Learning Generative Structure Prior for Blind Text Image Super-resolution

Xiaoming Li[1]   Wangmeng Zuo[2]   Chen Change Loy[1(✉)]

[1]S-Lab, Nanyang Technological University   [2]Harbin Institute of Technology

csxmli@gmail.com   wmzuo@hit.edu.cn   ccloy@ntu.edu.sg

(a) A real-world LR text image cropped from a Chinese invoice   (b) A real-world LR text image cropped from a signboard

Figure 1. Results obtained with TBSRN [7], TATT [41] and our approach on the segments of Chinese text image cropped from a real low-resolution invoice and a signboard. Both baselines are re-trained with elaborated degradations as in BSRGAN [75] and Real-ESRGAN [66].

Abstract

Blind text image super-resolution (SR) is challenging as one needs to cope with diverse font styles and unknown degradation. To address the problem, existing methods perform character recognition in parallel to regularize the SR task, either through a loss constraint or intermediate feature condition. Nonetheless, the high-level prior could still fail when encountering severe degradation. The problem is further compounded given characters of complex structures, e.g., Chinese characters that combine multiple pictographic or ideographic symbols into a single character. In this work, we present a novel prior that focuses more on the character structure. In particular, we learn to encapsulate rich and diverse structures in a StyleGAN and exploit such generative structure priors for restoration. To restrict the generative space of StyleGAN so that it obeys the structure of characters yet remains flexible in handling different font styles, we store the discrete features for each character in a codebook. The code subsequently drives the StyleGAN to generate high-resolution structural details to aid text SR. Compared to priors based on character recognition, the proposed structure prior exerts stronger character-specific guidance to restore faithful and precise strokes of a designated character. Extensive experiments on synthetic and real datasets demonstrate the compelling performance of the proposed generative structure prior in facilitating robust text SR. Our code is available at https://github.com/csxmli2016/MARCONet.

1. Introduction

Blind text super-resolution (SR) aims at restoring a high-resolution (HR) image from a low-resolution (LR) text image that is corrupted by unknown degradation. The problem is relatively less explored than SR of natural and face images.

This seemingly easy task actually possesses many unique challenges. In particular, each character owns its unique structure. A suboptimal restoration that destroys the structure with, *e.g.*, distorted, missing or additional strokes, is easily perceptible and may alter the meaning of the original word. The task becomes more difficult when dealing with writing systems such as Chinese, Kanji and Hanja characters, in which the patterns of thousands of characters vary in complex ways and are mostly composed of different subunits (or compound ideographs). A slight deviation in strokes, say '太' and '大' carries entirely different meanings. The problem becomes more intricate given the different typefaces. In addition, complex and unknown degradation make the data-driven convolutional neural networks (CNNs) incapable of generalizing well to real-world degraded observations.

To ameliorate the difficulties in blind text SR, especially in retaining the unique structure of each character, recent studies design new loss functions to constrain SR results semantically close to the high-resolution (HR) ground-truth [7,8,45,50,79], or incorporate the recognition prior into the intermediate features [40,41,44]. Figure 1 shows several representative results of these two types of methods, *i.e.*, TB-SRN [7] that applies content-aware constraint and TATT [41]

that embeds the recognition prior into the SR process. The results are not satisfactory despite the models were retrained with data synthesized with sophisticated degradation models, *i.e.*, BSRGAN [75] and Real-ESRGAN [66]. TBSRN and TATT perform well when the LR character has fewer strokes. However, they both fail to generate the desired structures when the character is composed of complex strokes, *e.g.*, the 4, 6, 8, 11-*th* characters in (a). The results suggest that high-level recognition prior cannot well benefit the SR task in restoring faithful and accurate structures.

When dealing with complex characters, we can identify some commonly used strokes as constituents of characters or their subparts. These obvious structural regularities have not been studied thoroughly in the literature. In particular, most existing text SR studies [7, 40, 41, 45, 50, 79] focus on scene text images that are presented in different view perspectives or arranged in irregular patterns. In contrast, our goal is to explore structural regularities specific to complex characters and investigate an effective way to exploit the prior. The investigation has practical values in many applications, including the restoration of old printed media (*e.g.*, newspapers or books) for digitalization and preservation, and the enhancement of captions in digital media.

The key idea of our solution is to mine structure prior from a generative network, and use the prior to facilitate blind SR of text images. To capture the structure prior of characters, we train a StyleGAN to generate characters of different styles. As our intent is not to generate infinite and non-repetitive results, we restrict the generative space of StyleGAN to obey the structure of characters by constraining the generation with a codebook. Specifically, the codebook stores the discrete code of each character, and each code serves as a constant to StyleGAN for generating a specific high-resolution character. The font style is governed by the $\mathcal{W}$ latent space of StyleGAN. Once the character StyleGAN is trained, we can extract its intermediate features serving as the generative structure prior, which encapsulates the intrinsic structure and style of the LR character.

Using the prior for text SR can be conveniently achieved through an additional Transformer-based encoder and a text SR network that accepts prior. In particular, given a text image that constitutes multiple characters (*e.g.*, Figure 1), we use a Transformer-based encoder to predict the font style, character bounding boxes, and their respective indexes in the codebook jointly. The codebook indexes drive the structure prior generation for each character. In the text SR network, each LR character is super-resolved, guided by the respective priors that are aligned using their bounding boxes.

Experiments demonstrate that our design can generate robust and consistent structure priors for diverse and severely degraded text input. With the embedded structure prior, our approach performs superior against state-of-the-art methods in generating photo-realistic text results and shows excellent generalization to real-world LR text images. While our study mainly employs Chinese characters as examples, the proposed method can be readily extended to characters of other languages. We call our approach as *MARCONet*. The main contributions are summarized as follows:

- We show that blind SR task, especially for characters with complex structures, can be restored by using their structure prior encapsulated in a generative network.
- We present a viable way of learning such generative structure prior through reformulating a StyleGAN by replacing its single constant with discrete codes that represent different characters.
- To retrieve the prior accurately, we propose a Transformer-based encoder to jointly predict the font styles, character bounding boxes and their indexes in codebook from the LR input.

