. No-reference metrics evaluate the image internal components in a way to describe how natural they appear to be. In most SR applications, no HR reference image is available, which explains the increased interest of no-reference evaluation metrics, such as in the Challenge on Perceptual Image Super Resolution (PIRM) (BLAU et al., 2018).

Ma et al. (2017) proposed a modern blind quality assessment of SR images by calculating low-level statistical features in both spatial and frequency domains to quantify super-resolved artifacts, being capable of an effective evaluation based on visual perception. The Natural Image Quality Evaluator (NIQE) (MITTAL; SOUNDARARAJAN; BOVIK, 2012) employs a collection of "quality-aware" features fitted as a multivariate Gaussian model, being remarkable for the non-exposure to distorted images during the training process. More recently, the Perceptual Index (PI) aggregates the proposals of Ma et al. (2017) and Mittal, Soundararajan e Bovik (2012) to be used as a benchmark in the

PIRM-2018 challenge, following the formulation

$$PI(I) = \frac{1}{2}((10 - Ma) + NIQE). \tag{2.18}$$

As a final comment, it is important to mention that some metrics measure the *similarity* between images while others try to assess *distance* values. For similarity metrics (such as PSNR, SSIM), higher scores are better. For distance metrics (such as MSE, LPIPS or PI), on the other hand, lower scores are desired.

2.6 Related work

In this section, we describe important studies about the core super resolution theme, while also citing multiple researches covering adjoint themes, such as joint learning, perceptual quality and semantic segmentation.

2.6.1 Single Image Super Resolution

Artificial Neural Networks (ANNs) represent the beginning of deep learning approaches to represent data. According to Yang et al. (2019), early ANNs can be traced back to the 1960s, when concepts such as layer perceptrons were introduced, with significant developments achieved in the 80's due to the first implementations of backpropagation algorithms (RUMELHART; HINTON; WILLIAMS, 1986). Convolutional Neural Networks (CNNs) also date back decades (LECUN et al., 1989), but it was only with the advance of powerful Graphics Processor Units (GPUs) that we noticed its explosive popularity on computer vision tasks. One of the pioneer networks using CNN is the LeNet-5 published by LeCun et al. (1998), where stacks of convolutional layers were employed to recognize handwritten characters.

CNNs are very powerful when used over structured data like images. Since they can operate directly on raw images, they are spatial-aware and stacks of convolutional layers are powerful feature extractors. Besides that, the kernel re-usability causes a substantial parameter reduction over fully-connected approaches. The first deep convolutional neural network aimed to tackle the super-resolution problem was presented by Dong et al. (2014). The proposed SRCNN has three convolutional layers and used Rectified Linear Units (NAIR; HINTON, 2010) as activation function, outperforming traditional methods

and demonstrating the strong learning capacity of CNNs in an end-to-end training scenario

From there on, numerous deep learning strategies were proposed to enhance the learning capability of DL models. He et al. (2016) proposed residual connections to improve the training of very deep networks, arguing that such connections ease the learning of identity functions. The author demonstrated that very deep residual networks have better reconstruction quality than non-residual models, which was a huge achievement in solving the vanishing gradient problem (that gets more serious as the model gets deeper). This strategy yielded huge boosts in the image recognition task for the ImageNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) dataset and is still very used in current methods. Other powerful concepts that increase the performance and ease the train of deep CNNs are batch normalization (IOFFE; SZEGEDY, 2015), which diminishes data internal variance shift by parameterizing normalized inputs, and skip connections (KIM; LEE; LEE, 2016), where layer outputs are directly fed to deeper regions of the model, allowing the network to register low-level signals.

It is common to find models that adopt a blend of attention mechanisms by whether using them separately (WOO et al., 2018) or jointly (GUO et al., 2022). Recent practices also exploit attention modules embedded in specific families of training strategies, as is the case of the Self-Attention Generative Adversarial Network (SAGAN) (ZHANG et al., 2019), which employs attention-driven modeling for GAN-based image generation tasks. In remote sensing applications, we notice the usage of both spatial and channel attention in CNNs, in most of the cases, but also in GANs, where the attention modules can be applied in both generator and/or discriminator depending on the targeted task (GHAFFARIAN et al., 2021).

In the super resolution field, state-of-the-art (SOTA) restoration quality is achieved by multiple solutions: from simpler residual-learning strategies like the Enhanced Deep Residual Networks (EDSR) (LIM et al., 2017), Residual Dense Network (RDN) (ZHANG et al., 2018c) and Densely Residual Laplacian Network (DRLN) (ANWAR; BARNES, 2020), to more complex strategies using Generative Adversarial Networks such as Super Resolution Generative Adversarial Network (SRGAN) (LEDIG et al., 2016) and Enhanced Super Resolution Genenative Adversarial Network (ESRGAN) (WANG et al., 2018c). Recent contributions also propose channel attention mechanisms to produce accurate super resolved images, like Residual Channel Attention Network (RCAN)(ZHANG et al., 2018b) and Cross-Scale Non-Local Attention (CSNLN) (MEI et al., 2020b), or

a mixed procedures (generally also using attention networks), such as back-projection based Deep Back-Projection Network (DBPN) (HARIS; SHAKHNAROVICH; UKITA, 2018a) and Attention Back Projection Network (ABPN) (LIU et al., 2019) or pyramidal nets like Pyramid Attention Networks (PAEDSR) (MEI et al., 2020a).

In fact, attention mechanisms are quite common in the super resolution literature, where models adopt a blend of attention procedures by whether using them separately (WOO et al., 2018) or jointly (GUO et al., 2022). In remote sensing applications, we notice the usage of both spatial and channel attention in CNNs, in most of the cases, but also in GANs, where the attention modules can be applied in both generator and/or discriminator depending on the targeted task (GHAFFARIAN et al., 2021).

2.6.2 Generative Adversarial Networks

GAN mechanisms are very powerful, thus backing up multiple state-of-the-art studies about generative models in tasks ranging from image translation (HUANG et al., 2018), image manipulation (WANG et al., 2018a) and image restoration (XU et al., 2017; TSAI et al., 2017). In the context of SISR, multiple SOTA applications employed adversarial training, such as the EnhanceNet (SAJJADI; SCHOLKOPF; HIRSCH, 2017), which used a fully convolutional net with learn-able "deconvolution" layers (or convolutional layers with fractional stride) and a perceptual loss; SRGAN, which also exploits the perceptual loss with a modified ResNet architecture, skip connections and ParametricReLU (HE et al., 2015) activations for the generator; and the ESRGAN (WANG et al., 2018c), operating on a new Residual-in-Residual Dense Block (RRDB) without batch normalization and a relativistic discriminator. Despite the ability to produce good visual results, GANs often suffer from instabilities in the training process, which might generate undesired artifacts (ZHAO et al., 2019; BLAU et al., 2018).

Recent practices also exploit attention modules embedded in specific families of training strategies, as is the case of the Self-Attention Generative Adversarial Network (SAGAN) (ZHANG et al., 2019), which employs attention-driven modeling for GAN-based image generation tasks.

2.6.3 Semantic Segmentation

Certain networks have made significant contributions to the field and have become widely known standards, such is the case of AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), VGG (SIMONYAN; ZISSERMAN, 2014), GoogLeNet (SZEGEDY et al., 2015) and ResNet (HE et al., 2016), winner models from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) competition in the years of 2012, 2013, 2014 and 2016, respectively. The UNet framework (RONNEBERGER; FISCHER; BROX, 2015) also achieved great popularity: its simple yet efficient fully convolutional architecture employing 3×3 convolutional filters and a set of max-pooling and upsample layers, is still used as backbone to multiple general-purpose DL applications.

In the remote sensing field, such models are commonly used as the backbone and later tailored for specific tasks to better handle complicated scenarios. That is the case of propositions of Zhang, Liu e Wang (2018), which employs a deep residual UNet for road extraction, and Zhang, Liu e Wang (2018), which used a ResNet-based architecture to classify urban land usage. Such tailored models still face difficulties in achieving very good segmentation results because of the difficulties in classifying complex aerial scenes, which are mostly represented under a large volume of data (YUAN; SHI; GU, 2021). Amongst SOTA models, architectures such as EfficientNets Tan e Le (2019) or YOLO adaptations (HURTIK et al., 2022) are commonly used as backbone, further enhanced with implementations of new pre and self-training techniques (ZOPH et al., 2020). Naive decoders are also implemented amongst top-performing models, as is the case of HRNet (WANG; CHEN; HOI, 2020), which uses four parallel stages of convolutional blocks on different resolutions that are further merged by an up-sampling process, bringing SOTA segmentation results for the CityScapes dataset (CORDTS et al., 2016).

2.6.4 Joint Learning

Even though the idea of using multiple helper functions to optimize a goal task is quite recent in the DL world, we notice several propositions of end-to-end training methodologies that ensemble multiple computer vision tasks. Haris, Shakhnarovich e Ukita (2018b) propose a task-driven framework by using a super-resolution component base on a Deep Back-Projection Network (DBPN) (HARIS; SHAKHNAROVICH; UKITA, 2018a) and a fixed Single Shot MultiBox Detector (SSD) (LIU et al., 2016) capable of de-

tecting multiple objects. They introduce a task-driven compound loss $L = \alpha L_{rec} + \beta L_{task}$, where L_{rec} is the Mean Squared Error reconstruction loss and L_{task} is the detection loss.

Optimization of object detection networks are also proposed on JCS-Net (PANG et al., 2019), which specify a joint classification and super-resolution network to improve small-scale pedestrian detection, and Joint-SRVDNet (MOSTOFA et al., 2020), who introduced an end-to-end joint training process to enhance vehicle detection. Rabbi et al. (2020) also use super-resolution along with object detection, but focus on recovering high frequency edge details by using a Edge-Enhanced GAN (JIANG et al., 2019).

2.6.5 Perceptual Quality

Objective evaluation of perceptual quality is a mature subject in the computer vision field. Studies of Wang, Sheikh e Bovik (2002) and Moorthy e Bovik (2011), for example, are some noticeable contributions, which evaluates natural scene statistics to propose a no-reference evaluation metric. More recently, the NIQE (MITTAL; SOUNDARARAJAN; BOVIK, 2012) and PI (BLAU et al., 2018) also evaluates the natural image quality in no-reference way, being employed in multiple perceptually-aware reconstruction metrics.

In other hand, deep learning procedures that evaluates (or enhances) perceptual quality is a very recent topic, mostly due to contemporary improvements in hardware and software. Its first core subject, the perceptual (or feature reconstruction) loss , was proposed by Johnson, Alahi e Fei-Fei (2016) to capture distances between intermediate feature maps on a pre-trained VGG-19 network. He also introduces a style loss that penalize differences in style, such as colors and textures, but such function was not employed in the SR problem. Another DL-oriented perceptual evaluator is the LPIPS (ZHANG et al., 2018a), which evaluates cossine distances between deep feature spaces using VGG-19 or AlexNet models. This work also compare the discrepancies between low-level metrics and classification networks, proving that trained networks learns a representation of the world that correlates well with perceptual judgments.

3 A JOINT-LEARN METHODOLOGY FOR IMAGE SUPER-RESOLUTION

This work aims to suggest a methodology able to improve the super-resolution perceptual quality of any neural network by employing a task-oriented approach. A semantic segmentation module was chosen as a support function to evaluate the super-resolved outputs, allowing it to be an evaluator proxy of the SR generative model. In other words, the segmentation module will force the generator to produce better SR images in regions where there is compatibility between the original and SR segmentation masks, and we believe that textured regions can benefit the most from the strategy. The segmentation loss (L_{seg}) provided by the comparison of SR output masks and original masks is not related to traditional per-pixel metrics, allowing to focus on class content on synthesized images and, therefore, create perceptually better images.

To assert the validity of this proposal, multiple experiments were run with a combination of super-resolution and segmentation modules, on different datasets. For the SR networks, we analyze the behavior of two GAN-based super-resolution modules: ESRGAN (WANG et al., 2018c) and SAGAN (ZHANG et al., 2019). Generative Adversarial modules were chosen because the adversarial training process is able to generate realistic textures, and GAN-based architectures are a to-go in SOTA perceptual-oriented networks (BLAU et al., 2018). ESRGAN was chosen as a direct enhancement of the already proposed SRGAN, the first GAN-based super-resolution method in the deep learning literature, while the SAGAN method was a pioneer when using attention mechanisms and generative adversarial nets.

For the segmentation module, two were chosen: the first one is the vanilla UNet (RONNEBERGER; FISCHER; BROX, 2015), widely used in multiple tasks and commonly chosen as a backbone for specific applications. The "vanilla" descriptor means that the UNet module used in this work is identical to the author's proposal. The second choice was HRNet (WANG et al., 2020), a powerful state-of-the-art multi-purpose procedure that keeps high-resolution feature maps through the training process and implements information exchange between lower-resolution fields. Both models will describe the hybrid task behavior on a common (and theoretically less performing) and actual (and perhaps more restrictive) models.

Regardless of the chosen SR and segmentation baselines, the objective loss of our training procedure is a compound of multiple individual loss functions. The first component is the MSE reconstruction loss (L_{mse}), widely explored in image restoration

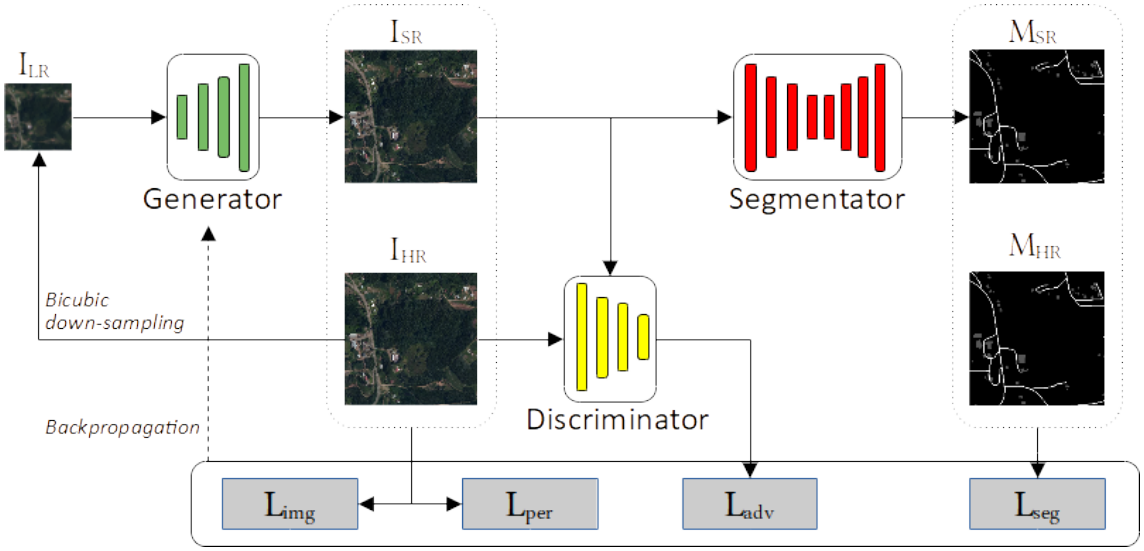


Figure 3.1 – Schematic representation of the proposed methodology. The prefixes I and M indicates the rgb image and the gray-scaled mask, respectively, while de prefix L describes the loss functions employed in this work.

techniques. As given in the Equation (2.4), the mean squared error objective between the ground-truth high-resolution image I_{HR} and the synthesized image I_{SR} is given by

$$L_{mse} = \sum_{i} \|I_{HR} - I_{SR}\|^2 \tag{3.1}$$

for every pixel i of the image in question (or batch of images, in which case the value is averaged by the amount of images). In other words, it's the sum of euclidean distance between I_{HR} and I_{LR} correspondent pixels, for all three RGB bands.

