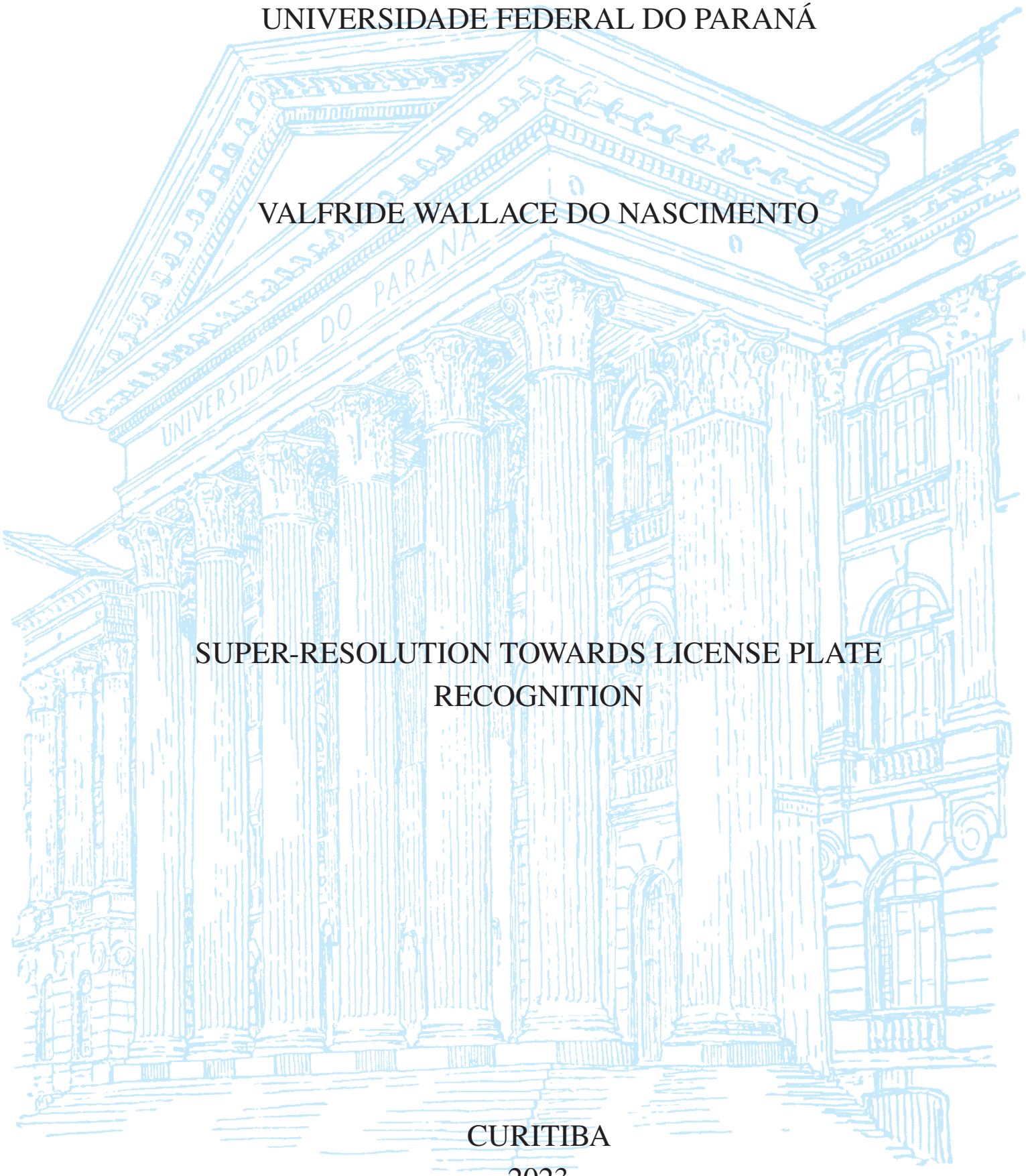
UNIVERSIDADE FEDERAL DO PARANÁ

VALFRIDE WALLACE DO NASCIMENTO




# SUPER-RESOLUTION TOWARDS LICENSE PLATE RECOGNITION

CURITIBA
2023

VALFRIDE WALLACE DO NASCIMENTO

# SUPER-RESOLUTION TOWARDS LICENSE PLATE RECOGNITION

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Informática, no Programa de Pós-Graduação em Informática, setor de Ciências Exatas, da Universidade Federal do Paraná..

Área de concentração: *Ciência da Computação*.

Orientador: David Menotti.

CURITIBA

2023

Catalogação na Fonte: Sistema de Bibliotecas, UFPR
Biblioteca de Ciência e Tecnologia

N244s Nascimento, Valfride Wallace do

Super-resolution towards license plate recognition [recurso eletrônico] / Valfride Wallace do Nascimento – Curitiba: UFPR, 2023.

Dissertação (Mestrado) apresentada como requisito parcial à obtenção do grau de Mestre no Programa de Pós-Graduação em Informática, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Orientador: Prof. Dr. David Menotti Gomes

1. Placa veicular. 2. Identificação de veículos. 3. Processamento de imagens I. Gomes, David Menotti. II. Universidade Federal do Paraná. III. Título.

Bibliotecária: Vilma Machado CRB-9/1563

# TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **VALFRIDE WALLACE DO NASCIMENTO** intitulada: **Super-Resolution Towards License Plate Recognition**, sob orientação do Prof. Dr. DAVID MENOTTI GOMES, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 24 de Abril de 2023.

Assinatura Eletrônica
24/04/2023 22:14:28.0
DAVID MENOTTI GOMES
Presidente da Banca Examinadora

Assinatura Eletrônica
25/04/2023 08:22:20.0
RAONI FLORENTINO DA SILVA TEIXEIRA
Avaliador Externo (UNIVERSIDADE FEDERAL DE MATO GROSSO)

Assinatura Eletrônica
27/04/2023 09:04:25.0
ANDRÉ GUSTAVO HOCHULI
Avaliador Externo (PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ)

Assinatura Eletrônica
26/04/2023 06:43:50.0
LUIZ EDUARDO SOARES DE OLIVEIRA
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

*To my parents and friends.*

# Acknowledgments

I am deeply grateful to my advisor, Prof. David Menotti, whose guidance and mentorship have been invaluable throughout my academic journey. With his unwavering support and encouragement, I have been able to pursue my passion for research and gain invaluable experience and knowledge under his tutelage. David provided me with the freedom to explore various ideas and methodologies while offering critical insights that helped shape my research. His dedication and passion for his work have been a constant source of inspiration, and I feel fortunate to have had the opportunity to work under his guidance.

I am also grateful to my research colleague and dear friend, Rayson Laroca, for his extensive experience, support, and innovative ideas. Rayson's contributions were instrumental in shaping my research. His dedication and perseverance were a constant source of motivation, and I feel truly fortunate to have had the opportunity to collaborate with him.

Many thanks to the members of my dissertation committee, Prof. Raoni F. Teixeira and Prof. Luiz Eduardo Soares de Oliveira, for their time and effort in organizing and working on my committee, which significantly improved the quality of my dissertation. Their feedback on my methodology and writing style was invaluable, and I am grateful for their support throughout the process.

I am deeply grateful to all those who inspired me to pursue an academic research career, particularly Prof. Raoni and everyone at Universidade Federal of Mato Grosso. Without their guidance, I would not have discovered this field and the life-changing experience it has provided me.

I would like to express my deepest gratitude to my parents, José Rubens Ramos do Nascimento and Ariadna Aparecida de Oliveira Nascimento, for their unwavering support, unconditional love, and encouragement throughout my life. Their guidance and emotional support have been invaluable in helping me reach this point in my academic journey. I would also like to extend my appreciation to my family and friends for their unwavering love and support during my academic journey. The experiences and knowledge that I have gained have prepared me for future challenges and opportunities. Finally, I am grateful to God for all the opportunities and experiences that have shaped my journey.

*"One year spent in artificial intelligence is enough to make someone believe in God."*
*Alan Perlis*

# RESUMO

Nos últimos anos, houve avanços significativos no campo de Reconhecimento de placas de veículares (LPR, do inglês *License Plate Recognition*) por meio da integração de técnicas de aprendizado profundo e do aumento da disponibilidade de dados para treinamento. No entanto, reconstruir placas veiculares a partir de imagens de sistemas de vigilância em baixa resolução ainda é um desafio. Para enfrentar essa dificuldade, apresentamos uma abordagem de Super Resolução de Imagem Única (SISR, do inglês *Single-Image Super-Resolution*) que integra módulos de atenção para aprimorar a detecção de característica estruturais e texturais em imagens de baixa resolução. Nossa abordagem utiliza camadas de convolução sub-pixel (também conhecidas como PixelShuffle) e uma função de perda que emprega um modelo de Reconhecimento Óptico de Caracteres (OCR, do inglês *Optical Character Recognition*) para extração de características. Treinamos a arquitetura proposta com imagens sintéticas criadas aplicando ruído gaussiano pesado à imagens de alta resolução de placas veiculares de dois conjuntos de dados públicos, seguido de redução de sua resolução com interpolação bicúbica. Como resultado, as imagens geradas têm um Índice de Similaridade Estrutural (SSIM, do inglês *Structural Similarity Index Measure*) inferior a 0,10. Nossos resultados experimentais mostram que a abordagem proposta para reconstruir essas imagens sintéticas de baixa resolução superou as existentes tanto em medidas quantitativas quanto qualitativas.

**Palavras-chave:** PixelShuffle, Reconstrução, Super-Resolução.

# ABSTRACT

Recent years have seen significant developments in the field of License Plate Recognition (LPR) through the integration of deep learning techniques and the increasing availability of training data. Nevertheless, reconstructing license plates (LPs) from low-resolution (LR) surveillance footage remains challenging. To address this issue, we introduce a Single-Image Super-Resolution (SISR) approach that integrates attention and transformer modules to enhance the detection of structural and textural features in LR images. Our approach incorporates *sub-pixel convolution layers* (also known as PixelShuffle) and a loss function that uses an Optical Character Recognition (OCR) model for feature extraction. We trained the proposed architecture on synthetic images created by applying heavy Gaussian noise to high-resolution LP images from two public datasets, followed by bicubic downsampling. As a result, the generated images have a Structural Similarity Index Measure (SSIM) of less than 0.10. Our results show that our approach for reconstructing these low-resolution synthesized images outperforms existing ones in both quantitative and qualitative measures.



**Keywords:** PixelShuffle, Reconstruction, Super-Resolution.

# List of Figures

# List of Tables

# Acronyms

**ARB** Adaptive Residual Block. 29, 41

**CA** Channel Unit. 31, 32, 40, 41, 46

**CCPD** Chinese City Parking Dataset. 35

**CNN** Convolutional Neural Network. 16, 23, 24, 25, 27, 33, 35

**DConv** depthwise-separable convolutional layer. 25, 40, 41, 48

**DL** Deep Learning. 33

**DPCA** Dual-Coordinate Direction Perception Attention. 34

**Dw** Depth-wise Convolution. 25, 26, 31

**ESRGAN** Enhanced Super-Resolution Generative Adiversarial Network. 35

**FM** Feature Module. 39

**GAN** Generative Adversarial Network. ix, 17, 18, 29, 30, 35

**GP** Geometrical Perception Unit. 40, 41, 46

**HR** high-resolution. x, 16, 18, 24, 28, 33, 34, 35, 38, 44, 45, 46, 47, 50

**HVS** Human Visual System. 20, 22

**LP** license plate. viii, x, xi, 18, 34, 35, 37, 38, 39, 40, 41, 42, 44, 45, 46, 47, 48, 49, 50

**LPR** License Plate Recognition. viii, 16, 17, 18, 19, 33, 34, 35, 36, 39, 40, 42, 46, 47, 48, 50

**LR** low-resolution. viii, ix, x, 16, 17, 18, 19, 24, 28, 29, 31, 33, 34, 35, 37, 38, 42, 44, 45, 46, 47, 48, 50

**LReLU** Leaky ReLU. 27

**MAP** Maximum a Posteriori. 34

**MISR** Multi-Image Super-Resolution. 16, 35

**MPRNet** Multi-Path Residual Network. 17, 18, 19, 29, 33, 39, 40, 46, 47, 50

**MSE** Mean Squared Error. ix, 20, 21, 23, 42, 43, 48, 50

**NN** Neural Network. 26, 27

**OCR** Optical Character Recognition. viii, 17, 18, 19, 35, 38, 39, 42, 46, 47, 50

**POS** Positional Unit. 31, 32, 40, 41, 46

**PReLU** Parametric ReLU. 27

**PS** *PixelShuffle*. x, 18, 28, 39, 40, 41, 48, 49, 50

**PSNR** Peak Signal-to-Noise Ratio. 16, 20, 21, 22, 35, 37, 46

**PTFAM** Pixelshuffle Three-fold Attention Module. x, 18, 39, 40, 41, 48, 50

**PU** *PixelUnshuffle*. x, 18, 39, 40, 48, 49, 50

**Pw** Point-wise Convolution. 25, 26, 31

**RDB** Residual Dense Block. 39, 41, 50

**RDL** Residual Dense Layer. 41

**ReLU** rectified linear function. 27

**RM** Reconstruction Module. 39

**SFE** Shallow Feature Extractor. 39

**SISR** Single-Image Super-Resolution. viii, 16, 33, 34

**SR** super-resolution. ix, 16, 19, 28, 29, 30, 33, 34, 35, 36, 38

**SRCNN** Super-Resolution Convolutional Neural Network. 33

**SSIM** Structural Similarity Index Measure. viii, 16, 18, 22, 35, 37, 45, 46

**Tanh** hyperbolic tangent. 27

**TCL** Traposed Convolution Layer. 24, 25

**TFAM** Two-fold Attention Module. x, 31, 32, 34, 39, 40, 41, 48, 49

**VSR** Video Super-Resolution. 16

# SUMÁRIO

# Chapter 1

# Introduction

Super-resolution (SR) is a crucial technology that enhances the quality of images and videos by increasing their resolution, enabling the retrieval of subtle details and textures from low-resolution (LR) images to generate their high-resolution (HR) counterparts [10, 11], as shown in Fig. 1.1. Its importance has grown in fields such as medical imaging and surveillance, where SR is extensively used [10, 11, 12]. Recent advancements in interpolation-based, example-based, and deep learning-based SR methods have made it possible to enhance LR images and videos in a way that was once deemed impossible. This is particularly important in surveillance applications such as License Plate Recognition (LPR), face, and object recognition, where image quality improvement is critical but challenging [13]. Moreover, it is desirable to store HR images in LR format and recover them when necessary [14, 15].

SR is a challenging problem due to its ill-posed nature, where there can be multiple solutions in the HR space. The difficulty of the problem increases as the upscale factor increases, and LR images may lack the necessary information to reconstruct the desired details. SR techniques can be broadly classified into three categories: Single-Image Super-Resolution (SISR), Multi-Image Super-Resolution (MISR), and Video Super-Resolution (VSR) [11, 16]. In this study, we focus on SISR for forensic license plate recognition, as low-cost cameras frequently used in surveillance systems produce LR images that make the characters on license plates barely recognizable.

As deep learning continues to show remarkable success in computer vision applications, the integration of Convolutional Neural Networks (CNNs) into SR techniques has become increasingly prevalent [6, 11, 17]. Although significant advances have been made, most existing SR approaches rely on very deep architectures, which not only increase overall computation but also prioritize achieving higher Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) metrics at the expense of the contextual information of the application at hand. However, in LPR, this approach falls short as a single model may produce highly realistic images yet fail to differentiate between visually similar characters (e.g., 'Q' and 'O', 'T' and '7', 'Z' and '2', among others) [6, 15, 17, 18, 19, 20]. In the context of LPR, it may not be the best approach to generate highly realistic images without taking into account the potential confusion between characters. Therefore, it is essential to consider the particular application at hand when proposing SR methods [11, 14].