2. Related Work

Blind Image SR. Blind image SR is challenging due to the complex mixture of unknown degradations. Recent studies address the problem from two aspects, *i.e.*, degradation estimation [3, 19, 38, 39, 62] and establishing more realistic training data [4, 28, 33, 66, 70, 75]. The former paradigm mainly focuses on estimating degradation model parameters and then applies non-blind SR methods, *e.g.*, ZSSR [57]. The latter builds training pairs either through capturing real-world LR and HR pairs [4, 70] or designing elaborate degradation models that imitate real-world degradation [28, 33, 66, 75]. Since characters in text images have specific and semantic structures, we show that a good restoration performance cannot be achieved by merely using elaborately designed degradation models.

Text Image SR. Text image SR has been studied for many years. In traditional methods, maximum a posterior (MAP) [5] and Bayesian framework [9] are exploited for super-resolving text images. These earlier approaches are incapable of generating high-quality results. Dong *et al*. [13] adopt CNNs (*i.e.*, SRCNN [12]) for text image SR and achieve promising results in the ICDAR 2015 competition [49]. Xu *et al*. [72] adopt a Generative Adversarial Network (GAN) [17] to learn category-specific prior for face and text images SR, along with the supervision from a multi-class GAN loss. Mou *et al*. [44] propose to plug a SR unit into the recognition process of degraded text images. Wang *et al*. [64] introduce the first real-world text SR pairs (*i.e.*, TextZoom), which are cropped from RealSR [4] and SRRAW [77]. They also present a sequential residual block by incorporating Bi-directional LSTM to capture the sequential information for low-level reconstruction. Both RealSR and SRRAW are captured by different digital cameras with different focal lengths, aiming to collect natural LR/HR pairs in real-world scenarios. The setting is not designed specifically for text images. We observe that most HR text images

in TextZoom are limited in clarity, which could limit the SR performance when taking them as ground-truth.

Text SR benefits from prior and auxiliary constraints. Quan *et al.* [51] recover text images in a cascade model by predicting the high-frequency information. Chen *et al.* [7] propose Transformer-based position-aware and content-aware modules to emphasize the position and the content of each character. Analogously, Nakaune *et al.* [45] and Qin *et al.* [50] respectively present the structure-aware loss and content-perceptual constraint for learning detailed structural skeletons. Zhao *et al.* [78] propose a parallel contextual attention network to learn sequence-dependent features for bringing more high-frequency information into text reconstruction. Zhao *et al.* [79] exploit linguistic, recognition and visual clues to jointly boost the SR performance. Chen *et al.* [8] propose a stroke-aware framework by concentrating on stroke-level internal structures. Ma *et al.* [41] introduce text recognition prior to text reconstruction with a Transformer-based module, leveraging the global attention mechanism. They also embed categorical text priors in the encoder and employ multi-stage refinement to progressively enhance low-resolution images [40].

Many methods above employ recognition information either as a loss function on the SR results [7, 8, 50, 79] or as intermediate SR features for providing high-level guidance [40, 41]. Albeit the high-level recognition prior helps improve the text recognition ability, it is limited in providing accurate structure and style guidance, especially for certain texts with complex structures. In this study, we show that generative structure prior benefits high-quality guidance to restore the faithful structure of LR characters.

Generative Structure Prior in Image SR. Image structure prior is proven effective in many low-level vision tasks, *e.g.*, depth image enhancement [21, 27, 37], image inpainting [11, 46, 52], and image restoration [10, 35, 36, 47]. Most recently, by using generative structure priors obtained from pre-trained StyleGANs [30, 31] or codebooks [15], blind face restoration has achieved tremendous improvement [6, 22, 34, 65, 69, 73], suggesting the apparent advantage of such prior over other methods in generating photo-realistic textures. Our study is inspired by the success of these approaches. Nevertheless, crafting a suitable structural prior for text images are harder than face images. This is because each character has its unique strokes, but may have a vast variety of font styles. Any distorted, missing or additional strokes easily change their semantic layout that is easily perceptible, and worsens their actual meaning (see Figure 1). All of these challenges aggravate the difficulties for learning a generative structure prior for the text SR task. In this study, we show an effective way of learning such prior through replacing the constant input of StyleGAN with a discrete codebook, while controlling the font style via the $\mathcal{W}$ space.
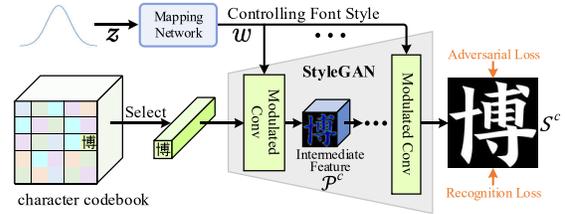

Figure 2. Pre-training generative structure prior for each character. The codebook stores the discrete code of each character, and each code serves as a constant to StyleGAN for generating a specific high-resolution character. The intermediate features encapsulate the generative structure prior and will be used for guiding text SR.

3. Methodology

A GAN model can well capture the intrinsic structure prior for a category by training the model on abundant images of the same category. Many image restoration tasks [6, 20, 42, 48, 65, 73] have shown the benefits of using such generative priors for restoring photo-realistic details from LR inputs despite severe and complex degradation. Previous studies mainly use the prior for face restoration, while its application to text SR is under-explored. In this paper, we propose MARCONet, the first attempt to e**M**bed gener**A**tive st**Ru**Cture pri**O**r for blind text SR. The proposed pipeline is depicted in Figure 3. It mainly contains three parts, 1) the prediction of font style, character bounding boxes and their indexes in codebook based on a given LR input, 2) the generation of the structure prior for each character, 3) the SR framework that takes the generative structure prior as guidance for restoration. Next, we first describe the approach to obtain the generative structure prior for each character, followed by the way to employ such prior for the text SR.

