Super-Resolution of License Plate Images Using Attention Modules and Sub-Pixel Convolution Layers

Valfride Nascimento*, Rayson Laroca*, Jorge de A. Lambert†, William Robson Schwartz‡, and David Menotti*
*Department of Informatics, Federal University of Paraná, Curitiba, Brazil
†Regional Superintendence at Bahia, Brazilian Federal Police, Salvador, Brazil
‡Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, Brazil
*{vwnascimento,rblsantos,menotti}@inf.ufpr.br †lambert.jal@pf.gov.br ‡william@dcc.ufmg.br

***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. Our code is publicly available at https://github.com/valfride/lpr-rsr-ext/.**

I. Introduction

Super-resolution is a method for enhancing the quality of an image or video by increasing its resolution. It has become a widespread technology in fields like medical imaging and surveillance [1], [2]. In recent times, there have been remarkable advancements in super-resolution techniques, particularly in interpolation-based, example-based, and deep learning-based methods [3]–[5]. These improvements have made it feasible to enhance low-resolution (LR) images and videos in a manner that was once considered impossible.

Despite advances in recent years, super-resolution remains a challenging issue due to its ill-posed nature, where there can be numerous solutions in the high-resolution (HR) space [2], [3]. Furthermore, the computational difficulty of the problem grows as the upscale factor increases, and LR images may lack sufficient information to reconstruct the desired details [2], [3]. Super-resolution can be classified into three main categories: Single-Image Super-Resolution (SISR), Multi-Image Super-Resolution (MISR), and video super-resolution [2], [6]. This study focuses on the application of SISR in the context of License Plate Recognition (LPR), as images from real-world surveillance systems are often characterized by low resolution and poor quality [7]–[9]. Although such challenging conditions are common in forensic applications, recent studies in LPR have mainly concentrated on scenarios where the license plates (LPs) are perfectly legible [10]–[13].

To address the super-resolution problem, many researchers have proposed approaches based on Convolutional Neural Networks (CNNs) [2], [14], [15]. These approaches have achieved exceptional results, but often rely on deep architectures that can be computationally expensive and focus on increasing the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) without considering the particular application at hand. In the context of LPR, we assert that such methods may not be effective in dealing with confusion between closely resembling characters, such as ‘Q’ and ‘O’, ‘T’ and ‘7’, ‘Z’ and ‘2’, and other similar pairs.

We present a novel approach for improving LP super-resolution through the use of *PixelShuffle* (PS) layers and a Three-Fold Attention Module. Our method extends the work of Mehri et al. [15] and Nascimento et al. [16] by taking into account not only the pixel intensity values, but also structural and textural information. To further enhance the performance, we incorporate an auto-encoder that extracts shallow features by squeezing and expanding the network constructed with PS and *PixelUnshuffle* (PU) layers. Additionally, we leverage a pre-trained Optical Character Recognition (OCR) model [17] to extract features from the LP images during the training phase, resulting in improved super-resolution performance and recognition rates. It is notable that the choice of the OCR model can be tailored to specific application requirements.

In summary, the main contributions of this work are:

- A super-resolution approach that builds upon MPR-Net [15] and the architecture we proposed in [16] (see the next paragraph) by incorporating subpixel-convolution layers (PS and PU) in combination with a Pixel Level Three-Fold Attention Module (PLTFAM);
- A novel perceptual loss that combines features extracted by an OCR model [17] with L1 loss to reconstruct characters with the most relevant characteristics. This loss function allows the use of any OCR model for LPR;
- The datasets we built for this work, as well as the source code, are publicly available to the research community.

A preliminary version of this study was presented at the 2022 Conference on Graphics, Patterns and Images (SIB-

This manuscript is a postprint of a paper accepted by *Computers & Graphics*. See the final version on *Science Direct* (DOI: 10.1016/j.cag.2023.05.005).

GRAPI) [16]. The approach described here differs from the previous version in several aspects. For example, we introduce novel approaches for LP super-resolution, such as an attention module architecture that considers vertical and horizontal lines to extract more structural and textural details of the LP font. We propose a new loss method that employs feature extraction through a pre-trained network for LP recognition. The images used for training and testing consist of paired low- and high-resolution LPs, with the LR samples degraded until their SSIM falls below 0.10. These improvements have enabled us to achieve better results than those reported in our previous work. In this work, we report the results of experiments performed on two datasets, collected in different regions under various conditions, instead of a single one. In the RodoSol-ALPR dataset [18], our approach recognizes at least five characters in 74.2% of LPs compared to 42.2% by our previous model trained and evaluated under the same conditions. In the PKU dataset [19], the improvement was even more significant, from 82.5% by the preliminary approach to 97.3% by the improved one (proposed in this work).