## 1.1 Motivations

Accurate LPR relies on high-quality images, and while many techniques have been proposed to enhance image quality, little research has been dedicated to using SR techniques to

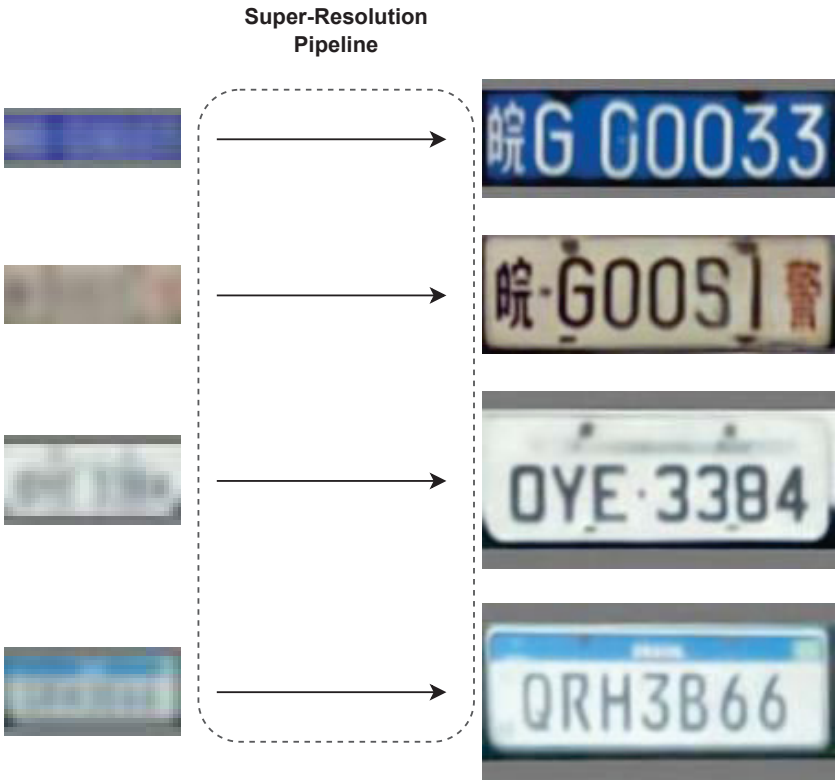



Figure 1.1: This figure illustrates the super-resolution pipeline. The left-hand side shows low-resolution/quality images, which are then super-resolved to the higher resolution images on the right-hand side. These resulting images were generated using the method proposed in this work.

improve LPR accuracy [21]. The limited effectiveness of existing interpolation methods, such as bicubic, is due to their inability to learn and improve over time.

To address this issue, we propose an extension of the Multi-Path Residual Network (MPRNet) developed by Mehri et al. [6] and our previous work in [7]. We incorporate sub-pixel convolution layers (also known as *PixelShuffle*) introduced by Shi et al. [22] and a Three-Fold Attention Module that considers not only pixel intensity values but also structural and textural information. Additionally, we propose a novel loss function that incorporates Optical Character Recognition (OCR) predictions, perceptual quality metrics, and Generative Adversarial Network (GAN) techniques to enhance character recognition in low-quality and LR scenarios.

Our research hypothesis is that *this improved version will outperform OCR systems that directly receive LR license plate images*. To support our research, we have created a publicly available synthetic dataset.

## 1.2 Objectives

This research aims to enhance the perceptual quality and resolution of LR license plate images by extending the MPRNet architecture with sub-pixel convolution layers [22] and a new loss function that prioritizes character reconstruction for recognition. Specifically, our objectives are:

- Develop a loss function that considers quality metrics and predictions from a pre-trained OCR model. This will allow us to focus on reconstructing important features in license plates to improve recognition accuracy.
- Design an encoder module that emphasizes crucial features in the license plate image by utilizing an autoencoder with sub-pixel convolution layers. This module will enable us to enhance the image resolution and quality.
- Create an Attention module capable of identifying the most critical aspects within an image for quality and resolution enhancement;
- Develop a loss function to guide the proposed architecture towards more realistic and higher-quality super-resolved license plates (LPs) using GAN techniques. This will ensure that the generated images are of high-resolution/quality with more realism;
- Compare the performance of the original MPRNet with our proposed approaches on LPR, and evaluate their ability to improve OCR accuracy on the created datasets.

## 1.3 Contributions

Our work presents the following contributions:

- A super-resolution approach that builds upon MPRNet [6] and the architecture we proposed in [7] (see Section 1.4) by incorporating subpixel-convolution layers (*PixelShuffle* (PS) and *PixelUnshuffle* (PU)) in combination with a Pixelshuffle Three-fold Attention Module (PTFAM);
- A novel perceptual loss that combines features extracted by an OCR model [23] with L1 loss to reconstruct characters with the most relevant characteristics. Note that this loss function allows the use of any OCR model for LPR;
- Datasets with paired HR and synthetic LR images generated by applying heavy Gaussian noise at different SSIM levels;
- The datasets we created for this research are publicly available upon request. [1]

## 1.4 Publications

Our study introduces a new approach to license plate super-resolution, which we first presented in a preliminary version at the 2022 SIBGRAPI conference titled "Combining Attention Module and Pixel Shuffle for License Plate Super-Resolution" [7]. Compared to our previous work, our current approach has several novel features that improve license plate reconstruction. For instance, we use a three-fold attention module architecture that considers vertical and horizontal lines to extract more structural and textural details of the license plate font. This module extends the concepts presented in our previous work [7] and leverages inter-channel feature relationships to enhance the reconstruction process.

To further enhance our approach, we introduce a new loss function that utilizes a pre-trained network for license plate recognition to extract features. For training and testing, we

[1] Interested parties must register by filling out a form and agreeing to the terms of use.

employed paired low- and high-resolution license plate images, with the low-resolution images degraded until their structural similarity index measure dropped below 0.10.

Our improvements resulted in better performance than our previous work, as demonstrated by experiments conducted on two datasets collected in different regions under various conditions. Unlike our previous work, which only used a single dataset, we tested our approach on both the RodoSol-ALPR and PKU datasets. The results showed a significant improvement in license plate recognition rates. In the RodoSol-ALPR dataset, our approach recognized at least five characters in 74.2% of the license plates, compared to 42.2% by our previous model, which was trained and evaluated under the same conditions. In the PKU dataset, our approach achieved a recognition rate of 97.3%, compared to 82.5% achieved by our preliminary approach.

We believe that the improvements we made over the preliminary version in [7] fulfill the objectives and contributions detailed in Section 1.2 and Section 1.3, respectively. Currently, our work is under review as a SIBGRAPI 2022 Post-Conference Special Section of the Computers & Graphics Journal.

## 1.5 Outline

The remainder of this dissertation is organized as follows. In Chapter 2, we present the theoretical foundation for the concepts discussed in this work. Additionally, in Chapter 3, we review related works on general SR techniques, LPR, OCR, and their application in improving LPR in LR and low-quality scenarios. In Chapter 4, we describe how we extended the MPRNet architecture proposed by [6] and present our improved loss function. In Chapter 5, we provide a detailed description of the experiments we conducted to validate our proposed approach, including the experimental setup, results, and analysis. Finally, in Chapter 6, we present our conclusions and discuss future work in the field.

# Chapter 2

# Theoretical Foundation

In this chapter, we will establish the theoretical foundation for the concepts and methods utilized in this study. We will begin by outlining the metrics commonly employed for assessing image quality in Section 2.1. In Section 2.2, we will delve into the details of the CNN layers used in the context of super resolution image enhancement. Section 2.3 will focus on the Generative Adversarial Networks (GANs) and their significance to this research. Lastly, in Section 2.4, we will introduce the Multi-Patch Relational Network (MPRNet) architecture from Mehri et al. [6] as the baseline for this research.

## 2.1 Evaluation Metrics

In this section, we will discuss the evaluation metrics commonly used for assessing the quality of digital images. These metrics are necessary due to the fact that the human perception of image quality can vary from person to person and is therefore subjective.

Several metrics have been developed in the literature to measure the visual quality of images. The simplest ones include the Mean Squared Error (MSE) and its variation, the PSNR. While these metrics are in line with physical principles and are easy to calculate and implement, they do not take into account the Human Visual System (HVS) [24].

To address this issue, considerable efforts have been made over the last three decades to develop methods that consider human visual perceptual qualities. One of the most prominent methods is the Structural SIMilarity index (SSIM) [25, 26, 27, 1].

In this section, we will provide a theoretical foundation for the commonly used quantitative metrics for image quality assessment, specifically MSE, PSNR, and SSIM. These metrics are widely used in the literature to quantify the level of quality within an image. For the following definitions, we consider two non-negative and perfectly aligned images: $G$ as the original high-quality image and its distorted version $P$.

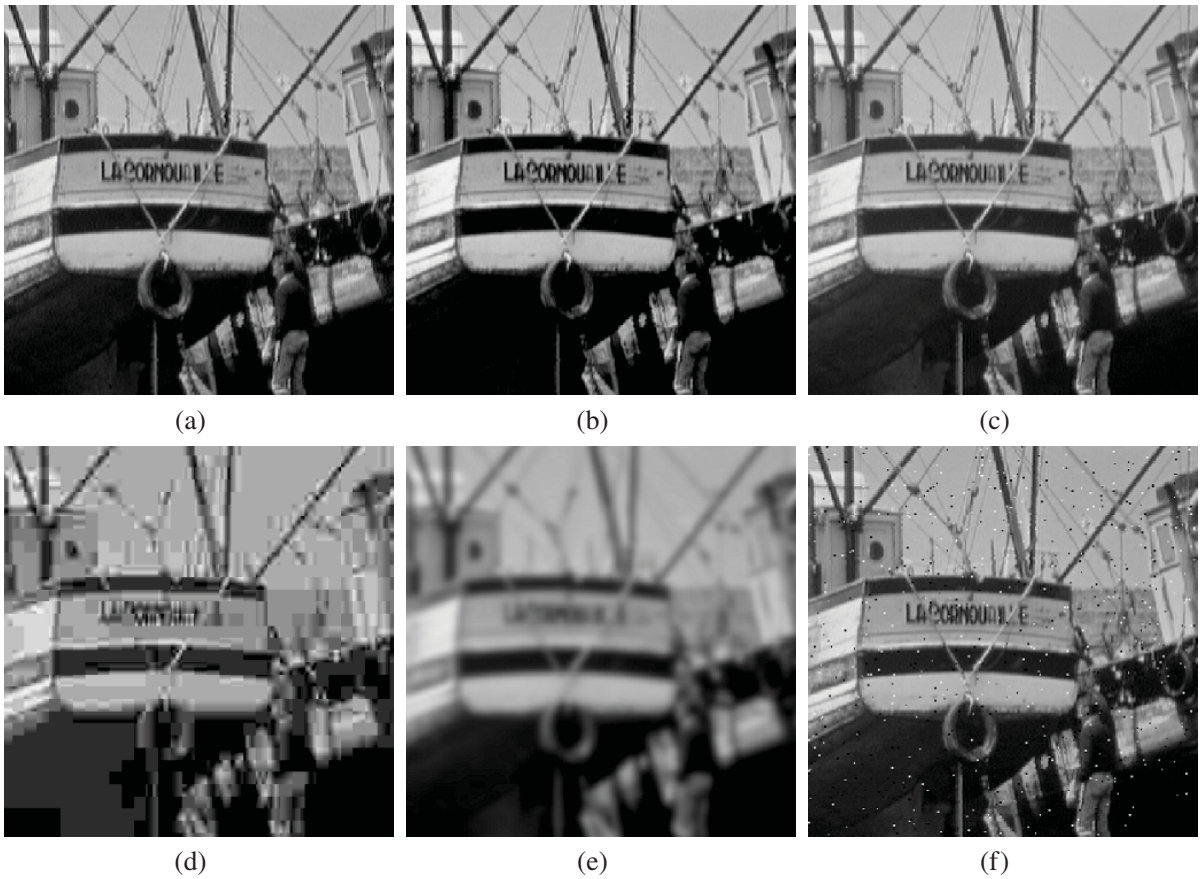
### 2.1.1 Mean Squared Error

The MSE is a commonly used metric for measuring the quality of images. It is always positive, and higher values indicate the presence of distortion or noise in the distorted image (P) when compared to the original image (G). As shown in Equation 2.1, it is simple to calculate and inexpensive to compute as it only involves multiplication and addition operations. Additionally, MSE has the desirable properties of convexity, symmetry, and differentiability, and its gradients are easy to compute.

$$MSE(G,P) = \frac{1}{ij}\sum^{i-1} i = 0 \sum^{j-1} j = 0 |G(i,j) - P(i,j)|^2 \tag{2.1}$$

However, the MSE does not take into account the structural information of the image during its calculation. This means that the value of MSE between two images will remain unchanged even if the pixels are randomly rearranged. This is illustrated in Figure 2.1.

Despite its limitations, MSE is widely used in the literature due to its simplicity and ease of computation.



(a) (b) (c)

(d) (e) (f)

Figure 2.1: The comparison shows an original image and its distorted versions with different types of noise and distortions, all having the same MSE value of 0.210. The original image is shown in (a), while (b) shows a contrast-stretched version, (c) is a mean-shifted version, (d) is a JPEG compressed version, (e) is a blurred version, and (f) is a salt-pepper impulsive noise contaminated version. The image is reproduced from [1].

The PSNR is a metric that quantifies the ratio between the maximum possible signal value and the value of corrupting noise that affects the quality or fidelity of an image representation. It is commonly used to measure the quality of reconstructed images subject to distorting noises such as those caused by compression methods or poor image acquisition devices.

PSNR is typically defined using MSE as in Equation 2.2, where $G$ is a ground truth noise-free image and its degraded counterpart $P$ (see Equation 2.1). $MAX_G$ represents the maximum possible pixel value in $G$ (e.g. for 8-bit samples, this is 255). Since signals have a wide changeable quantity between the largest $MAX$ and the smallest values, PSNR is usually defined in terms of the logarithmic decibel scale, where higher values are better. For monochrome images with 8-bit depth, good values range between 30 dB and 50 dB.

However, it has been shown that although PSNR has an intuitive natural physical definition in Equation 2.2, its values do not accurately represent human visual perception [28, 29, 1].

$$\begin{aligned} PSNR(G,P) &= 10\log_{10}\left(\frac{MAX_G^2}{MSE(G,P)}\right) \\ &= 20\log_{10}\left(\frac{MAX_G}{\sqrt{MSE(G,P)}}\right) \\ &= 20\log_{10}\left(MAX_G\right) - 10\log_{10}\left(MSE(G,P)\right) \end{aligned} \tag{2.2}$$

### 2.1.2 Structural Similarity Index Measure

Previous methods have clear physical meanings and simple formulations, but they are not able to match the human visual system's (HVS) assessment of perceptual quality. This is because these methods rely on assumptions and generalizations based on linear or quasi-linear models from the early days of computer vision, which do not accurately reflect the highly complex and non-linear nature of the HVS.

