Combining Attention Module and Pixel Shuffle for License Plate Super-Resolution

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*—The License Plate Recognition (LPR) field has made impressive advances in the last decade due to novel deep learning approaches combined with the increased availability of training data. However, it still has some open issues, especially when the data come from low-resolution (LR) and low-quality images/videos, as in surveillance systems. This work focuses on license plate (LP) reconstruction in LR and low-quality images. We present a Single-Image Super-Resolution (SISR) approach that extends the attention/transformer module concept by exploiting the capabilities of PixelShuffle layers and that has an improved loss function based on LPR predictions. For training the proposed architecture, we use synthetic images generated by applying heavy Gaussian noise in terms of Structural Similarity Index Measure (SSIM) to the original high-resolution (HR) images. In our experiments, the proposed method outperformed the baselines both quantitatively and qualitatively. The datasets we created for this work are publicly available to the research community at *https://github.com/valfride/lpr-rsr/*.**

I. Introduction

Image quality improvement is a recurrent yet challenging topic in computer vision due to its complexity and high practical value in many surveillance applications, for example, license plate (LP), face, and object recognition. In this regard, super-resolution (SR) [1] has been an important research subject within the last few decades due to its capacity to retrieve subtleties and textures from low-resolution (LR) images and generate their high-resolution (HR) counterparts. Moreover, the storage of HR images as LR versions and the ability to recover them when necessary is desirable [2], [3].

SR is typically divided into Single-Image Super-Resolution (SISR), multi-image super-resolution, and video super-resolution. Here, we focus on SISR. There are two main reasons why SR is still considered a challenging research topic. First, it is an inherently ill-posed problem since one LR image may have multiple plausible HR reconstructions [4]. Second, increasing the upscaling factor also increases the complexity of the problem [1].

With the advancement of deep learning methods and their chain of success in computer vision problems, the spread of Convolutional Neural Networks (CNNs) to deal with SR applications can be easily noticed [3], [5], [6]. Although significant advances have been made, most of the approaches to solving the SR problem are based on very deep architectures, which increase general computation operations, and focus only on achieving a higher Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), disregarding contextual information of the application at hand. In the License Plate Recognition (LPR) context, we consider this is not the best way to handle the problem as one approach may create very realistic images even if it fails to differentiate similar characters (e.g., 'Q' and 'O', '1' and 'I', among others).

In this work, *PixelShuffle* (PS) layers are employed in a new perspective to improve LP super-resolution by extending the Multi-Path Residual Network (MPRNet) [6] attention module. In addition, shallow features from input images are extracted by squeezing and expanding with an auto-encoder built with PS and *PixelUnshuffle* (PU) layers. Lastly, considering our target application (i.e., LPR), we propose a loss function that also considers the predictions returned by an Optical Character Recognition (OCR) model [7] on the reconstructed images.

In summary, the main contributions of our work are:

- An extension of the attention mechanism from MPRNet that uses PS layers for channels reorganization [8];
- A new loss function that incorporates both LPR predictions and quality metrics in its formulation;
- The datasets we built for this research (images degraded at different SSIM intervals) are publicly available[1].

II. Related Work

This section briefly describes related work. An overview of SISR approaches is given in Section II-A, while deep learning methods for LP super-resolution are described in Section II-B.

A. Single-Image Super-Resolution

Dong et al. [9] proposed one of the first deep learning-based methods, called Super-Resolution Convolutional Neural Network (SRCNN), to tackle the SR ill-posed characteristic. It showed to be faster and quantitatively better in restoration capabilities than previous example-based methods, with fewer pre- or post-processing steps.

Despite its success, SRCNN receives pre-upsampled LR images generated through interpolation methods, drastically increasing computational complexity without aggregating vital

[1] Access is granted *upon request*, i.e., interested parties must register by filling out a registration form and agreeing to the dataset's terms of use.

information to further image restoration [10], [11]. Later, Dong et al. [12] and Shi et al. [8] explored the upsample process near the end of the network as part of the architecture pipeline. This strategy resulted in an expressive reduction of run time, parameters, and computational cost. In [13], Dong et al. observed that bicubic interpolation is also a convolutional operation, so it can be formulated as a convolutional layer.

The importance of learnable upscaling was highlighted by Shi et al. [8]. They designed specialized convolution layers to learn upscaling filters, making it possible to learn more complex mappings from LR to HR images with increased performance compared to fixed-size interpolation methods.