The rest of this article is structured as follows. Section II provides a concise overview of relevant studies on SISR, as well as works that designed or applied super-resolution techniques specifically to LPR. In Section III, we elaborate on our proposed network architecture and the implementation of the new perceptual loss function that explores an OCR model as a feature extractor. Section IV presents the experiments performed and the results obtained. In Section V, we summarize the findings and their significance, concluding this study.

II. Related Work

This section provides a brief overview of related work. Some approaches used in SISR are discussed in greater detail in Section II-A, and the use of deep learning methods for LP super-resolution is discussed in Section 2.2.

A. Single-Image Super-Resolution

The field of SISR has experienced significant advancements in recent years, leading to its broad application in various domains [1], [2]. Early SISR methods were generally classified into four categories: prediction models, edge-based methods, image statistical methods, and example-based methods [20]–[24]. In 2016, Dong et al. [25] introduced the Super-Resolution Convolutional Neural Network (SRCNN), a deep learning-based approach to SISR, which demonstrated both superior quality and faster performance compared to previous methods.

Despite the success of SRCNN, some limitations were observed such as relying on pre-upsampled LR images, which drastically increased computational complexity without providing significant additional information for image restoration [26], [27]. To overcome these limitations, later studies by Dong et al. [28] and Shi et al. [29] incorporated the upsampling process near the end of the network architecture, leading to a substantial reduction in execution time, parameters, and computational cost.

Shi et al. [29] highlighted the importance of learnable upscaling and designed specialized sub-pixel convolution layers to optimize the learning of upscaling filters. This enabled the networks to learn complex mappings from LR to HR images, resulting in improved performance compared to using fixed filters from interpolation methods.

Recent research in the field of super-resolution has introduced attention mechanisms as a means of improving image reconstruction. Zhang et al. [30] were among the pioneers to introduce the use of first-order statistical attention mechanisms in this context. Afterwards, Dai et al. [31] presented an improved version that uses second-order statistics to extract more meaningful features. Huang et al. [32] proposed an attention network that preserves detail fidelity by using a divide-and-conquer strategy to progressively process smooth and detailed features.

Mehri et al. [15] introduced the Multi-Path Residual Network (MPRNet), which leverages information from both inner-channel and spatial features using a Two-fold Attention Module (TFAM). MPRNet has demonstrated superior or competitive performance compared to multiple state-of-the-art methods such as those presented in [33]–[35].

Recently, Zhang et al. [36] proposed a structure- and texture-preserving image super-resolution reconstruction method, known as the Dual-Coordinate Direction Perception Attention (DPCA) mechanism. This method effectively emphasizes structure and feature details, resulting in improved image quality compared to previous methods.

B. Super-Resolution for License Plate Recognition

The goal of LPR is to precisely extract and identify characters from each LP. Despite recent progress and successful outcomes in LPR [10], [12], [13], most of the models proposed have only been trained and evaluated on HR images, where the LP characters are clear and easily recognizable to the human eye. This does not reflect the typical conditions encountered in real-world surveillance scenarios, where images frequently have low resolution and poor quality [7]–[9].

The quality of LP images is closely linked to various factors, such as camera distance, motion blur, lighting conditions, and image compression techniques used for storage [7]. In commercial LPR systems, sharp images are typically captured using global shutter cameras. However, in surveillance systems, cost-effective cameras that use rolling shutter technology are often employed, leading to blurry images [37] with illegible LPs.

Super-resolution techniques have been proposed as a solution to address the issue of poor image quality in LPR. The first works to combine the idea of super-resolution and LP recognition date back to the 2000s, such as those proposed by Suresh et al. [38] and Yuan et al. [39], which rely on image processing or interpolation techniques to increase resolution. While some algorithms incorporate character semantics and segmentation for super-resolution, these methods tend to perform poorly on noisy images [40]. Considering space

limitations, the remainder of this section describes works published in recent years.

