Embedding Similarity Guided License Plate Super Resolution

Abderrezzaq Sendjasni, Mohamed-Chaker Larabi

Embedding Similarity Guided License Plate Super Resolution

Abderrezzaq Sendjasni*[a], Mohamed-Chaker Larabi[a]

[a] *CNRS, Univ. Poitiers, XLIM, UMR 7252, France*

Abstract

Super-resolution (SR) techniques play a pivotal role in enhancing the quality of low-resolution images, particularly for applications such as security and surveillance, where accurate license plate recognition is crucial. This study proposes a novel framework that combines pixel-based loss with embedding similarity learning to address the unique challenges of license plate super-resolution (LPSR). The introduced pixel and embedding consistency loss (PECL) integrates a Siamese network and applies contrastive loss to force embedding similarities to improve perceptual and structural fidelity. By effectively balancing pixel-wise accuracy with embedding-level consistency, the framework achieves superior alignment of fine-grained features between high-resolution (HR) and super-resolved (SR) license plates. Extensive experiments on the CCPD dataset validate the efficacy of the proposed framework, demonstrating consistent improvements over state-of-the-art methods in terms of $PSNR_{RGB}$, $PSNR_Y$ and optical character recognition (OCR) accuracy. These results highlight the potential of embedding similarity learning to advance both perceptual quality and task-specific performance in extreme super-resolution scenarios.

Keywords: Super-resolution, License plate, Convolutional neural networks, Embedding similarity, Contrastive learning.

1. Introduction

Single image super-resolution (SISR) is a well-known research field in computer vision focused on enhancing spatial resolution and visual fidelity of low-resolution images. Its significance lies in the ability to reconstruct high-resolution details from degraded visual data, thereby improving image quality across diverse applications, including digital photography [46, 23], medical imaging [44], and video surveillance [34]. In particular, SISR has been increasingly applied to enhance license plate (LP) images, where the clarity and legibility of such critical visual data are paramount for effective and reliable automated recognition systems.

*Corresponding author

License plate recognition (LPR) systems are integral components of modern surveillance, traffic management, and security applications [13]. However, the efficacy of these systems heavily relies on the quality of the captured images [37]. In real-world scenarios, LP images captured by surveillance cameras or other sources often suffer from visual degradations, such as low resolution, motion blur, and noise. As illustrated in Fig. 1, these issues make it challenging to accurately read the LPs, posing significant difficulties for LPR systems and compromising their accuracy and reliability. Among these challenges, the most critical is the distance at which the images are captured, leading to limited pixel resolution. When images are taken from a great distance, LPs appear smaller within the frame, drastically reducing the number of pixels representing the plate. This loss of resolution causes fine details to become invisible and unrecognizable upon zooming, making it difficult for both LPR systems and human users to interpret and recognize the characters accurately. To address this, advanced image enhancement techniques, such as super-resolution (SISR), are crucial for upscaling low-resolution images and preserving the key details necessary for reliable LPR.

Figure 1: Example of license plates taken under different conditions, showcasing difficulties to properly read the plates in some cases [22].

Despite recent advances in the field of SISR [35, 29, 26], LP super-resolution (LPSR) remains a significant challenge. The unique characteristics of LP images, such as small text, complex backgrounds, varying lighting conditions, and diverse fonts [10, 33, 36, 11], make it difficult for standard SISR models to perform robustly. Traditional SR methods often fail to reconstruct the fine details and sharp edges required to accurately recognizing characters on license plates. Deep learning-based methods, while more robust, often face challenges in striking the delicate balance between enhancing image quality and preserving critical textual information. Achieving high accuracy in LPR systems is crucial, particularly in applications such as law enforcement, where misrecognition or failure to identify a license plate can have serious consequences. This underscores the critical need to develop super-resolution (SR) methods specifically tailored to address these challenges.

Compared to the extensive research on SISR, studies specifically addressing LPSR remain limited. Most existing approaches rely on deep learning, primarily due to its ability to leverage inherent prior knowledge of natural scenes and pre-

serve image details more effectively than traditional methods like interpolation. For example, interpolation-based techniques such as bilinear and bicubic interpolation are simple and fast but often produce blurry images with a loss of fine details. These methods estimate new pixel values based on linear or cubic interpolation of neighboring pixels, resulting in smooth but less detailed outputs. In contrast, deep learning-based approaches, such as convolutional neural networks (CNNs) and generative adversarial networks (GANs), have demonstrated superior performance in LPSR [39]. Their ability to learn complex patterns enables them to reconstruct high-resolution images with finer details and enhance visual fidelity.

In this work, we address the limitations of existing LPSR methods by focusing on extreme super-resolution scenarios with a challenging scaling factor of x8. To tackle the reconstruction of severely degraded license plates with minimal pixel information, we propose a deep learning framework that combines residual dense blocks (RDBs) and channel attention mechanisms to enhance visual quality and textual detail recovery. Our approach integrates pixel-level and embedding-level losses, implemented through a Siamese network to align embeddings of high-resolution (HR) and super-resolution (SR) images. By doing so, the character recovery is improved along with the overall fidelity. Besides, the training strategy based on embedding similarity utilizes contrastive loss [12] to minimize the discrepancy between HR and SR embeddings. This strategy preserves fine details critical for character recognition. Extensive experiments conducted on the CCPD dataset [47], which includes real-world license plate images under diverse conditions, demonstrate the effectiveness and robustness of the proposed method. The primary contributions of this work can be summarized as:

- We developed a deep learning framework for extreme LPSR with a scaling factor of x8, leveraging residual dense blocks and channel attention mechanisms to enhance visual quality and recover fine details.
- We introduced the pixel and embedding consistency loss (PECL), which integrates pixel-level and embedding-level similarities. A Siamese network and contrastive loss are employed to align and constrain the similarity between HR and SR embeddings.
- We conducted a comprehensive evaluation of the proposed method on the CCPD dataset, demonstrating its robustness and effectiveness across diverse real-world conditions.

2. Related work

2.1. Single image super-resolution

Single image super-resolution (SISR) has seen extensive research in the past decade, primarily driven by deep learning techniques that aim to reconstruct high-resolution (HR) images from their low-resolution (LR) counterparts. Early

methods for SISR were based on interpolation techniques such as bilinear and bicubic interpolation, which, despite their simplicity, often resulted in blurred images with a loss of fine details [17]. These shortcomings prompted the development of more sophisticated methods like sparse coding-based models [48], and later, deep learning-based approaches such as convolutional neural networks (CNNs) and generative adversarial networks (GANs), which have demonstrated substantial improvements in both perceptual quality and quantitative performance metrics.

CNN-based models like SRCNN [6] introduced the concept of end-to-end learning for SISR, laying the groundwork for more complex architectures such as VDSR [18], EDSR [28], and RCAN [50], which leverage residual learning and attention mechanisms to improve the super-resolution performance. GAN-based methods, such as SRGAN [23], ESRGAN [46], and more recent works like Real-ESRGAN [45] and SwinIR [27], focus on enhancing the perceptual quality of the super-resolved images by employing adversarial learning and perceptual loss functions based on deep features [14] as well as vision transformers [7]. These methods have successfully generated SR images with sharper details and more visually appealing results compared to traditional interpolation-based methods.

Despite these advancements, standard SISR models often struggle when applied to domain-specific tasks such as license plate super-resolution (LPSR), where the primary goal is not just to improve image fidelity, but also to preserve critical textual and structural information that is vital for recognition tasks. Indeed, the challenges are more domain-specific and tied to the unique visual characteristics of license plate images, such as small fonts, varying lighting conditions, and complex backgrounds. Traditional SISR models, when directly applied to license plates, tend to fail in recovering the fine-grained details required for character recognition, particularly when dealing with extreme scaling factors such as x8 or higher.

2.2. License plate super-resolution

Several approaches have been proposed to address the previously mentioned challenges by incorporating domain-specific knowledge into the super-resolution pipeline. The work in [49] presented a multi-scale CNN tailored for LPSR, focusing on minimizing the mean squared error (MSE) between HR and super-resolved (SR) license plate images. While effective at enhancing the overall image quality, this approach still struggled with preserving fine textual details, which are critical for accurate LPR. Recent advances in deep learning have paved the way for more sophisticated LPSR models. The work in [31] introduced a GAN-based architecture that incorporates a gradient profile prior to emphasize character boundaries, thus improving the contrast between the characters and the background. Similarly, the authors in [39] extended the SRGAN framework by adding an optical character recognition (OCR)-based loss function, which directly computes the recognition error between HR and SR images, thus ensuring that the generated SR images retain the legibility of the text. This approach also utilizes perceptual loss functions based on VGG-19 [45, 15] to improve both the visual quality and recognition accuracy of the license plates.

Other notable works have explored the use of character-based perceptual losses, where the super-resolution process is guided by intermediate feature representations learned by an OCR network. For example, the work in [24] proposed a loss function based on character classification features, while [38] employed the Levenshtein distance to measure discrepancies between predicted and ground-truth characters. These methods highlight the importance of integrating recognition tasks directly into the loss function, allowing the models to not only enhance the visual quality but also optimize for accurate character recognition.