Introducing a Generative Adversarial Network as a learning mechanism usually implies the usage of a GAN generative loss (L_g, from Equation (2.6)). The min-max loss described by Goodfellow et al. (2014) provides an effective way of training the generator to produce more realistic images while improving the capacity of the discriminator to identify real or fake images. Whilst the discriminator loss function L_d is used only in the optimization of the discriminator, L_g is employed in the generator optimization in a slightly modified version from Equation (2.6):

$$L_{gan} = -\log(D_{gan}(G_{gan}(I_{LR})), \tag{3.2}$$

and it can be viewed as another way of framing the loss perspective, where the generator maximizes the probability of images being real, instead of minimizing the probability of an image being fake. This helps with the vanishing gradient problem often observed in the beginning of the training, since the loss expression normally would evaluate to small

values since G is not yet capable of creating good faithful images.

Since our main goal is to achieve visually plausible SR images, we also explicitly explore a perceptual loss (L_{per}) to train the model. As per Equation (2.15), it is given by

$$L_{per} = ||\phi_j(I_{SR}) - \phi_j(I_{HR})||_2^2, \tag{3.3}$$

such that the model is expected to generate textures that are similar to intermediate feature representations of an image. Computing distances into a feature space instead of the image space allows a better representation of high-frequency information, resulting in a photo-realistic images.

Finally, the fourth and final component of the joint loss is the task-driven segmentation loss (L_{seg}). The core idea of using a segmentation model to complement the joint loss is to employ an auxiliary evaluator that mixes the pixel-wise symbolism (since segmentation maps are calculated pixel-wise, such is the cross-entropy loss) and local texture detection (by using segmentation models), when translating the super-resolved image into the label space. This idea shares a similarity with (LIM et al., 2017), which uses the feature space instead, and obtains interesting visual results. Although multiple choices for loss functions related to segmentation can be used, we explored the categorical cross-entropy between the original and SR-inferred masks over N_c classes, given by

$$L_{seg} = \sum^{i} -t_i \log q_i, \tag{3.4}$$

where, for each pixel i of the image, t_i is equal to 1 if the analysed pixel is correctly classified, being equal to 0 otherwise, and q_i determinate the classification probability of that pixel for the analysed class. An ideal classification score ($L_{seg} = 0$) would mean that every pixel is correctly classified with 100 % confidence, and in a totally mislabeled case, L_{seg} would tend to infinite.

The four employed losses are balanced through parameters $(\alpha, \beta, \gamma, \delta)$ according to the following expression

$$L = \alpha L_{mse} + \beta L_{gan} + \gamma L_{per} + \delta L_{seg}, \tag{3.5}$$

and an overview of the proposed method can be readily verified in the Figure 3.1.

3.1 Datasets

Remote sensing data has been widely used as a way to monitor and assess land cover and land usage in natural resources management and change detection in urban and countryside areas (BOGUSZEWSKI et al., 2020). They are used in a large pool of applications, ranging from urban planning (ZHOU; HUANG; CADENASSO, 2011) to vegetation monitoring (BARBEDO, 2018) to military operations.

From the specific niche of satellite imagery, there are multiple datasets covering many computer vision tasks, such as object detection (XIA et al., 2017; LAM et al., 2018), scene classification (SUMBUL et al., 2019; LóPEZ-JIMéNEZ et al., 2019) and semantic segmentation (BOGUSZEWSKI et al., 2020; MOHAJERANI; SAEEDI, 2020). However, just a few of them provide images from high-resolution sensors. Furthermore, accurate multi-label datasets for semantic segmentation tasks are scarce as well.

Choosing and/or manipulating a dataset for machine learning applications is essential for the success of specific hand-tailored tasks. This happens because the dataset is expected to reproduce with fidelity the range of conditions expected to be found in a real-world scenario. Besides, building a comprehensive dataset that captures all visual characteristics of complex environments is often a labor-intensive, complex and error-prone process (BARBEDO, 2018). In this context, this work presents a high-resolution land cover image set that describes very well important features of the terrain in very diversified biomes. The proposed CGEO dataset is one of the results of the mapping project of the state of Rio Grande do Sul - BR, which aimed to refresh the state's cartographic database to follow a better government strategic planning. This project mapped, until now, more than 10.000 km² of the metropolitan area of Porto Alegre, the largest city of Rio Grande do Sul.

The CGEO dataset captures multiple contexts between mix of densely populated areas and extensive agriculture lands, appropriately describing the heterogeneous scene of the global surface. It contains aerial RGB orthophotos with resolution of 50cm/pixel, split in 25,000 patches of 512×512 pixels distributed in three folds: 80% train, 10% validation and 10% test. This collection contains annotations of five classes: natural soil, woodlands, watersheds, roads and buildings.

Aside from the proposed CGEO dataset, a second set of images is also used for the sake of comparison. The LandCoverAI dataset, proposed by Boguszewski et al. (2020), covers a total area of 216.27 km^2 from Poland and contains annotations of four classes:

Figure 3.2 – Sample of CGEO (on top) and LCAI (bottom) datasets. The right part represents the semantic map of such images.

exposed soil, buildings, woodlands and water. The raw images were pre-processed to obtain $10,674$ tiles with dimension 512×512 and a resolution of 50cm. The authors divided the image series into three splits: 70% train set, 15% validation set, and 15% test set. Samples of both datasets are shown in Figure 3.2.

3.2 Evaluation

As the main focus of this work glances on perceptual quality of super-resolved images, LPIPS and and PI (Perceptual Index) are used as the main evaluation metrics to assess the visual quality of super-resolved images: while the first one is a full-reference metric (it uses I_{HR} to calculate its score), the PI is a no-reference metric (since it only uses I_{SR}). Both metrics were used in the PIRM 2018 challenge (BLAU et al., 2018) to select the competition winner amongst various categories. However, we also use conventional reconstruction metrics such as PSNR (Equation (2.18)) and SSIM (Equation (2.17)) as complementary measures for two reasons: first, they are still widely used in multiple research papers involving super-resolution (WANG et al., 2018c; KIM; LEE; LEE, 2016; LEDIG et al., 2016); and second, to analyze discrepancies between per-pixel and per-

ceptual metrics and inferring, after visual analysis, which descriptors better represent the human perception of quality.

Since a segmentation-based loss is used, we expected that the joint approach would be able to yield better segmentation results than SR images produced by a baseline super resolution module. The "baseline" model is a network that did not use the segmentation loss as optimization tool. If confirmed, that hypothesis should reinforce the idea that a joint task could create a closer representation of the latent space described by the observed HR images. Hence, we analyze the segmentation maps of super-resolved images to check the mask outputs during the optimization process. For that, widely known segmentation metrics such as accuracy, IoU and its multi-class counterparts, given by Equations (2.10),(2.11),(2.12), and (2.13), were employed.

3.3 Training procedure

The complete training schedule is composed by two main parts: 1) pre-training the SR / Seg modules that will be used as baseline methods; and 2) fine-tuning of the baseline SR module. The first step, training the segmentation module, was judged necessary for many reasons: first, initial outputs from the Seg module will be dissonant when compared to M_{HR} data, thus contributing poorly (or even against) the learning procedure, which requires some sort of pre-train / warm-up. Second, a pre-trained module will greatly reduce training duration, providing a reusable component between every experiment that uses the same Seg model. Third, it will ensure that a common segmentator is shared between experiments, disclosing the same calculus of L_{seg}.

Both UNet and HRNet modules were trained for 200 epochs by using an Adam optimizer (KINGMA; BA, 2014) with default coefficients $(\alpha, \beta) = (0.9, 0.999)$ and a multi-step learning strategy which schedules the learning rate to to $1e^{-4}, 1e^{-5}$ and $1e^{-6}$ on the first, 100^{th} and 150^{th} epochs, respectively. Weights are initialized using a normal distribution since multiple training procedures report performance gains with such initialization (HE et al., 2015; WANG et al., 2018c). Random flip and rotation ($90°$, $180°$, $270°$) were employed as data augmentation policies. Due to memory restrictions, the chosen batch size varies accordingly to the baseline segmentation module: 16 when using the UNet, and 8 for the HRNet, since the latter has a larger memory footprint.

The baseline super-resolution network also passes through a pre-train procedure, since the fine-tuning of multiple experiments requires a way less time than re-training a

model from scratch. In such training scenarios, the GAN-based networks were trained for 200 epochs and the base learning rate was empirically set to $1e^{-4}$, with a decay factor of 0.1 in the middle of training. Adam optimizer was also employed with default coefficients $(0.9, 0.999)$. The fixed input size of 256×256 was obtained after using random crop, rotation and flip as data augment functions, and the batch size varied according to the choice of the GAN network: 12 for ESRGAN and 3 for SAGAN.

After pre-training the baselines, the SR and Seg modules can be both employed in the joint training process to fine-tune the super-resolution network. Since achieving stable training of GANs is a known problem (GOODFELLOW; BENGIO; COURVILLE, 2016), it was necessary to employ a two-stage sweep strategy to find the optimal hyper-parameters described in Equation (3.5). Initially, a grid search is performed for the SR module alone (i.e., we set $\delta = 0$) on hyper-parameters α, β, γ. To avoid overly large search spaces, the range for each parameter is a set of three values $\{1e^{-3}, 1e^{-2}, 1e^{-1}\}$, yielding 27 runs for each module-dataset combination. The optimal tuple $(\alpha_*, \beta_*, \gamma_*)$ was determined by the highest PSNR/SSIM combination in the validation test at the end of each train, after 50 epochs. In the second stage, the optimal set $(\alpha_*, \beta_*, \gamma_*)$ was frozen, and the full loss function (with the segmentation term) was tested by using grid search only for $\delta \in \{1e^{-3}, 1e^{-2}, 1e^{-1}\}$, yielding a total of 30 runs per module/dataset. For the four possible combinations between SAGAN and ESRGAN modules with LCAI and CGEO datasets, a total of 120 runs were made just to define the optimal hyper-parameters $(\alpha_*, \beta_*, \gamma_*, \delta_*)$ for the experiments. A simple overview of of training procedure is available as pseudo-code in Algorithm 1.

The joint learning experiments are named with a combination of three letters and a number, which indicates the dataset used, the SR module employed, the chosen segmentation module and an indicator of the δ value, respectively. The baseline super-resolution experiments, which do not use the segmentator, have a similar naming convention, but do not use the last two digits and have a 'B-' prefix. For example, CEH1 is a experiment run on **C**GEO dataset using **E**SRGAN and **H**RNET with a loss segmentation weight of $\delta = 1e^{\mathbf{-1}}$. Similarly, the B-LS indicates the **b**aseline **S**AGAN network for the **L**CAI dataset.

Experiments were performed using a single Tesla V100 GPU with 32Gb of memory, a PyTorch v1.7 backend, and Python 3.7.3. Helper top-level libraries such as Pytorch-Ignite (FOMIN et al., 2020), and Hydra (YADAN, 2019) were also employed to organize the training procedure and training metadata organization, respectively.

Algorithm 1 Training procedure

for Dataset in (LCAI,CGEO) **do**
  for Segmentation Module in (HRNET,UNET) **do**
    Pretrain Segmentation Module
  end for
  for Super-Resolution Module in (ESRGAN,SAGAN) **do**
    Pretrain Super-Resolution Module
  end for
  for Super-Resolution Module in (ESRGAN,SAGAN) **do**
    Load Super Resolution pretrained weights
    Sweep hyper-parameters for (α, β, γ)
    for Segmentation Module in (HRNET,UNET) **do**
      Load Segmentator pretrained weights
      Fine-tune the Super-Resolution Module using δ sweep: $(\alpha_*, \beta_*, \gamma_*, \delta)$
    end for
  end for
end for

4 EXPERIMENTAL RESULTS

The multiple combinations between different data sets, SR and segmentation networks, as described in the last chapter, yielded several experiments. As the results are very dependent of the training data, they were organized according to the image source used – either CGEO or LCAI –, resulting in two batches of experiments for each dataset: one for ESRGAN and another for SAGAN as the baseline SR approach. **Best** results per batch of experiments are shown in bold, while the best outcomes per segmentation module are highlighted in blue for the UNet model and in red for the HRNet one.

The generalization capability of the proposed methodology is also analyzed in Section 4.3, when networks trained in one data set were inferred in a different one. Then, Section 4.4 displays a comparison between the proposed method and multiple SOTA super resolution procedures.

4.1 Results for the CGEO dataset

The hyperparameter sweep process yielded the best baseline SR results when using the weights $(\alpha_*, \beta_*, \gamma_*) = (0.1, 0.01, 0.001)$. These values were used to tune the L_{seg}, when another grid search was performed to tune hyper-parameter δ, setting up six procedures (three for each segmentation network). Therefore there are, for each dataset/SR network combination, six runs to be analyzed, plus one regarding the baseline experiment (without the segmentator).

4.1.1 Super Resolution Results

The first set of experiments focus on evaluating the visual quality of super-resolved images using different strategies. Table 4.1 shows the results using ERSGAN as the baseline SR, and it is noticeable that the introduction of the segmentation loss produced better perceptual metrics in every experiment, peaking in the run CEU1, which produced $\approx 25\%$ improvement on LPIPS, better SSIM and even a comparable PSNR when faced against the baseline B-CE that did not use the segmentation module. In fact, almost every experiment performed better in all analyzed metrics: the few exceptions were the marginally lower scores of PSNR observed when using a greater contribution for L_{seg},

which is expected since the attenuation of pixel-wise participation on total loss L would deem it less PNSR-oriented.

Table 4.1 – PSNR, SSIM, LPIPS and PI metrics for SR outputs using CGEO dataset and ESRGAN module.

Experiment	δ	PSNR ↑	SSIM ↑	LPIPS ↓	PI ↓
B-CE	0	30.2424	0.6568	0.2702	7.1803
CEU1	$1e^{-1}$	30.1782	0.6747	**0.2085**	**6.5564**
CEU2	$1e^{-2}$	30.3724	0.6751	0.2148	6.8191
CEU3	$1e^{-3}$	30.5273	0.6792	0.2261	6.7887
CEH1	$1e^{-1}$	29.9092	0.6672	0.2396	6.6893
CEH2	$1e^{-2}$	30.2448	0.6779	0.2410	6.6905
CEH3	$1e^{-3}$	**30.5362**	**0.6806**	0.2511	6.8993

It is also worth noticing the uncorrelated nature between conventional metrics and perceptual quality: it is readily observed that better PSNR/SSIM scores are not always correlated to similar improvements on perceptual metrics, here represented by LPIPS/PI, when comparing CEH1 and B-CE runs, for example. This phenomenon can be explained by the perception-distortion trade-off described by Blau e Michaeli (2018): superior performance of perceptual metrics can come with the cost of distortion and vice versa. Overall, we observe that higher values of δ promoted better LPIPS/SSIM, supporting the theory that realistic textures can be enhanced with segmentation maps from an ensemble model.

For SAGAN-based experiments, reported in Table 4.2, lower values of δ provided better overall metrics, and the experiment running the UNet as the segmentation module yielded the best perceptual metrics. A balanced choice of $\delta = 1e^{-2}$ yielded the best LPIPS/PI results in the experiment CSU2, yet smaller weights also display perceptual improvements and better distortion metrics. For HRNet-based experiments, it is noticed that a larger value of δ in the CSH1 experiment already generates great boost in per-pixel scores, but worse LPIPS and slightly better PI.