Lin et al. [41] proposed a super-resolution approach for LPR using their Super-Resolution Generative Adversarial Networks (SRGAN) model and a perceptual OCR loss. Despite obtaining promising results, the experiments were limited to 100 images and excluded those with low brightness/contrast.

Hamdi et al. [42] proposed a GAN-based architecture named Double Generative Adversarial Networks for Image Enhancement and Super Resolution (D_GAN_ESR), which outperformed previous SRGAN methods [41]. The architecture consists of two networks concatenated together, with the first network responsible for denoising and deblurring and the second network performing super-resolution. The authors assessed the performance of their method in terms of PSNR and SSIM, but acknowledged that these metrics alone do not necessarily indicate superior image reconstruction. The model was trained using LR images downsampled from HR images.

Lee et al. [43] observed that previous super-resolution approaches did not take character recognition into account. Thus, they designed a GAN-based model that incorporates a perceptual loss composed of intermediate features extracted by a scene text recognition model. Specifically, the authors used an intermediate representation (*block4*) of ASTER [44]. While their method produced better results than the same GAN-based model trained with the original perceptual loss, the authors did not make the dataset used available, and the degradation method employed was not detailed.

Although the primary objective of enhancing LP images is to improve recognition accuracy, it is surprising that most works have primarily evaluated the quality of the reconstructed images through subjective visual evaluations or metrics such as PSNR and SSIM. It is well-known that these metrics have a limited correlation with human assessment of visual quality [45], [46]. Furthermore, we observed that most previous studies explored private datasets in the experiments [8], [42], [43], [47], which makes it challenging to accurately assess the reported results.

III. Proposed Approach

This section details our super-resolution approach that enhances the extraction of structural and textural features from low-resolution LPs. Our network extends the network proposed in our previous work [16], further expanding the MPRNet architecture and TFAM algorithm by Mehri et al. [15] while taking inspiration from [36] to improve the proposed attention module to enable the network to capture structural and textural information. The proposed approach leverages a novel perceptual loss function that uses an OCR model as a feature extractor.

A. Network Architecture Modifications

The proposed approach for super-resolution in LPR features a network architecture that builds upon the work of Mehri et al. [15] and Zhang et al. [36]. As illustrated in Fig. 1, the architecture comprises four key components: a Shallow Feature Extractor (SFE); Residual Dense Blocks (RDBs) (refer to [4] for more information); a Feature Module (FM) module; and a 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 paragraphs.

The design of the SFE block includes a convolutional layer with a 5×5 kernel, followed by an autoencoder that employs depthwise-separable convolutional layers (DConvs), PU and PS operations instead of conventional convolutional layers and pooling and upscale operations. The output of the layers is then combined with a skip connection from the initial convolutional layers. Finally, the resulting output is processed by the RDBs.

In Fig. 2, we show our modifications to the MPRNet's TFAM [15] and to the attention module by Nascimento et al. [16] to create the PLTFAM. The design of this module is based on the following insights: **(i)** images are composed of the relationship between channels, where each channel contributes unique characteristics to form the final image, therefore, the extraction of these features is crucial for proper image restoration; **(ii)** the positional information of these essential 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 able to learn a custom process for different tasks; **(iv)** the module captures salient structure from the character fonts of the LP, highlighting both structure and textural features in the image.

The Channel Unit (CA) module is designed to identify and retain the inter-channel relationship features while eliminating less relevant ones. This is accomplished by using two parallel convolutional layers, concatenating their outputs, and processing the combined output through a convolutional layer, a PU layer, a PS layer, and a DConv later. This effectively summarizes the inter-channel relationship features for enhanced image restoration.

The Positional Unit (POS) complements the CA module by identifying the location of significant features within the image. This is done by extracting first-order statistics through average and max pooling operations, concatenating the results, and processing them through DConvs and PS layers, restoring the original feature map dimension. This highlights the positions of the relevant inter-channel relationship features, resulting in further improvement of image restoration.

We incorporated a third branch named Geometrical Perception Unit (GP) to the network to enhance its ability to extract critical characteristics such as structural, textural, and geometric features from the LP. This approach was motivated by the work of [36]. The GP 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. The results from this layer are then aggregated through an element-wise multiplication to obtain the final output.