Despite these advances, several challenges remain in LPSR research. First, many existing approaches are designed for moderate upscaling factors (*i.e.* x2 or x4) and fail to generalize to extreme cases such as x8 and beyond, where the license plate details are severely degraded. Moreover, current LPSR models often struggle to balance the trade-off between improving perceptual quality and preserving critical textual information. While OCR-guided loss functions have shown promise in mitigating this issue, there is still room for improvement, particularly in cases where the input LR images suffer from extreme distortions, such as motion blur or severe compression artifacts.

This study addresses the limitations of existing LPSR approaches by introducing a novel training strategy that incorporates both pixel-wise and embedding-level losses. By leveraging a combination of perceptual and contrastive loss functions, the proposed method ensures robust super-resolution even at extreme scaling factors, achieving a balance between perceptual quality and the preservation of textual and structural details. The following sections detail the methodology and demonstrate its effectiveness compared to state-of-the-art techniques.

3. Proposed Methodology

3.1. Problem formulation

The task of SR aims to reconstruct an HR image $\mathbf{I}_{HR} \in \mathbb{R}^{h \times w \times 3}$ from its LR counterpart $\mathbf{I}_{LR} \in \mathbb{R}^{H \times W \times 3}$, where typically $h = s \cdot H$ and $w = s \cdot W$, and $s \in \mathbb{Z}^+$ is the upscaling factor. The goal of SR is to recover fine-scale details lost during image degradation processes such as downscaling, compression, or noise corruption. Thus, the problem we try to solve can be formulated as a learning task where the objective is to estimate a function $f_\theta : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{h \times w \times 3}$, parameterized by θ, that maps the LR image $\mathbf{I}_{LR}$ to its super-resolved version $\mathbf{I}_{SR}$:

$$\mathbf{I}_{SR} = f_\theta(\mathbf{I}_{LR}), \tag{1}$$

where the learning task can be framed as an optimization problem with the objective to minimize a composite loss function $\mathcal{L}_{\text{total}}$ over a set of parameters θ, subject to constraints imposed by the nature of the task. Formally, the goal is to find:

$$\theta^* = \arg\min_{\theta} \mathcal{L}_{\text{total}}(\mathbf{I}_{HR}, f_\theta(\mathbf{I}_{LR})), \tag{2}$$

where $f_\theta(\cdot)$ is the parameterized mapping from LR to SR images. The total loss function $\mathcal{L}_{\text{total}}$ is a weighted combination of several distinct loss terms:

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^{N} \lambda_i \mathcal{L}_i, \tag{3}$$

where $\mathcal{L}_i$ represents different loss components corresponding to specific properties or objectives that must be optimized, and $\lambda_i \in \mathbb{R}^+$ are scalar weights that control the contribution of each term.

In our method, we model the problem using a patch-based approach. Rather than processing entire HR and LR images, we operate on smaller patches extracted from these images. Formally, let $\mathbf{I}^{HR} \in \mathbb{R}^{H \times W \times C}$ and $\mathbf{I}^{LR} \in \mathbb{R}^{h \times w \times C}$ represent the HR and LR images respectively, where $H > h$ and $W > w$. Instead of directly mapping the whole LR image to its HR counterpart, we extract small overlapping patches $\mathbf{P}_i^{HR} \in \mathbb{R}^{p \times p \times C}$ and $\mathbf{P}_i^{LR} \in \mathbb{R}^{p' \times p' \times C}$ from both $\mathbf{I}^{HR}$ and $\mathcal{I}^{LR}$.

Thus, for a given LR patch $\mathbf{P}_i^{LR}$, the goal is to reconstruct its corresponding HR patch $\mathbf{P}_i^{SR} \in \mathbb{R}^{p \times p \times C}$ using a learned mapping f_θ, such that:

$$\mathbf{P}_i^{SR} = f_\theta(\mathbf{P}_i^{LR}), \tag{4}$$

where θ represents the parameters of our SR model. The final SR image $\mathbf{I}^{SR}$ is obtained by aggregating the predicted patches $\mathbf{P}_i^{SR}$ across the image domain.

The patch-based strategy allows for finer local structure preservation, a better handling of complex patterns, and efficient training on smaller receptive fields. It also enables the network to focus on local dependencies and details, which are crucial for reconstructing high-frequency information from low-resolution patches. Besides, this formulation ensures that the learned model can generalize better across varying image scales.

3.2. LPSR model overview

To achieve high-quality super-resolution, the proposed model is built upon foundational observations derived from state-of-the-art Single Image Super Resolution (SISR) techniques, specifically incorporating Residual Dense Blocks (RDBs) [51] and channel attention mechanisms [4]. The architecture is designed to progressively refine low-resolution inputs $\mathbf{P}_{\text{LR}} \in \mathbb{R}^{3 \times W \times H}$ into high-resolution outputs $\mathbf{P}_{\text{SR}} \in \mathbb{R}^{3 \times W \times H}$, where W and H denote the width and height, respectively, through a series of strategically implemented processing stages.

The input LR patch is first processed by a 3×3 convolutional layer to obtain shallow features:

$$\mathbf{F}_0 = \text{Conv}_{3\times3}(\mathbf{P}_{\text{LR}}), \tag{5}$$

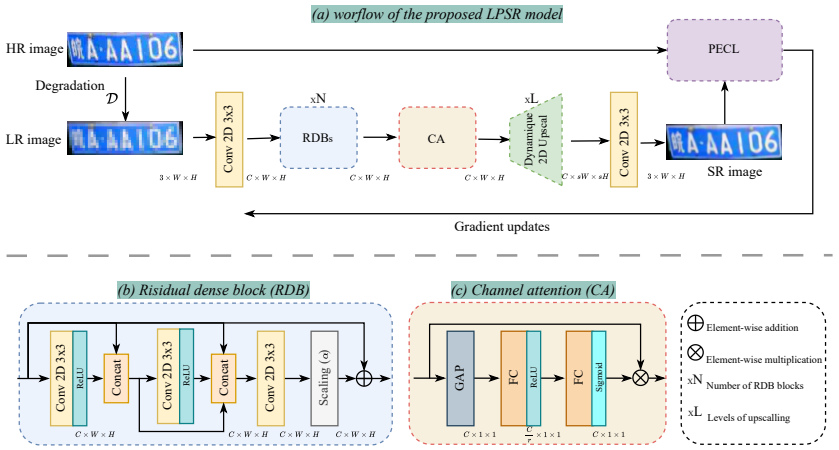

Figure 2: Workflow of the proposed LPSR model. Residual dense blocks (RDBs) capture complex hierarchical features using dense connections and residual learning. Channel attention (CA) focuses on the most informative feature channels. The pixel and embedding consistency loss (PECL) computes pixel-to-pixel and embeddings fidelity for the gradient updates.

where $\mathbf{F}_0$ represents the feature map produced by the convolution operation. The 3×3 kernel size is employed to balance local structure and computational efficiency, enabling the extraction of essential features such as edges and textures, which are critical for subsequent processing stages.

Following the initial feature extraction, the low-resolution features $\mathbf{F}_0$ are fed into a serial of N Residual Dense Blocks (RDBs), which are designed to capture complex hierarchical features through dense connections and residual learning, see Fig. 2 (b). The output of the i-th RDB can be expressed as:

$$\mathbf{F}_i = \mathbf{F}_{i-1} + \mathrm{RDB}(\mathbf{F}_{i-1}), \tag{6}$$

where $RDB(\cdot)$ represents the operations performed within the i-th RDB, including convolution, activation, and feature concatenation. A crucial component of each RDB is the scaling operation applied to the output before it is added back to the residual input:

$$\mathbf{F}_i = \alpha \cdot \mathrm{Conv}_{\mathrm{last}}(\mathbf{F}_{i-1}) + \mathbf{F}_{i-1}, \tag{7}$$

where $\mathrm{Conv}_{\mathrm{last}}(\cdot)$ denotes the last convolution operation within i-th RDB and α is a learnable scaling factor that adjusts the contribution of the RDB output relative to the residual input. This scaling mechanism enhances the ability of the model to control the influence of each residual dense learning output, allowing for adaptive learning of the feature importance during the training process. By incorporating this scaling operation, the RDBs effectively facilitate the retention of rich information and enhance gradient flow during training, addressing challenges such as degradation and the vanishing gradient problem.

The architecture of each RDB allows for the concatenation of feature maps from previous layers, facilitating rich information retention and enhancing the gradient flow during training [51]. Besides, such a design addresses the challenges associated with deep networks, such as the vanishing gradient problem, by maintaining a direct path for gradient propagation.