3.1. Pre-training of Generative Structure Prior

The original StyleGANs [30, 31] take a learnable constant as input, and control the style of output image through the $w \in \mathcal{W}$ space, which is a projection from the initial latent space $z \in \mathcal{Z}$. Layer-wise noise is also introduced to support stochastic variation on fine-grained details. To better capture the structure of text images, we remove the layer-wise noise and replace the single constant with discrete codes that represent different characters (see details in Figure 2).

We denote each character code as $c \in \{C_i\}_{i=1}^{M}$, where C is the codebook that stores all the character features. Each code is learnable, with a size of $1 \times 1 \times 512$. Following CRNN [56], the cardinality of codebook M is set to 6736, a size covering simplified Chinese characters, English letters, and numbers. The retrofitted StyleGAN is defined as:

$$\mathcal{S}^c = G(c, \mathcal{F}(z); \Theta_G) = G(c, w; \Theta_G), \quad (1)$$

where $\mathcal{F}$ represents the network that maps z to w, and Θ_G is the model parameters of StyleGAN. For generalizing to different scenarios, we simplify the structure image $\mathcal{S}^c$ with pixel values $\in \{0, 1\}$ (see the output in Figure 2). The struc-
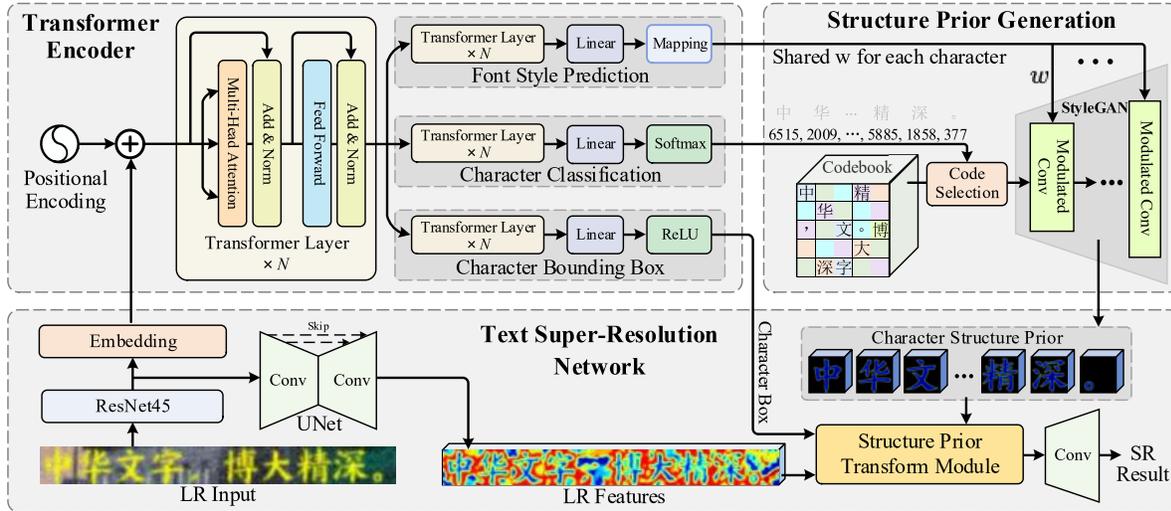
Figure 3. Overview of our MARCONet. It contains three parts, *i.e.*, (i) Transformer encoder for predicting the font style, classification and bounding boxes of each character from LR input, (ii) structure prior generation with pre-trained StyleGAN for generating reliable structure prior for each character, and (iii) the SR process for reconstructing the SR output with the incorporation of each characters' structure prior.

ture prior $\mathcal{P}^c$ used in this work is the intermediate features from G in Eqn. (1), and can be formulated as:

$$\mathcal{P}^c = G_i(c, w; \Theta_G), \qquad (2)$$

where G_i represents the output features from i-*th* layer of G.

The training of the character StyleGAN can be done on synthetic images yet with satisfying generalization ability to real-world texts. In particular, the PIL package[1] is adopted to synthesize high-resolution character images with hundreds of font styles (see examples in the suppl.). We also augment the diversity with random translations, font sizes and slight affine transformation. Different from the original StyleGANs that adopt only adversarial loss [17], we introduce an additional recognition loss derived from a pre-trained Transformer-based recognizer as regularization. Once learned, each learnable code c well captures the distinctive features of each character and $w \in \mathcal{W}$ can control the font styles of the output image. In the following, we will describe the approach to extract the structure prior for each character in a given LR text image.

3.2. Transformer Encoder

A LR text image usually is composed of several characters. To derive the generative structure prior for each character, first, we need to obtain the font style w of the LR characters. Second, the indexes of each character in the codebook and their bounding boxes are also necessary for aligning the structure prior with each LR character. To this end, we adopt a Transformer-based encoder to jointly predict the needed information. Transformer [61] is chosen so that we can better capture the dependency across different characters in the input image.

---
[1]https://python-pillow.org/

Learning to predict the font style w is similar to learning-based GAN inversion [2, 54, 60, 63], where the Transformer network serves as an encoder. It is observed that the Transformer-based encoder achieves satisfactory inversion, achieving reconstruction with low distortion and of high quality [25]. Note that characters in one text image generally have the same font style. So we use one linear layer to map the features of all characters to a single prediction (w), which is shared with all the characters in the same LR image. We also use a similar structure as w prediction to carry out the regression and classification of the bounding boxes and code indexes. These three sub-tasks share the same Transformer encoder (*i.e.*, ViT [14]) and a CNN backbone (*i.e.*, ResNet45 [24]). Positional encoding is incorporated to inject positions of features in the sequence.