The SSIM index, proposed by [1], addresses this limitation by exploiting the structural nature of image signals. It does this by analyzing local illuminance similarity in terms of luminance, contrast, and structural similarity within the image. These similarities are expressed as simple statistics that are easy to compute, which are then combined to form the SSIM index, as shown in Eq. (2.3) and Fig. 2.2.

$$\begin{aligned} SSIM(G,P) &= l(G,P)\cdot c(G,P)\cdot s(G,P) \\ &= \left(\frac{2\mu_G\mu_P + C_1}{\mu_G^2 + \mu_P^2 + C_1}\right)\cdot\left(\frac{2\sigma_G\sigma_P + C_2}{\sigma_G^2 + \sigma_P^2 + C_2}\right)\cdot\left(\frac{\sigma_{GP} + C_3}{\sigma_G\sigma_P + C_3}\right), \end{aligned} \tag{2.3}$$

where:

$$\mu(x) = \frac{1}{N}\sum_{i=0}^{N} x_i \quad \text{and} \quad \sigma(x) = \left(\frac{1}{N-1}\sum_{i=0}^{N}(x_i - \mu_x)\right)^{\frac{1}{2}} \tag{2.4}$$

$\mu_G$ stands for the local sample average and $\sigma_G$ stands for the local standard deviation, while $\sigma_{GP}$ is the local sample cross-correlation when removing the averages of $G$ and $P$.

The values $C_1$, $C_2$, and $C_3$ are small constants added to stabilize the terms, although even with $C_1 = C_2 = C_3 = 0$, the SSIM index works relatively well. The SSIM has the property of symmetry; therefore, it generates the same value regardless of the ordering of $G$ and $P$: $SSIM(G,P) = SSIM(P,G)$. The resulting SSIM index is bounded between 0 and 1, achieving 1 only when $G$ and $P$ are identical.

Despite this, the SSIM index performs well across a variety of images $P$ generated with different types of noise and distortions, as shown in [30]. Several experiments were conducted to compare its performance against the HVS.

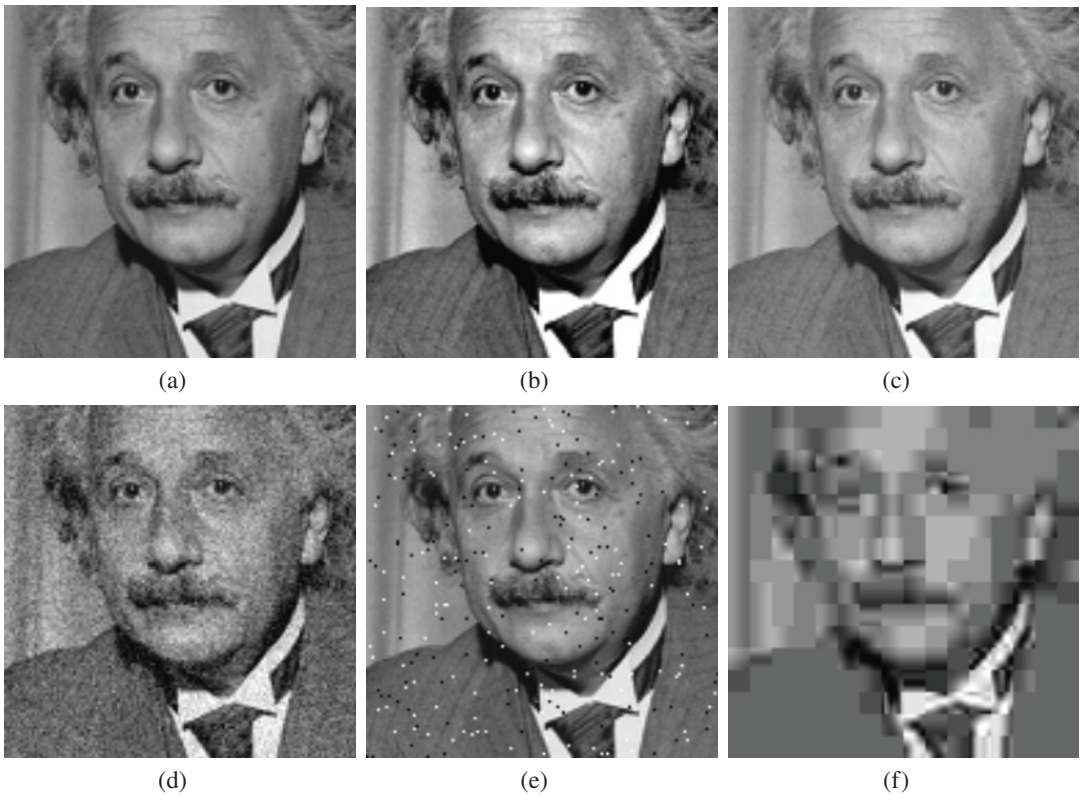



Figure 2.2: Comparison of image fidelity measures for the "Einstein" image altered with different types of distortions and an MSE of around 0.308. (a) Reference image. (b) Mean contrast stretch. (c) Luminance shift. (d) Gaussian noise contamination. (e) Impulsive noise contamination. (f) JPEG compression. The image is reproduced from [2].

## 2.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs), also known as ConvNets, are a widely-used type of neural network architecture in computer vision tasks. These networks are specifically designed to process data with a grid-like pattern, such as images, and are capable of learning hierarchical spatial features from low to high levels of abstraction.

The building blocks of a CNNs typically include convolutional and pooling layers to extract features, an activation function, and fully connected layers (similar to hidden layers of a Multilayer Perceptron) to map the extracted features to the final output.

Mathematically, a CNN can be described as a sequence of functions that take an image $I_i$ and a set of weights $W_i$ as input and produce a vector $O$ as output:

$$f(x_1) = f_l(\cdots f_2(f_1(x_1, W_1), W_2)\cdots), W_l)$$

where $f_l(\cdot)$ performs a convolution operation, which gives the architecture its name.

For the rest of this section, we will describe in more detail the convolutional layers, activation functions, and pooling operations that make up a CNNs.

### 2.2.1 Convolutional Layer

A convolutional neural network (CNN) is primarily composed of convolutional layers. These layers consist of units that are connected to feature maps from previous layers through sets

of weights called filter banks. Each unit in a feature map shares the same filter bank, but each layer has its own unique set of filter banks.

A filter is defined as a matrix of kernel size $n \times m$ that slides over the entire image in steps of a specific size $s$, resulting in a weighted sum. This process is mathematically similar to a discrete convolution operation [31, 5].

The convolutional layer is designed for two main reasons:

1. In images, nearby pixels often have a strong correlation and form distinct patterns that can be easily detected.

2. Patterns can appear anywhere in the image. By having units share the same set of weights, the same pattern can be identified in any part of the image.

As stated in the literature, convolutional layers are not affected by changes in the scale or position of the image. Therefore, other mechanisms such as non-linear functions (Section 2.2.4) and pooling operations (Section 2.2.5) are necessary to handle these issues [5, 31].

### 2.2.2 Deconvolution Layer

Often, CNNs build up HR images from LR feature map descriptors. This allows the network to extract relevant features from the rough image and fill in the missing details.

CNNs often reconstruct images from low-resolution (LR) feature maps, which allows the network to extract relevant features from the rough image and fill in missing details. To achieve this, an operation that goes in the opposite direction of a convolutional layer is needed, meaning going from an LR space to a higher dimensional one. This is achieved through an operation known as a Traposed Convolution Layer (TCL).

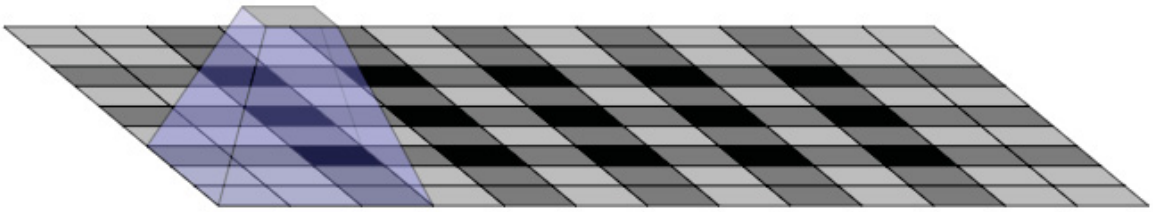


Figure 2.3: Deconvolution layers can generate checkboard patterns, as shown in the image reproduced from [3].

The TCL [1] concept was first introduced by[32] and later formalized in [33]. Convolutional operations are defined in terms of their kernel, but whether they behave as a transposed or direct convolution is determined by the order in which the forward and backward operations are performed. These operations can be represented as a sparse matrix $M$ and a kernel $k$. In a direct convolution, the forward and backward passes are defined by multiplying $M$ with $k$ and $M^T$, respectively. By swapping the order of this multiplication, we obtain the transposed convolution. Essentially, TCL allows the model to learn how to generate a set of pixels in high-resolution (HR) space based on a single pixel from LR space [3, 34].

[1] Also known as fractionally strided convolutions or deconvolutions (This term is misleading since deconvolution is mathematically described as the inverse of a convolution which, in fact, is not the operation performed)

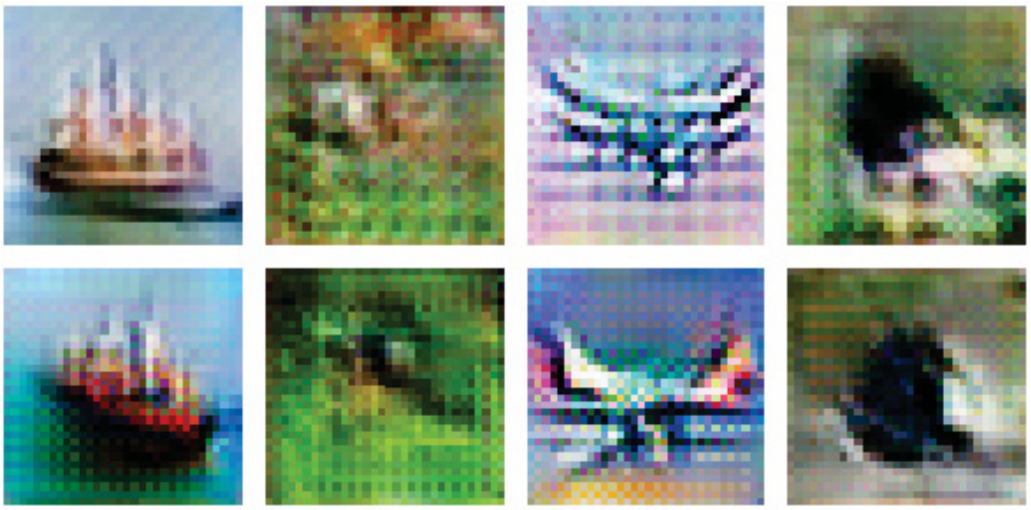


Figure 2.4: Pixel Overlapping in [3]. Image reproduced from [3].

After the success of [32]'s work, TCLs were adopted in many areas, including flow estimation [35] and generative modeling [36]. Despite their outstanding results, TCL operations can easily reproduce overlapping pixels, as highlighted in Figure 2.4, leading to checkerboard-like patterns, as exemplified in Figure 2.3 [37]. Although theoretically, CNNs can learn the correct weights to avoid such artifacts, in practice, doing so restricts possible filters and reduces the model's learning capacity.

### 2.2.3 Depth-wise Convolution Layer

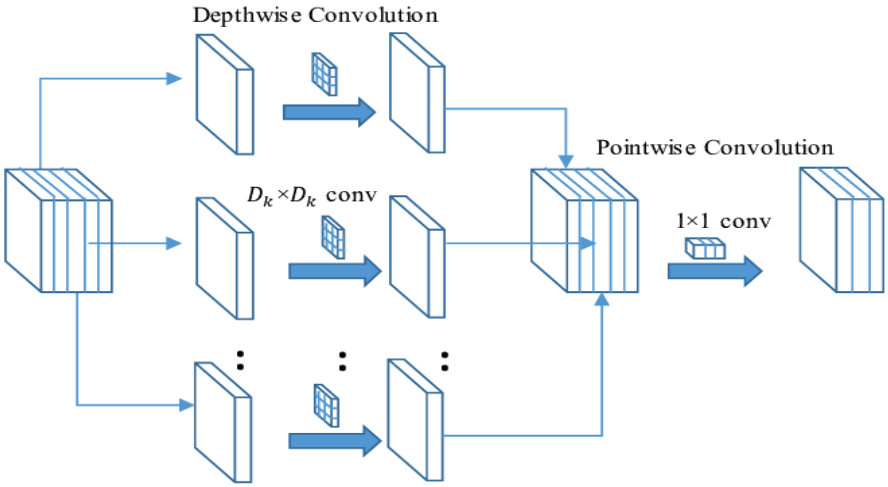



Figure 2.5: Depthwise-separable convolutional layer pipeline. Image reproduced from [4].

Depthwise-separable convolutional layers (DConvs) are a popular technique used in many efficient neural networks, such as ShuffleNet [38], MobileNets [39], and Xception [40]. The main idea of this method is to split a standard Convolution layer into two consecutive operations: Depth-wise Convolution (Dw) and Point-wise Convolution (Pw), as shown in Figure 2.5.

In the Dw operation, each channel of the input is convolved with a single filter. Then, in the Pw operation, a $1 \times 1$ convolution is applied to linearly combine the filtered outputs into new feature maps [39].

The following equation describes the Dw operation with one filter per input channel:

$$\hat{F}k, l, m = \sum n = i, j\hat{K}i, j, m \cdot \hat{G}k + i - j, l + j - 1, m \tag{2.5}$$

Here, $\hat{K}$ is a Dw kernel of size $D_k \times D_k \times M$, with the $m^{th}$ kernel applied to the $m^{th}$ channel of $\hat{G}$ to generate a filtered output feature map $\hat{F}$. The filtered outputs are then combined linearly in the Pw operation.

In comparison, a standard Convolution operation applies both operations in a single layer, taking into account the effects of all $K$ convolutional kernels of size $D_K \times D_K \times M \times N$, where $D_K$ is the spatial dimension with $M$ input and $N$ output channels. It generates a new representation $F$ by combining the features of $G$ as shown in the following equation:

$$F_{k,l,n} = \sum_{i,j,m} K_{i,j,m,n} \cdot G_{k+i-1,l+j-1,m} \tag{2.6}$$

The standard Convolution operation has a much higher computational cost compared to the Dw operation, which leads to a reduction in computation cost when using Dw instead. Specifically, the computational cost of a standard Convolution operation is:

$$D_K \times D_K \times M \times N \times D_F \times D_F \tag{2.7}$$

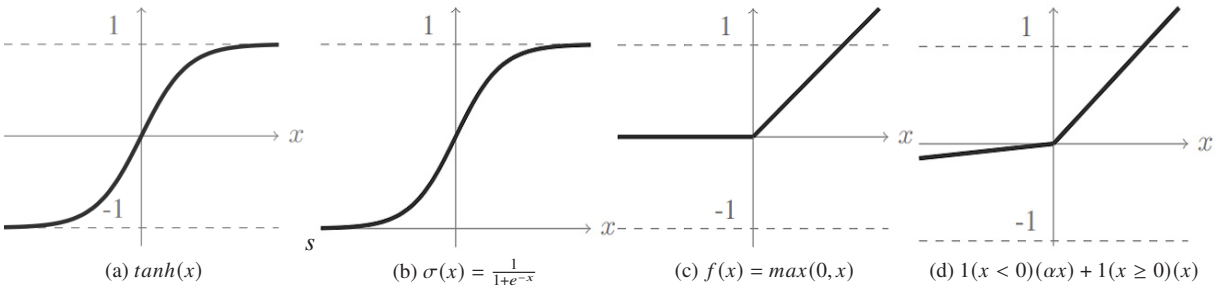
whereas the computational cost of using Dw and Pw operations is:

$$D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F \tag{2.8}$$

According to Howard et al. [39], their MobileNet architecture uses Dw layers, resulting in 8 to 9 times less computation with a small reduction in accuracy.