Notice that, in this set of experiments, the PSNR was the metric that most benefited from using the segmentation, even in high values of δ. Metrics improvements translated to better visual quality, since the more pleasant images had less jagged edges and smoother transitions between different image features.

Table 4.2 – PSNR, SSIM, LPIPS and PI metrics for SR outputs using CGEO dataset and SAGAN module.

Experiment	δ	PSNR ↑	SSIM ↑	LPIPS ↓	PI ↓
B-CS	0	23.542	0.6385	0.26	7.2329
CSU1	$1e^{-1}$	26.7678	0.6247	0.2536	7.5562
CSU2	$1e^{-2}$	27.9387	0.6522	**0.2294**	**6.6737**
CSU3	$1e^{-3}$	28.935	0.6675	0.2302	7.0847
CSH1	$1e^{-1}$	28.2939	0.6544	0.3117	7.1836
CSH2	$1e^{-2}$	28.806	0.6586	0.2453	6.8459
CSH3	$1e^{-3}$	**28.9896**	**0.6678**	0.2491	6.941

(a) B-CS sample (b) CSU2 sample (c) CSH1 sample

Figure 4.1 – Sample of baseline (B-CS), better LPIPS/PI (CSU2) and worst LPIPS (CSH1) experiments. Even though having worse LPIPS score, the high improvement of PSNR wields sharper and more pleasant super resolved images.

4.1.2 Segmentation Results

Using the segmentation module can also improve the segmentation results of super-resolved images. This is particularly useful in image synthesis tasks involving aerial imagery, where segmentation maps are often necessary for multiple applications. Experiments run using the joint methodology displayed better segmentation scores than the baseline SR method, suggesting that this SR task can be used as a pre-processing phase on other computer vision task, such as the semantic segmentation.

Improvements of segmentation metrics are vastly improved when using the segmentation loss L_{seg} when compared to the baseline experiments. The vanilla implementations of the SR task are not capable, face the segmentator, of producing loyal textures, but higher values of δ force the outputs to be more coherent to the class labels, which is directly translated into better perceptual scores according to Tables 4.1 and 4.2. For runs using the ESRGAN SR network, described on Table 4.3, it is noticeable that the

baseline experiments reveal a domain transfer problem that was also observed in the experiments running the HRNET module: the road region suffered a class reclassification, and its common signature (oftentimes similar to exposed soil) was identified as another class after the super-resolved image synthesis. This behavior was not observed using the UNET module, but the IoU values for the track region are still low. That is why mean class scores, like mean accuracy and mean IoU are better on UNet-based experiments; otherwise the performance of both segmentators would be very similar. Nonetheless, it is noticed great segmentation improvements in every experiment, especially when considering the watershed and building classes. The segmentation maps for runs B-CEU and CEU-1 (best MAcc / MIoU) are displayed on Figure 4.2.

Table 4.3 – Accuracy (Acc), Mean Accuracy (MAcc), Per-class IoU (in the following order: Woodland, Building, Watershed, Road) and Mean IoU (MIoU) for segmentation masks obtained from super-resolved images of CGEO dataset and ESRGAN network.

Experiment	δ	Acc	MAcc	IoU	MIoU
B-CEU	0	0.8282	0.4238	0.8328 / 0.5129 / 0.1619 / 0.0549	0.3906
B-CEH	0	0.8628	0.4162	0.8378 / 0.6223 / 0.2555 / 0	0.4289
CEU1	$1e^{-1}$	0.8998	**0.5803**	0.8818 / **0.7205** / **0.4019** / **0.2356**	**0.56**
CEU2	$1e^{-2}$	0.8999	0.5401	0.8809 / 0.7202 / 0.3977 / 0.1862	0.5464
CEU3	$1e^{-3}$	0.8891	0.5067	0.8670 / 0.6997 / 0.3277 / 0.1527	0.5118
CEH1	$1e^{-1}$	**0.901**	0.4767	**0.8825** / 0.7194 / 0.3980 / 0	0.5
CEH2	$1e^{-2}$	0.8974	0.4665	0.8788 / 0.7166 / 0.3488 / 0	0.4861
CEH3	$1e^{-3}$	0.8897	0.4551	0.8681 / 0.7027 / 0.3141 / 0	0.4712

(a) Original mask (b) B-CEU mask (c) CEU1 mask

Figure 4.2 – Original mask and sample masks after inference of baseline (B-CEU) and better MAcc/MIoU (CEU1) experiments. Notice that the road class is not captured on most experiments, due the domain transfer and thin dimensions of road parts.

Using the SAGAN network has shown similar segmentation results when compared to the ESRGAN model. The baseline experiments are remarkably comparable,

showing that both segmentators initially have similar inference capabilities on synthesized data. Procedures using the UNet observed better metrics on multiple classes, but buildings, watersheds and roads benefited the most, while other classes had smaller improvements. The HRNet model was still not able to detect the road class, nor in the vanilla SAGAN module in the joint methodology with the HRNet as segmentator, but still pleaded great segmentation results on woodland and building classes, as reported in Table 4.4

Table 4.4 – Accuracy (Acc), Mean Accuracy (MAcc), Per-class IoU (in the following order: Woodland, Building, Watershed, Road) and Mean IoU (MIoU) for segmentation masks obtained from super-resolved images of CGEO dataset and SAGAN network.

Experiment	δ	Acc	MAcc	IoU	MIoU
B-CSU	0	0.8485	0.4266	0.8141 / 0.6281 / 0.2239 / 0.0347	0.4252
B-CSH	0	0.8693	0.41	0.8391 / 0.6583 / 0.2370 / 0	0.4336
CSU1	$1e^{-1}$	0.8929	0.5318	0.8698 / 0.7065 / 0.3990 / **0.1856**	0.5402
CSU2	$1e^{-2}$	**0.8977**	**0.533**	**0.8758** / **0.7111** / **0.4280** / 0.1818	**0.5491**
CSU3	$1e^{-3}$	0.8874	0.4977	0.8629 / 0.6984 / 0.3259 / 0.1613	0.5125
CSH1	$1e^{-1}$	0.8859	0.444	0.8629 / 0.7061 / 0.2617 / 0	0.4576
CSH2	$1e^{-2}$	0.8895	0.4564	0.8687 / 0.7056 / 0.3017 / 0	0.469
CSH3	$1e^{-3}$	0.8861	0.4479	0.8630 / 0.6966 / 0.2982 / 0	0.4644

4.1.3 Visual Results

Visual results for the inference process of experiments in this chapter are displayed in the Appendix A. Figures A.1 and A.3 show inference results for SR outputs of experiments from Tables 4.1 and 4.2, while Figures A.2 and A.4 refer to experiments from Tables 4.3 and 4.4. Notice that experiments using the segmentation module are able to better replicate vegetation and soil textures in comparison with the baselines B-CE or B-CS experiments. Resampling the LR image using bicubic interpolation yields jagged linear features (such as roads and cart tracks) and unrecognizable buildings, while the joint learning approach can better reproduce the vegetation granularity and buildings edges.

For the SAGAN-based experiments, the non-optimal choice of δ is quickly noticeable in the SR outputs: the example displayed in Figure A.3 has uncanny granularity for $\delta = 0.1$, but more natural cart tracks on CSU3 and CSH3. It's also very important to notice that small visual artifacts do not seem to reasonably affect perceptually-aware

metrics since CSU2 and CSH2 experiments have the best results for the SAGAN/CGEO module/dataset combination and still showed artifacts.

Figures A.2 and A.4 exhibits how the SR outputs are observed face the segmentation networks. CEU3 run captured the best vegetation contour details and a small nuance of buildings on the top of the image, even when the original label did not assert it. The baseline runs show jagged vegetation contour, suggesting the use of SR task as pre-processing phase for the pixel classification. For the Figure A.4, a forest area on the left of the image is not retracted in the original label, but is still captured in every experiment. The UNet and HRNet modules, although, have slightly different perceptions of the terrain, since the central area, which is a mix of land and vegetation, is captured mostly as woodlands for the HRNet, while the UNet classifies it as land. In both experiments, a major problem present in large computer vision datasets is noticeable: the quality of annotated data. High-quality annotations are usually done manually, so very large aerial imagery datasets would require much manpower to create and review an "ideal" set of labels, which is often not feasible.

4.2 Results for the LCAI dataset

Like the experiments run on the CGEO dataset, the 27-run hyper-parameter sweep wielded optimal vales of $(\alpha_*, \beta_*, \gamma_*) = (0.1, 0.01, 0.01)$, which were used in the joint approach. Since the same optimal set of hyper-parameters was found, there is an indication of good cross-dataset generalization for the ESRGAN proposal.

4.2.1 Super Resolution Results

The baseline experiment dwelled on a vanilla ESRGAN implementation (B-LE) displayed higher values of PSNR and LPIPS if compared to the CGEO dataset, indicating different core characteristics between datasets. The segmentation-guided experiments reported in Table 4.5 show consistent improvements in perceptual metrics when compared to the baseline SR approach, supporting the hypothesis that using a segmentation module could improve super-resolution results. In particular, experiments LEU1 and LEH1 yielded the best overall perceptual scores, while LEU1 displayed improvements on every four metrics. Notice that the LEH1 experiment displayed better perceptual scores but

worse pixel-oriented metrics, which could be correlated to the distortion-perception trade-off already reported in the literature (BLAU; MICHAELI, 2018), which asserts the complementary (or oppose) behavior between pixel-wise observations and perceptual ones.

Table 4.5 – PSNR, SSIM, LPIPS and PI metrics SR outputs using LandCoverAI dataset and ESRGAN module

Experiment	δ	PSNR ↑	SSIM ↑	LPIPS ↓	PI ↓
B-LE	0	32.1645	0.7909	0.2608	7.1715
LEU1	$1e^{-1}$	32.1049	0.7929	**0.1912**	6.794
LEU2	$1e^{-2}$	32.2496	0.7954	0.2092	6.8183
LEU3	$1e^{-3}$	**32.2701**	0.7961	0.2105	6.8752
LEH1	$1e^{-1}$	29.7972	0.7774	0.2041	**6.4591**
LEH2	$1e^{-2}$	31.6353	0.7871	0.2153	6.7021
LEH3	$1e^{-3}$	32.2266	**0.7982**	0.2201	6.8829

The combination of SAGAN and segmentation proxy also yielded better metric measurements against comparison with the baseline approach. Almost every factor of δ, as seen in Table 4.6, reported better values, with the exception of experiments that had $\delta = 0.1$, which created some training instability and degradation of perceptual and pixel-wise metrics. Unbalanced values of L_{seg} in front of L_{mse}, L_{gan} and L_{per} is probably the reason for degradation in the image generation procedure, since incoherent values of L_{seg} are not expected to improve the SR module.

Larger values of δ in experiments LSU1 and LSH1 displayed distortions of PSNR, SSIM and LPIPS against the B-LS run, while the LSU3 and LSH3 had great improvements in all four metrics. As the SSIM value was not so affected between runs (opposed to the PSNR), it is plausible to affirm that the distortions originated by LSU1 and LSH1 are mostly in the pixel level and not on image structure, which is confirmed in Figure 4.3. Experiments LSU2 and LSH2 had an average performance: both have better LPIPS and worse PI values.

4.2.2 Segmentation Results

The segmentation results for the LandCoverAI dataset + ESRGAN module is displayed in Table 4.7. As it was observed in other combinations of segmentator/datasets, the segmentation results also benefit from the joint approach when compared to vanilla SR implementations. Classification improvements were mostly noticeable in the HRNet

Table 4.6 – PSNR, SSIM, LPIPS and PI metrics SR outputs using LandCoverAI dataset and SAGAN module

Experiment	δ	PSNR ↑	SSIM ↑	LPIPS ↓	PI ↓
B-LS	0	28.4893	0.7727	0.2782	6.7645
LSU1	$1e^{-1}$	26.5764	0.7475	0.3566	7.0317
LSU2	$1e^{-2}$	28.4868	0.7797	0.2581	6.813
LSU3	$1e^{-3}$	**30.4482**	**0.7883**	0.2319	**6.6583**
LSH1	$1e^{-1}$	24.5875	0.7507	0.3628	7.1098
LSH2	$1e^{-2}$	29.4145	0.7787	0.2591	6.7822
LSH3	$1e^{-3}$	30.0162	0.785	**0.2237**	6.7004

(a) HR image (b) LSH1 image (c) LSH3 image

Figure 4.3 – HR image, LSH1 and LSH3 generated images. The LSH1 low PSNR value was originated by a pixel translation value, but the image structure is still present.

experiments, since the super-resolved images from these runs improved the building IoU by over 20%. In contrast to the CGEO dataset, the watershed detection worked very well on the baseline experiments, especially using the UNet background and achieving around 90% of IoU. Besides the watershed class, the LCAI dataset is perceived similarly between the UNet and HRNet modules.

The watershed detection in the LCAI set achieved almost 0.95 value in the LEH2 run, which is a remarkable improvement of over 50% from the vanilla B-LEH experiment. For HRNet-based methodologies, segmentation metrics were enhanced in every experiment if compared to the B-LEH experiment suggesting, in consonance with 4.5, that the L_{seg} contribution is guiding a better texture generation face the improved segmentation perception. UNet-based experiment, on the other hand, displayed only minimal enhancements on segmentation maps, but such small improvements were already enough to establish perceptual improvements on the SR module. We also note that, from amongst the four segmentation metrics here displayed, the accuracy is the most correlated to perceptual scores: LEU1 and LEH1 have the best accuracy scores and best perceptual metrics,

which is expected, since higher accuracy means a smaller value of L_{seg} produced by the cross-entropy loss function.

Table 4.7 – Accuracy (Acc), Mean Accuracy (MAcc), Per-class IoU (in the following order: Soil,Building,Woodland,Watershed) and Mean IoU (MIoU) for segmentation masks obtained from super-resolved images of LCAI dataset and ESRGAN network.