The Channel Attention (CA) block is pivotal in enhancing the model's ability to prioritize the most informative feature channels [4], significantly boosting the overall performance of the super-resolution task. The CA mechanism operates on the feature maps generated by the final RDB block, as depicted in Fig. 2 (c). These feature maps are denoted as $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$, where C is the number of channels, and H and W represent the height and width of the feature maps, respectively. The CA block employs a global average pooling operation to capture the global context, resulting in a channel descriptor $z \in \mathbb{R}^{C \times 1 \times 1}$:

$$z = \mathrm{GAP}(\mathbf{F}) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{F}(:, i, j), \tag{8}$$

where GAP denotes the global average pooling operation. This descriptor is then passed through two fully connected (FC) layers to generate the attention weights. The first layer reduces the dimensionality of the channel descriptor:

$$z_1 = \mathrm{ReLU}(\theta_{\mathrm{FC1}} \cdot z + b_1), \tag{9}$$

where $\theta_{\mathrm{FC1}} \in \mathbb{R}^{\frac{C}{r} \times C}$ is the weight matrix of the first FC layer, b_1 is the bias term, and r is the reduction ratio. The output of this layer is then passed through the second fully connected layer to restore the original dimensionality:

$$z_2 = \sigma(\theta_{\mathrm{FC2}} \cdot z_1 + b_2), \tag{10}$$

where $\theta_{\mathrm{FC2}} \in \mathbb{R}^{C \times \frac{C}{r}}$ and b_2 are the weight matrix and bias of the second FC layer, respectively, and σ represent the sigmoid activation function. The resulting attention vector $z_2 \in \mathbb{R}^C$ is then reshaped and used to scale the original feature maps:

$$\mathbf{F}_{\mathrm{CA}} = \mathbf{F} \otimes z_2, \tag{11}$$

where $\mathbf{F}_{\mathrm{CA}}$ represents the output of the CA block after applying the attention weights. This block enhances the network's representational capacity by emphasizing informative channels and suppressing less relevant ones, resulting in improved feature extraction for subsequent processing stages. Integrating the CA block into the super-resolution architecture is crucial, as it aligns with the model's goal of generating high-quality images by selectively focusing on critical features that significantly impact perceptual quality.

Subsequently, the proposed model incorporates a dynamic upsampling mechanism through a series of 2D transposed convolution operations [8]. This approach incrementally doubles the spatial dimensions of the feature maps, effectively enhancing the spatial resolution while preserving essential details. The

multi-stage upsampling strategy enables the model to progressively refine and upscale the output patches across multiple levels, which is critical for mitigating artifacts and improving overall image fidelity. Therefore, the upscaled features $\mathbf{F}_{\mathrm{UP}} \in \mathbb{R}^{C \times sW \times sH}$ are obtained by:

$$\mathbf{F}_{\mathrm{UP}} = \mathrm{ReLU}\left(\mathrm{ConvT}^{L}(\mathbf{F}_{\mathrm{CA}})\right), \tag{12}$$

where $\mathbf{F}_{\mathrm{CA}}$ represents the feature maps obtained from the channel attention block. L denotes the number of upsampling stages required, calculated as $L = \log_2(s)$, with s being the scale factor. The use of transposed convolutions ConvT allows for the integration of learned features from the preceding layers, ensuring that the generated high-resolution output retains the rich structural and contextual information from the low-resolution input.

Finally, the model ends with a convolutional layer that further refines the output feature maps to produce the final SR patch. This layer utilizes a 3×3 convolution operation to seamlessly integrate the features extracted in the preceding stages into a coherent and high-fidelity output. The final SR patch is obtained as:

$$\mathbf{P}_{\mathrm{SR}} = \mathrm{Conv}(\mathbf{F}_{\mathrm{up}}), \tag{13}$$

where $\mathbf{P}_{\mathrm{SR}}$ denotes the super-resolved patch, and Conv represents the 3×3 convolution operation applied to the upsampled feature maps $\mathbf{F}_{\mathrm{up}}$.

This final convolution serves multiple purposes. It consolidates the features learned through the RDBs and the CA mechanism, effectively integrating high-level representations with spatial information. Additionally, it plays a crucial role in mitigating potential artifacts introduced during the upsampling stages, ensuring that the output not only achieves the target resolution but also maintains visual consistency and high quality.

3.3. Pixel and embedding consistency loss

To effectively enhance the performance of the LPSR model, we propose a comprehensive loss function that combines pixel-wise loss with embedding similarity learning through a Siamese network architecture [20]. The pixel and embedding consistency loss (PECL) function $\mathcal{L}_{\mathrm{PEC}}$ is formulated as a weighted sum of the mean squared error (MSE) loss $\mathcal{L}_{\mathrm{pixel}}$ and the Contrastive loss $\mathcal{L}_{\mathrm{contrastive}}$, which incorporates the embedding similarity component.

Pixel-wise loss: it quantifies the discrepancy between the super-resolved output $\mathbf{P}_{\mathrm{SR}}$ and the corresponding high-resolution target $\mathbf{P}_{\mathrm{HR}}$. This loss is computed using the MSE, expressed as:

$$\mathcal{L}_{\mathrm{pixel}} = \frac{1}{K}\sum_{i=1}^{K}(\mathbf{P}_{\mathrm{SR}}^{(i)} - \mathbf{P}_{\mathrm{HR}}^{(i)})^2, \tag{14}$$

where K represents the total number of pixels in the image patch, and $\mathbf{P}_{\mathrm{SR}}^{(i)}$ and $\mathbf{P}_{\mathrm{HR}}^{(i)}$ denote the pixel values at the i-th position for the super-resolved

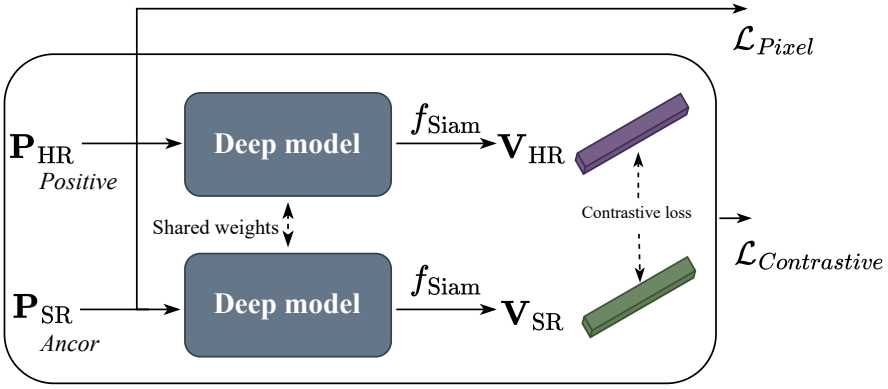

Figure 3: Illustration of the Siamese network architecture used for pixel and embedding consistency loss. The network comprises two identical sub-networks extracting embeddings from the super-resolved ($\mathbf{P}_{\mathrm{SR}}$) and high-resolution ($\mathbf{P}_{\mathrm{HR}}$) patches. The distance D between the embeddings $\mathbf{V}_{\mathrm{SR}}$ and $\mathbf{V}_{\mathrm{HR}}$ is computed in the embedding space, informing the Contrastive loss $\mathcal{L}_{\mathrm{Contrastive}} = \max(m - D, 0)^2$, which reinforces the model's capacity to preserve key features across resolutions.

and high-resolution patches, respectively. This loss function effectively captures the average squared differences between corresponding pixel values, promoting fine-grained accuracy in pixel representation.

Contrastive loss: To ensure that the super-resolved patches maintain key features characteristic of the high-resolution patches, we implement a Siamese network [20] to extract embeddings from both $\mathbf{P}_{\mathrm{SR}}$ and $\mathbf{P}_{\mathrm{HR}}$. A Siamese network comprises two identical sub-networks that share the same architecture and parameters, allowing for direct comparison of the generated embeddings.

The Siamese network architecture leverages a pre-trained ResNet-18 model, with the final fully connected layer replaced to yield embeddings of size 128. For an input pair $(\mathbf{P}_{\mathrm{SR}}, \mathbf{P}_{\mathrm{HR}})$ (as shown in Fig. 3), the network outputs two embeddings:

$$\mathbf{V}_{\mathrm{SR}}, \mathbf{V}_{\mathrm{HR}} = f_{\mathrm{Siam}}(\mathbf{P}_{\mathrm{SR}}, \mathbf{P}_{\mathrm{HR}}; \theta_{\mathrm{Siam}}, d), \tag{15}$$

where θ_{Siam} represents the parameters of the Siamese network and d the size of the embedding. To maintain consistent distance magnitudes across samples and prevent the embeddings from growing arbitrarily large, an L2 normalization step is applied to the output embedding. This normalization constrains the embeddings to a unit hypersphere, which improves convergence and training stability. Therefore, the normalized embedding $\tilde{\mathbf{V}}$ is given by:

$$\tilde{\mathbf{V}} = \frac{\mathbf{V}}{\|\mathbf{V}\|_2}, \tag{16}$$

where $\|\mathbf{V}\|_2$ is the L2 norm of $\mathbf{V}$. This ensures that all embeddings lie on a consistent scale.

The objective of this architecture is to generate similar embeddings for the super-resolved and high-resolution patches, reflecting their inherent similarity. To achieve this, we employ the Contrastive loss function $\mathcal{L}_{\text{Contrastive}}$, which encourages similarity in the embedding space. This loss is simplified to:

$$\mathcal{L}_{\text{Contrastive}} = \max(m - D, 0)^2, \tag{17}$$

where D is the Manhattan distance (ℓ_1-norm) between the embeddings of the super-resolved and high-resolution patches:

$$D = \|\tilde{\mathbf{V}}_{\text{SR}} - \tilde{\mathbf{V}}_{\text{HR}}\|_1. \tag{18}$$

In this context:

- m represents a margin, a hyperparameter that establishes a threshold distance between the embeddings, set to 2 in this study.
- The squaring operation $(m - D)^2$ imposes a heavier penalty for larger deviations, thereby reinforcing the model's capacity to minimize the distance D when it is below the margin m.