The font style prediction branch is optimized with the gradient from StyleGAN. Since the input text image has unconstrained numbers of characters, we adopt CTC loss [18] for recognition. The loss is proposed to solve the alignment between predicted and target labels and is widely used in scene text recognition tasks. As for the learning of character bounding boxes regression, we adopt a linear combination of the Smooth L_1 loss [16] and the generalized IoU loss [53]. The ground-truth boxes of each character can be obtained when synthesizing the training images with PIL package. The gradient from the latter text SR network can further benefit the learning process of the whole Transformer encoder.

3.3. Text Super-resolution Network

Network Architecture. With the Transformer encoder, we can obtain the font style w, classification label (*i.e.*, index in codebook), and the bounding box for each character in the LR input. Then, the corresponding high-quality generative
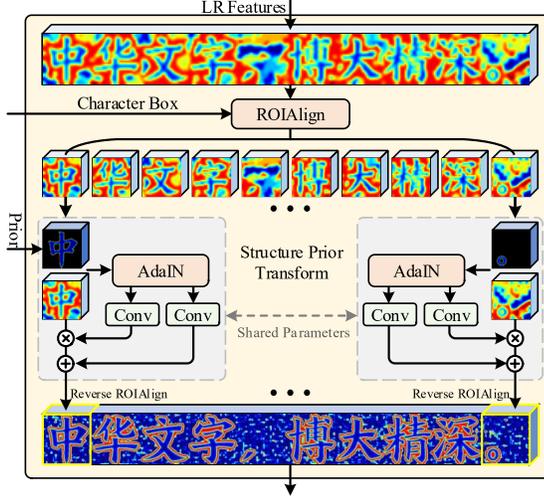
Figure 4. Details of the structure prior transform module. Each LR character is cropped and aligned by using the bounding boxes, and is super-resolved with the guidance of their structure prior.

structure prior $\mathcal{P}^c$ for each character can be generated with w and the selected code c through the pre-trained StyleGAN in Eqn. (2) (see the character structure prior in Figure 3). In this section, we describe how to use them for the text SR task. First, a simple UNet [55] is stacked upon the ResNet45 to extract the LR features. Then, the generative structure prior of each character is embedded into their LR characters through a structure prior transform module. The process is shown in Figure 4 (more details can be found in the suppl.). For the LR features, we adopt the detected bounding boxes to crop and align each LR character by RoIAlign operation [23]. For each character, AdaIN [26] is adopted to normalize the prior's distribution, and spatial feature transformation [67] is subsequently employed to predict the affine parameters, which are applied on the LR character features. The reverse RoIAlign is finally used to paste the enhanced features back to their original locations. The structure prior transform module is adopted at two scales $s \in \{32, 64\}$, allowing our MARCONet to remain high fidelity with different degradation. A Conv-ReLU-ResBlock based CNN module is stacked to generate the final SR result.

Learning Objective. We minimize the differences between the SR results $\hat{I}_{SR}$ and HR ground-truth I_{HR} on both the pixel and perceptual domains [29]:

$$\mathcal{L}_{rec} = \left\| \hat{I}_{SR} - I_{HR} \right\|_1 + \sum_{i=1}^{4} \frac{\lambda_{per}}{\mathcal{C}_i \mathcal{H}_i \mathcal{W}_i} \left\| \Phi_i(\hat{I}_{SR}) - \Phi_i(I_{HR}) \right\|_1 , \quad (3)$$

where $\mathcal{C}_i$, $\mathcal{H}_i$ and $\mathcal{W}_i$ are the feature dimensions from the i-th convolution layer of the pre-trained VGG-19 model Φ [58]. λ_{per} is the trade-off parameter and is set to 0.05. The loss $\mathcal{L}_{rec}$ is applied to the whole text image.

Adversarial loss [17] $\mathcal{L}_{adv}$ is also added to improve visual quality. Instead of constraining on the whole image as in Eqn. (3), $\mathcal{L}_{adv}$ is performed on each cropped character image

together with its corresponding structure image as additional conditions [43]. To be specific, the concatenation of HR character image I_{HR}^c and its structure image $\mathcal{S}_{HR}^c$ is expected to be classified as Real, while the concatenation of SR character image $\hat{I}_{SR}^c$ and its structure image $\mathcal{S}_{SR}^c$ is recognized as Fake. Note that $\mathcal{S}_{HR}^c$ is the binarized version of I_{HR}^c, which represents the ground-truth structure, and $\mathcal{S}_{SR}^c$ is obtained from Eqn. (1) with w and c predicted from LR input. Such design allows us to better constrain the embedding of generative structure prior into the SR results. The hinge version of adversarial loss [74] for each character is defined as:

$$\mathcal{L}_D = -\mathbb{E}[\min(0, -1 + D(I_{HR}^c, \mathcal{S}_{HR}^c))] - \mathbb{E}[\min(0, -1 - D(\hat{I}_{SR}^c, \mathcal{S}_{SR}^c))],$$
$$\mathcal{L}_G = -\mathbb{E}[D(\hat{I}_{SR}^c, \mathcal{S}_{SR}^c)]. \quad (4)$$

Finally, we use the L_1 loss on structure image $\mathcal{S}^c$ in the text SR training stage to further fine-tune w and code c in codebook. For each character, this constraint is defined as:

$$\mathcal{L}_{str} = \| \mathcal{S}_{SR}^c - \mathcal{S}_{HR}^c \|_1 . \quad (5)$$

With these learning objectives, the whole framework of MARCONet is optimized in an end-to-end manner.