The presence of "*dead*" filters was observed both in [13] and [14]. These dead filters may be seen as the network alone trying to "*discover*" what are the essential features from the input, resulting in a lack of learning and performance.

Accordingly, to better influence the network into allocating available computer resources to the most informative aspects of the input images, Zhang et al. [15] introduced the concept of first-order statistic channel attention mechanism for image reconstruction that uses only information across inner-channel features to image reconstruction. Afterward, Dai et al. [16] proposed the second-order version of the attention model to explore more meaningful features expression.

More recently, motivated by these works, Mehri et al. [6] presented the Two-fold Attention Module (TFAM) to exploit essential information – considering both inner-channel and spatial features – to boost the network's performance. Their results showed superior or competitive performance compared to multiple baselines [13], [17], [18]. MPRNet generated SR images with textures similar to the original HR images since it fully uses the abstract features within the LR input.

B. Super-Resolution for License Plate Recognition

The main goal of an LPR system is to extract the LP information from an image or sequence of images [19], [20]. These systems have been extensively researched due to their practical applications in security tasks such as traffic law enforcement, monitoring private areas, and criminal investigations [21].

Although impressive results have been reported in recent years in the LPR context [22]–[24], the datasets where the proposed models are being evaluated are mostly composed of HR images where all LP characters are pretty legible. In most surveillance scenarios, this is not in line with everyday reality.

The quality of LP images is intrinsically related to various factors, such as camera distance, motion blur, lighting conditions, and image compression technique used for storage [25]. While commercial LPR systems capture sharp images using *global shutter* cameras, cheaper cameras using *rolling shutter* technology are typically employed in surveillance systems, often resulting in blurry images [26] with illegible LPs.

While the first works to combine the idea of SR and LP recognition date from the 2000s [27], [28], this research area has received increasing attention in recent years given the rise of deep learning. Considering space limitations, the remainder of this section describes works published in recent years.

Svodoba et al. [29] demonstrated that CNNs trained on artificially generated blurry images provide superior quality enhancement on images with motion blur compared to traditional blind deconvolution methods. However, as they trained their model for a specific range of motion blur lengths and directions, the reconstruction quality degrades considerably for blurs outside the range of the blurs the network was trained for.

Lin et al. [30] exploited the high capability of a Generative Adversarial Network (GAN) for LP reconstruction. They reported promising results; nevertheless, their experiments were carried out on just 100 images. Moreover, their approach was compared with other methods only in terms of PSNR and SSIM, without exploring LP recognition at all.

In the same direction, Hamdi et al. [31] concatenated two GAN models for this task. The first was used for denoising and deblurring, while the second was applied to super-resolution. The authors compared their method with three baselines, but only in terms of PSNR and SSIM as well. Interestingly, after analyzing the results, they acknowledged that higher PSNR and SSIM do not necessarily mean better reconstruction.

Lee et al. [32] observed that previous SR approaches did not take character recognition into account. Thus, they designed a GAN-based model that relies on a perceptual loss composed of intermediate features extracted by a scene text recognition model. While their method reported better results than the same GAN-based model trained with the original perceptual loss, their dataset was not made available and the degradation method they used was not detailed as well.

While the final goal in enhancing LP images is to improve the recognition results, in current works (except [32]) the quality of the reconstructed images was evaluated either qualitatively or based on the PSNR and SSIM metrics, which are known not to correlate well with human assessment of visual quality [33], [34]. Also, in most related works the experiments were conducted exclusively on private datasets [29], [31], [32].

III. Methodology

This section describes the proposed approach. We first detail how we extend the MPRNet architecture proposed by Mehri et al. [6]. Then, we present our improved loss function.

A. Network Architecture Modifications

Inspired by the work of Mehri et al. [6], the proposed network architecture is presented in Fig. 1. It consists of four different modules, namely, Shallow Feature Extractor (SFE); Residual Concatenation Block (RCB); Feature Module (FM); and reconstruction module, which combines the output of the FM module with a long-skip connection from the end of the SFE module. We highlight our changes in the followings.