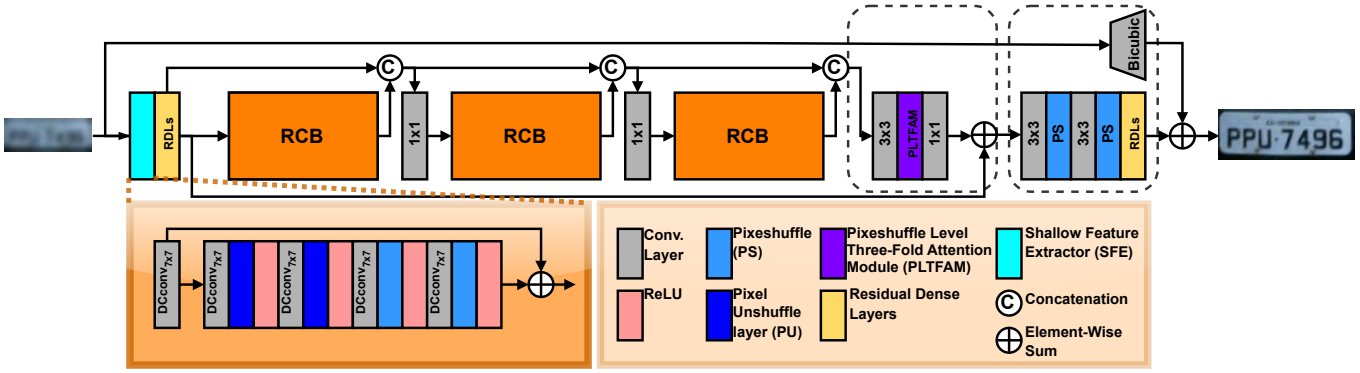

Fig. 1. The proposed architecture, which incorporates an autoencoder consisting of PS and PU layers for feature compression and expansion, respectively. The aim of this design is to eliminate less significant features. In addition, the TFAM modules in the original architecture were replaced with PLTFAM modules throughout the network. The legend inside the figure provides explanations for the acronyms used.

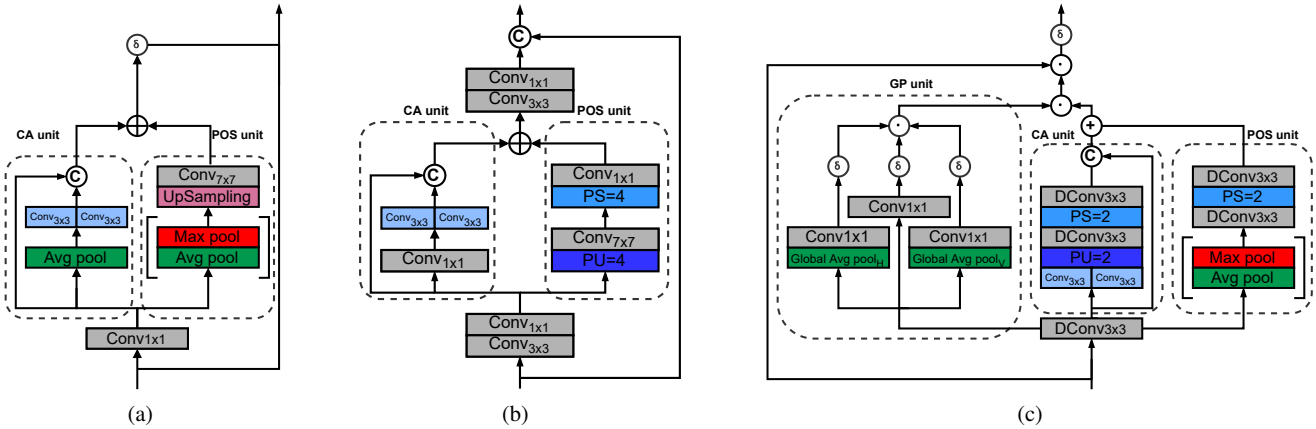

Fig. 2. Comparative illustration of the (a) Two-Fold Attention Module in MPRNet [15], (b) PixelShuffle Two-Fold Attention Module in Nascimento et al. [16], and (c) PixelShuffle Three-Fold Attention Module (ours).

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. This mask summarizes all relevant information extracted by the CA, POS and GP units, and is used to enhance the input to the PLTFAM 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.

The original Residual Concatenation Blocks (RCBs) were enhanced by incorporating the proposed PLTFAM instead of the TFAM and including dilated convolution layers in the bottleneck path of the Adaptive Residual Blocks (ARB). This modification retained the overall structure described in [15] while improving the network's capability to consider a broader context through an increased receptive field without adding extra parameters. Also, the use of dilated convolutions helped to reproduce fine details in LP images by avoiding the "smoothing" effect that can occur with traditional convolutions.