### 2.2.4 Activation Functions

An activation function is a tool used in Neural Networks (NNs) that helps determine which values generated by layers should be passed on to the output. These functions act like gates, deciding which information is useful and which is not. Without activation functions, the network's weights and biases would only be able to perform linear transformations, which are not powerful enough to learn complex patterns and mappings from input data [5]. Activation functions are typically applied after each layer in a NN, except for pooling layers, which only down-sample the input data and do not require activation functions.




(a) $tanh(x)$ (b) $\sigma(x) = \frac{1}{1+e^{-x}}$ (c) $f(x) = max(0, x)$ (d) $1(x < 0)(\alpha x) + 1(x \geq 0)(x)$

Figure 2.6: Activation Functions. (a) Hyperbolic Tangent. (b) Sigmoid. (c)Rectified Linear. (d) Parametric ReLU. Image reproduced from [5]

One of the most commonly used activation functions is the sigmoid function, which is named after its "S" shape on the y-axis in the $\mathbb{R}^2$ plane. It is a smoothing function that is easy to

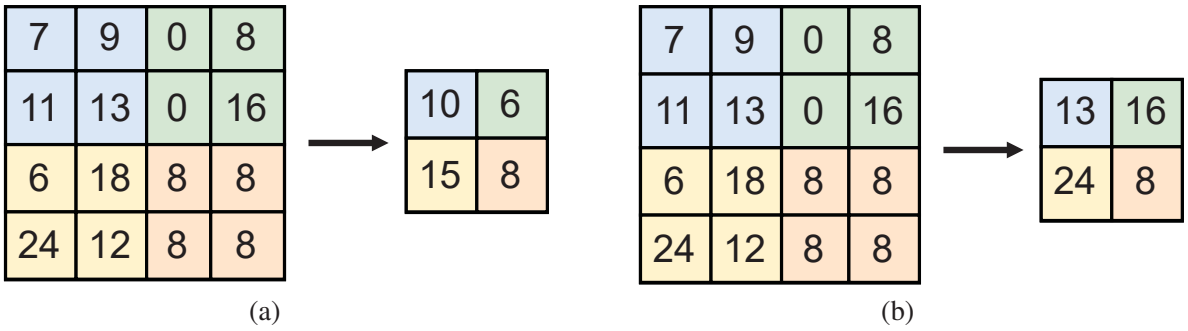



Figure 2.7: Comparative illustration of the (a) Average Pooling and (b) Max Pooling Operations.

derive and implement and has output values that are bounded between 0 and 1. However, the sigmoid function can slow down the training phase due to its non-zero centered characteristic and can also cause the vanishing gradient problem [41].

To address the problems with the sigmoid function, other activation functions have been proposed. The hyperbolic tangent (Tanh) function is zero-centered and gives output values between -1 and 1, but its output values may saturate. The rectified linear function (ReLU) function is often used in CNNs and cancels out all negative values, linearizing all positive values [42, 43], which solves the vanishing gradient problem [44]. Another popular activation function is the Parametric ReLU (PReLU), which allows a small negative slope and is parametrized by $0 \leq \alpha \leq 1$ [45]. The Leaky ReLU (LReLU) function is similar to PReLU but has a fixed $\alpha$ value.

The most common activation functions are illustrated in Fig. 2.6, which includes hyperbolic tangent, sigmoid, rectified linear unit, and parametric rectified linear unit.

### 2.2.5 Pooling Layer

Convolutional layers in a NN use learned filters to create feature maps, which summarize and identify features in an image. By stacking multiple convolutional layers in a deep NN, lower layers can learn low-level characteristics, such as edges and corners, while deeper layers can learn more abstract features, such as shapes and patterns. However, one drawback of convolutional layers is that they may not be able to precisely locate the position of features. This means that small changes in the input's position may result in different feature maps, making it harder for the model to learn spatial relationships between features. This issue is known as the "translation invariance" problem and can be addressed by using additional layers, such as pooling layers or spatial transformer networks, that allow the network to adjust to slight variations in the input's position [31].

To address this issue, a common approach is to down-sample the feature maps using pooling operations. Pooling operations summarize semantically important features into one while preserving essential structural information. This helps the convolutional layers to be less sensitive to small translations of the input, making them approximately translation-invariant. Pooling operations are typically specified using a pre-defined function, such as taking the maximum or average value in a neighborhood, as illustrated in Figure 2 [31, 46]. The process of down-sampling also results in improved performance and reduced memory requirements for storing parameters.

It is important to note that models used for image-to-image translation, such as super-resolution models, perform better when the network learns the down or up-sampling methods, rather than using pooling layers. This finding has been demonstrated in studies such as [22, 5].

### 2.2.6 Sub-pixel Convolution Layer

The sub-pixel convolution layer, also known as PS, is a technique used to improve the results of deep neural networks in image reconstruction. Traditionally, images are upscaled using handcrafted filters, such as bicubic interpolation, which can lead to suboptimal results and high computational overhead. The PS layer, introduced by the authors of [22], improves upon this by learning the filter weights to properly upscale LR images into HR images. This approach effectively replaces fixed handcrafted filters with filters that are specific to each feature map, reducing computational complexity.

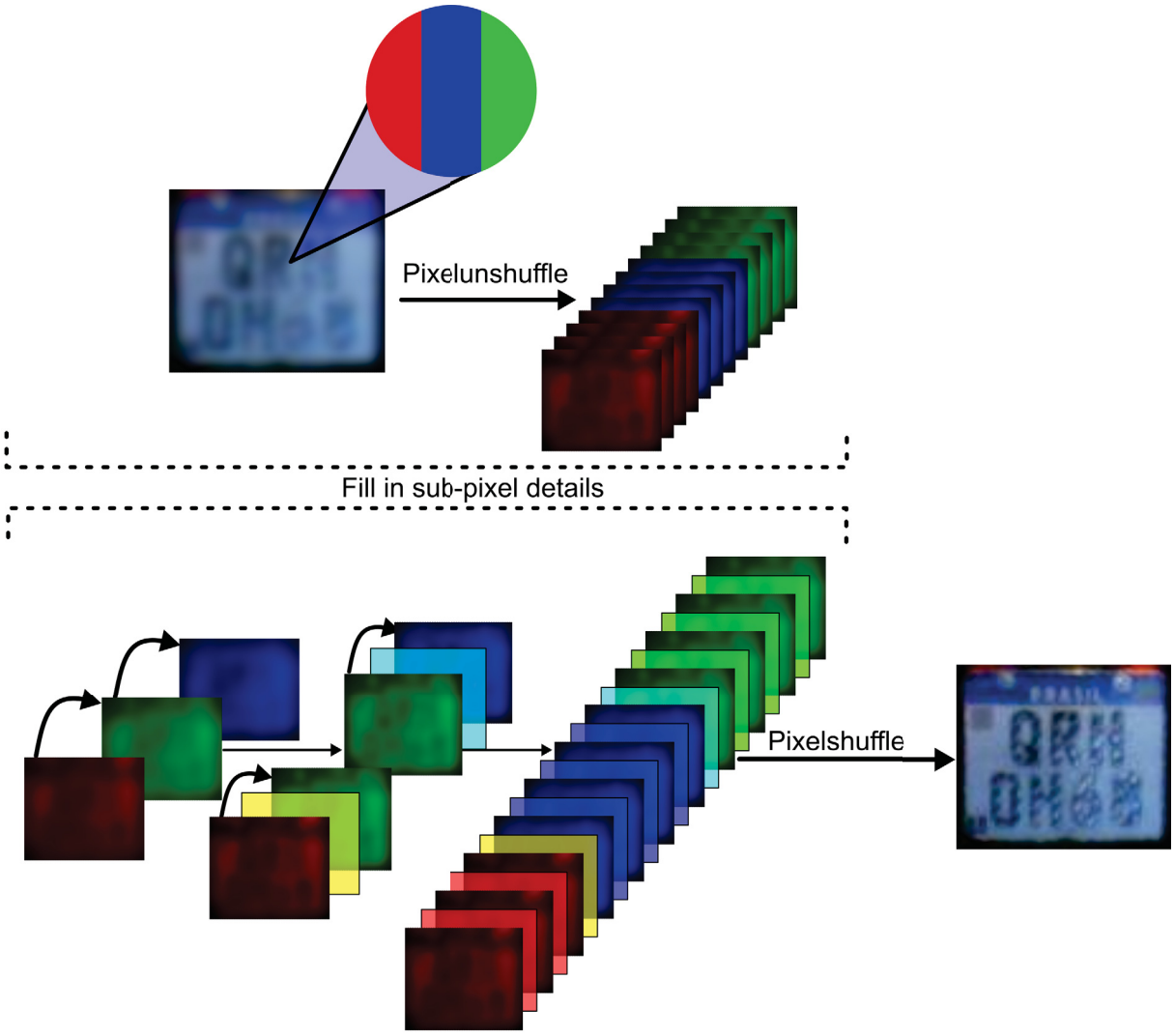



Figure 2.8: PixelShuffle layer aggregates the feature maps from the LR space and builds their SR version.

The PS layer rearranges the elements in an image of shape $(B, H, W, C \cdot r^2)$ to its upscaled version with dimensions of $(B, H \times r, W \times r, C)$, with a scale factor of $r$. By introducing the PS layer after a convolutional operation, which increases the number of channels using a set of filters, it is possible to reorganize the resulting pixels into a higher-resolution image. The PS layer is similar to a normal convolution with fractional stride or transposed convolution (i.e., deconvolution) layers. However, as pointed out by [22], the PS layer can learn filters with complex patterns for different feature maps, resulting in richer and more meaningful representations.

The mathematical operation performed by PS can be described as:

$$PS(T)x, y, c = T\lfloor x/r \rfloor, \lfloor y/r \rfloor, C \cdot r \cdot mod(y, r) + C \cdot mod(x, r) + c \tag{2.9}$$

Here, $x$ and $y$ represent pixel positions, $r$ is the scale factor, and $C$ represents the number of channels. The equation rearranges the elements in an LR feature map $T$ to produce an SR feature map by computing the new positions of each pixel in the output based on its position in the input.

## 2.3 Generative Adversarial Networks (GAN)

GANs were first introduced by [47]. GAN architectures are capable of generating new data that resembles the domain of a training set. GANs consist of two sub-models that work together in an adversarial manner: a Generative model and a Discriminative model. The Generative model is responsible for creating new data that is similar to the training set, while the Discriminative model is responsible for determining if the data generated by the Generative model is real or fake.

During training, both the Generative and Discriminative models engage in a game-like competition. The Generative model aims to improve its ability to create realistic data that can deceive the Discriminative model, while the Discriminative model aims to improve its ability to accurately distinguish between real and fake data. Similar to a game where players compete against each other, when the Generative model succeeds in deceiving the Discriminative model, it is rewarded and no updates are needed for the Generative model's weights. Conversely, if the Discriminative model fails, its weights are updated as a punishment.

As a result of this competition, both models in the GAN architecture improve with each other's help. The Generative model learns to create data that is increasingly similar to real data, while the Discriminative model becomes better at identifying fake data. A diagram of the GAN pipeline is shown in Fig. 2.9, where the Generator produces new data and the Discriminator judges if it is real or fake. If the Discriminator is successful, the Generator's weights are updated, otherwise, if the Discriminator fails, its weights are updated instead.

## 2.4 Multi-Path Residual Network

Adaptive Residual Blocks (ARBs) are a key component of the MPRNet model presented in [6]. ARBs were designed to address the gradient confusion problem that arises in other works such as [48] and [49] and to improve performance in single-image SR tasks.

Unlike traditional residual blocks, ARBs introduce multiple learning paths that are each responsible for extracting different types of information before the aggregation step. This allows the network to access more expressive spatial context information in noisy LR images. The different learning pathways and components of ARBs are detailed in Fig. 2.10.

### 2.4.1 Residual Path

The residual path is a technique that helps to avoid gradient confusion in a neural network. The basic idea is to bring the gradient flow from high-dimensional representations instead of narrow feature spaces between pathways. This approach was first proposed in [50]. By doing this, the network can more easily propagate gradients across multiple layers, which helps

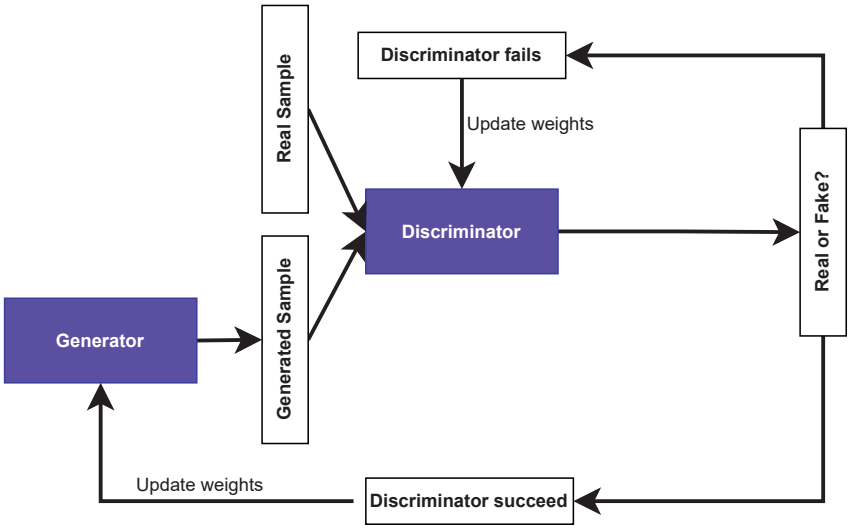



Figure 2.9: Diagram of the GAN pipeline. The generator produces new data, and the discriminator judges whether it is real or fake. If the discriminator successfully identifies fake data, the generator's weights are updated. On the other hand, if the discriminator fails to identify fake data, its weights are updated instead.

with optimization during training. This strategy has been shown to improve the performance of neural networks and make them converge faster.

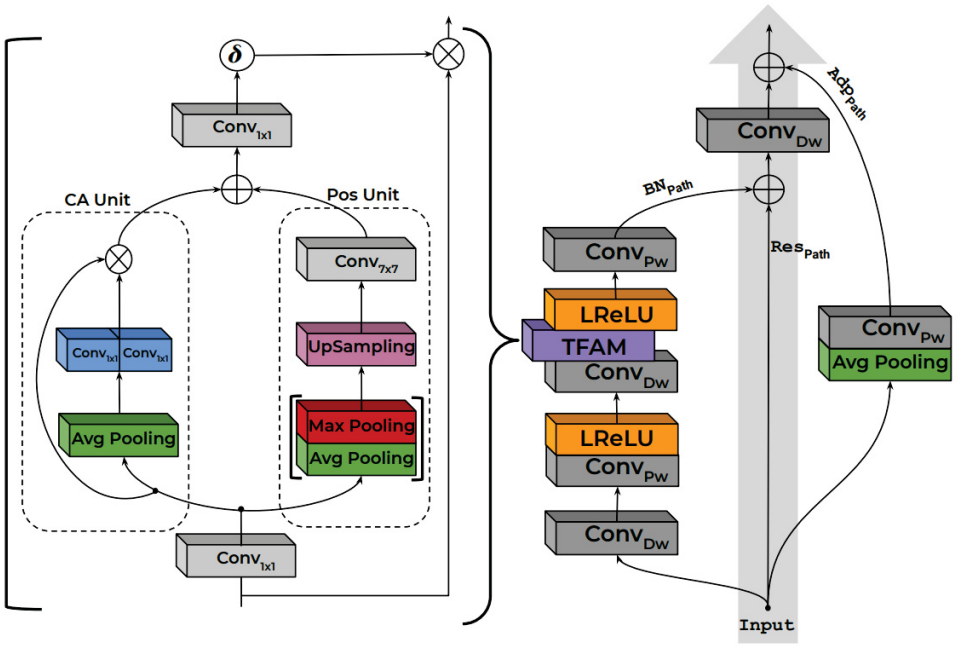



Figure 2.10: The Two-Fold Attention Module and Adaptive Residual Block are shown in the image reproduced from [6].