Experiment	δ	Acc	MAcc	IoU	MIoU
B-LEU	0	0.9378	0.8476	0.8906 / 0.4538 / 0.8720 / 0.9046	0.7803
B-LEH	0	0.9151	0.7499	0.8569 / 0.4514 / 0.8783 / 0.5944	0.6953
LEU1	$1e^{-1}$	0.9463	0.8441	0.9030 / 0.4681 / 0.8962 / 0.8850	0.7881
LEU2	$1e^{-2}$	0.9438	0.8408	0.9007 / 0.4667 / 0.8895 / 0.8849	0.7855
LEU3	$1e^{-3}$	0.9416	0.84	0.8992 / 0.4666 / 0.8801 / 0.8874	0.7833
LEH1	$1e^{-1}$	**0.9604**	**0.8829**	0.9280 / **0.5630** / 0.9147 / **0.9463**	**0.838**
LEH2	$1e^{-2}$	0.96	0.8797	0.9276 / 0.5504 / 0.9145 / 0.9439	0.8342
LEH3	$1e^{-3}$	0.9597	0.879	**0.9283** / 0.5562 / **0.9176** / 0.9387	0.8353

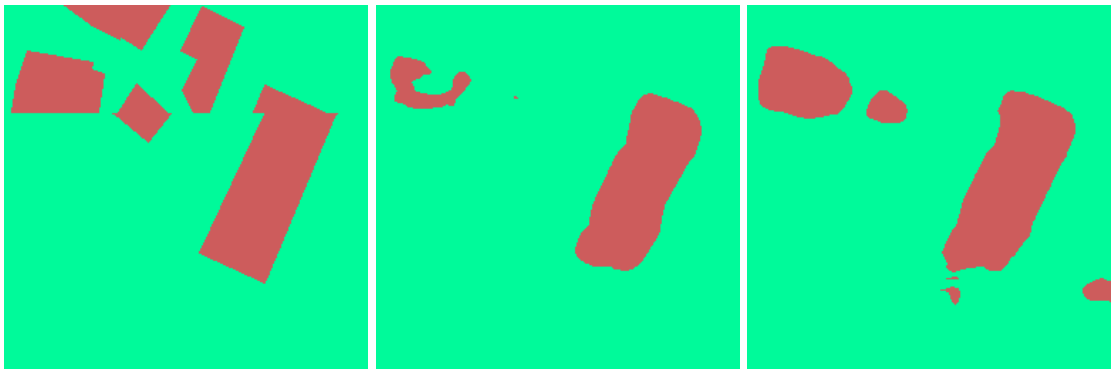
Following up on the SAGAN procedures, we noticed the best relative segmentation performance increase of any combination of dataset/segmentator. The watershed IoU, for example, surged from 0.298 from the B-LSU probe to 0.8483 in the LSU3 run, almost tripling its initial value; for the same tuple of experiments, the building detection improved from 0.0366 to 0.3615, an improvement of around ten times. UNet-based runs observed better results with lower values of δ, as also the HRNet-based ones, but the sensibility to σ adjustments was higher when employing the UNet segmentation module.

Optimization of the SR module behaves differently between different combinations of SR model / dataset. While using the SAGAN module over the CGEO data provided better segmentation metrics for $\delta = 0.01$, LCAI data displayed optimal scores when $\delta = 0.001$, while higher values of δ did not provide the best segmentation metrics (but still yielded great improvements when compared to baseline experiments B-LSU and B-LSH). ESRGAN experiments, on the other hand, usually yield the best scores for higher values of δ.

The UNet-related run witnessed huge IoU improvements on experiment LSU3, especially on buildings (from 0.0366 to 0.3615) and watershed (from 0.2980 to 0.8483) classes. Using the HRNet module also yielded similar improvements: the mIoU improved from 0.425 on B-LSH to 0.7076 on LSH3, also pushed up from enhancement on buildings, which is noticeable in Figure 4.4, and watershed classifications. The baseline experiments also have a similar segmentation capacity on synthesized data, with the UNet showing slightly better classification capacity on the watershed class. That is why the MIoU and MAcc scores are better on UNet runs.

Table 4.8 – Accuracy (Acc), Mean Accuracy (MAcc), Per-class IoU (in the following order: Soil,Building,Woodland,Watershed) and Mean IoU (MIoU) for segmentation masks obtained from super-resolved images of LCAI dataset and SAGAN network.

Experiment	δ	Acc	MAcc	IoU	MIoU
B-LSU	0	0.779	0.5883	0.6561 / 0.0366 / 0.7733 / 0.2980	0.4411
B-LSH	0	0.7576	0.5652	0.6267 / 0.0520 / 0.7898 / 0.2312	0.425
LSU1	$1e^{-1}$	0.8544	0.7067	0.7590 / 0.1948 / 0.8481 / 0.4329	0.5588
LSU2	$1e^{-2}$	0.8879	0.7269	0.8157 / 0.1096 / 0.7971 / 0.7450	0.6169
LSU3	$1e^{-3}$	**0.9317**	**0.8171**	0.8805 / 0.3615 / 0.8705 / **0.8483**	**0.7403**
LSH1	$1e^{-1}$	0.9166	0.7283	0.8714 / 0.2661 / 0.8645 / 0.6151	0.6543
LSH2	$1e^{-2}$	0.926	0.7425	0.8760 / 0.3401 / 0.8896 / 0.6542	0.69
LSH3	$1e^{-3}$	0.9309	0.7633	**0.8825** / **0.3829** / **0.8990** / 0.6659	0.7076



(a) Label (b) B-LSH image (c) LSH3 image

Figure 4.4 – Original label, B-LSH and LSH3 segmentation maps. The inference capability of super-resolved images face the segmentator is increased, as observed by the building classification gains.

4.2.3 Visual Results

Like the experiments conducted with CGEO dataset, more visual results are expressed in the Appendix A. Results related to LCAI are shown in Figures A.5, A.7, A.6 and A.8. The chosen picture to represent the combo ESRGAN/LCAI contains a mix of watershed, woodlands and exposed soil, and the results of Table 4.5 are very explainable from figures of A.5: near metrics are translated to similar images, but the PSNR/SSIM differences observed when using the HRNet module are readily seen from the slight coloration change on runs LEH1 and LEH2. This subtle color change did not affect the classification accuracy of the HRNet module, since experiments using the ESRGAN/HRNet combination wielded the best segmentation metrics over the other combinations. Even though the experiments in Figure A.6 share similar metrics, the segmentation maps can utterly change, a fact mostly observed in the tree class. Also, the small lake on the right

part of the image is most of the time “forgotten" by the UNet module but remembered by the HRNet one, a probable cause of higher values IoU values for LEH experiments for the watershed class.

The joint approach demonstrated a strong regularization capacity for low values of δ on SAGAN experiments, results that can be seen on Figure A.7. The color match problem on baseline and LSU1 and LSH1 runs explains the low PSNR/SSIM values, while the best looking images , from inquired on LSU3 and LSH3, also had the best perceptual scores. For the segmentation results displayed on Figure A.8, it’s also noticeable once more the regularization factor of L_{seg} since B-LSU and B-LSH demonstrated insalubrious masks, which got better on other experiments.

4.3 Generalization capability over datasets

Successful deployment of machine learning models often requires generalization capability, being able to accurately cover multiple data domains without hurting the model accuracy. While this is usually achieved by applying transfer learning techniques to the trained model, poor generalization and decreased accuracy are still observable due the domain shift. Besides transfer learning (model weights manipulation from multiple networks trained on different data), generalization can also be achieved by incrementing the amount of data (and annotations) from the target domain, a situation where augmentation techniques could create substantial datasets to tackle multiple domains.

Ideally, the trained model should present similar results on multiple unseen domains without further training, but limitations on training a generalized model issue poor performance of “unseen” data, which is a major roadblock for deploying models into real-world data. This is especially true for remote sensing imagery: good performance over a unique dataset is already difficult because of the very extensive set of possible object representations of the terrain, while the literature about aerial multi-dataset coverage is still scarce (NEUPANE; HORANONT; ARYAL, 2021). Therefore, to analyze the generalization capacity of the proposed joint methodology, the inference phase runs on a different dataset from the training procedure, so experiments trained on CGEO data will be inferred on LCAI data and vice-versa. The chosen models for this analysis are the versions that produced the best perceptual metrics on each dataset, namely CEU1 and LEH1, and the respective baselines without the segmentation proxy, B-CE and B-LE, respectively, as reported in Table 4.9.

Table 4.9 – PSNR, SSIM, LPIPS and PI metrics for models trained on CGEO data and inferred on LCAI (B-CE,CEU1) and trained on LCAI and inferred on CGEO (B-LE,LEH1) for the best perceptually-aware runs from Tables 4.1, 4.2, 4.5 and 4.6. Original runs follow the naming convention adopted before, while the inference on a module trained on a different dataset uses a "-CGEO" or "-LCAI" suffix to nominate on which dataset the inference occurred.

Experiment	PSNR ↑	SSIM ↑	LPIPS ↓	PI ↓
Trained on CGEO data				
B-CE	30.2424	0.6568	0.2702	7.1803
B-CE-LCAI	30.1805	0.7142	0.335	7.6286
CEU1	30.1782	0.6747	0.2085	6.5564
CEU1-LCAI	30.1458	0.7255	0.3134	7.2349
Trained on LCAI data				
B-LE	32.1645	0.7909	0.2608	7.1715
B-LE-CGEO	29.2101	0.6432	0.4382	6.5829
LEH1	29.7972	0.7774	0.2041	6.4591
LEH1-CGEO	26.1745	0.5949	0.5757	6.722

The generalization capacity of the chosen ESRGAN-based models is not great when it is inferred from data not seen during the training phase. While similar values of PSNR for runs trained on CGEO data, on the upper part of Table 4.9, could somehow suggest good delineation capacities, it is easily seen the performance degradation between baseline experiments: inference of LCAI data on runs that did not employ the segmentation approach display PSNR / SSIM values of around 32 / 0.79, respectively, while the same values of PSNR/SSIM on B-CE-LCAI experiment are around 8 % worse just because of the domain change. A similar scenario can be observed in the perceptual metrics: B-CE-LCAI displays worse LPIPS / PI values if compared to the vanilla B-LE implementation. This is due to many factors: 1) the dataset has multiple different core characteristics, such as resolution, land descriptors and RGB mean and standard deviation; 2) augmentation techniques, while still employed, were chosen to not change the core attributes of each dataset (for example, saturation, brightness and contrast were not used), reducing the generalization capacity on diverse data, and 3) different classes could especially affect performance on segmentation-oriented experiments, since models trained on specific class data may synthesize unrealistic images if train/test data do not present the same classes, which is actually the case of CGEO and LCAI.

A different behavior is found for models trained on LCAI dataset. The baseline experiment, B-LE-CGEO, had a similar SSIM value if compared to B-CE and slightly smaller PSNR, but a huge deterioration for LPIPS. Metrics degradation were more prominent on the LEH1-CGEO run, but the combination between ESRGAN module and HR-

Net segmentator already yielded relative worse values amongst values from Table 4.1. Nonetheless, PSNR, SSIM, and LPIPS values were deeply depreciated, which can be expected when using the SR + segmentation approach because of the class domain problem between both datasets, therefore hurting the optimization realized by L_{seg} on G. Perceptual indexes were not as damaged as other metrics, such as the experiment LEH1-CGEO, but still did not display similar results for data inferred on different datasets, such as the pair B-CE-LCAI and B-LE, for example.

It's also worth noticing that, for the model trained on CGEO data and inferred on LCAI, the perceptual metrics were improved while using the segmentation proxy, but we observed the inverse trend when reversing the datasets. In special, the LPIPS results between experiments LEH-1 / LEH-1-CGEO were disruptive for the cross-dataset run, which can be explained by differences in class annotations between datasets and the domain transfer problem. The latter is quite perceived in both segmentation and super resolution tasks (TANG et al., 2020), but it's even more noticeable in the B-LE-CGEO and LEH1-CGEO experiments since both SR and segmentation networks will be negatively impacted by different class data distributions and absence of classes in the LCAI dataset, in comparison to the CGEO data.

Results of visual inspection in Figure 4.5 show that LCAI images generated on models trained on CGEO data were more realistic than the other combination, but still far from pleasant visual textures, as it would suggest the metrics on "-LCAI" experiments. More visual results are available on Appendix B.

(a) LEH1 image (b) B-CE-LCAI image (c) CEU1-LCAI image

Figure 4.5 – Images generated from the LEH1 (trained only on LCAI data), B-CE-LCAI and CEU1-LCAI (both trained on CGEO data) for a image from the LCAI dataset. Notice that the last two models fails to replicate the same levels of detail from the LEH1 run.

4.4 Comparison against other super-resolution methods

To compare the performance gains of the proposed joint super-resolution method in a real-world scenario, we used multiple recent supervised super-resolution methodologies using deep learning for a benchmark comparison:

- Attention-based Back Projection Network (ABPN) (LIU et al., 2019), where back-projection blocks are suggested to iteratively update low and high-resolution feature residues, and spatial attention blocks learn correlations between features at different layers;
- Cross-Scale Non-Local Attention (CSNLN) (MEI et al., 2020b), which analyses cross-scale feature correlation by using intra- and inter-scale attention modules;
- Deep Back-Projection Network (DBPN) (HARIS; SHAKHNAROVICH; UKITA, 2018a), which calculates correlations between mutually connected up- and down-sampling stages by using a back-projection (IRANI; PELEG, 1991) based mechanism;
- Densely Residual Laplacian (DRLN) (ANWAR; BARNES, 2020), proposing a cascading residual on the residual structure to learn information from high and mid-level features via densely concatenated residual blocks and a Laplacian attention model;
- Enhanced Deep Residual Networks (EDSR) (LIM et al., 2017), displaying a multi-scale model that shares most of the parameters across different scales;
- Pyramid Attention Network (PAEDSR) (MEI et al., 2020a), which analyses self-similarity relative to image priors on a multi-scale level by using self scale-agnostic attention modules;
- Residual Channel Attention Networks (RCAN) (ZHANG et al., 2018b), consisting of several residual groups with skip connections and a channel attention mechanism to adaptively rescale channel-wise features;
- Residual Dense Network (RDN) (ZHANG et al., 2018c), proposing residual dense blocks to extract local features via dense connected convolutional layers and a global feature fusion method to learn global hierarchical features; and
- Single Image Super-Resolution Using a Generative Adversarial Network (SRGAN) (LEDIG et al., 2016), a pioneer GAN-based architecture to generate realistic super-resolved images by employing adversarial and a content losses.

It is noticeable that recent SOTA methods rely on self-attention mechanisms, stimulating a fair comparison against the joint segmentation methodology with the SAGAN module. Attention modules are a way to go to produce optimal PSNR/SSIM values, although they may not be able to redeem good perceptual scores. For a fair comparison, all models were trained using the default architecture settings suggested by the authors for the same number of epochs (100).

Besides the aforementioned experiments, we also compare the results of baseline runs, so we can clearly notice the impact of adopting the joint methodology on the top of baseline experiments.

Table 4.10 displays the values of PSNR, SSIM, LPIPS and PI for the nine competitive methodologies and the best perceptual-oriented experiment for the CGEO dataset on CEU1 run. The proposed methodology generates better results for perceptual LPIPS/PI metrics, with LPIPS value being approximately 40% better than the second-best model, namely ABPN. Perceptual values also were the best amongst all competitors, with the second place being the baseline B-CE experiment, and the third place is comprised by the SRGAN module, which is expected since both (B-CE and SRGAN) employs a perceptual loss in its optimization. However, the bad (second-worst) value of LPIPS value in SRGAN was unexpected for the same reason (sole experiment, besides B-CE and our proposal, to use a perceptual loss). One possible answer is a non-optimal combination of original hyper-parameters for the datasets used in this work.

Best PSNR/SSIM values are observed in the ABPN run, which has the second-worst PI. The proposed method has lower but still competitive values of pixel-wise metrics. The comparison against SRGAN method is especially interesting: both have similar values of PSNR/SSIM, but the joint method yields very good improvements over the perceptual metrics, proving the great potential of the joint method as a super-resolution technique.