Returning our attention to Fig. 1, the reconstruction module was added as an output block for better aggregating fine details. It consists of two PS with a scale factor value of 2 for pixel reorganization, each followed by a DConv layer and by consecutive RDBs.

B. Perceptual Loss

To further enhance the accuracy of LPR, we propose incorporating a perceptual loss function in our super-resolution approach. This loss function, shown in Eq. (1), is specifically designed to improve the accuracy of the system by considering the features that an OCR model typically expects.

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

In Eq. (1), H_i and S_i denote the high-resolution and super-resolved LP images, respectively, and $f_{OCR}(\cdot)$ represents the feature extraction process performed by the OCR model.

It is worth noting that the loss function allows the use of any OCR model for LPR. This flexibility is particularly appealing since novel models can be readily incorporated as they become available. In this work, we explored the multi-task model proposed by Gonçalves et al. [17], as it is quite efficient and has achieved remarkable outcomes in prior research [7], [16].

The Mean Squared Error (MSE) is used to compute the difference between the expected and generated pixel values,

with more significant errors being penalized more than minor errors. This approach is beneficial in enhancing the overall quality of the image. Also, the MSE effectively preserves the structural information in the image, which is essential in the super-resolution task. By contrast, the L1 loss ensures robustness to noise and outliers and helps to preserve sharp edges in the generated images by considering the expected features. Combining MSE and L1 loss allows for a more comprehensive evaluation of the generated images and helps achieve a balance between preserving structural information and minimizing errors.

IV. 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.

A. Setup

We made use of LP images obtained from the RodoSol-ALPR [18] and PKU [19] 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 opted for these two datasets since they provide a wide range of scenarios under 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]. Observe in Fig. 3 the diversity of this dataset regarding several factors such as LP colors, lighting conditions, and character fonts. Here, we follow the standard protocol (defined in [18]) that involves using 40% of the images for training, 20% for validation, and 40% for testing.

Fig. 3. Some LP images from the RodoSol-ALPR dataset [18]. The first two rows show Brazilian LPs, while the last two rows show Mercosur LPs. For scope reasons, in this work, we conduct experiments on LPs that have all characters arranged in a single row (i.e., 10K images).

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 [49]). Despite the diverse settings, the LP images have good quality and are perfectly legible (see some examples in Fig. 4). Following [48], [49], we use 60% of the images for training/validation, while the remaining 40% are used for testing. Laroca et al. [50] recently revealed that the PKU dataset (as well as several other datasets but not RodoSol-ALPR) 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.

Fig. 4. Examples of LP images from the PKU dataset [19]. 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 [51], 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 apply padding before resizing them to 20×40 pixels, resulting in an output shape of 80×160 pixels for an upscale factor of 4. Fig. 5 and Fig. 6 show examples of the LP images generated for the RodoSol-ALPR and PKU datasets, respectively.

Fig. 5. Some HR-LR image pairs created from the RodoSol-ALPR dataset.

Fig. 6. Examples of HR-LR image pairs created from the PKU dataset.

Our experiments were conducted using the PyTorch and Keras frameworks on a high-performance computer that is

[1]Following [12], [16], [48], we use the term "Brazilian" to refer to the layout used in Brazil prior to the adoption of the Mercosur layout.

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.

B. Experimental Results

In the LPR literature, models are usually evaluated in terms of the number of correctly recognized LPs divided by the number of LPs in the test set [12], [13], [50]. A correctly recognized LP means that all characters on the LP were correctly recognized. Considering our focus on low-resolution LPs, which are very common in forensic applications, we also report the recognition results considering partial matches (when at least 5 or 6 of the 7 characters are correctly recognized) as they may be useful in narrowing down the list of candidate LPs by incorporating additional information such as the vehicle's make and model.

The results of the LPR experiment are shown in Table I. The table demonstrates the recognition accuracy of HR and LR LP 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 0.10, as illustrated in Fig. 5, where the LP characters are barely distinguishable.