The Manhattan distance (ℓ_1-norm) offers several advantages for measuring embedding similarity, particularly in high-dimensional feature spaces. Unlike the Euclidean distance for instance, which squares differences and can amplify the influence of outliers, the Manhattan distance computes the sum of absolute differences, making it more robust to noisy or extreme feature values. This property is beneficial when embeddings exhibit sparsity or when certain dimensions dominate due to variability in the data. Additionally, the Manhattan distance treats each feature dimension independently, which aligns well with many neural embedding spaces where feature contributions vary. Providing stable gradients also facilitates smoother optimization during training, improving alignment and generalization [1, 9].

The focus on embedding similarity loss is particularly relevant for applications such as optical character recognition (OCR) [3], automatic number plate recognition (ANPR) [16], and vehicle identification. By forcing the model to minimize the distance between embeddings of super-resolved and high-resolution images, the embedding similarity loss ensures that the reconstructed images align more closely with their high-resolution counterparts in the feature space. This alignment is crucial for preserving distinctive features necessary for accurate recognition and identification, as it guarantees that critical details are reconstructed, enhancing fidelity and perceptual quality.

Total loss: The total loss function, $\mathcal{L}_{\text{PECL}}$ is designed as a weighted sum of two complementary components: the pixel-wise loss (L_{pixel}) and the contrastive loss ($L_{\text{contrastive}}$). The pixel-wise loss ensures fidelity at the pixel level by minimizing the MSE. In contrast, the contrastive loss focuses on aligning the embeddings of the predicted super-resolved image and the high-resolution target in the feature space, minimizing the distance between their representations. Together, these losses enforce both pixel-level accuracy and feature-level

consistency, improving the perceptual and semantic quality of the reconstructed images. The total loss is expressed as:

$$\mathcal{L}_{\text{PECL}} = w_{\text{pixel}} \cdot \mathcal{L}_{\text{pixel}} + w_{\text{contrastive}} \cdot \mathcal{L}_{\text{contrastive}}, \tag{19}$$

where w_{pixel} and $w_{\text{contrastive}}$ are the weights that govern the contributions of the pixel-wise and contrastive losses, respectively. These weights are learnable parameters constrained within the range $0, 1$ to ensure balanced optimization. Furthermore, they are designed to satisfy $w_{\text{pixel}} + w_{\text{contrastive}} = 1$, ensuring that the total weight is dynamically distributed between the two loss components. To maintain valid weight ranges, $\mathcal{L}_{\text{PECL}}$ clips w_{pixel} and $w_{\text{contrastive}}$ within $0, 1$ during training. This ensures stability and prevents either loss component from dominating excessively. By enforcing this balance, the total loss function effectively combines the strengths of pixel-level fidelity and feature-level alignment, enabling the model to achieve sharper reconstructions while preserving semantic consistency.

4. Experiments

4.1. Datasets and Implementation Details

Dataset: We use the Chinese city parking dataset (CCPD) [47], which contains over 200k images of license plates captured in various real-world conditions, including different angles, distances, and lighting scenarios. This dataset is particularly suitable for our study as it provides diverse examples of license plates that challenge traditional recognition systems. We select 100k from the CCPD dataset to train the model and 1k for testing and validation. This selection ensures the model has sufficient data for learning, while maintaining a separate subset for unbiased performance evaluation. Each image is segmented into patches of size 64×64 pixels.

The training dataset exhibits diverse capturing conditions, as illustrated in Fig. 4, including variations in angles, distances, and lighting. Additionally, many of the HR images suffer from significant distortions, further complicating the super-resolution task and emphasizing the need for robust models capable of handling degraded inputs.

Figure 4: Examples from the training dataset showcasing diverse capturing conditions, including variations in angles, distances, and lighting.

As described in Sec.3.2, the LPSR model processes input patches instead of full images. This patch-based design is a widely adopted practice in super-resolution (SR) tasks[6, 27, 45], as it enhances computational efficiency while

enabling the model to focus on localized features. Moreover, it effectively handles the inherent variability in resolution and aspect ratios of real-world license plate images, ensuring robust generalization across diverse input conditions. Fig. 5 showcases sample pairs of high-resolution (HR) and low-resolution (LR) patches. Given that the original images already exhibit visual degradations, the degradation process $\mathcal{D}$ is limited to downscaling. In this work, we address the challenging task of x8 scaling, categorized as extreme super-resolution due to the significant loss of visual information in LR inputs.

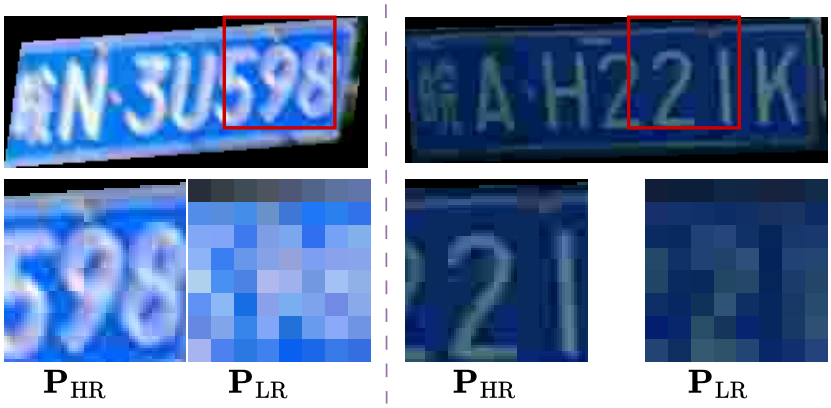

Figure 5: Examples of training patches and the degradation process applied, highlighting the downscaling operation used to simulate low-resolution inputs.

Implementation Details: The LPSR model is implemented using the PyTorch library [2] and trained on a server equipped with an Intel Xeon Silver 4208 2.1GHz CPU, 192GB of RAM, and an Nvidia Tesla V100S GPU with 32GB of memory. We train the model for 1000k iterations, a batch size of 128. We employ the Adam optimizer [19] to update the parameters of the model, with an initial learning rate $1e-4$. To adaptively adjust the learning rate throughout the training process, we utilize a cosine annealing learning rate scheduler [30]. The scheduler gradually decreases the learning rate over the iterations following a cosine curve. This dynamic adjustment aids in fine-tuning the model's parameters as training progresses.

Evaluation criteria: The evaluation of LPSR models focuses on two key aspects: visual quality and character recognition performance. Visual quality assesses how closely the SR images resemble the ground truth HR images, ensuring the restored images are perceptually faithful to the originals. For this purpose, peak signal-to-noise ratio (PSNR) is measured on both the full-color RGB channels ($PSNR_{RGB}$) and the luminance channel ($PSNR_Y$), with the latter being particularly significant due to its strong correlation with human visual perception. In addition to visual fidelity, the effectiveness of the super-resolution model in enhancing character readability is evaluated through OCR. PaddleOCR [43], a widely-used open-source OCR system [40, 41, 32, 21], is

employed to read characters on the SR images. The performance is quantified using the confidence scores provided by PaddleOCR, reflecting the accuracy and reliability of character recognition. This dual evaluation framework ensures a rebust assessment of the efficiency of the LPSR model.

4.2. Results and discussion

4.2.1. Performance comparison

The performance of the proposed SR model is evaluated in comparison to the baseline Bicubic and SOTA methods, including SRCNN [6], MSRN [25], ESPCN [42], ESRGAN [46], TBSRN [5], and SwinIR [27] using PSNR, PSNRy, and OCRρ. Table 1 presents the results of this comparison, where the median values, along with the standard deviations, are reported over the testing set. PSNR and PSNRy provide a quantitative measure of image fidelity, with higher values indicating better image reconstruction quality. OCRρ reflects the ability to improve text recognition accuracy, with higher values signifying better recognition performance.

Table 1: Performance comparison of the proposed model in terms of PSNR, PSNRy, and OCRρ with various SR methods. The median (± standard deviation) over the testing set is reported. The best and second best performance are respectively highlighted in **bold red** and **bold blue**.