4. Experiments

Training Data. In this work, we take Chinese characters as our initial exploration for super-resolving complex characters but with regularity. The same method can be extended to other languages by retraining the framework with specific data. Since a text image is usually composed of several characters, here we use the Chinese corpus from Xu *et al.* [71], which contains tens of millions of common items. We also collect 182 font families to introduce diverse structures. PIL toolbox is adopted to synthesize these text images, rendered with random RGB values, font sizes and locations. The text background is obtained from DIV2K [1] and Flick2K [59] datasets, in which each image is randomly cropped and up-sampled to $\times 4 \sim 16$ to synthesize the complex background. With this toolbox, we can obtain the HR text images I_{HR}, together with the classification label, bounding box and the ground-truth structure image $\mathcal{S}_{HR}^c$ of each character. The degradation pipeline presented in BSRGAN [75] and Real-ESRGAN [66] are applied online to degrade the HR image to LR input (see input in Figure 3). Since each text image usually contains varying numbers of characters, we pad zeros to the width dimension to keep the structure unchanged.

Implementation Details. The training of the whole MAR-CONet and the pre-training for generative structure prior are conducted on a server with four Tesla V100 GPUs. The batch size is set to 2 and 16, respectively. We employ Adam [32] as the optimizer. The initial learning rate lr is set to $1e-4$ and decreased by 0.5 when $\mathcal{L}_{rec}$ reaches a stable range on the validation set. The lr for StyleGAN in the text SR training stage is set to $1e-6$ for fine-tuning. The height of HR images is set to 128. Color jittering [80] is used to increase

Figure 5. Visual comparison on ×2 (1~2 columns) and ×4 (3~4 columns) SR. The stroke details are best viewed by zooming in.

Table 1. Quantitative comparison on ×2 and ×4 SR.

Methods	×2				×4			
	PSNR↑	SSIM↑	LPIPS↓	Acc.↑	PSNR↑	SSIM↑	LPIPS↓	Acc.↑
SRCNN [12]	23.34	.868	.095	82.17	19.56	.750	.218	71.41
ESRGAN [68]	23.72	.891	.085	83.25	19.86	.778	.204	72.12
TSRN [64]	24.27	.893	.094	85.82	20.67	.782	.201	76.94
TBSRN [7]	25.62	.906	.074	92.37	21.07	.794	.186	84.32
TATT [41]	25.84	.907	.073	93.26	21.49	.800	.185	85.17
Ours	**28.32**	**.941**	**.033**	**99.16**	**24.04**	**.873**	**.096**	**92.53**

image diversity. It takes two days for pre-training StyleGAN and nearly four days for training the whole MARCONet.

Baselines. Since there are few works studying the problem of these types of text images on SR task, we compare with two types of methods following TATT [41], *i.e.*, general image SR (*i.e.*, SRCNN [12], ESRGAN [68]) and recent text image SR (*i.e.*, TSRN [64], TBSRN [7] and TATT [41]). For a fair comparison, we modify their codes to handle ×2 and ×4 SR tasks, and carefully retrain them with the same training data as Ours. In particular, the retrained TBSRN adopts the ground-truth text and bounding boxes for its pre-trained content-aware and position-aware modules. As for TATT, we also use the ground-truth text to pre-train its TPG module. We report the quantitative and qualitative results on our synthetic dataset, which contains 1,000 LR/HR pairs for ×2 and ×4 tasks, respectively. These LR inputs are injected with random noise, blurring and JPEG compression, to simulate real-world degradation. We also provide the results on real-world text images collected from different sources. More results can be found in the suppl.

4.1. Quantitative Comparison

Table 1 summarizes the PSNR, SSIM and LPIPS [76] of different methods on the synthetic dataset. An additional pre-trained CRNN [56] is used to compute the recognition accuracy of SR results (Acc.), serving as an auxiliary indicator for the restoration performance. Note that the synthetic data is non-trivial given its low quality. Without specifically designed components for text images, general SR methods

perform poorly as expected. By incorporating the recognition information, both TBSRN [7] and TATT [41] show obvious improvement, especially in retaining the semantic layout of character for recognition (see Acc. columns). Our method achieves the best quantitative results and outperforms others with a large margin (*e.g.*, 2.5 dB higher than the second-best method on average). Even with severe degradation (*i.e.*, ×4), our method still attains favorable performance, which can be ascribed to the effectiveness of the learned generative structure prior for each character.

4.2. Qualitative Comparison

Figure 5 shows the visual comparison on our synthetic LR images for ×2 and ×4 tasks. With careful retraining, competing methods perform favorably when the LR input has slight degradation, but they are not able to recover the unique structure of characters with complex strokes. When the degradation turns more severe, all of them fail to restore satisfactory results (see the last column). The results yielded by TBSRN [7] and TATT [41] suggest that high-level recognition constraints bring limited benefit to the SR task, especially in coping with diverse and complex font structures and styles. With the guidance of characters' generative structure prior, our method shows compelling results that exhibit more consistent structures with ground-truth (GT). It is interesting to observe the strong performance of MARCONet in restoring heavily degraded LR input that is challenging for humans to recognize (last column of Figure 5).