SFE block internal design consists mainly of a Pre-shallow Feature Extractor (PSFE) with a 5×5 kernel and a convolution layer followed by an autoencoder with PU and PS layers in its composition instead of classic pooling and upscale operations. Finally, the autoencoder output is combined with a skip connection from PSFE. The main idea of this design is to learn and emphasize the most important characteristics by

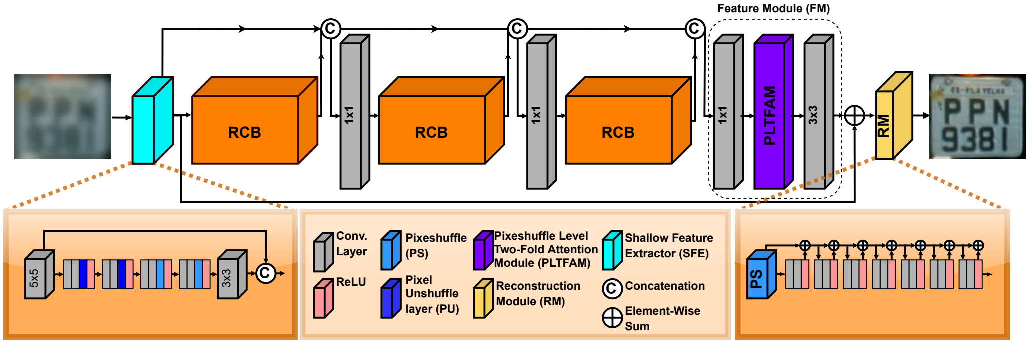
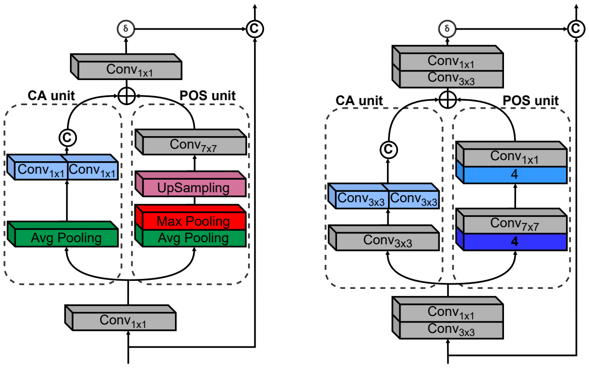

Feature Module (FM)
RCB
RCB
RCB
1x1
1x1
1x1
PLTFAM
3x3
RM
5x5
3x3
PS
Conv. Layer
ReLU
Pixeshuffle (PS)
Pixel Unshuffle layer (PU)
Pixeshuffle Level Two-Fold Attention Module (PLTFAM)
Reconstruction Module (RM)
Shallow Feature Extractor (SFE)
Concatenation
Element-Wise Sum

Fig. 1. The proposed architecture; acronyms are described in the internal legend. To the network was introduced an autoencoder composed of PU and PS layers for squeeze and expansion, respectively, aiming to exclude less relevant features. PLTFAM replaced the original TFAM modules along the network.



(a)

(b)

Fig. 2. Comparative illustration of the (a) two-fold attention module in MPRNet [6] and (b) Pixelshuffle two-fold attention module (ours).

squeezing (PU) and expanding (PS) to the original dimensions with an aggregation 3×3 convolution layer at the end. Less informative features are not lost thanks to the skip connection.

Fig. 2 highlights the improvements we made from MPRNet's TFAM [6] to produce our proposed PLTFAM (the purple block in Fig. 1), which was built following the insights: (i) images are composed of inter-channel relationships since each channel contributes with unique characteristics to compose the final image. Therefore, the extraction of such key features is essential; (ii) learning positional information of the key characteristics from each channel that compose the inter-channel relationships is required; and (iii) as stated by Shi et al. [8] and Dong et al. [13], downscale and upscale layers rely on Translational Invariance (e.g., *MaxPool* and *AvgPool*) and interpolation (e.g., bicubic) general techniques, thus, both layers are not able to learn a custom process for different tasks.

Channel Unit (CA) is concerned about summarizing inter-channel relationship features and excluding less relevant ones. For such an aim, it starts with a PS layer to better explore the most important features by optimally rearranging pixels to emphasize meaningful features. Later, two convolution layers – working side by side – receive half the input followed by the concatenation of the outputs to leverage the previous operation.