SR Method	#p	PSNR [dB]	PSNRy [dB]	OCRρ
Bicubic	-	18.53 (± 2.78)	18.70 (± 2.82)	0.697 (± 0.19)
SRCNN [6]	57K	19.57 (± 2.72)	19.84 (± 2.78)	0.665 (± 0.21)
MSRN [25]	6M	19.78 (± 2.74)	19.84 (± 2.82)	0.656 (± 0.20)
ESPCN [42]	800K	23.52 (± 2.47)	24.11 (± 0.61)	0.848 (± 0.14)
ESRGAN [46]	16M	18.76 (± 2.26)	19.01 (± 2.33)	0.812 (± 0.17)
TBSRN [5]	12M	**23.76** (± 2.48)	**24.42** (± 2.63)	**0.855** (± 0.15)
SwinIR [27]	11M	23.56 (± 2.50)	24.18 (± 2.65)	0.850 (± 0.14)
Ours	2M	**25.13** (± 2.46)	**25.92** (± 2.62)	**0.852** (± 0.13)

The performance trends in Table 1 reflect the characteristics and design choices of each SR model. The proposed model outperforms others in PSNR (25.13 dB) and PSNRy (25.92 dB) due to its integration of pixel-wise and embedding-level losses, which balance the reconstruction of high-frequency details with perceptual quality. This dual-focus approach allows the model to deliver superior visual fidelity and structural accuracy, particularly in luminance-sensitive regions, as compared to other state-of-the-art methods. Furthermore, its competitive OCRρ score (0.852) indicates that it effectively preserves textual information, which is critical for license plate recognition tasks.

TBSRN [5], the second-best performer in PSNR and PSNRy, scores the best OCR accuracy (0.855) due to its architecture tailored for scene text super-resolution, which emphasizes text preservation. However, its marginally lower PSNR compared to the proposed model suggests that while it is optimized for text readability, it lags for the overall image fidelity. Similarly, SwinIR [27],

leveraging the hierarchical transformer-based structure, achieves balanced performance across metrics, with strong OCRρ (0.850) and PSNR (23.56 dB) but slightly lags behind in reconstructing fine luminance details, as reflected in its PSNRy (24.18 dB).

ESPCN [42], designed for real-time applications with lightweight upsampling, demonstrates competitive OCR accuracy (0.848) and improved PSNR (23.52 dB) compared to earlier methods like SRCNN [6] and MSRN [25]. However, its simpler architecture limits its ability to fully reconstruct complex textures or high-frequency details, resulting in performance that falls short of more advanced models. ESRGAN [46], despite employing adversarial loss to enhance perceptual quality, achieves only moderate OCRρ (0.812) and low PSNR (18.76 dB), likely due to its emphasis on producing visually pleasing textures at the expense of structural and text-based details.

Traditional methods, such as Bicubic, SRCNN, and MSRN, exhibit the weakest performance across all metrics. Bicubic interpolation's simplicity leads to significant oversmoothing, which degrades both image fidelity and text clarity. SRCNN and MSRN, though early milestones in deep-learning-based SR, lack advanced features like attention mechanisms or perceptual losses, making them less effective for domain-specific tasks like license plate recognition.

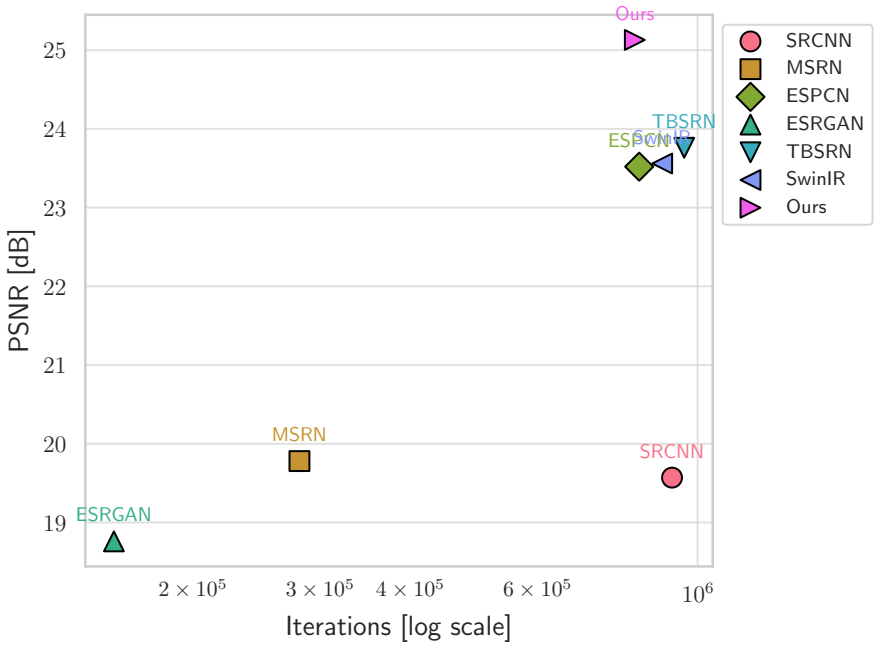

Figure 6: Performance in terms of PSNR *vs.* the number of iterations (on a logarithmic scale) required for convergence during training.

The superiority of the proposed model stems from its ability to dynamically

balance pixel-level fidelity and embedding similarity, effectively addressing the trade-off between visual quality and text preservation. Its design demonstrates the importance of combining domain-specific considerations with state-of-the-art architectural and loss functions to achieve robust performance in challenging SR scenarios.

We further analyzed the performances in terms of PSNR on the testing set versus the number of iterations required for each model to converge. All models were trained for 1000k iterations, with the best version saved at the highest PSNR during training. The plot in Fig. 6 illustrates the PSNR *vs.* number of iterations (on a logarithmic scale). A clear correlation between the number of iterations and the resulting PSNR values can be observed, except for SRCNN.

Models that required more iterations to converge generally delivered higher PSNR on the testing set. This trend highlights the importance of longer training for deeper models. For instance, ESPCN, TBSRN, and SwinIR required approximately 900k iterations to converge, achieving PSNR values of 23.52 dB and 23.76 dB, respectively. These models' performances indicate their ability to capture complex features from the dataset directly correlates with their extended training duration. Similar behavior can be seen for the proposed model. MSRN and ESRGAN converged more quickly with fewer iterations (approximately 281k and 155k iterations, respectively), and achieved lower PSNR values (19.78 dB and 18.76 dB, respectively). This suggests that despite their faster convergence, these models struggled to learn the necessary low-level features for the LPSR task from the training set, which resulted in a suboptimal generalization on the testing set. Regarding SRCNN behavior, it simple architecture, which lacks the capacity for learning complex features for extreme SR, limits its performance, even with prolonged training

These observations emphasize the trade-off between convergence speed and model performance in image super-resolution tasks. Models that require longer training tend to produce better results on unseen data, confirming the need for adequate training to fully exploit the potential of more complex architectures. In contrast, quicker-converging models may face challenges in terms of learning the finer details, potentially leading to a degradation in performance.

To further compare the performance of the proposed model with state-of-the-art methods, we provide a qualitative comparison in Fig. 7. This comparison showcases the visual differences in SR outputs across various models compared to the proposed one, including Bicubic interpolation, SRCNN, MSRN, ESPCN, ESRGAN, TBSRN, and SwinIR. The results indicate significant variability in the restoration of fine details and overall visual fidelity.

Bicubic interpolation serves as a baseline, delivering overly smooth images with a lack of texture recovery, underscoring its inability to reconstruct high-frequency details. SRCNN, being one of the earliest SR models, provides slightly sharper results but still struggles with generating realistic textures, leading to noticeable artifacts in regions with fine details. MSRN and ESPCN demonstrate improved performance, with MSRN showing better preservation of structural information and ESPCN excelling in edge sharpness. However, both models exhibit some limitations in producing natural textures.

Figure 7: Qualitative comparison with state-of-the-art methods on various samples, taken under different conditions.

ESRGAN's outputs, known for its GAN-based approach, introduce more realistic details and textures but often at the cost of over-enhanced or unnatural artifacts in certain areas, especially on less complex regions. TBSRN and SwinIR deliver a strong balance between detail recovery and artifact suppression, with SwinIR slightly outperforming TBSRN in preserving structural consistency across diverse regions of the images. The proposed model achieves the best results overall, demonstrating superior texture restoration, edge sharpness, and fidelity to high-resolution ground truths, with minimal artifacts and a more natural appearance.

4.2.2. Ablation experiment

We conduct an ablation study to verify the effectiveness of the main components of the proposed method, focusing on three key aspects: (1) the comparison

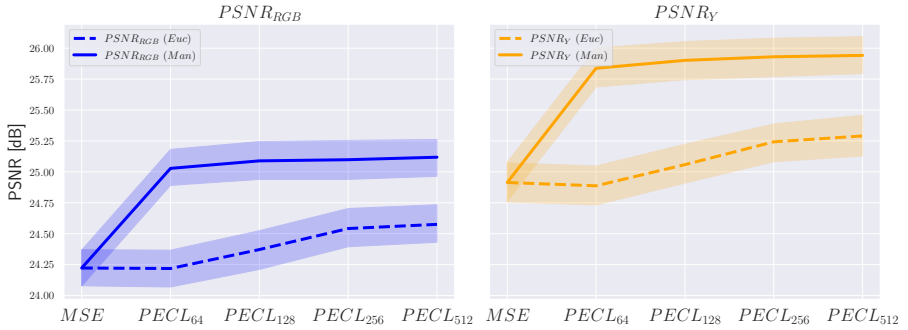

Figure 8: Performance trends of $PSNR_{RGB}$ and $PSNR_Y$ across different configurations: embedding distance measures (Euc: Euclidean, Man: Manhattan) and dimensionality $PECL_d$ with d the size of the embeddings. The solid lines represent the mean values, while the shaded areas around them indicate the 95% confidence intervals, reflecting the variability or uncertainty of the measurements for each configuration.

between the traditional MSE loss and the proposed PECL, (2) the impact of the embedding dimensionality d used for the contrastive loss, and (3) the choice of distance metrics (Euclidean vs. Manhattan) for measuring embedding similarity.