### Bottleneck Path

The bottleneck path is a technique designed to extract important information for single-image SR tasks while avoiding unnecessary computational costs. It is based on three key insights:

i) High-frequency information and spatial information are key to super-resolution tasks, so the bottleneck path focuses on extracting these types of information. ii) Wide feature maps with irrelevant information can slow down the network, so the bottleneck path is designed to prevent this. iii) To avoid unnecessary computational costs, the bottleneck path uses Dw layers with small kernels to generate meaningful information, and Pw layers to encode inter-channel information and reduce computational costs.

Moreover, to enable the path to function with high-dimensional feature spaces in a low-cost and spatially focused context, the bottleneck path uses Two-Fold Attention Modules (Two-fold Attention Module (TFAM)) which are incorporated into the bottleneck path. These modules assist in the efficient and effective extraction and aggregation of crucial features from high-dimensional feature maps.

#### Adaptive Path

The adaptive path is a technique that aims to extract important information from the LR image by using first-order statistics, such as the mean, retrieved through an average pooling layer. This process eliminates noise present in the LR image and reduces the dimensionality of feature maps while preserving important information for quality enhancement. Afterward, it uses a Pw layer to encode the information across the channels. These steps help the network generate more detailed and sharper super-resolved images. The adaptive path helps extract important information from the LR image while reducing noise, which can improve the quality of the final super-resolved image.

### 2.4.2 Two-Fold Attention Module

The Two-Fold Attention Module (TFAM), as shown in Fig. 2.10, is a crucial component that addresses the problem of allocating available computational resources to the most important features and informative regions within an input image for better reconstruction. The TFAM was proposed in [6] to emphasize the relevant features within the LR input image by focusing on both channel and spatial information simultaneously.

The architecture of the TFAM has two branches, one of which focuses on the features present in the channels, using the Channel Unit (CA) mechanism, and the other on the location of these features, using the Positional Unit (POS). This allows the network to learn "what" and "where" to focus its attention on, the channel and spatial axes respectively, to emphasize relevant features and suppress irrelevant ones.

Experiments conducted in [6] have shown that the TFAM outperforms other state-of-the-art attention mechanisms, such as Squeeze-and-Excitation (SE) [51], Channel Attention (CA) [52], Residual Attention (RA) [53], and Convolutional Block Attention Module (CBAM) [54], in terms of both performance and image reconstruction quality.

#### Channel Unit

The CA unit is a part of the TFAM that focuses on the features present in the channels. It starts with an average pooling operation to extract first-order statistics of the input, such as the mean. This step helps to reduce noise in the input image and extract important information. The CA unit then uses two group-wise convolutional layers. Each layer receives half of the input channels and outputs half of the feature maps. These feature maps are then concatenated to generate the final output. By using group-wise convolutional layers, the CA unit is able to compute a summary of meaningful features while reducing the influence of redundant or useless

information. This approach allows the CA unit to extract important information in a low-cost way.

**Positional Unit**

The POS unit of the TFAM focuses on the location of the features generated by the CA. The POS unit performs an average pooling operation and a max pooling operation, followed by a concatenation operation, to generate an efficient feature descriptor that describes the position of the features.

Next, an up-sampling layer is used to restore the original shape of the feature maps. Then, a convolutional layer aggregates the resulting information. The output of the POS unit is concatenated with the output of the CA unit, and this concatenated output is processed by a $1 \times 1$ convolutional layer, which is activated by a sigmoid function to generate the final mask.

Additionally, a residual connection is used to transfer the input features of the TFAM to its output. This helps to preserve important information from the input and improve the performance of the network. Overall, the POS unit improves the quality of the final super-resolved image by allowing the network to focus on the location of important features.

# Chapter 3

# Related Works

In this chapter, we will provide an overview of recent research in the field of sparse-code and Deep Learning (DL) methods, with a focus on image SR and LPR. In section 3.1, we will discuss approaches for general image SR, such as those used for landscapes and urban scenes. In section 3.2, we will delve into the use of deep learning methods for super-resolution of license plates. Finally, we will conclude this chapter with some final remarks and references to the papers that have been reviewed. A comprehensive list of SR methods for general images can be found in Table 3.1, while Table 3.2 lists the LPR-specific SR methods that have been reviewed in this chapter.

## 3.1 Single-Image Super-Resolution

SISR has seen significant advancements in recent years, making it applicable to various domains [10, 55]. Early SISR methods fell into four categories: prediction, edge-based, image statistical, and example-based [56, 57, 58, 59, 60]. In 2016, Dong et al. [61] introduced Super-Resolution Convolutional Neural Network (SRCNN), a deep learning-based approach that outperformed previous methods in terms of both quality and speed.

Dong et al. [61] proposed one of the first deep learning-based methods, called SRCNN, to tackle the SISR problem. They found that deep CNNs were superior to previous methods, providing better quality reconstruction without the limitations of prior assumptions. This approach is also faster and demonstrates superior restoration capabilities compared to previous example-based methods, with fewer pre- or post-processing steps.

Although SRCNN was successful, some limitations were observed, such as its reliance on pre-upsampling of LR images, which increased computational complexity without providing significant additional information for image restoration [62, 63]. To address these limitations, Dong et al.[64] and Shi et al [22] later incorporated upsampling near the end of the network architecture, significantly reducing execution time, parameters, and computational cost.

Shi et al. [22] emphasized the significance of learnable upscaling and developed specialized convolution layers for learning upscaling filters. This technique allows for more intricate mappings from LR to HR images, leading to improved performance compared to fixed-size interpolation methods.

Attention mechanisms have been introduced in recent super-resolution research to improve image reconstruction. Zhang et al. [65] pioneered the use of first-order statistical attention mechanisms, followed by Dai et al. [66], who presented an improved version using second-order statistics. Huang et al. [67] proposed an attention network that preserves detail fidelity using a divide-and-conquer strategy. Mehri et al. [6] introduced MPRNet, which leverages

information from both inner-channel and spatial features using a TFAM, outperforming multiple state-of-the-art methods such as those presented in [68, 69, 70].

Recently, Zhang et al. [71] proposed the Dual-Coordinate Direction Perception Attention (DPCA) mechanism, a structure- and texture-preserving image super-resolution reconstruction method. This method emphasizes structure and feature details, resulting in improved image quality compared to previous methods.

In summary, SISR methods based on sparse-code and example-based techniques have limitations because they rely on prior assumptions about the data model. In contrast, deep learning-based methods have been shown to be more effective and efficient in image restoration without these limitations. However, these methods also have their own set of challenges, including the use of pre-upsampling of LR images and the need for computational efficiency.

To address these challenges, researchers have proposed various solutions, such as incorporating the upsampling process near the end of the network architecture, using specialized convolution layers for upscaling, and introducing attention mechanisms for better feature extraction. These solutions have shown promising results in improving the quality and efficiency of SISR methods.

## 3.2 Super-Resolution for License Plate Recognition

The primary objective of an LPR system is to extract information from an image or series of images, as reported in literature such as Laroca et al. [72] and Du et al. [73]. LPR systems have a wide range of practical applications in security tasks, such as enforcing traffic laws, monitoring private areas, and criminal investigations [74]. According to Menotti et al. [75], LPR typically involves three main stages: license plate detection, character segmentation, and character recognition. Detection of the LP is the most crucial stage as it sets the foundation for the success of the next stages. However, not all detected LPs result in high-quality images for recognition. Despite recent advancements in LPR, as reported in studies such as Laroca et al. [76], Wang et al. [77], and Silva et al. [78], the datasets used to evaluate these proposed models often comprise only of HR images where all characters on the LP are clearly legible. This does not align with the reality of most surveillance scenarios.

The quality of LP images is closely related to several factors, such as the camera's distance, motion blur, lighting conditions, and image compression techniques used for storage [79]. While commercial LPR systems tend to capture sharp images with the use of global shutter cameras, cheaper cameras that employ rolling shutter technology are often used in surveillance systems, resulting in blurry images [80] with illegible LP characters. In summary, improving LP character quality is a significant challenge because many factors that cause poor image quality are often unknown beforehand in real-world scenarios.

The concept of combining SR and LP recognition has been around since the early 2000s [81, 82, 83], but this area of research has gained more attention in recent years with the advancement of deep learning techniques. One of the earliest works that applied this concept in actual traffic conditions is presented in [81] using a Maximum a Posteriori (MAP) based method. While this approach was innovative, it was found to be computationally demanding and impractical for real-time applications due to its high computational requirements. Tanaka and Okutomi [82] later proposed a faster MAP version for general SR on images. Yuan et al. [83] further reduced the computational cost when applied specifically for LP recognition. However, MAP-based approaches rely on prior information about the desired output, which may not guarantee the best solution in a problem with insufficient information to uniquely determine the desired output, such as SR.

Seibel et al. [84] developed a MISR method that uses projecting and selecting k-nearest neighboring pixels on a HR grid. Despite achieving impressive results, their SR algorithm relies on accurately aligning multiple images and correctly selecting the HR-grid size. This method may not perform well on blurred images, such as those affected by motion blur.

Svoboda et al. [85] demonstrated that CNNs trained on artificially generated blurry images can provide superior quality enhancement for images with motion blur compared to traditional blind deconvolution methods. However, as the model was trained for a specific range of motion blur lengths and directions, the reconstruction quality deteriorates significantly for blurs that fall outside the range for which the network was trained.

Lin et al. [86] used the high capability of a GAN for LP reconstruction and reported promising results. However, their experiments were conducted on only 100 images, which may not be representative of the general performance of the method. Additionally, their approach was only evaluated in terms of PSNR and SSIM without assessing the LPR performance. Despite the positive results, the authors filtered out images with poor brightness and contrast from the Chinese City Parking Dataset (CCPD) dataset [87] as input for testing, and no explicit degradation methods were used.

Kabiraj et al. [88] proposed to use a Enhanced Super-Resolution Generative Adiversarial Network (ESRGAN) paired with an OCR to recover character information in LR and poor quality LPs, building on the promising results from prior work [86]. The experiments suggest that the OCR used achieved superior results with a small training dataset and minimal pre- or post-processing after the ESRGAN stage. The model was trained using a dataset of 1000 high-resolution images with their corresponding low-resolution versions, artificially generated, but evaluated using 182 real-world images.

In the same vein, Hamdi et al. [21] concatenated two GAN models for this task, with the first one used for denoising and deblurring, and the second for super-resolution. The authors compared their method to three baselines, but only in terms of PSNR and SSIM as their evaluation metrics as well. Notably, they acknowledged that higher PSNR and SSIM values do not necessarily indicate a better reconstruction.

Lee et al. [89] proposed a GAN-based super-resolution model that incorporates a perceptual loss based on intermediate features extracted from a scene text recognition model[90]. Their method reportedly achieved better results than the same GAN-based model trained with the original perceptual loss. However, the authors did not make their dataset publicly available, and the degradation method used was not specified.

Maier et al. [91] introduced a Bayesian neural network for LPR that can express uncertainty within a single frame. They also incorporated a reliability score that considers predictive uncertainty, entropy, and prior information, enabling the network to detect and mitigate unreliable predictions. The findings indicate that their approach outperformed traditional softmax statistics. Nevertheless, the authors did not make the datasets used in the experiments publicly available.

Similarly, Moussa et al. [92] proposed a parameter-efficient Transformer model for LPR and evaluated it on real-world data. They showed that Transformers can be effectively used for LPR and highlighted the importance of incorporating compression levels as prior knowledge. Their approach achieved better results than existing LPR methods for low-quality data while requiring fewer parameters and matching the performance on medium and high-quality data.

Despite the primary objective of enhancing LP images to improve recognition accuracy, it is surprising that most previous studies have primarily evaluated the quality of the reconstructed images through subjective visual evaluations or metrics such as PSNR and SSIM. These metrics have limited correlation with human assessment of visual quality [93, 94]. Furthermore, in

most previous studies, private datasets were used in the experiments [85, 89, 21, 91], making it challenging to accurately assess the reported results.

Table 3.1: Papers for bibliographical review on general super-resolution methods:

| **Super-Resolution** | | | |
|---|---|---|---|
| Citation Number | Author Name | Year of Publish | Topic |
| [6] | Mehri et al. | 2021 | MPRNet: Multi-path residual network for lightweight image super-resolution |
| [95] | Dong et al. | 2014 | Learning a deep convolutional network for image super-resolution |
| [22] | Shi et al. | 2016 | Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network |
| [96] | Farsiu et al. | 2004 | Fast and robust multi-frame super resolution |
| [97] | Chang et al. | 2004 | Super-resolution through neighbor embedding |
| [98] | Vanderwalle et al. | 2007 | Super-resolution from unregistered and totally aliased signals using sub-space methods |
| [99] | Yang et al. | 2008 | Image super-resolution as sparse representation of raw image patches |
| [100] | Yang et al. | 2010 | Image super-resolution via sparse representation |
| [101] | Shah et al. | 2012 | Image super-resolution: a survey |
| [61] | Dong et al. | 2016 | Image super-resolution using deep convolutional networks |
| [102] | Chen Y. and Pock T. | 2017 | Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration |
| [61] | Dong et al. | 2016 | Accelerating the super-resolution convolutional neural network |
| [103] | Krizhevsky et al. | 2012 | ImageNet classification with deep convolutional neural networks |
| [59] | Yang et al. | 2013 | Fast image super-resolution based on in-place example regression |
| [65] | Zhang et al. | 2018 | Image super-resolution using very deep residual channel attention networks |

## 3.3 Final Remarks

In summary, many papers addressing LPR with SR to improve recognition accuracy have performed their experiments in controlled noise scenarios and with artificially generated

Table 3.2: Papers for bibliographical review on super-resolution license plate recognition methods

| **Super-Resolution for License Plate Recognition** | | | |
|---|---|---|---|
| [104] | Nasrollani et al. | 2014 | Super-resolution of license plates in real traffic videos |
| [72] | Laroca et al. | 2018 | A robust real-time automatic license plate recognition based on the YOLO detector |
| [73] | Du et al. | 2013 | Automatic license plate recognition: A state-of-the-art review. |
| [74] | Weihong W. and Jiaoyang T. | 2020 | Research on license plate recognition algorithms based on deep learning in complex environment |
| [75] | Menotti et al. | 2014 | Vehicle license plate recognition with random convolutional networks |
| [76] | Laroca et al. | 2021 | An efficient and layout-independent automatic license plate recognition system based on the YOLO detector |
| [77] | Wang et al. | 2021 | Rethinking and designing a high-performing automatic license plate recognition approach |
| [78] | Silva S. and Jung C. | 2022 | A flexible approach for automatic license plate recognition in unconstrained scenarios |
| [80] | Liang et al. | 2008 | Analysis and compensation of rolling shutter effect |
| [82] | Tanaka M. and Okutomi M. | 2006 | fast MAP-based super-resolution algorithm for general motion |
| [83] | Yuan et al. | 2008 | Fast super-resolution for license plate image reconstruction |
| [84] | Seibel et al. | 2017 | Eyes on the target: Super-resolution and license-plate recognition in low-quality surveillance videos |
| [85] | Svodoba et al. | 2016 | CNN for license plate motion deblurring |
| [86] | Lin et al. | 2021 | License plate image reconstruction based on generative adversarial networks |
| [87] | Xu et al. | 2018 | Towards end-to-end license plate detection and recognition: A large dataset and baseline |
| [89] | Lee et al. | 2020 | Super-resolution of license plate images via character-based perceptual loss |
| [90] | Shi et al. | 2019 | An attentional scene text recognizer with flexible rectification. |
| [93] | Johnson et al. | 2016 | Perceptual losses for real-time style transfer and super-resolution |
| [94] | Zhang et al. | 2018 | The unreasonable effectiveness of deep features as a perceptual metric |
| [88] | Kabiraj et al. | 2021 | Number plate recognition from enhanced super-resolution using generative adversarial network |
| [91] | Maier et al. | 2022 | Reliability scoring for the recognition of degraded license plates |
| [92] | Moussa et al. | 2022 | Forensic license plate recognition with compression-informed transformer |

LR images, which do not reflect real-world situations where surveillance cameras are often of poor quality and affected by environmental conditions. Despite the main objective being the improvement of recognition results, most related works only evaluate LP reconstruction qualitatively or quantitatively based on PSNR and SSIM. Additionally, the majority of experiments

were conducted exclusively on private datasets, and no real LR paired with HR images are publicly available to the best of our knowledge.