Results over the LCAI dataset, shown in Table 4.11, indicate that the proposed methodology has the best LPIPS/PI amongst all competitors, displaying LPIPS improvements of over 50% from the second-best model (CSNLN), 20% over the baseline B-LE model and more than one point of PI over the SRGAN method. The CSNLN approach had the best PSNR/SSIM values, but all four metrics had very similar values to the ABPN run. The pyramid-based method PAEDSR displayed the worst LPIPS and PSNR and second-worst SSIM, and similar bad values were observed when running the CGEO data, probably explained by the lack of network adherence to both datasets due to non-optimal

Table 4.10 – PSNR, SSIM, LPIPS and PI metrics for other networks when using the CGEO dataset

Experiment	PSNR ↑	SSIM ↑	LPIPS ↓	PI ↓
Proposed	30.1782	0.6747	**0.2085**	**6.5564**
ABPN	**31.5132**	**0.7115**	0.3516	10.6724
CSNLN	30.9173	0.7029	0.3534	10.4000
DBPN	30.3087	0.6746	0.4794	9.8185
DRLN	31.1385	0.7064	0.3769	10.0155
EDSR	30.7681	0.6914	0.3974	9.0428
PAEDSR	28.8526	0.6818	0.3915	10.7262
RCAN	30.9879	0.6995	0.3757	8.2211
RDN	31.2054	0.7073	0.3653	7.9088
SRGAN	30.2158	0.6710	0.4558	7.5756
B-CE	30.2424	0.6568	0.2702	7.1803

hyper-parameters. It is important to mention that we selected our representative approach (LEH1) based on the perceptual metrics at the cost of having smaller PSNR/SSIM values. As analyzed in Chapter 4.2, selecting smaller values for δ alleviate the PSNR/SSIM degradation. For example, the LEH3 run has more competitive PSNR (32.2266) and SSIM (0.7982) scores while still providing the best LPIPS (0.2201) and PI (6.8829) when compared to competitive approaches.

Table 4.11 – PSNR, SSIM and LPIPS and PI metrics for other networks - LCAI dataset

Experiment	PSNR ↑	SSIM ↑	LPIPS ↓	PI ↓
Proposed	29.7972	0.7774	**0.2041**	**6.4591**
ABPN	33.2417	0.8236	0.3288	8.4332
CSNLN	**33.2694**	**0.8259**	0.3231	8.4518
DBPN	31.4422	0.7778	0.3968	7.8218
DRLN	32.6290	0.8073	0.3752	7.4282
EDSR	32.2913	0.7980	0.4000	7.5908
PAEDSR	28.2293	0.7571	0.4334	7.5296
RCAN	32.9592	0.8167	0.3521	7.7881
RDN	32.3606	0.8018	0.3852	7.4369
SRGAN	30.8207	0.7543	0.3511	7.3551
B-LE	32.1645	0.7909	0.2608	7.1715

Visual outputs for the proposed methodology, the nine corresponding models and the original high-resolution image are displayed in Appendix C. These outputs give us a visual representation of how better perceptual metrics are translated to texture synthesis and, therefore, better better human assessment of the scene. Figure C.1, representing the CGEO dataset, displays over-smoothed outputs very noticeable in most of the methods, not only in texture but also in shape, especially in non-uniform areas (rooftops and veg-

etation, for example). Most of the methods also find it hard to produce sharp building edges, producing irregular edges, explained by the texture discontinuity between buildings and soil on LR images and the ill-posed nature of the super-resolution task. The proposed method is able to produce jagged edges at the cost of blurriness, especially in densely populated areas: class transitions are blurry, being otherwise faithful to the HR image, particularly in heterogeneous patterns, such as rooftops and vegetation. The same issue is verified in Figure C.2: no method is very close to the same sharpness details of the HR image, but a middle term between over-smoothed (such as CSNLN and ABPN) and granulated (SRGAN) terms made the proposed model a good alternative to generate more realistic images.

5 CONCLUSION

Whilst being one of the oldest and most important problems of computer vision, super resolution only noticed considerable performance gains and rising popularity after the adoption of approaches based on deep learning. In this context, multiple proposals achieved good reconstruction quality when analyzing solely pixel-wise metrics, such as PSNR and SSIM, but failed to reconstruct realistic textures since commonly used pixel-oriented metrics like the MSE do not consider either shallow or deep features contexts from the SR module, producing mostly blurred images guided by the minimized distance in the SR's image pixel space.

Perceptually-based super resolution is, therefore, a recent alternative to produce more appealing and realistic images, particularly regarding textured regions. This is specially applicable when dealing with land cover imagery, since visual coherence is essential to synthesize natural-like images from a huge diversity of terrain textures.

This work proposed a novel approach to produce more realistic textures in class-aware manner. By introducing a loss function provided by a semantic segmentation module, the optimization of the super resolution module gets a feedback about the synthesis of faithful class-wise textures. Such input does not rely on simple comparisons about HR and SR images: it takes in consideration whole different outputs, the real and inferred segmentation maps, to introduce a perceptually-aware loss component on the SR module optimization.

Employing this joint training methodology proved successful to create better images on a human comprehension scale in a complex and multi-class domain created by aerial imagery. Even though pixel-wise metrics were marginally affected by the introduction of a new training strategy, perceptually-aware metrics, both full-reference (LPIPS) and no-reference (PI) metrics observed great improvements over vanilla SR methods that did not employ the segmentation strategy. This was translated to better visual results in four sets of experiments, running two different datasets and two distinct super resolution backbones, and also better segmentation results for the super-resolved image over vanilla SR methods.

Such methodology has potential to be used as an enhancement tool in any DL-based SR framework. This is reinforced by improvements on reconstruction metrics even when using different combinations of SR and segmentation models. Besides, this joint methodology is easy to customize, understand and implement, while supporting a range

of adaptations, either in super resolution or segmentation modules, either in the loss components. However, the proposed methodology possesses some pitfalls, such is the difficulty in finding optimal hyperparameters, which can be solved by implementing optimal parameter-searching algorithms, and the higher memory consumption, which are very notable when using heavy segmentation modules.

Comparisons against state of the art methodologies also confirm the capacity of the joint approach in synthesizing images with high perceptual indexes. Even when choosing older super resolution methods, the joint methodology yields better perceptual metrics (and in some cases, even pixel-wise metrics too) than its vanilla implementations, bringing much space to improve the latest SR techniques by introducing a trained segmentation module on the SR pipeline at a relatively low training cost.

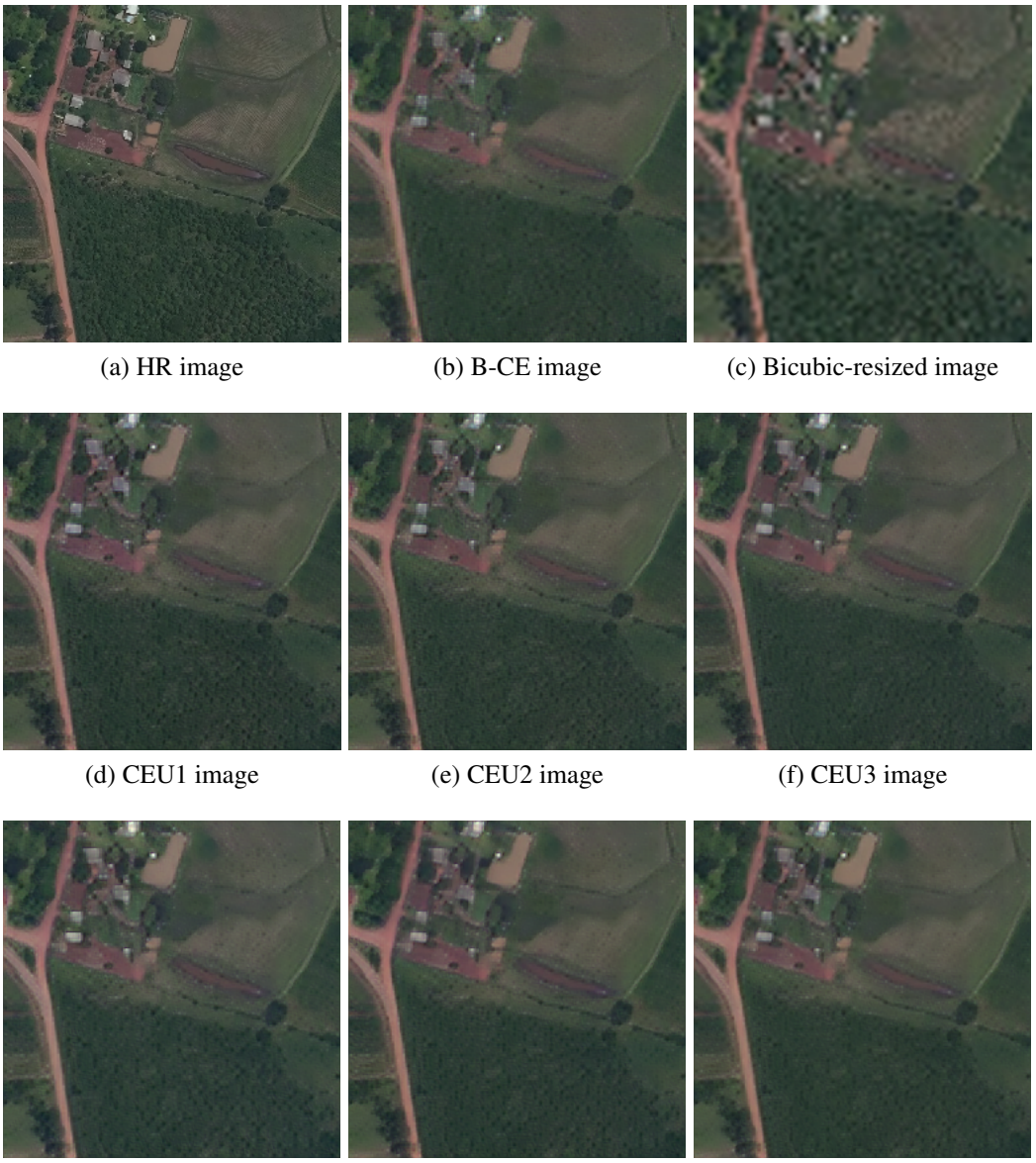
5.1 Future work

The are several avenues for extending this work. One option would be to widen the choices of datasets, super-resolution and segmentation modules, including an analysis on newer methodologies on our joint proposal, such Graph Neural Networks or Multi-head Attention mechanisms on vision transformers. Analysis of multiple super resolution factors (x2, x3, x8) or even fractional SR scales could also be tackled in future versions of this research.

Another possibility of improvement is the development of a better strategy for tuning the hyper-parameters, since grid search brings a big computational burden. The joint refinement of both the SR and the segmentation modules could be an interesting idea, but the number of learnable parameters would be high and care should be taken with memory issues. For addressing the generalization problem, possible solutions could be the use of different augmentation strategies, a mix of datasets or even using a self-training strategy, which has already been proven successful in the literature (ZOPH et al., 2020).

Finally, one last direction worth mentioning is the inclusion of other tasks in the learning process. For example, one could simultaneously explore semantic segmentation and object detection to guide the super resolution task.

APPENDIX A — EXPERIMENT OUTPUTS

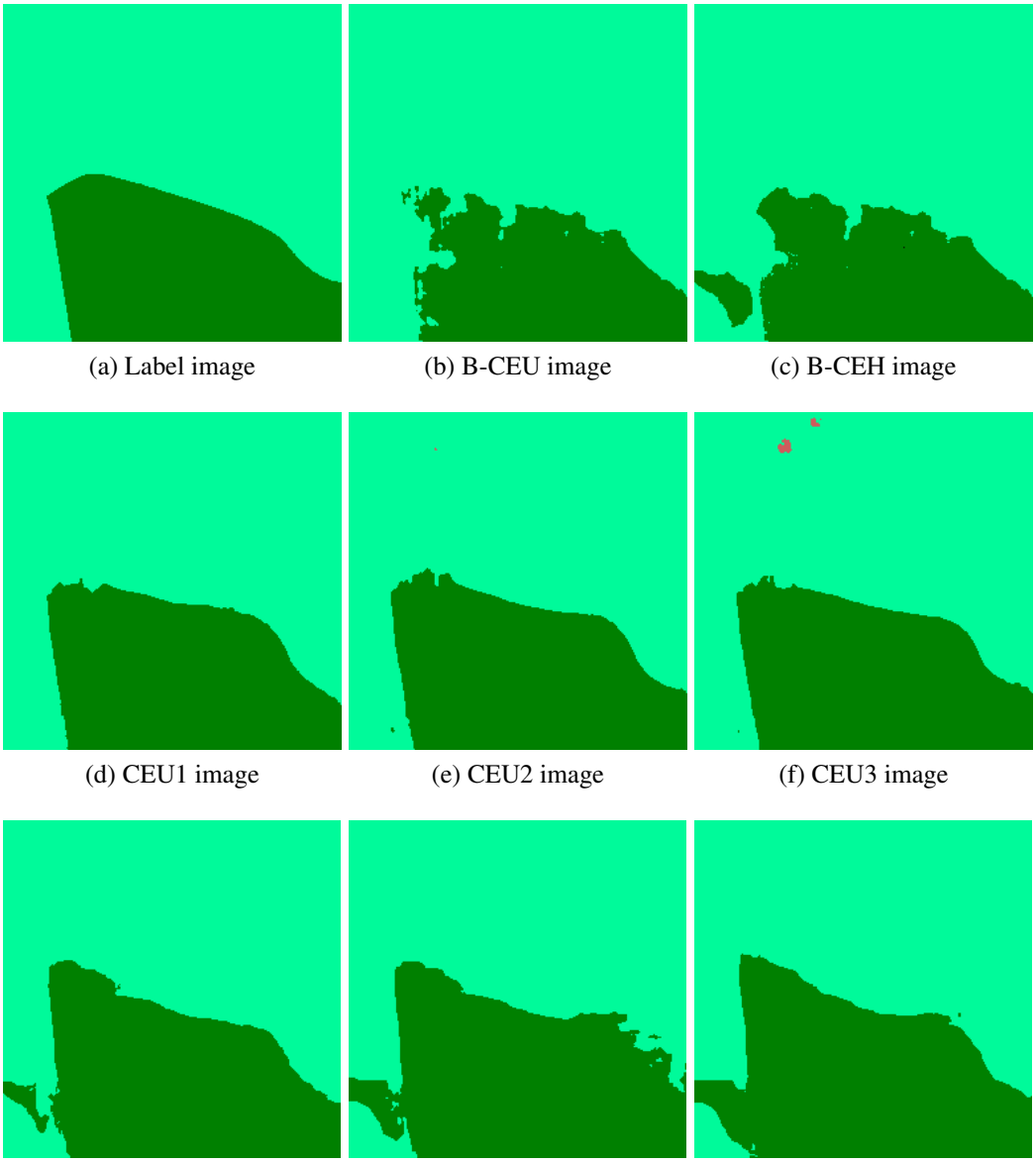


(a) HR image (b) B-CE image (c) Bicubic-resized image

(d) CEU1 image (e) CEU2 image (f) CEU3 image

(g) CEH1 image (h) CEH2 image (i) CEH3 image

Figure A.1 – HR image, bicubic re-sampled image and and inference results of Table 4.1 for the CGEO dataset using an ESRGAN-based SR module and UNet or HRNet segmentation netwoks.