To evaluate the impact of embedding dimensionality and distance measures on the LPSR model performance, we conducted a comparative analysis using PSNR and PSNRy. The evaluation spans embedding dimensions of 64, 128, 256, and 512, trained with Euclidean and Manhattan distances. These metrics provide insights into both the fidelity and perceptual quality of the reconstructed images. The results, depicted in Fig. 8, highlight the influence of these factors on model performance on the testing set.

From the PSNR curves, the first observation that emerges is that the MSE-based loss lags behind the proposed loss function despite its natural alignment with PSNR, which inherently measures pixel-to-pixel fidelity. This is true independently from the used embedding distance measure. The proposed loss function combines pixel fidelity with embedding similarity, offering a more holistic representation. The adaptive learning mechanism, which dynamically balances the weights of MSE and contrastive losses during training, further boosts model performance. This dual-focus scheme enables the model to effectively capture both fine-grained image details and meaningful feature relationships, resulting in superior super-resolution outcomes.

By comparing among the used distances, the curves highlight clear trends. For PSNR, the Manhattan distance achieves consistently higher values compared to Euclidean across all dimensionalities, with a peak at d=512 where Manhattan reaches approximately 25.0 dB compared to 24.0 dB for Euclidean. Similarly, for PSNRy, Manhattan distance outperforms Euclidean, achieving a maximum of approximately 26.0 dB, compared to 25.0 dB for Euclidean. The gains are evident across all embedding dimensions, particularly in the lower

sizes, such as $d = 64$ and $d = 128$, where Manhattan shows a sharper improvement over the baseline MSE loss function. These results suggest that Manhattan distance offers a more robust representation for feature matching, translating to better perceptual and quantitative image quality during inference. Such behavior can be attributed to its robustness to outliers and its suitability for high-dimensional embedding spaces. Unlike Euclidean distance, Manhattan evaluates differences independently across dimensions, making it less sensitive to extreme values and better aligned with sparse feature representations. This robustness allows Manhattan distance to more effectively capture embedding relationships, leading to improved perceptual and quantitative image quality. The trends also indicate that increasing embedding dimensionality enhances performance for both metrics, with the best results achieved at $d = 512$, emphasizing the importance of rich feature representations for accurate embedding similarity measurement.

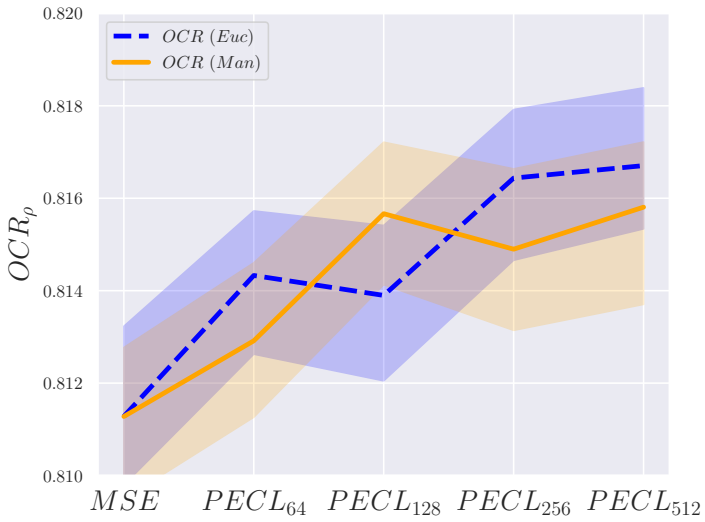

Figure 9: OCR confidence (OCR_ρ) trend across different configurations: $PECL_d$ with d the size of the embeddings.

By analyzing the accuracy of the detection and readability from the OCR confidence (OCR_ρ) provided in Fig. 9, we observe that the proposed loss function consistently outperforms the baseline MSE. This comparison highlights the limitations of MSE as a pixel-focused loss function, which primarily optimizes for fidelity at the pixel level but fails to capture feature relationships critical for OCR performance. Additionally, the steady increase in OCR confidence with larger embedding dimensions suggests that richer feature representations enhance OCR accuracy, with the highest performance achieved at d=512. Overall, the training strategy by incorporating embedding similarity with pixel-to-pixel fidelity demonstrates significant advantages over MSE in preserving both

perceptual and structural features critical for OCR recognition.

$$\text{Contrast}(\text{PECL}, MSE) = \frac{\text{PSNR}_{\text{PECL}} - \text{PSNR}_{\text{MSE}}}{\text{PSNR}_{\text{MSE}}}. \tag{20}$$

To evaluate the training dynamics of the proposed PECL, we compare its PSNR performance to the baseline MSE across various embedding sizes (64, 128, 256, and 512). The analysis aims to quantify the relative improvement introduced by PECL using the contrast metric defined in Eq.20. This metric normalizes the difference in PSNR values by the baseline, providing a more interpretable measure of performance gains during training. The results are depicted in Fig.10 and Fig. 11, offering insights into the learning behavior of the proposed PECL with Euclidean and Manhattan distances, respectively. The x-axis, presented on a logarithmic scale, captures training iterations, while the y-axis shows the contrast in PSNR values (in dB). Positive contrast values indicate that the proposed PECL outperforms MSE, while negative values suggest underperformance. The baseline version, MSE, favors high PSNR due to its pixel-wise loss objective. In contrast, the PECL combines MSE with a contrastive loss, which focuses on embedding similarity and perceptual quality, making it less focused on maximizing PSNR directly.

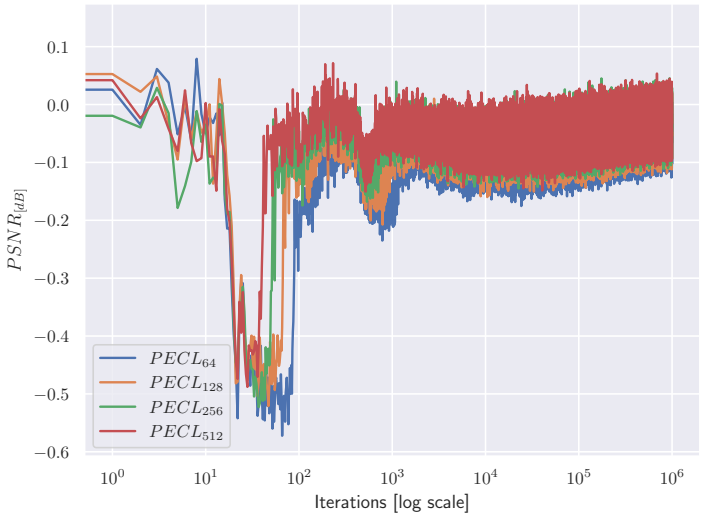

Figure 10: Contrast of PSNR values for PECL with Euclidean distance and embedding sizes of 64, 128, 256, and 512 compared to the MSE during training. The contrast, defined as the relative improvement in PSNR over the MSE, is plotted against the training iterations on a logarithmic scale. Positive contrast values indicate better performance of the PECL models compared to MSE, while negative values indicate worse performance.

Upon closer inspection, the contrast curves for Manhattan distance are more centered around the zero axis, with values fluctuating between -0.22 and 0.25. In contrast, the Euclidean distance curves exhibit a wider spread, ranging from -0.6

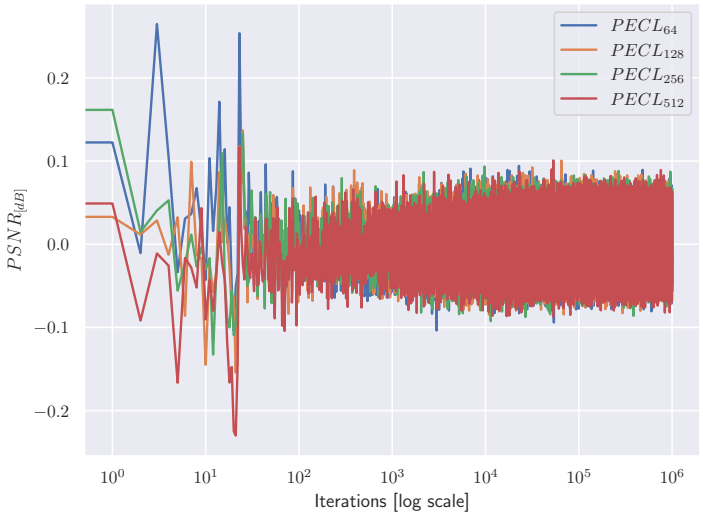

Figure 11: Contrast of PSNR values for PECL with Manhattan distance and embedding sizes of 64, 128, 256, and 512 compared to the MSE during training. The contrast, defined as the relative improvement in PSNR over the MSE, is plotted against the training iterations on a logarithmic scale. Positive contrast values indicate better performance of the PECL models compared to MSE, while negative values indicate worse performance.

to 0.13. This difference highlights that, while both distances show a tendency for improvement as training progresses, the Euclidean curves are more influenced by negative contrast values, indicating that Euclidean-based PECL start off with a larger disparity in performance. The more centered Manhattan curves suggest that its impact on the PSNR performance is more stable and balanced throughout training.