TABLE I
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 [17] — 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 [17] — with super-resolution						
Proposed	**39.0**	**59.9**	**74.2**	**72.0**	**90.3**	**97.3**
Nascimento et al. [16]	10.5	25.4	42.2	35.5	65.3	82.5
Mehri et al. [15]	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. [16]	**21.3**	0.52		18.1	0.54	
Mehri et al. [15]	16.8	0.38		16.4	0.41	

The proposed super-resolution network achieved superior performance compared to the two baseline models [15], [16], as presented in the second section of Table I. The multi-task OCR model [17] 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 method [16] and a 26.7% higher accuracy compared to MPRNet [15] for LPs with more than five correct characters.

For completeness, we detail in Table I the PSNR and SSIM obtained by each approach. Similar to what was observed in [41], [42], [46], the PSNR metric seems inappropriate for this particular application, as our approach and the one proposed in [16] 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 OCR network showed these improved results because of the effective extraction of textural and structural information by the proposed GP unit, along with the CA and POS units. These units were designed with pyramid and PixelShuffle layers to optimize channel scaling and reorganization within the image.

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, with less variation in the environmental conditions under which the images were collected.

Finally, the results of the LPR experiments are further substantiated by a visual contrast of the super-resolution images produced by our technique and the baseline methods [15], [16]. Fig. 7 and Fig. 8 show four LR images alongside their corresponding super-resolution counterparts, in addition to the original HR image as a reference. It is evident that the proposed approach outperforms both its preliminary version [16] and MPRNet [15] in terms of perceptual quality.

Fig. 7. Typical examples of the images generated by the proposed approach and baselines in the RodoSol-ALPR dataset [18]. GT = ground truth.

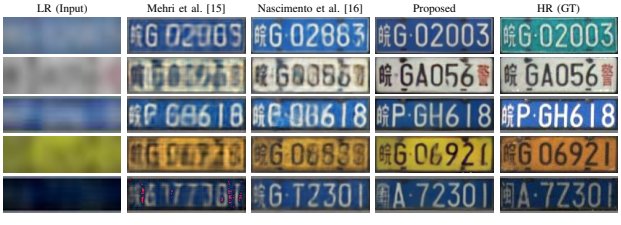

Fig. 8. Representative samples of the images generated by the proposed approach and baselines in the PKU dataset [19]. GT = ground truth.

In general, the images produced by MPRNet [15] exhibit a common issue of blurriness, where the character edges blend into the LP background, resulting in visible artifacts. This blurriness can also cause the edges of multiple characters

to blend together, leading to further visual distortions. The architecture proposed in our previous work [16] 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. 7). Conversely, the proposed model generates clear character edges and consistently reconstructs the original font, without any missing characters or incomplete lines.

When our model is uncertain about which character to reconstruct, it tends to hallucinate with characters that are most congruent with the LR input, as evident in the last row of Fig. 7 and Fig. 8, where the character "3" is reconstructed as "J", and the character "Z" is reconstructed as "2", respectively. This issue can be mitigated by incorporating a lexicon or vocabulary into the network's learning process to track the character type (letter, digit, or either) that can occupy each position on LPs of a specific layout.

Furthermore, the network tends to generate nearly identical background colors for different images. This behavior can be observed in the third row of Fig. 7 and the first row of Fig. 8. However, it is noteworthy that, based on our analysis, this does not considerably impact the recognition results achieved.

1) Ablation Study: As our approach integrates multiple concepts into a single architecture, we conducted an ablation study to validate the contribution of each incorporated unit to the results obtained. The study involved removing the autoencoder, TFAM, PS and PU layers and training the network without the perceptual loss (one modification at a time).

Four baselines were established for the experiments. The first baseline replaced the autoencoder with a DConv layer with a 5×5 kernel for shallow feature extraction [15]. The second baseline removed the TFAM 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 [29]. Finally, in the fourth baseline, the perceptual loss was replaced by MSE, which is commonly used in super-resolution research [2], [3]. Table II presents the results.

TABLE II
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 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 the autoencoder unit was removed, resulting in a recognition rate of 32.7% for all characters recognized. 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. 3 and Fig. 4). 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 [13], [49], [52].