In light of this, we propose to use an OCR as a fundamental part of the LRLP reconstruction pipeline to create a robust and efficient LPSR network. Additionally, we aim to build a publicly available dataset composed of real-world scenarios paired LR/HR images.

# Chapter 4

# Proposal

In this chapter, we present our super-resolution approach that enhances the extraction of both structural and textural features from low-resolution LP. Our proposed network is an extension of the architecture proposed in our previous work [7], which builds on the MPRNet and TFAM algorithm developed by Mehri et al. [6]. Drawing inspiration from the attention module proposed in [71], we have further improved our network's ability to capture both structural and textural information. We leverage a perceptual loss function that uses an OCR model as a feature extractor to enhance the performance of our network.

## 4.1 Network Architecture Modifications

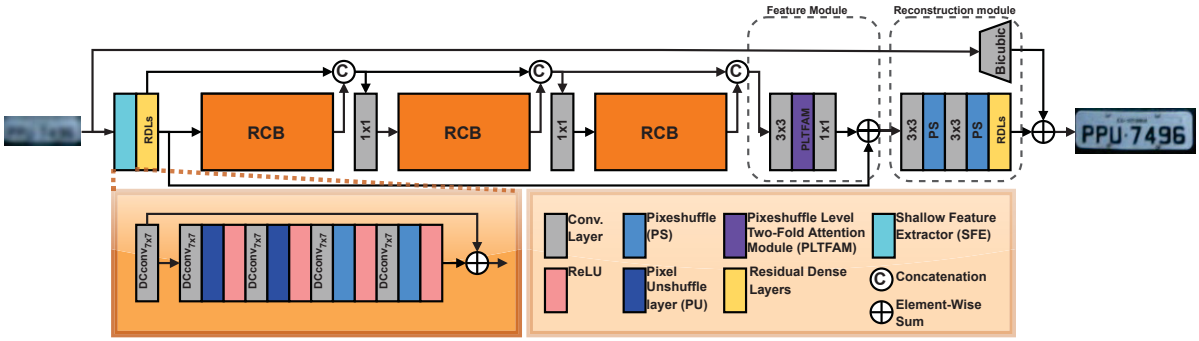



Figure 4.1: The proposed architecture incorporates an autoencoder consisting of PS and PU layers for feature compression and expansion, respectively, with the aim of eliminating less significant features. In addition, the TFAM modules in the original architecture were replaced with PTFAM modules throughout the network. The legend inside the figure provides explanations for the acronyms used.

The proposed approach for super-resolution in LPR features a network architecture that builds upon the work of Mehri et al.[6] and Zhang et al.[71]. As illustrated in Fig. 4.1, the architecture comprises four key components: an Shallow Feature Extractor (SFE); Residual Dense Blocks (RDBs) (refer to [105] for more information); an Feature Module (FM) module; and an Reconstruction Module (RM). The RM combines the output of the FM module with two long-skip connections, one from the end of the SFE module and the other from the input image, to produce the final high-resolution output. Our specific modifications are discussed in the following sections.

**Shallow Feature Extractor**

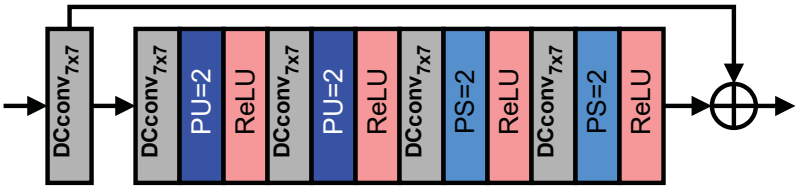



Figure 4.2: Shallow Feature Extractor block: It uses a $7 \times 7$ kernel depth-wise convolutional layer, an autoencoder with PU and PS layers, and depth-wise separable convolutional layers. The resulting mask emphasizes important features for image reconstruction, and a skip connection prevents the loss of information.

The proposed modifications to the network architecture include a modified Shallow Feature Extractor block, shown in Figure 4.2. This block consists of a $7 \times 7$ kernel Depth-wise convolutional layer and an autoencoder that utilizes PU and PS operations instead of conventional pooling and upscaling operations. The autoencoder employs PU layers to downscale the image by a factor of 2 and PS layers to upscale the image by a factor of 2. Following this, Depth-wise separable convolutional layers are applied to reduce the computational burden by decreasing the number of parameters and preventing network overfitting. The process of squeezing and expanding is utilized to highlight the most significant features for image reconstruction and reduce the irrelevant information through squeezing. The result is a mask that emphasizes important features for image reconstruction in a specific application. Additionally, to avoid the loss of information during the autoencoder process, a skip connection is added from the initial convolutional layer to the output, allowing the rest of the network to benefit from a general feature map generated by the first convolutional layer and enhanced by the autoencoder mask.

**PixelShuffle Three-Fold Attention Module**

To obtain super-resolved images that closely resemble the ground truth HR image for LPR, attention mechanisms can effectively allocate computer resources to the most informative and relevant input features for a given application [54, 38, 106, 66, 6]. In our approach, we propose a modified version of the TFAM algorithm in MPRNet[6] that combines an attention module developed by Nascimento et al. [7], called PTFAM (shown in Fig. 4.3). We rely on the following insights to design our approach: **(i)** extracting the relationship between channels is crucial for proper image restoration; **(ii)** the positional information of these features from the channels composing the images is required; **(iii)** traditional downscale and upscale operations rely on translational invariance and interpolation techniques, which are not suitable for learning a customized process for different tasks; and **(iv)** the module captures salient structure from the character fonts of the license plate, highlighting both structural and textural features in the image. The PTFAM is specifically designed to focus on the inter-channel relationship features via the CA unit, pinpoint the position of these features via POS, and enhance the network's ability to retrieve textural and structural information concerned with the character's shape against the LP background via Geometrical Perception Unit (GP).

**Channel Unit**. The purpose of the CA module is to identify and preserve significant inter-channel relationship features while discarding less important ones. To achieve this, the module utilizes two parallel convolutional layers, concatenates their outputs, and processes the concatenated output using a convolutional layer, a PU layer, a PS layer, and a DConv layer. This

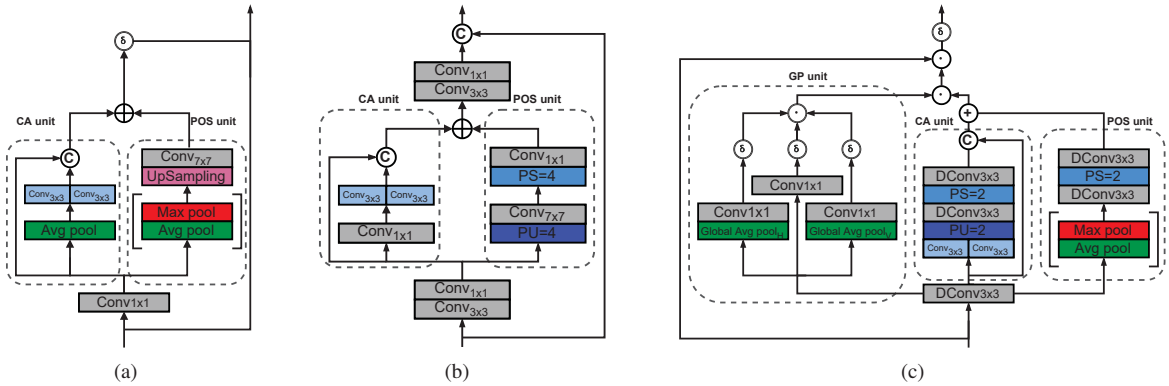



Figure 4.3: Comparison of the (a) Two-Fold Attention Module in MPRNet [6], (b) PixelShuffle Two-Fold Attention Module in Nascimento et al. [7], and (c) PixelShuffle Three-Fold Attention Module (ours).

approach effectively summarizes the inter-channel relationship features, leading to improved image restoration.

**Positional Unit**. The purpose of the POS module is to enhance the CA module by identifying the important features' locations in the image. It achieves this by using average and max pooling operations to extract first-order statistics of the image, concatenating the outputs, and then processing them through DConvs and PS layers to restore the original feature map dimension. By highlighting the positions of the relevant inter-channel relationship features, the POS module further improves the quality of image restoration.

**Geometrical Perception Unit**. We incorporated a third branch called GP to enhance the network's ability to extract critical characteristics, such as structural, textural, and geometric features from the LP, which was motivated by the work of [71]. This module utilizes global average pooling in both the vertical and horizontal directions of the input image. The output from this layer is then subjected to a point-wise convolutional layer followed by the sigmoid function to ensure the right channel dimensions. Finally, the results from this layer are aggregated through an element-wise multiplication to obtain the final output.

Finally, the outputs from the CA, POS and GP units are combined through an element-wise sum and multiplication to generate the final attention mask, which is then used to enhance the input to the PTFAM module through a DConv layer and a sigmoid function. This process effectively emphasizes the key features of the image, including the inter-channel relationships, positional information, and structural information, resulting in improved image restoration.

### Overall Network Architectural Modifications

We modified the original ARBs to improve the network's performance, as illustrated in Fig. 4.4. Specifically, we replaced the TFAM with our proposed PTFAM module and substituted the traditional convolution layers in the bottleneck path with dilated convolution layers. This modification enables the network to consider a broader context by increasing the receptive field without introducing extra parameters. Nonetheless, the overall network structure remained similar to the one described in [6].

Returning to Fig. 4.1, we have incorporated Residual Dense Layers (RDLs) based on the RDB introduced in Zhang et al. [106] to enhance the network's representational ability. These layers leverage both local and global feature fusion to generate texture patterns likely to be learned from the training data. We added RDLs after the autoencoder to improve the aggregation

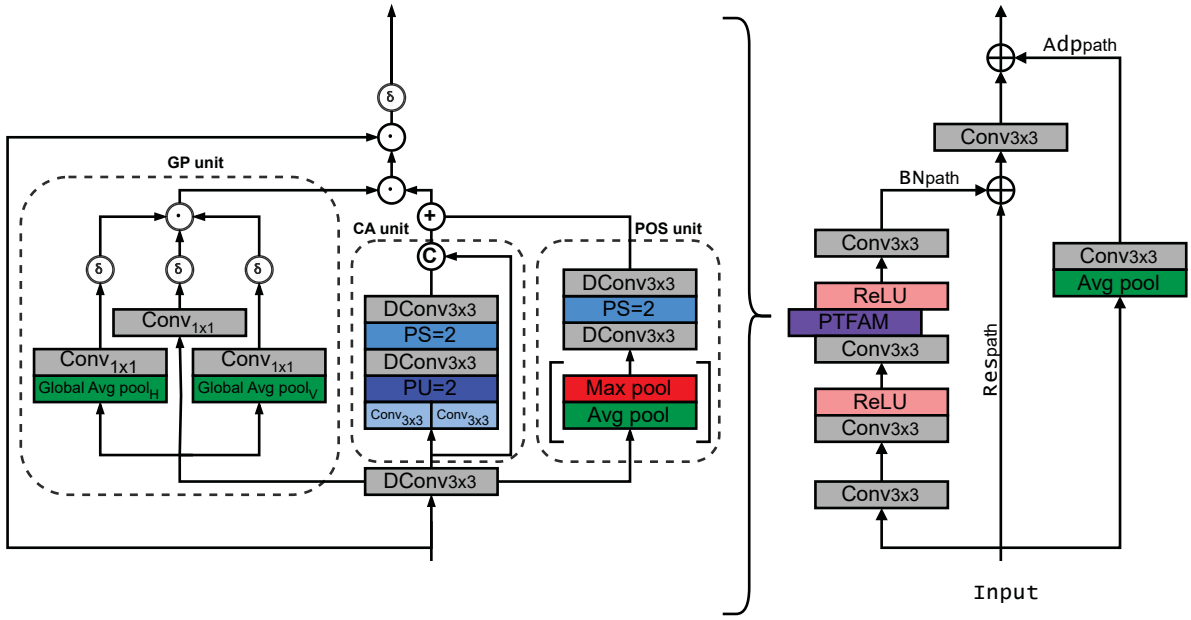



Figure 4.4: Proposed Adaptive Residual Block with Dilated Convolutions along the Bottleneck Path ($BN_{path}$) and Pixel-Level Three-Fold Attention Module.

of shallow features and in the reconstruction module to enhance output quality for tasks such as compression artifact reduction, image deblurring, and Gaussian denoising. Additionally, to extract hierarchical features and improve the image based on the original LR image, we added a global skip connection from the input for global residual learning, as proposed by Zhang et al. [106].

## 4.2 Perceptual Loss

We propose enhancing the accuracy of our super-resolution approach by incorporating a perceptual loss function that considers the features expected by an OCR model. The function, presented in Eq. (4.1), aims to improve the accuracy of the system. $H_i$ and $S_i$ represent the high-resolution and super-resolved LP images, respectively, and $f_{OCR}(\cdot)$ represents the feature extraction process performed by the OCR model. In Eq. (4.1), we calculate the mean of the squared differences between $H_i$ and $S_i$, as well as the absolute differences between the feature representations of $H_i$ and $S_i$ obtained from $f_{OCR}(\cdot)$. These two terms are combined and normalized by the number of images $n$ in the dataset.

$$PL = \frac{1}{n}\Big(\sum_{i=1}^{n}(H_i - S_i)^2 + \sum_{i=1}^{n}|f_{OCR}(H_i) - f_{OCR}(S_i)|\Big) \tag{4.1}$$

It is worth noting that the loss function allows the use of any OCR model for LPR, which provides flexibility and the ability to incorporate novel models as they become available. In this work, we explore the multi-task model proposed by Gonçalves et al.[23], which is efficient and has achieved remarkable outcomes in prior research [79, 7].

In addition, to enhance the overall quality of the image, we use the MSE to compute the difference between the expected and generated pixel values, penalizing significant errors more than minor errors. The MSE is effective in preserving the structural information in the image, which is essential in the super-resolution task. On the other hand, the L1 loss ensures robustness

to noise and outliers, and helps to preserve sharp edges in the generated images by considering the expected features. This is particularly important in the early stages of training when there may be a significant discrepancy between the expected and actual features produced by the network. Combining MSE and L1 loss provides a more comprehensive evaluation of the generated images, achieving a balance between preserving structural information and minimizing errors.