The positional unit learns where the significant features within the image are located. It decreases and increases the input by the same scale factor with PS and PU layers in sequence. As a result of this process, the positions of the relevant inter-channel relationship features are highlighted. Lastly, the results from both CA and Positional Unit (POS) proceed to an element-wise sum, followed by a 3×3 and 1×1 kernel convolution layers and a sigmoid function to generate the final attention mask, which aggregates all the relevant information extracted by CA and POS units.

The RCBs had their original TFAM replaced by the PLTFAM, but its main structure remained the same adopted in [6].

Returning our attention to Fig. 1, the reconstruction module was added as an output block for better aggregating fine details. It comprises one PS for pixel reorganization, followed by seven recurrent blocks with two 3×3 kernel size convolution layers and ReLU activation function. A sigmoid activation function was added at the end of the reconstruction module to limit the values to $[0, 1]$ and thus smooth the output.

B. Perceptual Loss

Considering that the final goal of an LPR system is to achieve recognition rates as high as possible, we propose the following perceptual loss:

$$PL = \frac{1}{n}\sum_{i=0}^{n}(H_i - S_i)^2(1 + \alpha \cdot D(H_i, S_i)), \tag{1}$$

where the details term $D(H_i, S_i)$ for the i-th high-resolution image H_i and its respective super-resolution S_i stands for:

$$D(H_i, S_i) = Lev(H_i, S_i)/7 + (1 - SSIM(H_i, S_i)). \tag{2}$$

Here, the loss is weighted by the trade-off between visual quality and recognition rate using $D(H_i, S_i)$. This task is accomplished by measuring the Levenshtein distance (also known as edit distance) and comparing the SSIM score between the SR and HR images. The resulting value is scaled to the same magnitude as the squared error between the SR and HR images using α, thus avoiding dominance from either term.

Notably, any OCR model can be employed for LP recognition in this loss function. Such flexibility is attractive since we can straightforwardly use novel models as they are introduced. In this work, we explored the multi-task model proposed by Gonçalves et al. [7], as it was specifically designed for recognizing Brazilian LPs and has high efficiency.

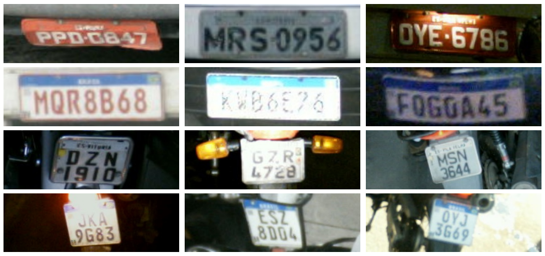


Fig. 3. Examples of cropped LPs from the RodoSol-ALPR dataset [20].

IV. Experiments

This section describes the experiments performed to validate the proposed approach. We first present our experimental setup and then report the results obtained.

A. Setup

We conducted our experiments using PyTorch on a computer with an AMD Ryzen 9 5950X CPU, 128 GB of RAM, and an NVIDIA Quadro RTX 8000 GPU (48 GB).

1) Dataset: we performed our experiments on LP images extracted from the RodoSol-ALPR dataset [20]. This dataset comprises 20,000 images taken by static cameras at pay rolls located in the Brazilian state of Espírito Santo.

Among the 20,000 images, there are 5,000 images of each of the following combinations of vehicle type and LP layout: (i) cars with Brazilian LPs, (ii) cars with Mercosur LPs, (iii) motorcycles with Brazilian LPs, and (iv) motorcycles with Mercosur LPs[2]. While all Brazilian LPs consist of three letters followed by four digits, the initial pattern adopted in Brazil for Mercosur LPs consists of three letters, one digit, one letter and two digits, in that order [20] (this is the pattern adopted on all Mercosur LPs in the RodoSol-ALPR dataset).

As far as we know, RodoSol-ALPR is the largest public dataset in terms of both Brazilian and Mercosur LPs. Fig. 3 shows some LPs cropped from the RodoSol-ALPR dataset. Observe the diversity of this dataset regarding several factors such as LP colors, lighting conditions, and character fonts.

The HR images used in our experiments were generated as follows. For each image from the RodoSol-ALPR dataset, we first cropped the LP region using the annotations provided by the authors. Afterward, considering the scope of this work, we used the same annotations to rectify each LP image so that it becomes more horizontal, tightly bounded, and easier to recognize [35]. The rectified image is the HR image.