V. CONCLUSIONS

This article proposes a new super-resolution approach to improve the recognition of low-resolution LPs. Our method builds upon the existing MPRNet [15] and the architecture proposed in our previous work [16] by incorporating subpixel-convolution layers (PS and PU) in combination with a PLT-FAM. We also 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 conducted experiments on two publicly available datasets from Brazil and mainland China, which contain a diverse range of LP images. The results showed better recognition rates being achieved in the images reconstructed by the proposed method than in those reconstructed by the baselines. More specifically, for the RodoSol-ALPR dataset, our method led to a recognition rate of 39.0% being achieved by the OCR model, while the methods proposed in [16] and [15] led to recognition rates of 31.3% and 4.0%, respectively. Similarly, for the PKU dataset, our approach outperformed both baselines, with the OCR model reaching a recognition rate of 72.0%, compared to 35.5% and 22.5% for [16] and [15], respectively. We have made available all datasets used in our experiments (i.e., the LR–HR image pairs), as well as the source code, in order to encourage further research and development in the field of LPR super-resolution.

In the future, our plans include integrating a lexicon or vocabulary into the network's learning process to track the character type that can occupy each position on LPs of a specific layout. Additionally, we intend to create a large-scale dataset for LP super-resolution, consisting of thousands of LR and HR image pairs. We aim to collect videos in which the LP is legible in one frame but not in another, enabling us to assess existing methods in real-world scenarios and develop novel methods.

ACKNOWLEDGMENTS

This work was supported in part by the Coordination for the Improvement of Higher Education Personnel (CAPES) (*Programa de Cooperação Acadêmica em Segurança Pública e Ciências Forenses # 88881.516265/2020-01*), in part by the National Council for Scientific and Technological Development (CNPq) (# 309953/2019-7 and # 308879/2020-1), and also in part by the Minas Gerais Research Foundation (FAPEMIG) (Grant PPM-00540-17). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Quadro RTX 8000 GPU used for this research.

REFERENCES

[1] L. Yue, H. Shen, J. Li, Q. Yuan, H. Zhang, and L. Zhang, "Image super-resolution: The techniques, applications, and future," *Signal Processing*, vol. 128, pp. 389–408, 2016.

[2] A. Liu, Y. Liu, J. Gu, Y. Qiao, and C. Dong, "Blind image super-resolution: A survey and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5461–5480, 2023.

[3] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365–3387, 2021.

[4] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 7, pp. 2480–2495, 2021.

[5] M. Santos *et al.*, "Face super-resolution using stochastic differential equations," in *Conference on Graphics, Patterns and Images (SIBGRAPI)*, Oct 2022, pp. 216–221.

[6] G. Guarnieri, M. Fontani, F. Guzzi, S. Carrato, and M. Jerian, "Perspective registration and multi-frame super-resolution of license plates in surveillance videos," *Forensic Science International: Digital Investigation*, vol. 36, p. 301087, 2021.

[7] G. R. Gonçalves *et al.*, "Multi-task learning for low-resolution license plate recognition," in *Iberoamerican Congress on Pattern Recognition (CIARP)*, Oct 2019, pp. 251–261.

[8] A. Maier, D. Moussa, A. Spruck, J. Seiler, and C. Riess, "Reliability scoring for the recognition of degraded license plates," in *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2022, pp. 1–8.

[9] D. Moussa *et al.*, "Forensic license plate recognition with compression-informed transformers," in *IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 406–410.

[10] 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*, vol. 15, no. 4, pp. 483–503, 2021.

[11] Y. Gong, L. Deng, S. Tao, X. Lu, P. Wu, Z. Xie, Z. Ma, and M. Xie, "Unified Chinese license plate detection and recognition with high efficiency," *Journal of Visual Communication and Image Representation*, vol. 86, p. 103541, 2022.

[12] S. M. Silva and C. R. Jung, "A flexible approach for automatic license plate recognition in unconstrained scenarios," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 5693–5703, 2022.

[13] Y. Wang *et al.*, "Rethinking and designing a high-performing automatic license plate recognition approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 8868–8880, 2022.

[14] A. Lucas *et al.*, "Generative adversarial networks and perceptual losses for video super-resolution," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3312–3327, 2019.

[15] A. Mehri, P. B. Ardakani, and A. D. Sappa, "MPRNet: Multi-path residual network for lightweight image super resolution," in *IEEE Winter Conference on Applications of Computer Vision*, 2021, pp. 2703–2712.

[16] 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)*, Oct 2022, pp. 228–233.

[17] G. R. Gonçalves *et al.*, "Real-time automatic license plate recognition through deep multi-task networks," in *Conference on Graphics, Patterns and Images (SIBGRAPI)*, Oct 2018, pp. 110–117.