During the initial training phase, the contrast values for both Euclidean and Manhattan exhibit fluctuations, reflecting the combined influence of MSE and contrastive loss. As training progresses, the curves stabilize and shift toward positive values, consistently exceeding zero and converging around 0.1. This indicates that PECL, despite improving both MSE and embedding-based similarities, achieves a higher PSNR than MSE. This reflects its ability to produce better visual quality. The positive contrast suggests that the contrastive loss contributes to improving perceptual performance, making the images visually more accurate while still maintaining relatively similar embeddings.

Larger embedding dimensionality exhibits smoother and more stable contrast curves, particularly for Manhattan, suggesting that higher-dimensional embeddings contribute to more robust feature representation and performance stability throughout training. This highlights the role of embedding dimensionality in learning distinctive features, which enhances the model's ability to generalize. The observed trends emphasize the trade-off managed by PECL, effectively balancing pixel-wise accuracy with perceptual quality to achieve better

overall performance compared to the baseline MSE.

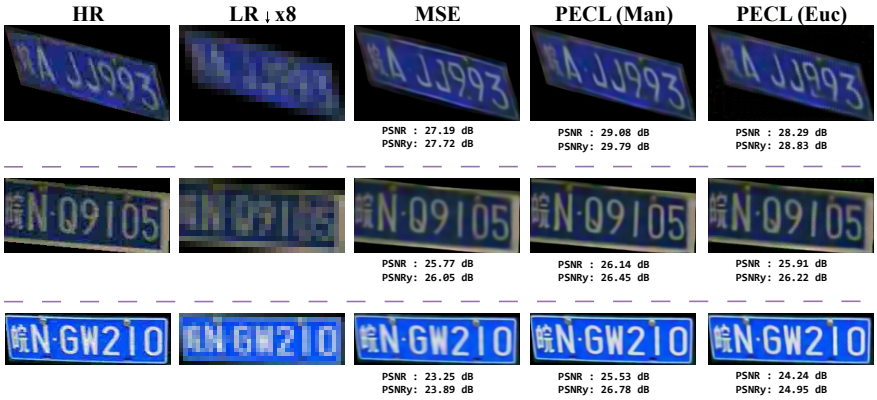

Figure 12: Qualitative comparison of license plates under varying acquisition conditions using the proposed PECL with Manhattan and Euclidean distances, compared to the baseline MSE.

The qualitative comparison, as shown in Fig. 12, provides a visual assessment of the performance of the proposed PECL framework with both Manhattan and Euclidean distances, benchmarked against the baseline MSE. The figure presents the high-resolution (HR) reference image alongside the low-resolution (LR) input, downsized by a factor of 8, and the super-resolved (SR) results obtained using MSE, PECL (Man), and PECL (Euc). This layout allows for a direct evaluation of the visual quality achieved by each approach, highlighting the differences in detail preservation, texture sharpness, and perceptual fidelity. As can be seen, the HR images exhibit varying acquisition conditions, including differences in lighting and angles. Additionally, their visual quality appears degraded, with noticeable noise and artifacts. After downsizing them by a factor of 8 to produce the LR inputs, the fine details, such as text and edges, become significantly unclear and, in some cases, unreadable. This highlights the challenge of reconstructing accurate and perceptually meaningful SR outputs.

In the first example, PECL (Man) achieves a PSNR of 29.08 dB and PSNRy of 29.79 dB, outperforming PECL (Euc) and MSE. This improvement is visually apparent, with PECL (Man) generating sharper edges and more detailed textures compared to both MSE and PECL (Euc), which exhibit noticeable smoothing and blurring, particularly in high-frequency regions such as the text. In the second example, PECL (Man) again outperforms both PECL (Euc) and MSE, with a PSNR of 26.14 dB and PSNRy of 26.45 dB. The HR image shows fine details and sharp edges, which are largely preserved in the PECL (Man) reconstruction, while MSE produces a more blurred and less detailed result. This trend is consistent in the third example, where PECL (Man) achieves 25.53 dB in PSNR and 26.78 dB in PSNRy, significantly improving the perceptual quality over PECL (Euc) and MSE, both of which still suffer from substantial blurring

and loss of fine details. Across all examples, PECL (Man) consistently restores sharper edges and clearer textures, indicating its ability to recover fine-grained details effectively.

The improved sharpness and perceptual quality in the PECL (Man) results demonstrate the ability of the proposed loss function to better align image features through the embedding consistency enforced by the contrastive loss. In comparison, while PECL (Euc) also improves upon MSE, it still cannot match the sharpness or perceptual quality achieved by PECL (Man), particularly in high-frequency regions. These findings demonstrate that pixel-to-pixel fidelity along with embedding similarity enhances perceptual quality by focusing on feature-level consistency.

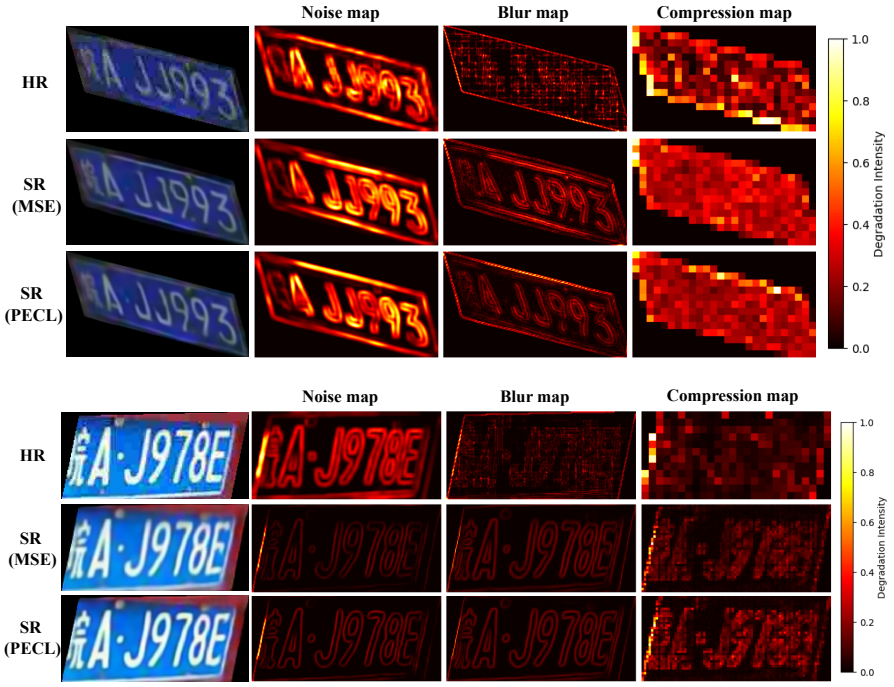

Figure 13: Qualitative comparison of HR and SR license plate images, with degradation maps showing noise, blur, and compression artifacts. The maps illustrate the intensity of degradation in both the HR and SR images, with each map providing insights into how the visual quality is affected by the SR model trained with the proposed PECL over MSE.

By analyzing the distortion maps (Noise, Blur, and Compression) in Fig. 13 we can highlight the differences in performance when training the model with MSE compared to PECL with respect to specific degradations. The blur map for PECL demonstrates significantly reduced blur compared to the MSE model, indicating that PECL can restore sharper edges and finer details. In contrast, with MSE, higher blur intensity can be observed around high-frequency regions such as edges. This reflects its tendency to produce over-smoothed outputs.

The noise map reveals that PECL maintains moderate and localized noise levels, comparable to MSE. The compression map further demonstrates that PECL introduces fewer compression artifacts, particularly in textured regions, resulting in perceptually superior reconstructions. Compared to the HR image, the proposed method appears to smooth the images, resulting in distribution depicted with the blur maps, regardless of the training loss.

Overall, the PECL model balances sharpness, controlled noise, and minimal compression artifacts, demonstrating its effectiveness in producing detailed and high-quality SR images while avoiding significant distortions. This highlights the strength of the embedding consistency enforced by the PECL loss, which aligns features and enhances sharpness generation without compromising overall image quality.

5. Conclusion

This article presents a novel framework for license plate super-resolution (LPSR) that achieves significant advancements compared to state-of-the-art methods. By introducing the pixel and embedding consistency loss (PECL), the proposed model effectively balances pixel-level fidelity with embedding-level similarity, resulting in consistent improvements across PSNR, PSNRy, and OCRρ. With a Siamese network and contrastive loss, PECL enhances the SR model, improving perceptual quality and task-specific performance while preserving fine-grained details.

The analysis of distortion maps underscores the model's capability to mitigate blurring and compression artifacts while maintaining sharpness, a critical factor in generating high-quality SR images. By enforcing embedding distance minimization through the proposed PECL, the SR model achieves superior alignment between SR and HR license plates. This results in semantic consistency and improved recognition accuracy. Additionally, the loss function dynamically optimizes the contributions of pixel-wise and embedding-based losses during training, ensuring stability, adaptability, and robustness against varying degradation levels.