For each HR image, we generate multiple LR images. Inspired by [18], we simulated the effect of an optical system with a lower resolution by iteratively applying random Gaussian noise to each HR image until the desired degradation level for a given LR image was reached. Intuitively, we measure the level of degradation of an LR image considering the SSIM score between it and the respective HR image.

2) Training: in the training stage, the LR and HR images were first padded to preserve the aspect ratio and then resized to 120 × 60 pixels with no upsample step on the subsequent SR outputs. We created four subsets at $]0.00, 0.10]$, $]0.10, 0.25]$, $]0.25, 0.50]$ and $]0.50, 0.75]$ SSIM intervals each with 8,000 and 4,000 images for training and validation, respectively. We also trained the proposed network and MPRNet [6] using the union of all subsets (i.e., $]0.00, 0.75]$ SSIM interval).

We used the Adam optimizer with an underlying learning rate of 10^{-4}, which decreases by a factor of 0.8 (up to 10^{-7}) when no improvement in the loss function is observed. The training stage stops after 5 epochs without loss improvement.

3) Testing: the proposed models were tested on the remaining images: 8,000 images from each of the $]0.00, 0.10]$, $]0.10, 0.25]$, $]0.25, 0.50]$, and $]0.50, 0.75]$ subsets. For the model and baseline trained with the $]0.00, 0.75]$ SSIM interval, the above subsets were combined. For each experiment, we report the number of correctly recognized LPs divided by the number of LPs in the test set. A correctly recognized LP means that all characters on the LP were correctly recognized.

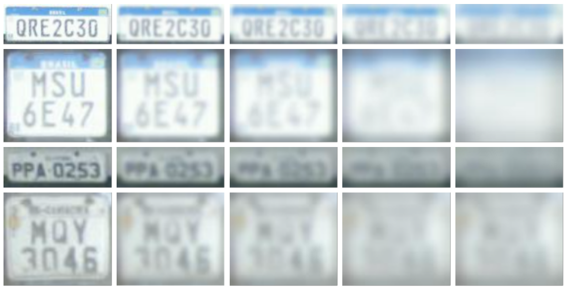


Fig. 4. Representative samples of the subsets used in our experiments. From left to right: 1 (original image), 0.75, 0.5, 0.25 and 0.1 SSIM scores.

B. Results

To illustrate how challenging is the problem at hand, the first section of Table I presents the recognition rates for the HR image and the respective LR images degraded by recursive Gaussian noise aiming for different SSIM interval scores, i.e., $]0.00, 0.10]$, $]0.10, 0.25]$, $]0.25, 0.50]$ and $]0.50, 0.75]$. Fig. 4 shows some representative examples of the images used in our experiments. Note that the LP characters are visually hard to distinguish in images with SSIM scores lower than 0.50. Observe that the recognition rates achieved on motorcycle LPs are higher than those achieved on car LPs in less degraded images ($]0.25, 0.50]$ and $]0.50, 0.75]$), but not in considerably degraded images ($]0.00, 0.10]$ and $]0.10, 0.25]$). This occurs because motorcycle LPs are generally smaller in size (having less space between the characters) and are often tilted [20]. Consequently, through the degradation process, the characters are mixed together, affecting the recognition performance.

In the second and third sections of Table I, we present the recognition rates obtained in SR images generated by the proposed network when trained on LR images with SSIM in the $]0.00, 0.10]$ and $]0.10, 0.25]$ intervals, respectively. As expected, considerably better recognition rates are achieved in

[2] Following previous works [7], [22], [24], we refer to "Brazilian" as the layout used in Brazil before the adoption of the Mercosur layout.

TABLE I
RECOGNITION RATES (%) ACHIEVED IN OUR EXPERIMENTS.