# Chapter 5

# Experiments

In this section, we detail the steps taken to validate the effectiveness of our proposed method for LP super-resolution. We first describe our experimental setup and then proceed to provide a comprehensive analysis of the results obtained.

## 5.1 Setup

In our experiments, we used LP images obtained from the RodoSol-ALPR [8] and PKU [9] datasets. To the best of our knowledge, there is currently no public dataset that provides paired LR and HR images from real-world settings. Hence, we chose these two datasets because they provide a wide range of scenarios in which the images were acquired.

RodoSol-ALPR is the largest public dataset acquired in Brazil. It comprises 20,000 images, with 10,000 showing vehicles with Brazilian LPs and 10,000 featuring vehicles with Mercosur LPs[1]. As shown in Fig. 5.1, the diversity of this dataset with respect to several factors such as LP colors, lighting conditions, and character fonts is significant. In this work, we follow the standard protocol (defined in [8]) that involves using 40% of the images for training, 20% for validation, and 40% for testing.


Figure 5.1: Some LP images from the RodoSol-ALPR dataset [8]. The first two rows show Brazilian LPs, while the last two rows show Mercosur LPs.

[1]In accordance with prior literature [107, 108, 78], we use the term "Brazilian" to refer to the layout used in Brazil prior to the adoption of the Mercosur layout.

The PKU dataset comprises images categorized into five distinct groups, namely G1 through G5, each representing a specific scenario. For instance, the images in G1 were captured on highways during the day and depict a single vehicle. On the other hand, the images in G5 were taken at crosswalk intersections, either during the day or night and have multiple vehicles. All images were collected in mainland China. We perform experiments using the 2,253 images in groups G1-G3, as they have labels regarding the LP text (these annotations were provided in [109]). Despite the diverse settings, the LP images have good quality and are perfectly legible (see some examples in Fig. 5.2). Following [109, 108], we use 60% of the images for training/validation, while the remaining 40% are used for testing. Laroca et al. [110] recently revealed that the PKU dataset (as well as other datasets) has multiple images of the same vehicle/LP. They referred to such images as *near-duplicates*. Accordingly, to prevent bias in our experiments, we ensured that all images showing the same LP were grouped in the same subset.


Figure 5.2: Examples of LP images from the PKU dataset [9]. Although the LPs in this dataset have varying layouts, they all have seven characters.

The HR images used in our experiments were generated as follows. For each image from the chosen datasets, we first cropped the LP region using the annotations provided by the authors. Afterward, we used the same annotations to rectify each LP image so that it becomes more horizontal, tightly bounded, and easier to recognize. The rectified image is the HR image.

Inspired by [106], we generated LR versions of each HR image by simulating the effects of an optical system with lower resolution. This was achieved by iteratively applying random Gaussian noise to each HR image until we reached the desired degradation level for a given LR image (i.e., SSIM $< 0.1$). To maintain the aspect ratio of the LR and HR images, we performed a padding prior to resizing them to $20 \times 40$ pixels, resulting in an output shape of $80 \times 160$ pixels for a magnification factor of 4. Fig. 5.3 and Fig. 5.4 show examples of the LP images generated for the RodoSol-ALPR and PKU datasets, respectively.


Figure 5.3: Some HR-LR image pairs created from the RodoSol-ALPR dataset.

Figure 5.4: Examples of HR-LR image pairs created from the PKU dataset.

Our experiments were conducted using the PyTorch framework on a high-performance computer that is equipped with an AMD Ryzen 9 5950X CPU, 128 GB of RAM, and an NVIDIA Quadro RTX 8000 GPU (48 GB).

We used the Adam optimizer with a learning rate of $10^{-4}$, which decreases by a factor of 0.3 (up to $10^{-7}$) when no improvement in the loss function is observed. The training process stops after 20 epochs without a decrease in the loss function.

## 5.2 Experimental Results

In the field of LPR, models are typically evaluated based on the ratio of correctly recognized LPs to the total number of LPs in the test set [76, 111, 78]. A LP is considered correctly recognized only if all characters on it are recognized accurately. As our focus is on low-resolution LPs that are commonly used in forensic applications, we also report partial match results where at least 5 or 6 out of the 7 characters are recognized correctly. These partial matches may be useful in narrowing down the list of potential LPs by incorporating additional information such as the make and model of the vehicle.

The results of the LPR experiment are shown in Table 5.1. The table shows the recognition accuracy of HR and LR license plate images degraded by bicubic downsampling and recursive Gaussian noise. The difficulty of the task can be seen from the SSIM score, which ranges from 0 to 0.1, as illustrated in Fig. 5.3, where the LP characters are barely distinguishable.

The proposed super-resolution network achieved superior performance compared to the two baseline models, as presented in the second section of Table 5.1. The multi-task OCR model [23] demonstrated remarkable improvement when applied to images reconstructed by our super-resolution approach in both datasets, particularly in the PKU dataset, with a 14.8% higher recognition rate compared to the method proposed in our preliminary work [7] and a 26.7% higher accuracy compared to MPRNet [6] for LPs with more than five correct characters.

For completeness, we detail in Table 5.1 the PSNR and SSIM obtained by each approach. Similar to what was observed in [94, 21, 86], the PSNR metric seems inappropriate for this particular application, as our approach and the one proposed in [7] reached comparable values, despite ours leading to significantly better results achieved by the OCR model. The SSIM metric, on the other hand, seems to better represent the quality of reconstruction of LP images, as the proposed method achieved considerably better SSIM values in both datasets.

The variation in accuracy between the two datasets can be attributed to the diversity present in the RodoSol-ALPR dataset, which includes a range of layouts, lighting conditions, and character fonts, while the *PKU* dataset largely comprises LPs with a uniform layout and less variation in the environmental conditions under which the images were collected.

The OCR network demonstrated improved results due to the effective extraction of textural and structural information by the proposed GP unit, in addition to the CA and POS units. These units were designed using pyramid and PixelShuffle layers to optimize channel scaling and reorganization within the image.

Table 5.1: Recognition rates (%) achieved in our experiments. "All" refers to LPs where all characters were recognized correctly; ≥ 6 and ≥ 5 refer to LPs where at least 6 or 5 characters were recognized correctly, respectively.

| | RodoSol-ALPR | | | PKU | | |
|---|---|---|---|---|---|---|
| | All | ≥ 6 | ≥ 5 | All | ≥ 6 | ≥ 5 |
| OCR [23] – no super-resolution | | | | | | |
| HR | 96.6 | 98.6 | 99.0 | 99.4 | 99.9 | 99.9 |
| LR | 0.8 | 4.6 | 12.7 | 0.0 | 0.0 | 0.0 |
| OCR [23] – with super-resolution | | | | | | |
| **Proposed** | **39.0** | **59.9** | **74.2** | **72.0** | **90.3** | **97.3** |
| Nascimento et al. [7] | 10.5 | 25.4 | 42.2 | 35.5 | 65.3 | 82.5 |
| Mehri et al. [6] | 1.45 | 7.0 | 17.4 | 22.5 | 49.2 | 70.6 |
| Average PSNR (dB) and SSIM | | | | | | |
| | PSNR | SSIM | | PSNR | SSIM | |
| **Proposed** | **21.2** | **0.59** | | **18.3** | **0.61** | |
| Nascimento et al. [7] | 21.3 | 0.52 | | 18.1 | 0.54 | |
| Mehri et al. [6] | 16.8 | 0.38 | | 16.4 | 0.41 | |

Finally, we can further confirm the results of the LPR experiments by visually comparing the super-resolution images produced by our technique with those generated by the baseline methods [6, 7]. Fig. 5.5 and Fig. 5.6 depict four LR images along with their corresponding super-resolution counterparts, and the original HR image is included as a reference. From the images, it is evident that our proposed approach outperforms both its preliminary version [7] and MPRNet[6] in terms of perceptual quality.


Figure 5.5: Representative examples of the images generated by the proposed approach and baselines in the RodoSol-ALPR dataset [8].

In general, the images produced by MPRNet [6] exhibit a common issue of blurriness, where the edges of the characters blend into the LP background, resulting in visible artifacts. This

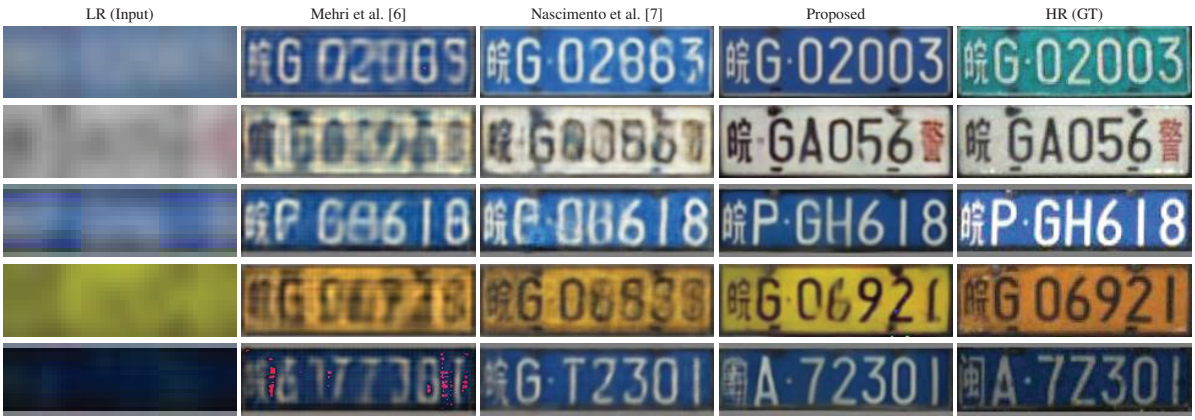



Figure 5.6: Representative examples of the images generated by the proposed approach and baselines in the PKU dataset [9].

blurriness can also cause the edges of multiple characters to blend together, leading to further visual distortions. The architecture proposed in our previous work[7] manages to reconstruct the characters but distorts them with strong undulations, making them appear as part of the LP background in some cases (see the first row of Fig. 5.5). Conversely, the proposed model generates clear character edges and consistently reconstructs the original font, without any missing characters or incomplete lines.

It is notable that when our model is uncertain about which character to reconstruct, it tends to hallucinate characters that are more congruent with the LR input, as seen in the last row of Fig. 5.5 and Fig. 5.6, where the character "3" is reconstructed as "J" and the character "Z" is reconstructed as "2," respectively. We believe that this issue could be mitigated by incorporating a lexicon or vocabulary into the network's learning process to identify the character types (letter, digit, or either) that can appear at each position on LPRs with specific layouts. Additionally, the network tends to generate nearly identical background colors for different images, as can be observed in the third row of Fig. 5.5 and the first row of Fig. 5.6. However, it is noteworthy that, based on our analysis, this does not significantly impact the recognition results achieved.

## 5.3 Ablation Study

As the proposed approach integrates multiple concepts into a single architecture, we conducted an ablation study to validate the contribution of each incorporated unit to the obtained results. The study involved removing one module at a time, such as the autoencoder, TFAM, PS, and PU layers, and training the network without the perceptual loss.

Four baselines were established for the experiments. The first baseline replaced the autoencoder with a DConv layer with a $5 \times 5$ kernel for shallow feature extraction as shown in [6]. The second baseline removed the PTFAM module and adjusted the output of the previous layer to match the input shape of the following layers. The third baseline replaced the PS and PU layers with transposed and strided convolution layers, respectively, as they are analogous [22]. Finally, in the fourth baseline, the perceptual loss was replaced by MSE, which is commonly used in the super-resolution field [13, 55]. Table 5.2 presents the results.

The results of the experiments on the RodoSol-ALPR dataset demonstrate that each of the units included in the proposed system significantly contributes to its overall performance. The complete system attained a recognition rate of 39.0%, while the best version without one of the components reached a recognition rate of 35.6%. The worst-case scenario was when

Table 5.2: Recognition rates (%) achieved in the ablation study. "All" refers to LPs where all characters were recognized correctly; ≥ 6 and ≥ 5 refer to LPs where at least 6 or 5 characters were recognized correctly, respectively.

| Approach | RodoSol-ALPR | | | PKU | | |
|---|---|---|---|---|---|---|
| | All | ≥ 6 | ≥ 5 | All | ≥ 6 | ≥ 5 |
| Proposed (w/o autoencoder) | 32.7 | 55.0 | 70.1 | **73.8** | 90.2 | 96.6 |
| Proposed (w/o TFAM) | 33.3 | 55.0 | 69.6 | 73.1 | 90.1 | 96.6 |
| Proposed (w/o PS and PU layers) | 34.3 | 54.8 | 68.5 | 70.4 | 89.9 | 96.7 |
| Proposed (w/o perceptual loss) | 35.6 | 57.3 | 71.9 | 72.4 | **91.4** | 97.1 |
| Proposed | **39.0** | **59.9** | **74.2** | 72.0 | 90.3 | **97.3** |

the autoencoder unit was removed, resulting in a recognition rate of 32.7% for all recognized characters. This is because the autoencoder module plays a vital role in facilitating the extraction of shallow features. Specifically, the autoencoder generates a mask by squeezing and expanding the input image, highlighting the most critical areas for reconstruction by the rest of the network. Without this mask, the network struggles to identify the relevant features, resulting in poor performance.

In contrast, the recognition rates in the PKU dataset were only enhanced with the incorporation of PS and PU layers. We conjecture that the other units are not required for this dataset due to its images being considerably less complex than those in the RodoSol-ALPR (as evidenced by the images in Fig. 5.1 and Fig. 5.2). This could explain why several authors opted to conduct ablation studies solely on the largest and most diverse dataset among those used in their experiments [109, 112, 111].

# Chapter 6

# Conclusions

This work proposes a new super-resolution approach to improve the recognition of low-resolution LPs. Our method adds to the existing MPRNet [6] and the architecture proposed in our previous work [7] by incorporating subpixel-convolution layers (PS and PU) in combination with a PTFAM. Moreover, we introduce a novel perceptual loss that combines features extracted from an OCR model with L1 loss to reconstruct characters with the most relevant characteristics, while also incorporating MSE to enhance overall image quality.

Our approach capitalizes on both structural and textural features by using the PS and PU layers for custom scale operations, rather than relying on conventional translational invariance and interpolation techniques. An autoencoder with PS and PU layers was integrated to extract shallow features and generate an attention mask that is added to the original input. The output of the autoencoder is processed by a RDB to identify regions of interest for reconstruction, optimizing computational resources, and producing super-resolution images that emphasize relevant information.

We evaluated the proposed method on two publicly available datasets containing a diverse range of LP images from Brazil and mainland China. The experimental results demonstrate that our method outperformed the baselines in terms of recognition rates. Specifically, on the RodoSol-ALPR dataset, our method achieved a recognition rate of 39.0% for the OCR model, compared to 31.3% and 4.0% for the methods proposed in [7] and [6], respectively. On the PKU dataset, our approach achieved a recognition rate of 72.0% for the OCR model, compared to 35.5% and 22.5% for [7] and [6], respectively. We have also made the LR-HR image pairs used in our experiments and the source code publicly available to encourage further research and development in the field of LPR super-resolution.

In the future, we plan to integrate a lexicon or vocabulary into the network's learning process, which would allow the network to learn the character types that can occupy each position on specific LPs layouts. Additionally, we intend to create a large-scale dataset for LP super-resolution, consisting of thousands of LR and HR image pairs. Our aim is to collect videos in which the LP is legible in one frame but not in another, enabling us to evaluate existing methods in real-world scenarios and develop novel methods.

# Bibliography


[1] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[2] Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009.

[3] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016.

[4] Delwar Hossain, Masudul Haider Imtiaz, Tonmoy Ghosh, Viprav Bhaskar, and Edward Sazonov. Real-time food intake monitoring using wearable egocnetric camera. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4191–4195. IEEE, 2020.