Apart from the evaluation on synthetic LR input, we also evaluate our method on real-world LR text images. We collect real-world LR text images from different sources, *e.g.*, invoices, license plates, scanned documents, road signs, plaques, video captions, and scene texts from TextZoom [64]. Each text region is cropped by CnSTD[2]. We compare with two competing methods, *i.e.*, TBSRN [7] and TATT [41], which are specifically designed for text images and achieve the top quantitative results in Table 1. The results are shown
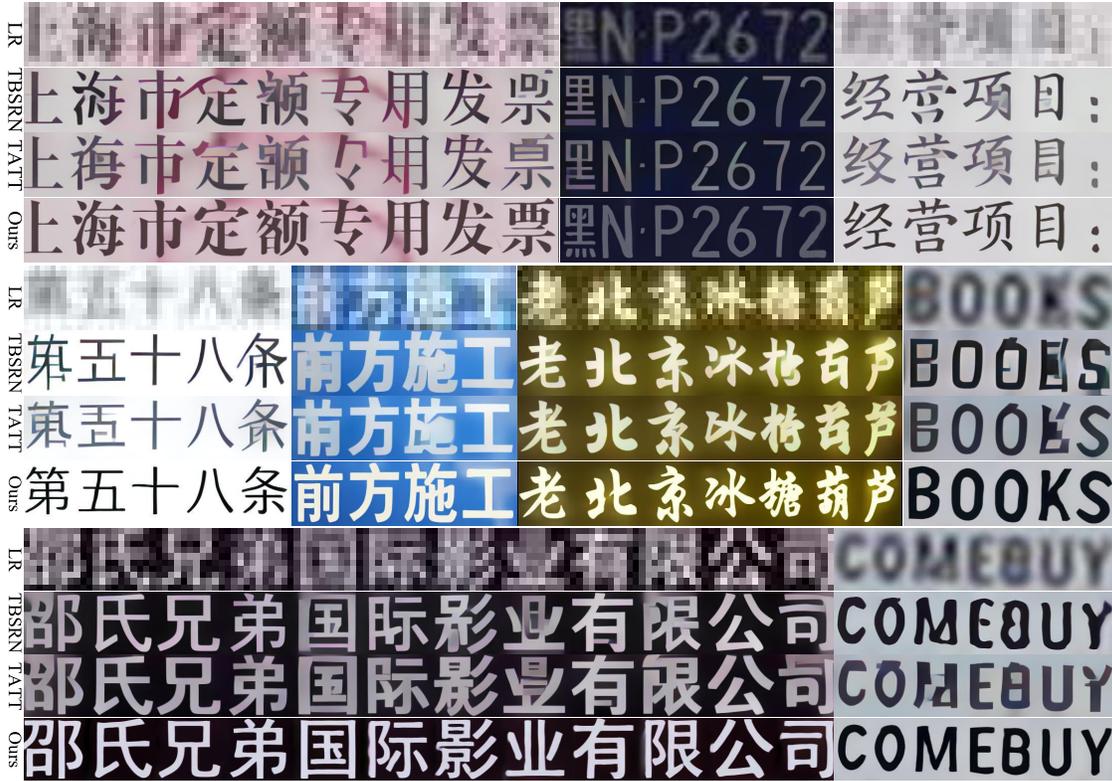
---

[2]https://github.com/breezedeus/CnSTD

Figure 6. Visual comparison on real-world LR text images, including invoice, license plate, scanned document, road sign, plaque, *etc.*

in Figure 6. Benefiting from our synthetic training data, the competing methods achieve comparable performance on these characters with simple structures, even for LR inputs corrupted by unknown and complex degradation. The positive results suggest that our synthetic HR text images are of high quality and well-suited as ground-truth for the SR task. The usage of degradation models from BSRGAN [75] and Real-ESRGAN [66] further benefits these methods in synthesizing realistic LR inputs. However, both of these two methods fail to generate plausible results when the LR character is degraded severely or contains complex structures. In contrast, with the proposed generative structure prior, our method achieves the best performance, not only in visual quality, but also in the generation of accurate structures.

4.3. Ablation Study

Analyses of $\mathcal{W}$ Space. The $\mathcal{W}$ space controls the style of a character's structure prior. Figure 7 shows the structure image $\mathcal{S}_{SR}^c$ from our StyleGAN with $w \in \mathcal{W}$ and character indexes predicted from the LR input. One can observe that w can well capture the manifold of font styles, *e.g.*, the thickness of different font families (1-*st* column) and different font sizes and locations (2-*nd* column). We also interpolate two w from different LR inputs and demonstrate their results in the 3∼5 rows. One can see that the generated structures change smoothly, indicating that w in our framework also has the editable ability like the original StyleGAN. Finally,

we use the predicted w from a real-world LR video frame (*The Dream of the Red Chamber*) and explore font style transfer. Even though the LR input contains a traditional Chinese character and the font family does not exist in our training data, our predicted w can well capture its style and successfully transfer it to other characters. Besides, the w latent and character code c are disentangled, judging from the good separation of roles between the w and c in style control and storing the unique structure for each character.

Analyses of Variants. Here we consider the following variants to evaluate each part of our MARCONet, 1) Ours (*UNet*) & Ours (*UNet†*): only taking ResNet45 and UNet to perform the SR task and increasing its model parameters, respectively, 2) Ours (*CRNN*) & Ours (*PSP*): replacing our Transformer-based character classification and w inversion branches with CRNN [56] and PSP encoder [54], respectively, 3) Ours (*D−*): only taking the SR and HR images as input of discriminator without concatenating their structure images, 4) Ours (*w/o S*): removing StyleGAN and directly incorporating the code c on each LR character features through spatial affine transformation, 5) Ours (*w/o C*): removing codebook and directly taking each LR character features into a pre-trained StyleGAN that produces colored output like GFPGAN [65], 6) Ours (*#32*) & Ours (*#64*): only using the structure prior transform module on feature sizes of either 32×32 or 64×64. Their performance on our synthetic test set is shown in Table 2 and Figure 8. We can observe
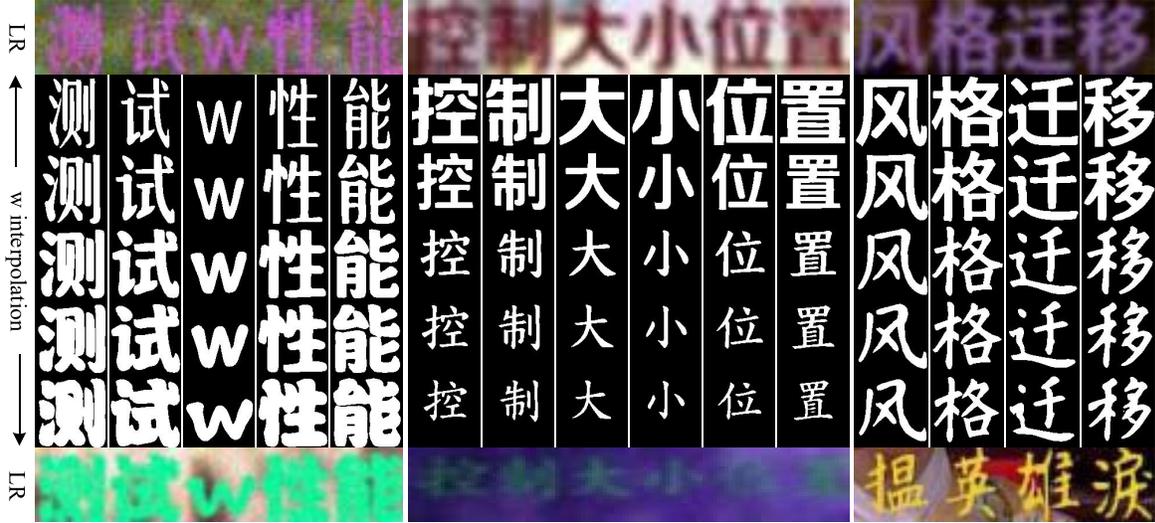
Figure 7. The $\mathcal{W}$ space in StyleGAN controls the font style of characters. The 1-*st* and 7-*th* rows are LR input. The 2-*nd* and 6-*th* rows are output of StyleGAN with w and code indexes from their LR input. The remaining rows are the interpolation results of w from two LR inputs.