SSIM	Cars			Motorcycles			Cars & Motor.		
	All	≤ 6	≤ 5	All	≤ 6	≤ 5	All	≤ 6	≤ 5
No super-resolution									
HR	90.9	97.5	98.9	95.2	99.5	99.9	92.8	98.4	99.4
$]0.50, 0.75]$	88.7	96.6	98.6	93.6	99.2	99.8	90.9	97.7	99.1
$]0.25, 0.50]$	73.5	88.3	94.3	81.7	95.3	98.3	77.2	91.4	96.1
$]0.10, 0.25]$	18.4	35.9	53.4	17.7	39.1	59.0	18.1	37.3	55.9
$]0.00, 0.10]$	0.3	1.3	4.8	0.1	0.9	3.7	0.2	1.1	4.3
Proposed model trained with $]0.00, 0.10]$ SSIM images									
$]0.50, 0.75]$	36.7	62.0	78.5	29.7	52.0	68.5	33.5	57.5	74.0
$]0.25, 0.50]$	31.2	55.7	72.6	24.8	46.1	62.2	28.3	51.4	68.0
$]0.10, 0.25]$	26.9	50.2	66.2	21.7	43.7	62.0	24.5	47.3	64.3
$]0.00, 0.10]$	17.4	35.1	50.5	8.5	20.7	36.1	13.4	28.7	44.0
Proposed model trained with $]0.10, 0.25]$ SSIM images									
$]0.50, 0.75]$	56.0	74.7	85.6	55.0	74.8	84.7	55.6	74.7	85.2
$]0.25, 0.50]$	56.4	75.6	85.3	59.2	78.5	88.6	57.7	76.9	86.8
$]0.10, 0.25]$	60.3	79.1	87.4	52.6	76.7	88.6	56.8	78.0	88.0
$]0.00, 0.10]$	13.4	27.3	40.0	5.9	14.8	25.5	10.0	21.7	33.5
Proposed model trained with $]0.00, 0.75]$ SSIM images									
$]0.50, 0.75]$	90.6	97.4	98.9	93.8	98.9	99.7	92.0	98.0	99.3
$]0.25, 0.50]$	87.3	96.1	98.5	91.2	98.0	99.5	89.0	96.9	98.9
$]0.10, 0.25]$	69.3	85.7	92.9	65.5	85.5	93.6	67.6	85.6	93.2
$]0.00, 0.10]$	32.1	51.1	65.1	13.9	30.9	47.6	23.9	42.1	57.3
Proposed model & baselines trained and tested with $]0, 0.75]$ SSIM images									
Proposed	**69.8**	**82.6**	**88.9**	**66.1**	**78.3**	**85.1**	**68.1**	**80.7**	**87.2**
LR-LPR (no SR) [25]	61.4	78.0	86.5	47.0	68.8	80.4	54.9	73.9	83.7
MPRNet [6]	48.2	66.1	75.7	50.0	65.0	74.6	49.0	65.6	75.2
Average PSNR (dB) and SSIM for tests with $]0, 0.75]$ SSIM images									
			PSNR			SSIM			
Proposed			**26.4**			**0.89**			
MPRNet [6]			19.7			0.79			

SR images generated by the model trained on images from the same SSIM interval. However, the OCR network did not perform well on SR images generated by models trained on images in different SSIM intervals, especially those with better quality indices, i.e., $]0.25, 0.50]$ and $]0.50, 0.75]$. Based on the recognition results shown in the fourth section of Table I, we can state that the SR model trained on images with SSIM in the $]0.00, 0.75]$ interval generalized much better than models trained exclusively on images from $]0.00, 0.10]$ or $]0.10, 0.25]$.

In the next-to-last section of Table I, we compare the proposed architecture with MPRNet [6][3]. The OCR network [7] achieved considerably better results on images reconstructed by our SR model than MPRNet. We believe this is due to the capabilities of PS and PU in learning the best way to scale and reorganize channels within the image. We also report the recognition rates obtained by the OCR model proposed in [25], henceforth called LR-LPR, as it was specifically designed for recognizing low-resolution LPs. Note that we trained and tested LR-LPR as in [25], that is, using the degraded images (in the $]0.00, 0.75]$ SSIM interval) and not reconstructed ones. Although it achieved better recognition rates than we expected, the proposed approach (i.e., first reconstructing the degraded LP images using our SR model and then feeding the resulting images into an OCR network trained on HR LPs) still achieved significantly superior results.

As can be seen at the bottom of Table I, the proposed architecture achieved significantly better results than MPRNet [6] also considering the SSIM and PSNR (dB) metrics, reaching 0.89 and 26.4 dB against 0.79 and 19.7 dB, respectively.

Finally, Fig. 5 shows four LR images and the respective SR images generated by our architecture and by MPRNet [6]. The original image is also shown for better comparison.