[5] Moacir Antonelli Ponti, Leonardo Sampaio Ferraz Ribeiro, Tiago Santana Nazare, Tu Bui, and John Collomosse. Everything you wanted to know about deep learning for computer vision but were afraid to ask. In *2017 30th SIBGRAPI conference on graphics, patterns and images tutorials (SIBGRAPI-T)*, pages 17–41. IEEE, 2017.

[6] Armin Mehri, Parichehr B. Ardakani, and Angel D. Sappa. MPRNet: Multi-path residual network for lightweight image super resolution. In *IEEE Winter Conference on Applications of Computer Vision*, pages 2703–2712, 2021.

[7] V. Nascimento, R. Laroca, J. A. Lambert, W. R. Schwartz, and D. Menotti. Combining attention module and pixel shuffle for license plate super-resolution. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 228–233, Oct 2022.

[8] R. Laroca, E. V. Cardoso, D. R. Lucio, V. Estevam, and D. Menotti. On the cross-dataset generalization in license plate recognition. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 166–178, Feb 2022.

[9] Y. Yuan, W. Zou, Y. Zhao, X. Wang, X. Hu, and N. Komodakis. A robust and efficient approach to license plate detection. *IEEE Transactions on Image Processing*, 26(3):1102–1114, March 2017.

[10] Linwei Yue, Huanfeng Shen, Jie Li, Qiangqiang Yuan, Hongyan Zhang, and Liangpei Zhang. Image super-resolution: The techniques, applications, and future. *Signal Processing*, 128:389–408, 2016.

[11] Anran Liu, Yihao Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Blind image super-resolution: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[12] M. Santos, R. Laroca, R. O. Ribeiro, J. Neves, H. Proença, and D. Menotti. Face super-resolution using stochastic differential equations. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 216–221, Oct 2022.

[13] Zhihao Wang, Jian Chen, and Steven C. H. Hoi. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3365–3387, 2021.

[14] Gabriele Guarnieri et al. Perspective registration and multi-frame super-resolution of license plates in surveillance videos. *Forensic Science International: Digital Investigation*, 36:301087, 2021.

[15] Anran Liu, Yihao Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Blind image super-resolution: A survey and beyond. *arXiv preprint*, arXiv:2107.03055:1–20, 2021.

[16] Gabriele Guarnieri, Marco Fontani, Francesco Guzzi, Sergio Carrato, and Martino Jerian. Perspective registration and multi-frame super-resolution of license plates in surveillance videos. *Forensic Science International: Digital Investigation*, 36:301087, 2021.

[17] Alice Lucas et al. Generative adversarial networks and perceptual losses for video super-resolution. *IEEE Transactions on Image Processing*, 28(7):3312–3327, 2019.

[18] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4482–4490, 2017.

[19] Bee Lim and Kyoung Mu Lee. Deep recurrent resnet for video super-resolution. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1452–1455, 2017.

[20] Renjie Liao, Xin Tao, Ruiyu Li, Ziyang Ma, and Jiaya Jia. Video super-resolution via deep draft-ensemble learning. In *IEEE International Conference on Computer Vision (ICCV)*, pages 531–539, 2015.

[21] Abdelsalam Hamdi, Yee Kit Chan, and Voon Chet Koo. A new image enhancement and super resolution technique for license plate recognition. *Heliyon*, 7(11):e08341, 2021.

[22] Wenzhe Shi et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016.

[23] G. R. Gonçalves, M. A. Diniz, R. Laroca, D. Menotti, and W. R. Schwartz. Real-time automatic license plate recognition through deep multi-task networks. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 110–117, Oct 2018.

[24] Bernd Girod. What's wrong with mean-squared error? *Digital images and human vision*, pages 207–220, 1993.

[25] Ahmet M Eskicioglu and Paul S Fisher. Image quality measures and their performance. *IEEE Transactions on communications*, 43(12):2959–2965, 1995.

[26] Stefan Winkler. Perceptual distortion metric for digital color video. In *Human Vision and Electronic Imaging IV*, volume 3644, pages 175–184. SPIE, 1999.

[27] Zhou Wang, Alan C Bovik, and Ligang Lu. Why is image quality assessment so difficult? In *2002 IEEE International conference on acoustics, speech, and signal processing*, volume 4, pages IV–3313. IEEE, 2002.

[28] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

[29] Quan Huynh-Thu and Mohammed Ghanbari. The accuracy of psnr in predicting video quality for different video scenes and frame rates. *Telecommunication Systems*, 49(1):35–48, 2012.

[30] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006.

[31] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[32] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. In *2010 IEEE Computer Society Conference on computer vision and pattern recognition*, pages 2528–2535. IEEE, 2010.

[33] Matthew D Zeiler, Graham W Taylor, and Rob Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *2011 international conference on computer vision*, pages 2018–2025. IEEE, 2011.

[34] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

[35] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[36] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[37] Jon Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class project for Stanford CS231N: convolutional neural networks for visual recognition, Winter semester*, 2014(5):2, 2014.

[38] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.

[39] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[40] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[41] Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and trends® in signal processing*, 7(3–4):197–387, 2014.

[42] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[43] Moacir Ponti, Elias S Helou, Paulo Jorge SG Ferreira, and Nelson DA Mascarenhas. Image restoration using gradient iteration and constraints for band extrapolation. *IEEE Journal of Selected Topics in Signal Processing*, 10(1):71–80, 2015.

[44] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[46] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[47] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[48] Daquan Zhou, Qibin Hou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. Rethinking bottleneck structure for efficient mobile network design. In *European Conference on Computer Vision*, pages 680–697. Springer, 2020.

[49] Duo Li, Aojun Zhou, and Anbang Yao. Hbonet: Harmonious bottleneck on two orthogonal dimensions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3316–3325, 2019.

[50] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[51] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):2011–2023, 2020.

[52] Yanting Hu, Jie Li, Yuanfei Huang, and Xinbo Gao. Channel-wise and spatial feature modulation network for single image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11):3911–3927, 2019.

[53] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2356–2365, 2020.

[54] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[55] Anran Liu, Yihao Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Blind image super-resolution: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5461–5480, 2023.

[56] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *2009 IEEE 12th international conference on computer vision*, pages 349–356. IEEE, 2009.

[57] Kwang In Kim and Younghee Kwon. Single-image super-resolution using sparse regression and natural image prior. *IEEE transactions on pattern analysis and machine intelligence*, 32(6):1127–1133, 2010.

[58] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 1920–1927, 2013.

[59] Jianchao Yang, Zhe Lin, and Scott Cohen. Fast image super-resolution based on in-place example regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1059–1066, June 2013.

[60] Chih-Yuan Yang, Chao Ma, and Ming-Hsuan Yang. Single-image super-resolution: A benchmark. In *European conference on computer vision*, pages 372–386. Springer, 2014.

[61] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016.

[62] Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang. Deep networks for image super-resolution with sparse prior. *arXiv preprint*, arXiv:1507.08905:1–10, 2015.

[63] Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE Trans. on Pattern Analysis and Machine Intel.*, 39:1256–1272, 2017.

[64] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *European Conf. on Computer Vision (ECCV)*, pages 391–407, 2016.

[65] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision (ECCV)*, pages 294–310, 2018.

[66] Tao Dai et al. Second-order attention network for single image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11057–11066, 2019.

[67] Yuanfei Huang, Jie Li, Xinbo Gao, Yanting Hu, and Wen Lu. Interpretable detail-fidelity attention network for single image super-resolution. *IEEE Transactions on Image Processing*, 30:2325–2339, 2021.

[68] Xiangyu He, Zitao Mo, Peisong Wang, Yang Liu, Mingyuan Yang, and Jian Cheng. Ode-inspired network design for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1732–1741, 2019.

[69] Xiaotong Luo, Yuan Xie, Yulun Zhang, Yanyun Qu, Cuihua Li, and Yun Fu. LatticeNet: Towards lightweight image super-resolution with lattice block. In *European Conference on Computer Vision (ECCV)*, pages 272–289, 2020.

[70] Abdul Muqeet, Jiwon Hwang, Subin Yang, JungHeum Kang, Yongwoo Kim, and Sung-Ho Bae. Multi-attention based ultra lightweight image super-resolution. In *European Conference on Computer Vision Workshops (ECCVW)*, pages 103–118, 2020.

[71] Yafei Zhang, Yuqing Huang, Kaizheng Wang, Guanqiu Qi, and Jinting Zhu. Single image super-resolution reconstruction with preservation of structure and texture details. *Mathematics*, 11:216, 01 2023.

[72] R. Laroca, E. Severo, L. A. Zanlorensi, L. S. Oliveira, G. R. Gonçalves, W. R. Schwartz, and D. Menotti. A robust real-time automatic license plate recognition based on the YOLO detector. In *International Joint Conference on Neural Networks*, pages 1–10, July 2018.

[73] Shan Du, Mahmoud Ibrahim, Mohamed Shehata, and Wael Badawy. Automatic license plate recognition (ALPR): A state-of-the-art review. *IEEE Trans. on Circuits and Systems for Video Technology*, 23:311–325, 2013.

[74] W. Weihong and T. Jiaoyang. Research on license plate recognition algorithms based on deep learning in complex environment. *IEEE Access*, 8:91661–91675, 2020.

[75] D. Menotti, G. Chiachia, A. X. Falcão, and V. J. O. Neto. Vehicle license plate recognition with random convolutional networks. In *Conf. on Graphics, Patterns and Images (SIBGRAPI)*, pages 298–303, 2014.

[76] R. Laroca, L. A. Zanlorensi, G. R. Gonçalves, E. Todt, W. R. Schwartz, and D. Menotti. An efficient and layout-independent automatic license plate recognition system based on the YOLO detector. *IET Intelligent Transport Systems*, 15(4):483–503, 2021.

[77] Yi Wang, Zhen-Peng Bian, Yunhao Zhou, and Lap-Pui Chau. Rethinking and designing a high-performing automatic license plate recognition approach. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–13, 2021.

[78] Sergio M. Silva and Cláudio Rosito Jung. A flexible approach for automatic license plate recognition in unconstrained scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):5693–5703, 2022.

[79] G. R. Gonçalves, M. A. Diniz, R. Laroca, D. Menotti, and W. R. Schwartz. Multi-task learning for low-resolution license plate recognition. In *Iberoamerican Congress on Pattern Recognition (CIARP)*, pages 251–261, Oct 2019.

[80] Chia-Kai Liang, Li-Wen Chang, and Homer H. Chen. Analysis and compensation of rolling shutter effect. *IEEE Transactions on Image Processing*, 17(8):1323–1330, 2008.

[81] K. V. Suresh, G. Mahesh Kumar, and A. N. Rajagopalan. Superresolution of license plates in real traffic videos. *IEEE Transactions on Intelligent Transportation Systems*, 8(2):321–331, 2007.

[82] Masayuki Tanaka and Masatoshi Okutomi. A fast MAP-based super-resolution algorithm for general motion. In *Computational Imaging IV*, volume 6065, pages 404–415, 2006.

[83] Jie Yuan, Si-Dan Du, and Xiang Zhu. Fast super-resolution for license plate image reconstruction. In *International Conference on Pattern Recognition (ICPR)*, pages 1–4, 2008.

[84] H. Seibel, S. Goldenstein, and A. Rocha. Eyes on the target: Super-resolution and license-plate recognition in low-quality surveillance videos. *IEEE Access*, 5:20020–20035, 2017.

[85] P. Svoboda, M. Hradiš, L. Maršík, and P. Zemcík. CNN for license plate motion deblurring. In *IEEE International Conference on Image Processing (ICIP)*, pages 3832–3836, Sept 2016.

[86] Mianfen Lin, Liangxin Liu, Fei Wang, Jingcong Li, and Jiahui Pan. License plate image reconstruction based on generative adversarial networks. *Remote Sensing*, 13(15):3018, 2021.

[87] Zhenbo Xu, Wei Yang, Ajin Meng, Nanxue Lu, Huan Huang, Changchun Ying, and Liusheng Huang. Towards end-to-end license plate detection and recognition: A large dataset and baseline. In *Proceedings of the European conference on computer vision (ECCV)*, pages 255–271, 2018.

[88] Anwesh Kabiraj, Debojyoti Pal, Debayan Ganguly, Kingshuk Chatterjee, and Sudipta Roy. Number plate recognition from enhanced super-resolution using generative adversarial network. *Multimedia Tools and Applications*, pages 1–17, 2022.

[89] Seyun Lee, Ji-Hwan Kim, and Jae-Pil Heo. Super-resolution of license plate images via character-based perceptual loss. In *IEEE International Conference on Big Data and Smart Computing*, pages 560–563, 2020.

[90] Baoguang Shi et al. ASTER: An attentional scene text recognizer with flexible rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2035–2048, 2019.

[91] Anatol Maier, Denise Moussa, Andreas Spruck, Jürgen Seiler, and Christian Riess. Reliability scoring for the recognition of degraded license plates. In *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2022.

[92] Denise Moussa, Anatol Maier, Andreas Spruck, Jürgen Seiler, and Christian Riess. Forensic license plate recognition with compression-informed transformers. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 406–410. IEEE, 2022.

[93] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711, 2016.

[94] Richard Zhang et al. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018.

[95] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 184–199, 2014.

[96] Sina Farsiu, M Dirk Robinson, Michael Elad, and Peyman Milanfar. Fast and robust multiframe super resolution. *IEEE transactions on image processing*, 13(10):1327–1344, 2004.

[97] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004.

[98] Patrick Vandewalle, Luciano Sbaiz, Joos Vandewalle, and Martin Vetterli. Super-resolution from unregistered and totally aliased signals using subspace methods. *IEEE Transactions on Signal Processing*, 55(7):3687–3703, 2007.

[99] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

[100] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.

[101] Amisha J Shah and Suryakant B Gupta. Image super resolution-a survey. In *2012 1st International Conference on Emerging Technology Trends in Electronics, Communication and Networking*, pages 1–6. IEEE, 2012.

[102] Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE Trans. on Pattern Analysis and Machine Intel.*, 39:1256–1272, 2017.

[103] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems*, pages 1097–1105, 2012.

[104] Kamal Nasrollahi and Thomas B. Moeslund. Super-resolution: a comprehensive survey. *Machine Vision and Applications*, 25:1423–1468, 2014.

[105] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2480–2495, 2021.

[106] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018.

[107] I. O. Oliveira, R. Laroca, D. Menotti, K. V. O. Fonseca, and R. Minetto. Vehicle-Rear: A new dataset to explore feature fusion for vehicle identification using convolutional neural networks. *IEEE Access*, 9:101065–101077, 2021.

[108] R. Laroca, M. Santos, V. Estevam, E. Luz, and D. Menotti. A first look at dataset bias in license plate recognition. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 234–239, Oct 2022.

[109] Linjiang Zhang, Peng Wang, Hui Li, Zhen Li, Chunhua Shen, and Yanning Zhang. A robust attentional framework for license plate recognition in the wild. *IEEE Transactions on Intelligent Transportation Systems*, 22(11):6967–6976, 2021.

[110] R. Laroca, V. Estevam, A. S. Britto Jr., R. Minetto, and D. Menotti. Do we train on test data? The impact of near-duplicates on license plate recognition. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Jun 2023. In press.

[111] Yi Wang et al. Rethinking and designing a high-performing automatic license plate recognition approach. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):8868–8880, 2022.

[112] Shuxin Qin and Sijiang Liu. Towards end-to-end car license plate location and recognition in unconstrained scenarios. *Neural Computing and Applications*, 34:21551–21566, 2022.