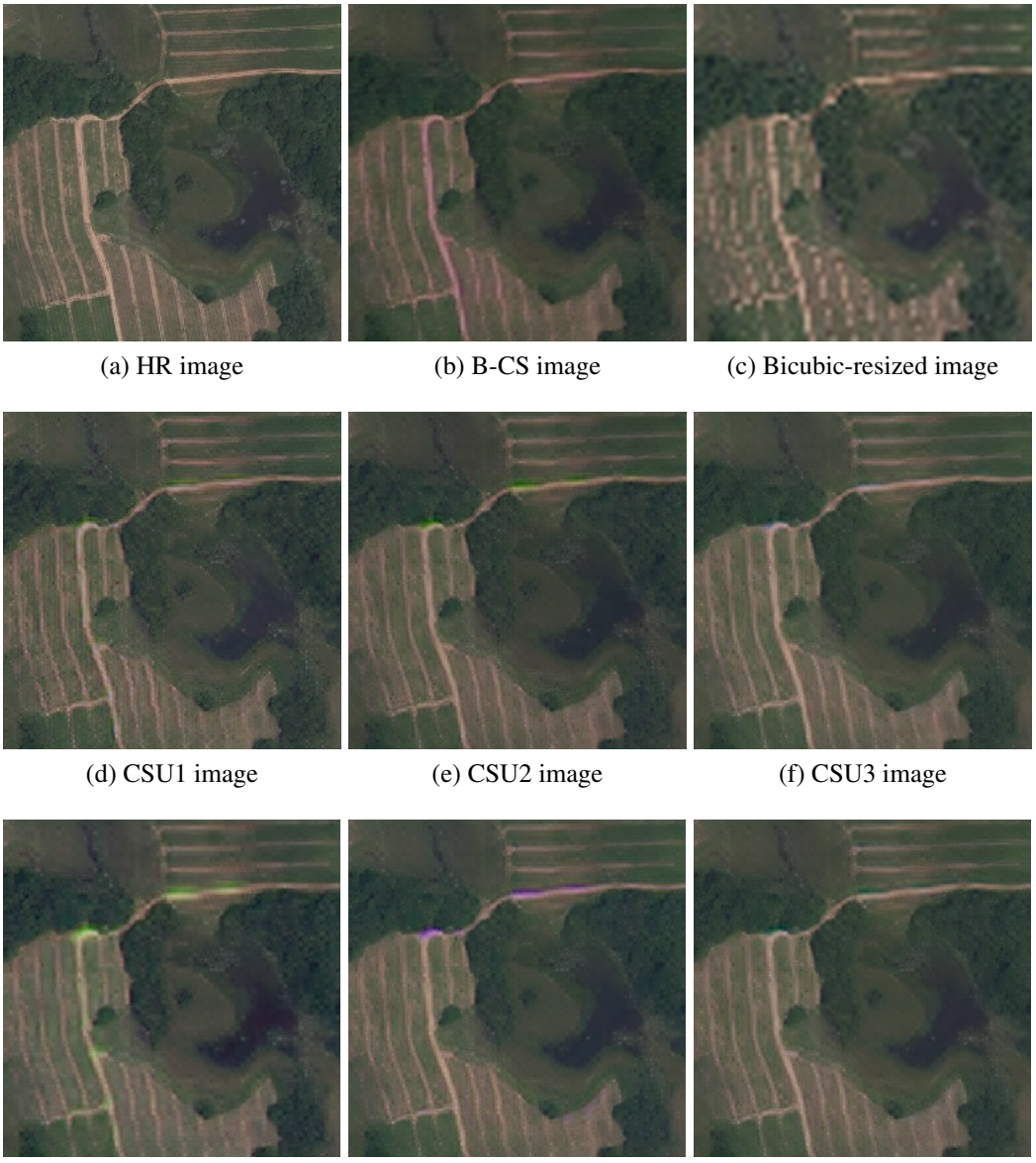


(a) Label image (b) B-CEU image (c) B-CEH image

(d) CEU1 image (e) CEU2 image (f) CEU3 image

(g) CEH1 image (h) CEH2 image (i) CEH3 image

Figure A.2 – Groud truth label and segmentation outputs of experiments from of Table 4.3 for the CGEO dataset using an ESRGAN-based SR module and UNet or HRNet segmentation netwoks.



(a) HR image (b) B-CS image (c) Bicubic-resized image

(d) CSU1 image (e) CSU2 image (f) CSU3 image

(g) CSH1 image (h) CSH2 image (i) CSH3 image

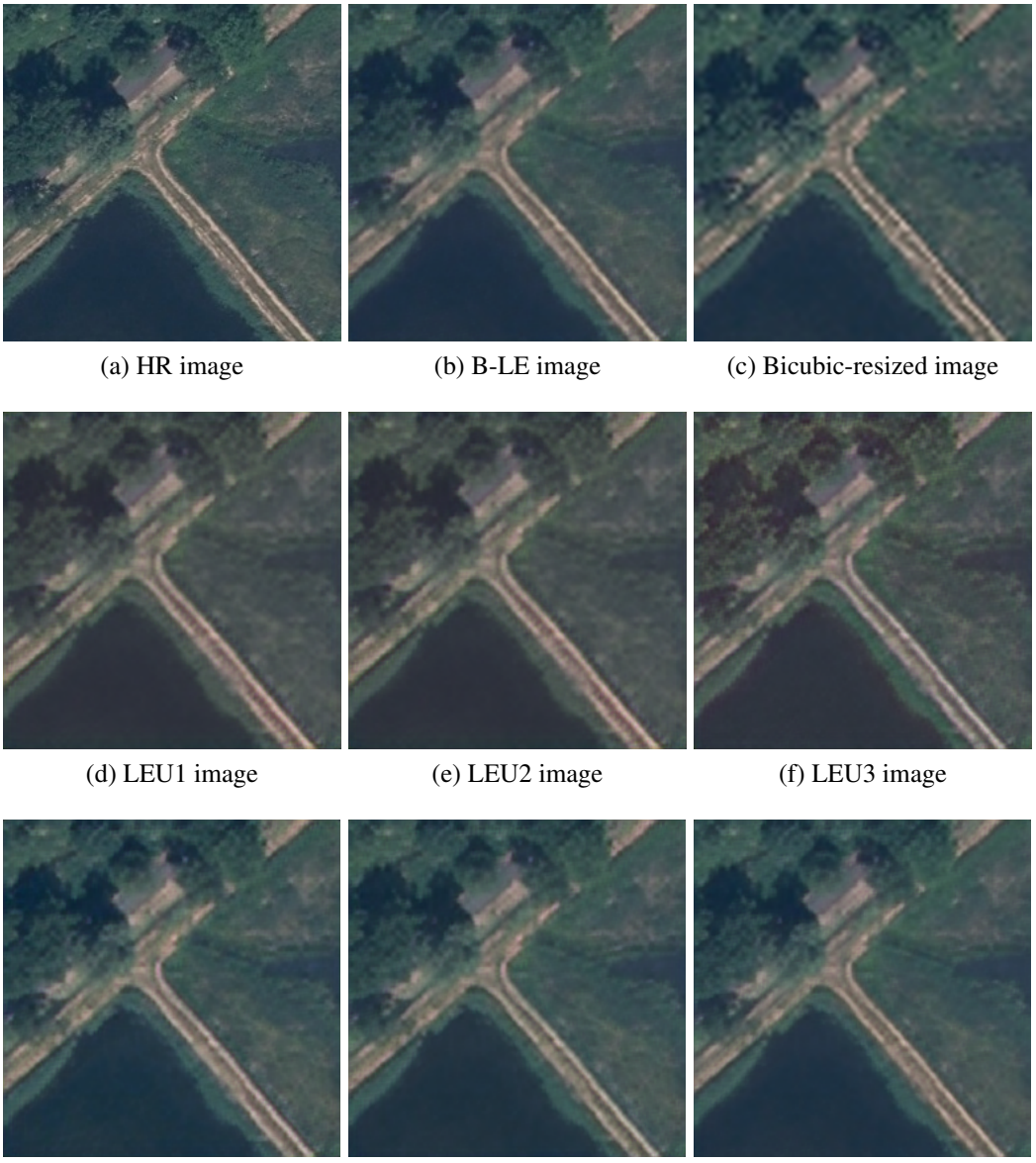
Figure A.3 – HR image, bicubic re-sampled image and and inference results of Table 4.2 for the CGEO dataset using an SAGAN-based SR module and UNet or HRNet segmentation netwoks.

(a) Label image (b) B-CSU image (c) B-CSH image

(d) CSU1 image (e) CSU2 image (f) CSU3 image

(g) CSH1 image (h) CSH2 image (i) CSH3 image

Figure A.4 – Groud truth label and segmentation outputs of experiments from of Table 4.4 for the CGEO dataset using an SAGAN-based SR module and UNet or HRNet segmentation netwoks.

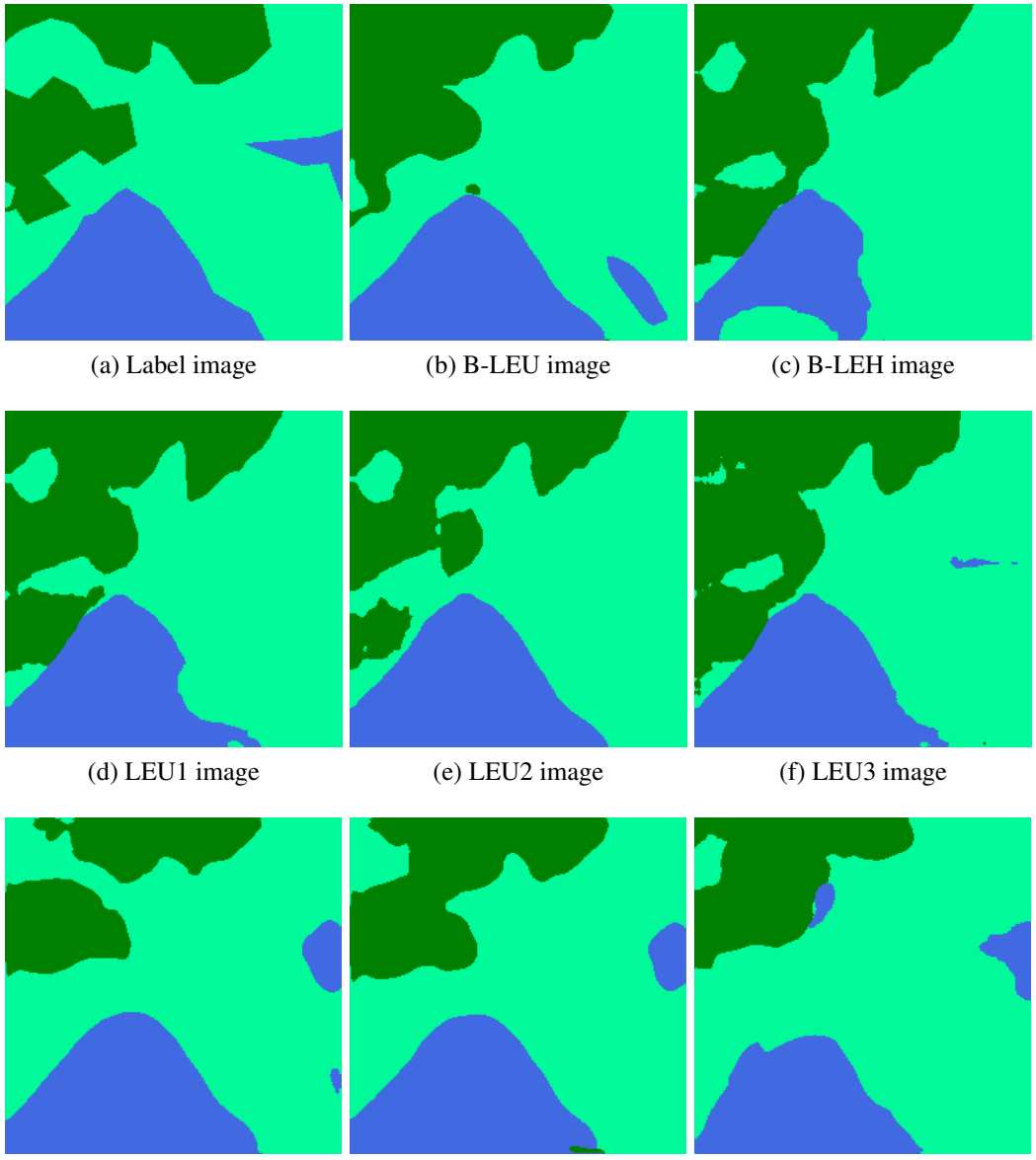


(a) HR image (b) B-LE image (c) Bicubic-resized image

(d) LEU1 image (e) LEU2 image (f) LEU3 image

(g) LEH1 image (h) LEH2 image (i) LEH3 image

Figure A.5 – HR image, bicubic re-sampled image and and inference results of Table 4.5 for the LCAI dataset using an ESRGAN-based SR module and UNet or HRNet segmentation netwoks.

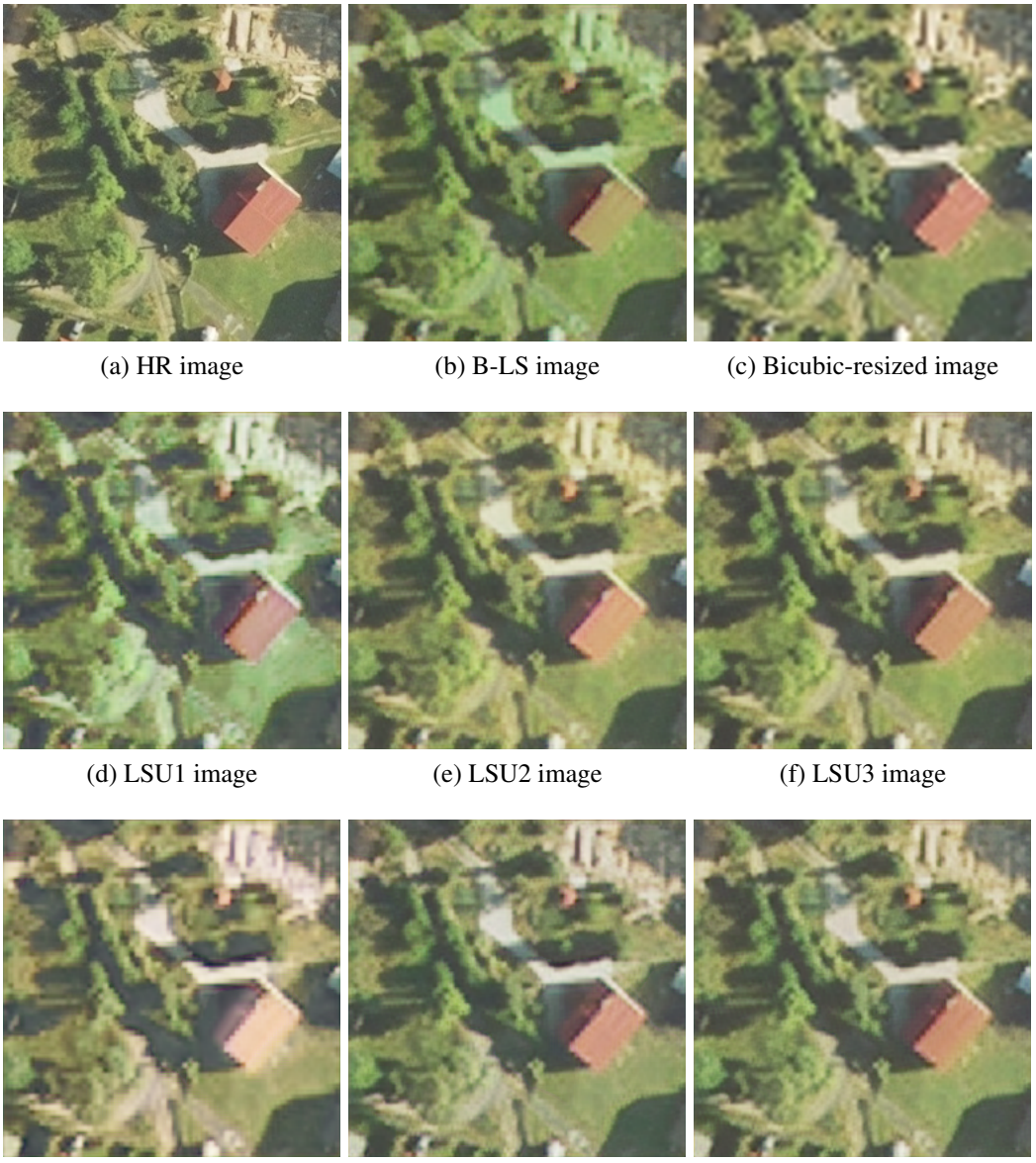


(a) Label image (b) B-LEU image (c) B-LEH image

(d) LEU1 image (e) LEU2 image (f) LEU3 image

(g) LEH1 image (h) LEH2 image (i) LEH3 image

Figure A.6 – Groud truth label and segmentation outputs of experiments from of Table 4.7 for the LCAI dataset using an ESRGAN-based SR module and UNet or HRNet segmentation netwoks.