[18] 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)*, Feb 2022, pp. 166–178.

[19] 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*, vol. 26, no. 3, pp. 1102–1114, 2017.

[20] D. Glasner *et al.*, "Super-resolution from a single image," in *International Conference on Computer Vision (ICCV)*, 2009, pp. 349–356.

[21] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 1127–1133, 2010.

[22] R. Timofte, V. De, and L. V. Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *IEEE International Conference on Computer Vision*, 2013, pp. 1920–1927.

[23] J. Yang, Z. Lin, and S. Cohen, "Fast image super-resolution based on in-place example regression," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 1059–1066.

[24] C.-Y. Yang *et al.*, "Single-image super-resolution: A benchmark," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 372–386.

[25] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[26] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 370–378.

[27] Y. Chen and T. Pock, "Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration," *IEEE Trans. on Pattern Analysis and Machine Intel.*, vol. 39, pp. 1256–1272, 2017.

[28] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 391–407.

[29] W. 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*, 2016, pp. 1874–1883.

[30] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 294–310.

[31] T. Dai *et al.*, "Second-order attention network for single image super-resolution," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 057–11 066.

[32] Y. Huang, J. Li, X. Gao, Y. Hu, and W. Lu, "Interpretable detail-fidelity attention network for single image super-resolution," *IEEE Transactions on Image Processing*, vol. 30, pp. 2325–2339, 2021.

[33] X. He, Z. Mo, P. Wang, Y. Liu, M. Yang, and J. Cheng, "Ode-inspired network design for single image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1732–1741.

[34] X. Luo, Y. Xie, Y. Zhang, Y. Qu, C. Li, and Y. Fu, "LatticeNet: Towards lightweight image super-resolution with lattice block," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 272–289.

[35] A. Muqeet, J. Hwang, S. Yang, J. Kang, Y. Kim, and S.-H. Bae, "Multi-attention based ultra lightweight image super-resolution," in *European Conference on Computer Vision Workshops*, 2020, pp. 103–118.

[36] Y. Zhang, Y. Huang, K. Wang, G. Qi, and J. Zhu, "Single image super-resolution reconstruction with preservation of structure and texture details," *Mathematics*, vol. 11, p. 216, 01 2023.

[37] C.-K. Liang, L.-W. Chang, and H. H. Chen, "Analysis and compensation of rolling shutter effect," *IEEE Transactions on Image Processing*, vol. 17, no. 8, pp. 1323–1330, 2008.

[38] K. V. Suresh, G. M. Kumar, and A. N. Rajagopalan, "Superresolution of license plates in real traffic videos," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 2, pp. 321–331, 2007.

[39] J. Yuan, S.-D. Du, and X. Zhu, "Fast super-resolution for license plate image reconstruction," in *International Conference on Pattern Recognition (ICPR)*, 2008, pp. 1–4.

[40] Y. Zou, Y. Wang, W. Guan, and W. Wang, "Semantic super-resolution for extremely low-resolution vehicle license plate," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3772–3776.

[41] M. Lin, L. Liu, F. Wang, J. Li, and J. Pan, "License plate image reconstruction based on generative adversarial networks," *Remote Sensing*, vol. 13, no. 15, p. 3018, 2021.

[42] A. Hamdi, Y. K. Chan, and V. C. Koo, "A new image enhancement and super resolution technique for license plate recognition," *Heliyon*, vol. 7, no. 11, p. e08341, 2021.

[43] S. Lee, J.-H. Kim, and J.-P. Heo, "Super-resolution of license plate images via character-based perceptual loss," in *IEEE International Conference on Big Data and Smart Computing*, 2020, pp. 560–563.

[44] B. Shi *et al.*, "ASTER: An attentional scene text recognizer with flexible rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2035–2048, 2019.

[45] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 694–711.

[46] R. Zhang *et al.*, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

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

[48] 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)*, Oct 2022, pp. 234–239.

[49] L. Zhang, P. Wang, H. Li, Z. Li, C. Shen, and Y. Zhang, "A robust attentional framework for license plate recognition in the wild," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 11, pp. 6967–6976, 2021.

[50] 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)*, Jun 2023, pp. 1–8.

[51] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2472–2481.

[52] S. Qin and S. Liu, "Towards end-to-end car license plate location and recognition in unconstrained scenarios," *Neural Computing and Applications*, vol. 34, p. 21551–21566, 2022.