Table 2. Comparison of different variants of the proposed method.

Variants	×2				×4			
	PSNR↑	SSIM↑	LPIPS↓	Acc.↑	PSNR↑	SSIM↑	LPIPS↓	Acc.↑
Ours (*UNet*)	23.69	.890	.094	82.63	19.84	.777	.209	71.96
Ours (*UNet†*)	23.71	.891	.092	82.74	19.87	.778	.203	72.16
Ours (*CRNN*)	28.00	.931	.034	98.87	23.98	.871	.101	92.41
Ours (*PSP*)	28.04	.932	.035	99.07	23.99	.871	.099	92.49
Ours (*D⁻*)	28.29	.940	.034	99.11	23.94	.870	.098	92.48
Ours (*w/o S*)	25.86	.907	.074	94.64	21.57	.802	.182	86.63
Ours (*w/o C*)	25.53	.905	.041	92.25	21.39	.798	.128	83.71
Ours (*#32*)	28.16	.937	.039	98.86	23.96	.870	.101	92.16
Ours (*#64*)	28.30	.940	.035	99.10	24.02	.872	.099	92.33
Ours (*Full*)	**28.32**	**.941**	**.033**	**99.16**	**24.04**	**.873**	**.096**	**92.53**

that a) Ours (*UNet*) and Ours (*UNet†*) perform on par with general SR methods, and indicate that a mere increase in model capacity would not bring significant performance improvement; b) the Transformer-based encoder is more effective than CRNN [56] and PSP encoder [54], which may be caused by the global attention and thus contribute better to the classification and GAN inversion tasks. It is observed that Ours (*CRNN*) sometimes makes wrong predictions on character. While the SR leads to a clear image, the character loses its actual semantic layout (see red box in Figure 8); c) the concatenation of structure image on discriminator is beneficial for emphasizing the structure prior more on LR input and thus boosting the final SR result; d) by removing StyleGAN and codebook, Ours (*w/o S*) and Ours (*w/o C*) show obvious inferior results, indicating that codebook can benefit the recognition accuracy while SyleGAN contributes more on the visual quality (*e.g.*, LPIPS); e) Ours (*#64*) has slightly better results than Ours (*#32*) but both of them are inferior to Ours (*Full*) that deploys a multi-scale structure prior transform module. From these analyses, we conclude that the Transformer-based encoder, codebook, StyleGAN and multi-scale structure prior transform module are all crucial
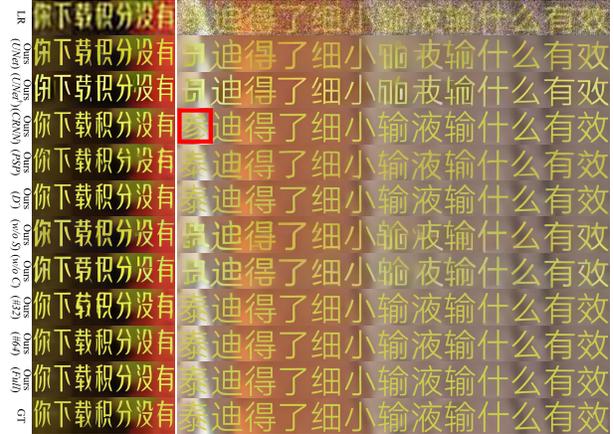

Figure 8. Visual comparison of different variants.

for achieving good performance in Ours (*Full*).

5. Conclusion

In this work, we made the first attempt to embed the generative structure prior for blind SR of text images. The combination of a codebook for storing distinctive character-specific codes and a retrofitted StyleGAN for controlling font style cope well with complicated structures, high similarity between characters, and a great variety of font styles. We have shown that such structure prior is beneficial for super-resolving LR text images, even for those with severe degradation. We can potentially extend it to other text-related tasks, *e.g.*, text image completion for ancient documents, font style transformation, and few-shot font generation.