In summary, this work contributes to advancing LPSR by combining an innovative loss function with a robust architectural design. The proposed model demonstrates competitive performance, delivering visually accurate results compared to state-of-the-art methods, such as SwinIR and TBSRN, while enhancing the text recognition task.

Future work will explore multi-image super-resolution to leverage temporal and spatial information for enhanced reconstruction. Besides, the proposed model may benefit from knowledge distillation to create a lightweight and more robust version.

6. Akgnowledement

This work is supported by The French Research Funding Agency (ANR) under project IMPROVED ANR-22-CE39-0006.

References

[1] Aggarwal, C., Hinneburg, A., Keim, D., 2001. On the surprising behavior of distance metrics in high dimensional space, in: Database theory—ICDT, Springer. pp. 420–434.

[2] Ansel, J., Yang, E., He, H., Gimelshein, N., et al., 2024. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation, in: ACM ICASPL, p. 929–947.

[3] Chaudhuri, A., Mandaviya, K., Badelia, P., Ghosh, S., 2017. Optical Character Recognition Systems. Springer. pp. 9–41.

[4] Chen, H., Gu, J., Zhang, Z., 2021a. Attention in attention network for image super-resolution. arXiv preprint arXiv:2104.09497 .

[5] Chen, J., Li, B., Xue, X., 2021b. Scene text telescope: Text-focused scene image super-resolution, in: IEEE/CVF CVPR, pp. 12026–12035.

[6] Dong, C., Loy, C., He, K., Xiaoou, X., 2015. Image super-resolution using deep convolutional networks. IEEE TPAMI 38, 295–307.

[7] Dosovitskiy, A., et al., 2021. An image is worth 16x16 words: Transformers for image recognition at scale, in: ICLR. URL: `https://openreview.net/forum?id=YicbFdNTTy`.

[8] Gao, H., Yuan, H., Wang, Z., Ji, S., 2020. Pixel transposed convolutional networks. IEEE TPAMI 42, 1218–1227.

[9] Gohrani, K., 2019. Different types of distance metrics used in machine learning.

[10] Gonçalves, G.R., Diniz, M.A., Laroca, R., Menotti, D., Schwartz, W.R., 2019. Multi-task learning for low-resolution license plate recognition, in: Progress in PRIACVA, Springer. pp. 251–261.

[11] Gong, Y., Deng, L., et al., 2022. Unified chinese license plate detection and recognition with high efficiency. JVCIR 86, 103541.

[12] Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping, in: IEEE CVPR), pp. 1735–1742.

[13] Hsu, G., Chen, J., Chung, Y., 2013. Application-oriented license plate recognition. IEEE TVT 62, 552–561.

[14] Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution, in: ECCV, Springer. pp. 694–711.

[15] Karen, S., Andrew, Z., 2015. Very deep convolutional networks for large-scale image recognition, in: ICLR, pp. 1–14.

[16] Kashyap, A., Suresh, B., Patil, A., Sharma, S., Jaiswal, A., 2018. Automatic number plate recognition, in: IEEE ICACCCN, Greater Noida, India. pp. 838–843.

[17] Keys, R., 1981. Cubic convolution interpolation for digital image processing. IEEE TASSP 29, 1153–1160.

[18] Kim, J., Lee, J., Lee, K., 2016. Accurate image super-resolution using very deep convolutional networks, in: IEEE CVPR, pp. 1646–1654.

[19] Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

[20] Koch, G., Zemel, R., Salakhutdinov, R., et al., 2015. Siamese neural networks for one-shot image recognition, in: ICML deep learning workshop.

[21] Kulkarniet, U., et al., 2023. Helmet detection using yolo-v5 and paddle ocr for embedded systems, in: International Conference on Recent Trends in Machine Learning, IOT, Smart Cities & Applications, Springer. pp. 233–244.

[22] Laroca, R., Cardoso, E.V., Lucio, D.R., Estevam, V., Menotti, D., 2022. On the cross-dataset generalization in license plate recognition, in: ICVISAPP, pp. 166–178.

[23] Ledig, C., Theis, L., et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network, in: IEEE CVPR.

[24] Lee, S., Kim, J.H., Heo, J.P., 2020. Super-resolution of license plate images via character-based perceptual loss, in: IEEE BigComp, pp. 560–563.

[25] Li, J., Fang, F., Mei, K., Zhang, G., 2018. Multi-scale residual network for image super-resolution, in: ECCV, pp. 517–532.

[26] Li, Y., Zhang, Y., Timofte, R., Van Gool, L., Yu, L., et al., 2023. Ntire 2023 challenge on efficient super-resolution: Methods and results, in: IEEE/CVF CVPR, pp. 1921–1959.

[27] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R., 2021. Swinir: Image restoration using swin transformer, in: IEEE/CVF CVPR, pp. 1833–1844.

[28] Lim, B., Son, S., Kim, H., Nah, S., Lee, K., 2017. Enhanced deep residual networks for single image super-resolution, in: IEEE CVPRw, pp. 136–144.

[29] Liu, A., Liu, Y., Gu, J., Qiao, Y., Dong, C., 2023. Blind image super-resolution: A survey and beyond. IEEE TPAMI 45, 5461–5480.

[30] Loshchilov, I., Hutter, F., 2017. SGDR: Stochastic gradient descent with warm restarts, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=Skq89Scxx.

[31] Lu, Y., Gu, Y., Wang, B., 2023. License plate recognition in wild with super-resolution, in: ICNNICE, pp. 523–526.

[32] Ma, C., Li, F., Wang, Z., Zhang, Y., Li, F., 2024. Application and practice of paddleocr based image recognition technology in business license, in: SPIC, IEEE. pp. 741–745.

[33] Maier, A., Moussa, D., Spruck, A., Seiler, J., Riess, C., 2022. Reliability scoring for the recognition of degraded license plates, in: IEEE AVSS, pp. 1–8.

[34] Mathur, P., Singh, A.K., Azeemuddin, S., Adoni, J., Adireddy, P., 2021. A real-time super-resolution for surveillance thermal cameras using optimized pipeline on embedded edge device, in: IEEE AVSS, pp. 1–7.

[35] Moser, B.B., Raue, F., Frolov, S., Palacio, S., Hees, J., Dengel, A., 2023. Hitchhiker's guide to super-resolution: Introduction and recent advances. IEEE TPAMI 45, 9862–9882.

[36] Moussa, D., Maier, A., Spruck, A., Seiler, J., Riess, C., 2022. Forensic license plate recognition with compression-informed transformers, in: IEEE ICIP, pp. 406–410.

[37] Nascimento, V., Laroca, R., de A. Lambert, J., Schwartz, W.R., Menotti, D., 2023. Super-resolution of license plate images using attention modules and sub-pixel convolution layers. Computers and Graphics 113, 69–76.

[38] Nascimento, V., Laroca, R., Lambert, J.d.A., Schwartz, W.R., Menotti, D., 2022. Combining attention module and pixel shuffle for license plate super-resolution, in: SIBGRAPI, pp. 228–233.

[39] Pan, Y., Tang, J., Tjahjadi, T., 2024. Lpsrgan: Generative adversarial networks for super-resolution of license plate image. Neurocomputing , 127426.

[40] Peng, Q., Tu, L., 2024. Paddle-ocr-based real-time online recognition system for steel plate slab spray marking characters. Journal of Control, Automation and Electrical Systems 35, 221–233.

[41] Sarkar, O., et al., 2024. Automatic number plate character recognition using paddle-ocr, in: 2024 ICICET, IEEE. pp. 1–7.

[42] Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, in: IEEE/CVF CVPR, pp. 1874–1883.

[43] Team, P., 2024. Paddleocr documentation. URL: `https://paddlepaddle.github.io/PaddleOCR/main/en/index.html`. accessed: 2024-12-04.

[44] Van Reeth, E., Tham, I.W., Tan, C.H., Poh, C.L., 2012. Super-resolution in magnetic resonance imaging: a review. CMR 40, 306–325.

[45] Wang, X., Xie, L., Dong, C., Shan, Y., 2021. Real-esrgan: Training real-world blind super-resolution with pure synthetic data, in: IEEE/CVF ICCV, pp. 1905–1914.

[46] Wang, X., Yu, K., Wu, S., et al., 2018. ESRGAN: Enhanced super-resolution generative adversarial networks, in: IEEE ECCVw, pp. 0–0.

[47] Xu, Z., Yang, W., Meng, A., Lu, N., Huang, H., 2018. Towards end-to-end license plate detection and recognition: A large dataset and baseline, in: ECCV, pp. 255–271.

[48] Yang, J., Wright, J., Huang, T., Ma, Y., 2010. Image super-resolution via sparse representation. IEEE TIP 19, 2861–2873.

[49] Yang, Y., Bi, P., Liu, Y., 2018. License plate image super-resolution based on convolutional neural network, in: IEEE ICIVC, pp. 723–727.

[50] Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y., 2018a. Image super-resolution using very deep residual channel attention networks, in: ECCV, pp. 286–301.

[51] Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y., 2018b. Residual dense network for image super-resolution, in: IEEE CVPR, pp. 2472–2481.