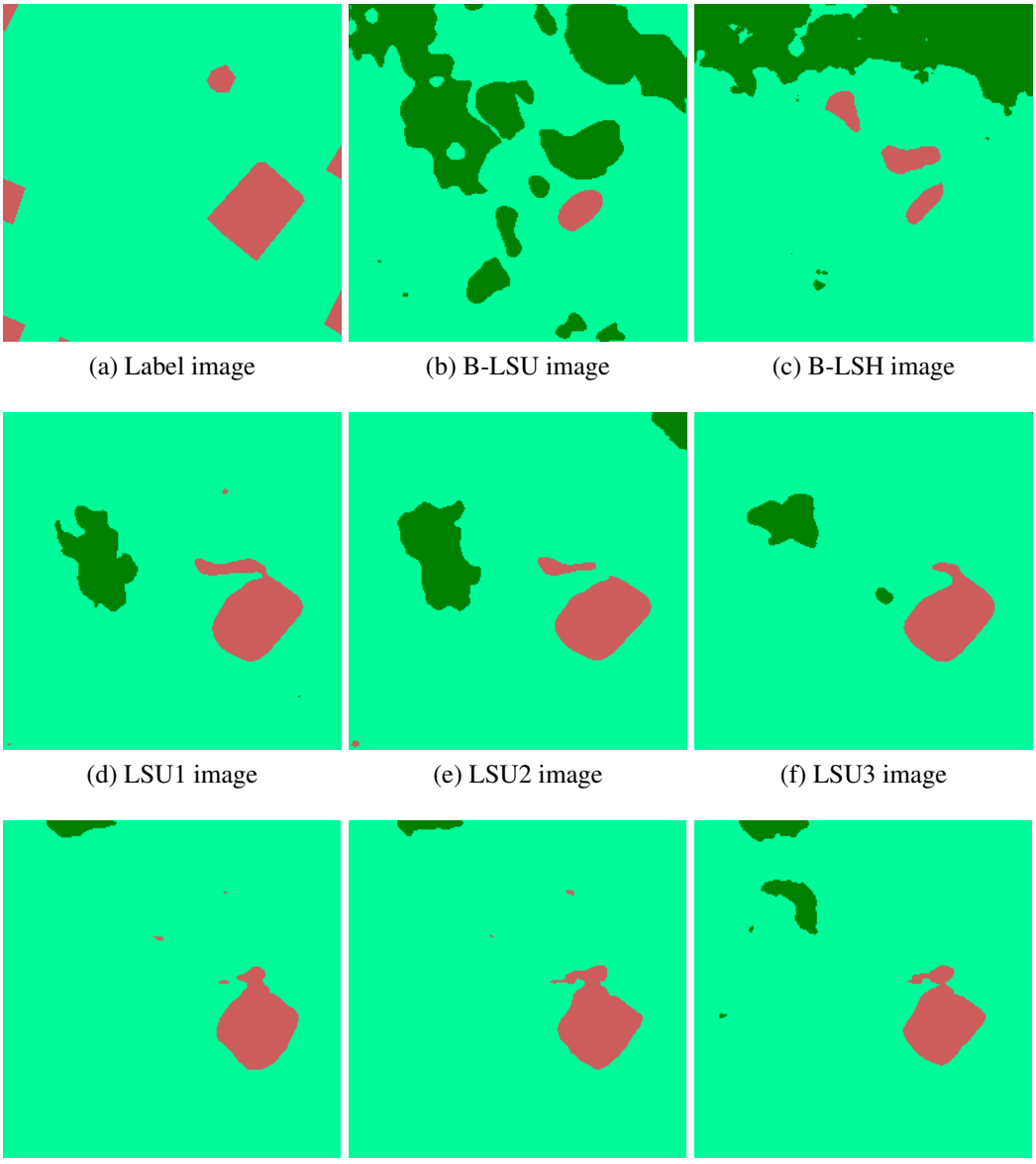


(a) HR image (b) B-LS image (c) Bicubic-resized image

(d) LSU1 image (e) LSU2 image (f) LSU3 image

(g) LSH1 image (h) LSH2 image (i) LSH3 image

Figure A.7 – HR image, bicubic re-sampled image and and inference results of Table 4.6 for the LCAI dataset using an SAGAN-based SR module and UNet or HRNet segmentation netwoks.



(a) Label image (b) B-LSU image (c) B-LSH image

(d) LSU1 image (e) LSU2 image (f) LSU3 image

(g) LSH1 image (h) LSH2 image (i) LSH3 image

Figure A.8 – Groud truth label and segmentation outputs of experiments from of Table 4.8 for the LCAI dataset using an SAGAN-based SR module and UNet or HRNet segmentation netwoks.

APPENDIX B — GENERALIZATION ON DIFFERENT TRAIN/TEST SETS

(a) HR Image (b) B-CE Image (c) B-LE-CGEO Image

(d) CEU1 Image (e) LEH1-CGEO

Figure B.1 – HR image and inference results for the CGEO dataset on networks trained on CGEO (B-CE,CEU1) or LCAI (B-LE-CGEO,LEH1-CGEO) data.

(a) HR Image (b) B-CE-LCAI Image (c) B-LE Image

(d) CEU1-LCAI Image (e) LEH1

Figure B.2 – HR image and inference results for the LCAI dataset on networks trained on CGEO (B-CE-LCAI,CEU1-LCAI) or LCAI (B-LE,LEH1) data.

APPENDIX C — COMPARISON BETWEEN SUPER RESOLUTION METHODS

(a) HR Image (b) Proposed (c) ABPN

(d) CSNLN (e) DBPN (f) DRNLN

(g) EDSR (h) EDSR (i) PAEDSR

(j) RCAN (k) RDN (l) SRGAN

Figure C.1 – Comparison between proposed method and multiple SOTA super-resolution methodologies for x4 enhancement on the CGEO dataset.

(a) HR Image (b) Proposed (c) ABPN

(d) CSNLN (e) DBPN (f) DRNLN

(g) EDSR (h) EDSR (i) PAEDSR

(j) RCAN (k) RDN (l) SRGAN

Figure C.2 – Comparison between proposed method and multiple SOTA super-resolution methodologies for x4 enhancement on the LCAI dataset.

APPENDIX D — RESUMO EXPANDIDO

Super resolução (SR) é um problema bastante conhecido no mundo da Visão Computacional (VC) que visa reconstruir uma uma imagem de alta resolução (HR) a partir de uma imagem de baixa resolução (LR). Este prolema é bastante desafiador por vários motivos, dentre eles a dificuldade em sintetizar texturas e estruturas não existentes nas imagens de baixa resolução, especialmente quando tratamos de altos fatores de reconstrução. Essas dificuldades são agravadas quando utilizamos imagens aéreas, visto que algumas características intrínsecas a este domínio dificultam a tarefa de super resolução, tais como a grande variedade de texturas na superfície terrestre, a pequena extensão espacial de algumas feições (sendo praticamente irreconhecíveis em imagens LR), a invariabilidade de objetos à rotação e orientação e a alta disponibilidade de dados brutos. Entretanto, a síntese de imagens aéreas super resolvidas gera um impacto positivo em aplicações que utilizam imagens aéreas, uma vez que o maior nível de detalhe e a síntese de cenas a partir de imagens mais recentes são bem-vindas no contexto de aplicações em sensoriamento remoto.

Dentre as múltiplas soluções para o problema de super resolução, vemos a crescente adoção de métodos baseados em Redes de Aprendizado Profundo, capazes de superar a performance de algoritmos clássicos de SR. Mesmo apresentando melhores estatísticas de reconstrução de imagem, essas redes ainda apresentam dificuldades na geração de texturas, e a supressão de artefatos ou texturas indesejáveis ainda são, no contexto da Aprendizagem Profunda, problemas a serem resolvidos. Uma das causas desses problemas é a diferença entre a percepção de qualidade entre o homem e a máquina: enquanto o computador utiliza modelos matemáticos para definir a qualidade de uma imagem, o ser humano emprega um modelo subjetivo que não é descritível por máquinas. A definição de uma métrica de percepção "ideal" que seja similar à opinião humana é ainda um desafio.

Nesse contexto, o nosso objetivo é gerar imagens aéreas super resolvidas que sejam próximas do critério de qualidade humana. A síntese de imagens com bons níveis de qualidade perceptual resultará em cenas mais realísticas, especialmente em regiões com textura. Para isso, nós utilizamos diversas técnicas:

- Aprendizagem conjunta utilizando duas redes de aprendizado profundo. Nós utilizamos uma rede de super resolução e uma rede de segmentação semântica que será responsável por gerar um parecer sobre a qualidade dos mapas de segmentação de imagens super resolvidas. A rede de segmentação, treinada em imagens HR,

quantifica a performance da rede de super resolução face à capacidade de geração de texturas que são coerentes às verdadeiras classes de cobertura terrestre. Em outras palavras, o módulo de super resolução é guiado pelo modelo de segmentação a gerar texturas conscientes à classe (como florestas, edificações ou rodovias, por exemplo). A consciência sobre a classe permite a síntese de texturas mais realistas.

- Módulo de SR baseado em Redes Adversariais Generativas (GANs). O método de aprendizagem utilizado por GANs, que utiliza duas redes (uma generativa, para a criação de dados sintéticos, e uma discriminativa, que analisa a veracidade do dado gerado), permite a geração de imagens mais naturais.
- Função de custo baseada em qualidade perceptiva. Para isso, utilizamos duas funções de custo com a intenção de gerar imagens com melhores qualidades perceptivas: a primeira, chamada custo de reconstrução de feições, ou simplesmente função de perda perceptiva, que analisa a distância de mapas de ativação em redes profundas pré-treinadas, e a segunda, uma função de entropia cruzada, que analisa a qualidade de reconstrução de texturas de imagens super resolvidas.

Para comprovar a nossa hipótese, reproduzimos experimentos utilizando múltiplas combinações de redes de super resolução (ESRGAN, “Enhanced Super Resolution Generative Adversarial Network”, e SAGAN, “Self Attention Generative Adversarial Network”), de segmentação semântica (UNet, e “High Resolution Network”, HRNet) e de conjuntos de imagens aéreas (CGEO e “Land Cover AI”) para analisar os resultados da metodologia proposta em diferentes condições de treino. Utilizamos métricas perceptivas para analisar a qualidade da super resolução, além das métricas convencionais de análise de reconstrução por pixel. Além disso, verificamos os resultados da segmentação semântica dos dados super resolvidos como uma forma de confirmar que a metodologia de treinamento conjunto está influenciando a geração de texturas coerentes à classe.

Nossos resultados amparam a hipótese de que a metodologia proposta gera imagens com melhores índices perceptivos. Quase a totalidade dos experimentos geram melhores índices para as duas métricas perceptivas utilizadas, o “Learned Perceptual Image Patch Similarity”, LPIPS, e o “Perceptual Index”, PI, para diferentes pesos de participação da função entropia cruzada fornecida pelo módulo de segmentação. Em alguns casos, também notamos melhorias nas métricas de reconstrução convencionais utilizadas nesse estudo, o “Peak Signal-to-Noise Ratio”, PSNR, e o “Structural Similarity”, SSIM.

Ao analisar os resultados das máscaras de segmentação produzidas por imagens SR, confirmamos que a metodologia proposta auxilia a geração de texturas coerentes à

classe, uma vez que métricas de segmentação semânticas tradicionais, como acurácia e interseção sobre união, apresentam melhorias após o uso conjunto dos módulos de super resolução e segmentação. Logo o método proposto não só auxilia a gerar imagens com melhores índices perceptivos, mas também gera imagens de maior resolução que performam melhor na tarefa de segmentação semântica que as imagens de baixa resolução correlacionadas.

A comparação entre diversas metodologias de super resolução no estado da arte comprovam que a metodologia proposta neste trabalho produz as melhores métricas perceptivas na síntese de imagens aéreas, gerando também valores competitivos de PSNR e SSIM.

Por fim, este trabalho propôs uma metodologia para a geração de texturas realísticas e coerentes à classe pela introdução de um módulo de segmentação que gera um parecer sobre a reconstrução de texturas coesivas à classe. Essa metodologia é de fácil entendimento, customização e capaz de melhorar a qualidade perceptual de imagens super resolvidas utilizando um custo computacional extra relativamente baixo.

APPENDIX — REFERENCES

ANWAR, S.; BARNES, N. Densely residual laplacian super-resolution. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, 2020.

BARBEDO, J. G. Factors influencing the use of deep learning for plant disease recognition. **Biosystems engineering**, Elsevier, v. 172, p. 84–91, 2018.

BLAU, Y. et al. The 2018 pirm challenge on perceptual image super-resolution. In: **Proceedings of the European Conference on Computer Vision (ECCV)**. [S.l.: s.n.], 2018. p. 0–0.

BLAU, Y.; MICHAELI, T. The perception-distortion tradeoff. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2018. p. 6228–6237.

BOGUSZEWSKI, A. et al. **LandCover.ai: Dataset for Automatic Mapping of Buildings, Woodlands and Water from Aerial Imagery**. 2020.

CHEN, Z. et al. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: PMLR. **International Conference on Machine Learning**. [S.l.], 2018. p. 794–803.

CORDTS, M. et al. The cityscapes dataset for semantic urban scene understanding. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 3213–3223.

DAI, D. et al. Is image super-resolution helpful for other vision tasks? In: IEEE. **2016 IEEE Winter Conference on Applications of Computer Vision (WACV)**. [S.l.], 2016. p. 1–9.

DONG, C. et al. Learning a deep convolutional network for image super-resolution. In: SPRINGER. **European conference on computer vision**. [S.l.], 2014. p. 184–199.

DUCHON, C. E. Lanczos filtering in one and two dimensions. **Journal of applied meteorology**, v. 18, n. 8, p. 1016–1022, 1979.

ETTEN, A. V. Satellite imagery multiscale rapid detection with windowed networks. In: IEEE. **2019 IEEE Winter Conference on Applications of Computer Vision (WACV)**. [S.l.], 2019. p. 735–743.

FANG, F. et al. Urban land-use classification from photographs. **IEEE Geoscience and Remote Sensing Letters**, v. 15, n. 12, p. 1927–1931, 2018.

FANG, Y. et al. Perceptual quality assessment of smartphone photography. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2020. p. 3677–3686.

FARSIU, S. et al. Fast and robust multiframe super resolution. **IEEE transactions on image processing**, IEEE, v. 13, n. 10, p. 1327–1344, 2004.

FOMIN, V. et al. **High-level library to help with training neural networks in PyTorch**. [S.l.]: GitHub, 2020. <https://github.com/pytorch/ignite>.