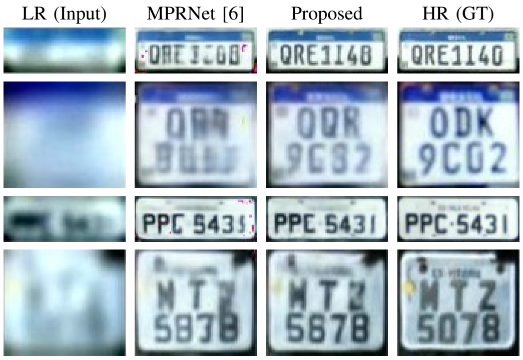


Fig. 5. Qualitative results on models trained with]0, 0.75] SSIM images.

The LR images shown in Fig. 5 are in the range of $]0, 10]$ SSIM; thus, the LPs are heavily degraded with little to no visible characters, representing a challenging reconstruction scenario for any SR model. In general, the proposed architecture performs better than MPRNet [6] on perceptual reconstruction. Observe that MPRNet tends to produce blurred edges, which in most cases mixes the characters with each other or with the LP background; this is very noticeable in the second row's image, where most LP characters are considerably blurred. In contrast, our model generates sharper character edges, which promotes better differentiation from the LP background. Also, its character reconstruction is not inconsistent, with incomplete lines or missing characters. When the SR model does not know which character to represent, it tends to hallucinate for the most congruent ones with respect to the LR input.

V. CONCLUSIONS

This paper proposes an extension to the MPRNet [6] architecture that achieves better performance on LP super-resolution by combining PS layers and attention modules. We proposed a new perceptual loss that combines the recognition results achieved by an OCR model with the SSIM metric between the LR and HR images for better LP reconstruction.

The main intuition behind our approach is to exploit channel reorganization and learning capabilities from the PS and PU layers for custom scale operations instead of translational invariance and interpolation methods. An autoencoder with PS and PU layers for shallow feature extraction was added as an input block to generate an attention mask with regions of interest within the image for reconstruction. Thus, by aggregating the mask and original input, we optimized computational resources and generated SR images with emphasis on relevant information. PLTFAM was proposed to better explore inter-channel relationship features and their position within the

[3] We implemented MPRNet ourselves, as its code has not been made public.

image. We exploited the PS and PU layers instead of the original TFAM *MaxPool*, *AvgPool* and upscaling ones.

All of our experiments were conducted on a public dataset with a wide variety of LP images. The results showed that recognition rates higher than those achieved by two baselines are achieved in images reconstructed by the proposed method.

In future work, we plan to build a large-scale dataset for LP super-resolution containing thousands of pairs of LR and HR images. More specifically, we aim to collect videos where the LP is perfectly legible on one frame but illegible on another. In this way, it would be possible to assess current approaches in real-world scenarios, as well as to develop new methods. We also intend to carry out experiments in cross-dataset setups to assess and eliminate the impact of dataset bias [36], [37].

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) (# 308879/2020-1 and # 309953/2019-7), 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] 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.

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

[3] 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*, pp. 1–19, 2022.

[4] K. Nasrollahi and T. B. Moeslund, "Super-resolution: a comprehensive survey," *Machine Vision and Applications*, vol. 25, pp. 1423–1468, 2014.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 184–199.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] L. F. Zeni and C. Jung, "Weakly supervised character detection for license plate recognition," in *Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2020, pp. 218–225.

[20] R. Laroca *et al.*, "On the cross-dataset generalization in license plate recognition," in *International Conference on Computer Vision Theory and Applications (VISAPP)*, Feb 2022, pp. 166–178.

[21] W. Weihong and T. Jiaoyang, "Research on license plate recognition algorithms based on deep learning in complex environment," *IEEE Access*, vol. 8, pp. 91 661–91 675, 2020.

[22] R. Laroca *et al.*, "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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] R. Laroca *et al.*, "Towards image-based automatic meter reading in unconstrained scenarios: A robust and efficient approach," *IEEE Access*, vol. 9, pp. 67 569–67 584, 2021.

[36] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp. 1521–1528.

[37] R. Laroca, M. Santos, V. Estevam, E. Luz, and D. Menotti, "A first look at dataset bias in license plate recognition," *arXiv preprint*, vol. arXiv:2208.10657, pp. 1–6, 2022, Accepted for presentation at the *Conference on Graphics, Patterns and Images (SIBGRAPI) 2022*.