FREEMAN, W. T.; JONES, T. R.; PASZTOR, E. C. Example-based super-resolution. **IEEE Computer graphics and Applications**, IEEE, v. 22, n. 2, p. 56–65, 2002.

GARCIA-GARCIA, A. et al. A review on deep learning techniques applied to semantic segmentation. **arXiv preprint arXiv:1704.06857**, 2017.

GEMAN, S.; GEMAN, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. **IEEE Transactions on pattern analysis and machine intelligence**, IEEE, n. 6, p. 721–741, 1984.

GHAFFARIAN, S. et al. Effect of attention mechanism in deep learning-based remote sensing image processing: A systematic literature review. **Remote Sensing**, Multidisciplinary Digital Publishing Institute, v. 13, n. 15, p. 2965, 2021.

GLASNER, D.; BAGON, S.; IRANI, M. Super-resolution from a single image. In: IEEE. **2009 IEEE 12th international conference on computer vision**. [S.l.], 2009. p. 349–356.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>.

GOODFELLOW, I. et al. Generative adversarial nets. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2014. p. 2672–2680.

GUO, M. et al. Dynamic task prioritization for multitask learning. In: **Proceedings of the European conference on computer vision (ECCV)**. [S.l.: s.n.], 2018. p. 270–287.

GUO, M.-H. et al. Visual attention network. **arXiv preprint arXiv:2202.09741**, 2022.

HARIS, M.; SHAKHNAROVICH, G.; UKITA, N. Deep back-projection networks for super-resolution. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2018. p. 1664–1673.

HARIS, M.; SHAKHNAROVICH, G.; UKITA, N. Task-driven super resolution: Object detection in low-resolution images. **arXiv preprint arXiv:1803.11316**, 2018.

HE, K. et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: **Proceedings of the IEEE international conference on computer vision**. [S.l.: s.n.], 2015. p. 1026–1034.

HE, K. et al. Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 770–778.

HUANG, X. et al. Multimodal unsupervised image-to-image translation. In: **Proceedings of the European conference on computer vision (ECCV)**. [S.l.: s.n.], 2018. p. 172–189.

HURTIK, P. et al. Poly-yolo: higher speed, more precise detection and instance segmentation for yolov3. **Neural Computing and Applications**, Springer, v. 34, n. 10, p. 8275–8290, 2022.

IOFFE, S.; SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. **arXiv preprint arXiv:1502.03167**, 2015.

IRANI, M.; PELEG, S. Improving resolution by image registration. **CVGIP: Graphical models and image processing**, Elsevier, v. 53, n. 3, p. 231–239, 1991.

JIANG, K. et al. Edge-enhanced gan for remote sensing image superresolution. **IEEE Transactions on Geoscience and Remote Sensing**, IEEE, v. 57, n. 8, p. 5799–5812, 2019.

JOHNSON, J.; ALAHI, A.; FEI-FEI, L. Perceptual losses for real-time style transfer and super-resolution. In: SPRINGER. **European conference on computer vision**. [S.l.], 2016. p. 694–711.

KIM, J.; LEE, J. K.; LEE, K. M. Deeply-recursive convolutional network for image super-resolution. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 1637–1645.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. **Advances in neural information processing systems**, v. 25, 2012.

LAM, D. et al. **xView: Objects in Context in Overhead Imagery**. 2018.

LATEEF, F.; RUICHEK, Y. Survey on semantic segmentation using deep learning techniques. **Neurocomputing**, Elsevier, v. 338, p. 321–348, 2019.

LECUN, Y. et al. Backpropagation applied to handwritten zip code recognition. **Neural computation**, MIT Press, v. 1, n. 4, p. 541–551, 1989.

LECUN, Y. et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, Ieee, v. 86, n. 11, p. 2278–2324, 1998.

LEDIG, C. et al. Photo-realistic single image super-resolution using a generative adversarial network. **CoRR**, abs/1609.04802, 2016. Disponível em: <http://arxiv.org/abs/1609.04802>.

LI, X. et al. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. **IEEE transactions on medical imaging**, IEEE, v. 37, n. 12, p. 2663–2674, 2018.

LIANG, Y. et al. Single-image super-resolution-when model adaptation matters. **Pattern Recognition**, Elsevier, v. 116, p. 107931, 2021.

LIM, B. et al. Enhanced deep residual networks for single image super-resolution. In: **Proceedings of the IEEE conference on computer vision and pattern recognition workshops**. [S.l.: s.n.], 2017. p. 136–144.

LIN, T.-Y. et al. Focal loss for dense object detection. In: **Proceedings of the IEEE international conference on computer vision**. [S.l.: s.n.], 2017. p. 2980–2988.

LIU, W. et al. Ssd: Single shot multibox detector. In: SPRINGER. **European conference on computer vision**. [S.l.], 2016. p. 21–37.

LIU, Z.-S. et al. Image super-resolution via attention based back projection networks. arXiv, 2019. Disponível em: <https://arxiv.org/abs/1910.04476>.

LóPEZ-JIMéNEZ, E. et al. Columnar cactus recognition in aerial images using a deep learning approach. **Ecological Informatics**, v. 52, p. 131 – 138, 2019. ISSN 1574-9541. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1574954119300895>.

MA, C. et al. Learning a no-reference quality metric for single-image super-resolution. **Computer Vision and Image Understanding**, Elsevier, v. 158, p. 1–16, 2017.

MEI, Y. et al. Pyramid attention networks for image restoration. **arXiv preprint arXiv:2004.13824**, 2020.

MEI, Y. et al. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2020. p. 5690–5699.

MILLETARI, F.; NAVAB, N.; AHMADI, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: IEEE. **2016 fourth international conference on 3D vision (3DV)**. [S.l.], 2016. p. 565–571.

MITTAL, A.; SOUNDARARAJAN, R.; BOVIK, A. C. Making a "completely blind" image quality analyzer. **IEEE Signal processing letters**, IEEE, v. 20, n. 3, p. 209–212, 2012.

MOHAJERANI, S.; SAEEDI, P. **Cloud-Net+: A Cloud Segmentation CNN for Landsat 8 Remote Sensing Imagery Optimized with Filtered Jaccard Loss Function**. 2020.

MOORTHY, A. K.; BOVIK, A. C. Blind image quality assessment: From natural scene statistics to perceptual quality. **IEEE transactions on Image Processing**, IEEE, v. 20, n. 12, p. 3350–3364, 2011.

MOSTOFA, M. et al. Joint-srvdnet: Joint super resolution and vehicle detection network. **IEEE Access**, IEEE, v. 8, p. 82306–82319, 2020.

NAIR, V.; HINTON, G. E. Rectified linear units improve restricted boltzmann machines. In: **Icml**. [S.l.: s.n.], 2010.

NEUPANE, B.; HORANONT, T.; ARYAL, J. Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis. **Remote Sensing**, Multidisciplinary Digital Publishing Institute, v. 13, n. 4, p. 808, 2021.

OUAHABI, A.; TALEB-AHMED, A. Deep learning for real-time semantic segmentation: Application in ultrasound imaging. **Pattern Recognition Letters**, Elsevier, v. 144, p. 27–34, 2021.

PANG, Y. et al. Jcs-net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images. **IEEE Transactions on Information Forensics and Security**, IEEE, v. 14, n. 12, p. 3322–3331, 2019.

PEREIRA, M. B.; SANTOS, J. A. dos. An end-to-end framework for low-resolution remote sensing semantic segmentation. In: IEEE. **2020 IEEE Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS)**. [S.l.], 2020. p. 6–11.

RABBI, J. et al. Small-object detection in remote sensing images with end-to-end edge-enhanced gan and object detector network. **Remote Sensing**, Multidisciplinary Digital Publishing Institute, v. 12, n. 9, p. 1432, 2020.

REETH, E. V. et al. Super-resolution in magnetic resonance imaging: a review. **Concepts in Magnetic Resonance Part A**, Wiley Online Library, v. 40, n. 6, p. 306–325, 2012.

REN, X.; MALIK, J. Learning a classification model for segmentation. In: IEEE COMPUTER SOCIETY. **Computer Vision, IEEE International Conference on**. [S.l.], 2003. v. 2, p. 10–10.

ROHITH, G.; KUMAR, L. S. Super-resolution based deep learning techniques for panchromatic satellite images in application to pansharpening. **IEEE Access**, IEEE, v. 8, p. 162099–162121, 2020.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. **International Conference on Medical image computing and computer-assisted intervention**. [S.l.], 2015. p. 234–241.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **nature**, Nature Publishing Group, v. 323, n. 6088, p. 533–536, 1986.

RUSSAKOVSKY, O. et al. Imagenet large scale visual recognition challenge. **CoRR**, abs/1409.0575, 2014. Disponível em: <http://arxiv.org/abs/1409.0575>.

SAJJADI, M. S.; SCHOLKOPF, B.; HIRSCH, M. Enhancenet: Single image super-resolution through automated texture synthesis. In: **Proceedings of the IEEE international conference on computer vision**. [S.l.: s.n.], 2017. p. 4491–4500.

SCHMIDHUBER, J. Deep learning in neural networks: An overview. **Neural networks**, Elsevier, v. 61, p. 85–117, 2015.

SENER, O.; KOLTUN, V. Multi-task learning as multi-objective optimization. **Advances in neural information processing systems**, v. 31, 2018.

SHERMEYER, J.; ETTEN, A. V. The effects of super-resolution on object detection performance in satellite imagery. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops**. [S.l.: s.n.], 2019. p. 0–0.

SHI, J.; MALIK, J. Normalized cuts and image segmentation. **IEEE Transactions on pattern analysis and machine intelligence**, Ieee, v. 22, n. 8, p. 888–905, 2000.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, 2014.

SMITH, M. **Enhancedview news not so rosy for GeoEye**. SpacePolicyOnline, 2012. Disponível em: <https://spacepolicyonline.com/news/enhancedview-news-not-so-rosy-for-geoeye/>.

SUMBUL, G. et al. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. **IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium**, IEEE, Jul 2019. Disponível em: <http://dx.doi.org/10.1109/IGARSS.2019.8900532>.

SZEGEDY, C. et al. Going deeper with convolutions. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2015. p. 1–9.

TAN, M.; LE, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: PMLR. **International conference on machine learning**. [S.l.], 2019. p. 6105–6114.

TANG, Z. et al. Srda-net: Super-resolution domain adaptation networks for semantic segmentation. **arXiv preprint arXiv:2005.06382**, 2020.

THORNTON, M. W.; ATKINSON, P. M.; HOLLAND, D. Sub-pixel mapping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping. **International Journal of Remote Sensing**, Taylor & Francis, v. 27, n. 3, p. 473–491, 2006.

TIMOFTE, R.; ROTHE, R.; GOOL, L. V. Seven ways to improve example-based single image super resolution. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 1865–1873.

TSAI, Y.-H. et al. Deep image harmonization. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2017. p. 3789–3797.

VASU, S.; MADAM, N. T.; RAJAGOPALAN, A. Analyzing perception-distortion tradeoff using enhanced perceptual super-resolution network. In: **Proceedings of the European Conference on Computer Vision (ECCV) Workshops**. [S.l.: s.n.], 2018. p. 0–0.

WANG, J. et al. Deep high-resolution representation learning for visual recognition. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 43, n. 10, p. 3349–3364, 2020.

WANG, T.-C. et al. High-resolution image synthesis and semantic manipulation with conditional gans. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2018. p. 8798–8807.

WANG, X. et al. Recovering realistic texture in image super-resolution by deep spatial feature transform. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2018. p. 606–615.

WANG, X. et al. Esrgan: Enhanced super-resolution generative adversarial networks. In: **The European Conference on Computer Vision Workshops (ECCVW)**. [S.l.: s.n.], 2018.

WANG, Z. et al. Image quality assessment: from error visibility to structural similarity. **IEEE transactions on image processing**, IEEE, v. 13, n. 4, p. 600–612, 2004.

WANG, Z.; CHEN, J.; HOI, S. C. Deep learning for image super-resolution: A survey. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 43, n. 10, p. 3365–3387, 2020.

WANG, Z.; SHEIKH, H. R.; BOVIK, A. C. No-reference perceptual quality assessment of jpeg compressed images. In: IEEE. **Proceedings. International conference on image processing**. [S.l.], 2002. v. 1, p. I–I.

WOO, S. et al. Cbam: Convolutional block attention module. In: **Proceedings of the European conference on computer vision (ECCV)**. [S.l.: s.n.], 2018. p. 3–19.

XIA, G. et al. DOTA: A large-scale dataset for object detection in aerial images. **CoRR**, abs/1711.10398, 2017. Disponível em: <http://arxiv.org/abs/1711.10398>.

XU, X. et al. Learning to super-resolve blurry face and text images. In: **Proceedings of the IEEE international conference on computer vision**. [S.l.: s.n.], 2017. p. 251–260.

XU, Y.; LIN, L.; MENG, D. Learning-based sub-pixel change detection using coarse resolution satellite imagery. **Remote Sensing**, Multidisciplinary Digital Publishing Institute, v. 9, n. 7, p. 709, 2017.

YADAN, O. **Hydra - A framework for elegantly configuring complex applications**. 2019. Github. Disponível em: <https://github.com/facebookresearch/hydra>.

YANG, C.-Y.; HUANG, J.-B.; YANG, M.-H. Exploiting self-similarities for single frame super-resolution. In: SPRINGER. **Asian conference on computer vision**. [S.l.], 2010. p. 497–510.

YANG, J.; HUANG, T. Image super-resolution: Historical overview and future challenges. In: **Super-resolution imaging**. [S.l.]: CRC Press, 2017. p. 1–34.

YANG, J. et al. Coupled dictionary training for image super-resolution. **IEEE transactions on image processing**, IEEE, v. 21, n. 8, p. 3467–3478, 2012.

YANG, W. et al. Deep learning for single image super-resolution: A brief review. **IEEE Transactions on Multimedia**, IEEE, v. 21, n. 12, p. 3106–3121, 2019.

YUAN, X.; SHI, J.; GU, L. A review of deep learning methods for semantic segmentation of remote sensing imagery. **Expert Systems with Applications**, Elsevier, v. 169, p. 114417, 2021.

ZHANG, H. et al. Self-attention generative adversarial networks. In: PMLR. **International conference on machine learning**. [S.l.], 2019. p. 7354–7363.

ZHANG, R. et al. The unreasonable effectiveness of deep features as a perceptual metric. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2018. p. 586–595.

ZHANG, Y. et al. Image super-resolution using very deep residual channel attention networks. In: **Proceedings of the European Conference on Computer Vision (ECCV)**. [S.l.: s.n.], 2018. p. 286–301.

ZHANG, Y. et al. Residual dense network for image super-resolution. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2018. p. 2472–2481.

ZHANG, Z.; LIU, Q.; WANG, Y. Road extraction by deep residual u-net. **IEEE Geoscience and Remote Sensing Letters**, v. 15, n. 5, p. 749–753, 2018.

ZHAO, Z. et al. Compression artifacts reduction by improved generative adversarial networks. **EURASIP Journal on Image and Video Processing**, SpringerOpen, v. 2019, n. 1, p. 1–7, 2019.

ZHOU, W.; HUANG, G.; CADENASSO, M. L. Does spatial configuration matter? understanding the effects of land cover pattern on land surface temperature in urban landscapes. **Landscape and urban planning**, Elsevier, v. 102, n. 1, p. 54–63, 2011.

ZOPH, B. et al. Rethinking pre-training and self-training. **Advances in neural information processing systems**, v. 33, p. 3833–3845, 2020.