References

[1] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 2017. 5

[2] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: StyleGAN inversion with hypernetworks for real image editing. In *CVPR*, 2022. 4

[3] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-GAN. In *NeurIPS*, 2019. 2

[4] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, 2019. 2

[5] David Capel and Andrew Zisserman. Super-resolution enhancement of text image sequences. In *ICPR*, 2000. 2

[6] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. GLEAN: Generative latent bank for large-factor image super-resolution. In *CVPR*, 2021. 3

[7] Jingye Chen, Bin Li, and Xiangyang Xue. Scene text telescope: Text-focused scene image super-resolution. In *CVPR*, 2021. 1, 2, 3, 6

[8] Jingye Chen, Haiyang Yu, Jianqi Ma, Bin Li, and Xiangyang Xue. Text gestalt: Stroke-aware scene text image super-resolution. In *AAAI*, 2022. 1, 3

[9] Gerald Dalley, Bill Freeman, and Joe Marks. Single-frame text super-resolution: A bayesian approach. In *ICIP*, 2004. 2

[10] Berk Dogan, Shuhang Gu, and Radu Timofte. Exemplar guided face image super-resolution without facial landmarks. In *CVPRW*, 2019. 3

[11] Brian Dolhansky and Cristian Canton Ferrer. Eye in-painting with exemplar generative adversarial networks. In *CVPR*, 2018. 3

[12] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE TPAMI*, 2015. 2, 6

[13] Chao Dong, Ximei Zhu, Yubin Deng, Chen Change Loy, and Yu Qiao. Boosting optical character recognition: A super-resolution approach. *arXiv preprint arXiv:1506.02211*, 2015. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 4

[15] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 3

[16] Ross Girshick. Fast R-CNN. In *ICCV*, 2015. 4

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2, 4, 5

[18] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 2006. 4

[19] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *CVPR*, 2019. 2

[20] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code GAN prior. In *CVPR*, 2020. 3

[21] Shuhang Gu, Wangmeng Zuo, Shi Guo, Yunjin Chen, Chongyu Chen, and Lei Zhang. Learning dynamic guidance for depth image enhancement. In *CVPR*, 2017. 3

[22] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. VQFR: Blind face restoration with vector-quantized dictionary and parallel decoder. In *ECCV*, 2022. 3

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 5

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[25] Xueqi Hu, Qiusheng Huang, Zhengyi Shi, Siyuan Li, Changxin Gao, Li Sun, and Qingli Li. Style transformer for image inversion and editing. In *CVPR*, 2022. 4

[26] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 5

[27] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. Depth map super-resolution by deep multi-scale guidance. In *ECCV*, 2016. 3

[28] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *CVPRW*, 2020. 2

[29] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 5

[30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3

[31] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 3

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[33] Xiaoming Li, Chaofeng Chen, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. From face to natural image: Learning real degradation for blind image super-resolution. In *ECCV*, 2022. 2

[34] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *ECCV*, 2020. 3

[35] Xiaoming Li, Wenyu Li, Dongwei Ren, Hongzhi Zhang, Meng Wang, and Wangmeng Zuo. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *CVPR*, 2020. 3

[36] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *ECCV*, 2018. 3

[37] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep joint image filtering. In *ECCV*, 2016. 3

[38] Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *ECCV*, 2022. 2

[39] Zhengxiong Luo, Yan Huang, Shang Li, Liang Wang, and Tieniu Tan. Unfolding the alternating optimization for blind super resolution. In *NeurIPS*, 2020. 2

[40] Jianqi Ma, Shi Guo, and Lei Zhang. Text prior guided scene text image super-resolution. *arXiv preprint arXiv:2106.15368*, 2021. 1, 2, 3

[41] Jianqi Ma, Zhetong Liang, and Lei Zhang. A text attention network for spatial deformation robust scene text image super-resolution. In *CVPR*, 2022. 1, 2, 3, 6

[42] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. PULSE: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, 2020. 3

[43] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 5

[44] Yongqiang Mou, Lei Tan, Hui Yang, Jingying Chen, Leyuan Liu, Rui Yan, and Yaohong Huang. Plugnet: Degradation aware scene text recognition supervised by a pluggable super-resolution unit. In *ECCV*, 2020. 1, 2

[45] Shimon Nakaune, Satoshi Iizuka, and Kazuhiro Fukui. Skeleton-aware text image super-resolution. In *BMVC*, 2021. 1, 2, 3

[46] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *ICCVW*, 2019. 3

[47] Jinshan Pan, Zhe Hu, Zhixun Su, and Ming-Hsuan Yang. Deblurring face images with exemplars. In *ECCV*, 2014. 3

[48] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE TPAMI*, 2021. 3

[49] Clément Peyrard, Moez Baccouche, Franck Mamalet, and Christophe Garcia. ICDAR2015 competition on text image super-resolution. In *ICDAR*, 2015. 2

[50] Rui Qin, Bin Wang, and Yu-Wing Tai. Scene text image super-resolution via content perceptual loss and criss-cross transformer blocks. *arXiv preprint arXiv:2210.06924*, 2022. 1, 2, 3

[51] Yuhui Quan, Jieting Yang, Yixin Chen, Yong Xu, and Hui Ji. Collaborative deep learning for super-resolving blurry text images. *IEEE Transactions on Computational Imaging*, 2020. 3

[52] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *ICCV*, 2019. 3

[53] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 4

[54] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a StyleGAN encoder for image-to-image translation. In *CVPR*, 2021. 4, 7, 8

[55] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 5

[56] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE TPAMI*, 2016. 3, 6, 7, 8

[57] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *CVPR*, 2018. 2

[58] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 5

[59] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. NTIRE 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, 2017. 5

[60] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for StyleGAN image manipulation. *ACM TOG*, 2021. 4

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 4

[62] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *CVPR*, 2021. 2

[63] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity GAN inversion for image attribute editing. In *CVPR*, 2022. 4

[64] Wenjia Wang, Enze Xie, Xuebo Liu, Wenhai Wang, Ding Liang, Chunhua Shen, and Xiang Bai. Scene text image super-resolution in the wild. In *ECCV*, 2020. 2, 6

[65] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021. 3, 7

[66] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *ICCVW*, 2021. 1, 2, 5, 7

[67] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 2018. 5

[68] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *ECCVW*, 2018. 6

[69] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. RestoreFormer: High-quality blind face restoration from undegraded key-value pairs. In *CVPR*, 2022. 3

[70] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *ECCV*, 2020. 2

[71] Bright Xu. Nlp chinese corpus: Large scale chinese corpus for nlp, 2019. 5

[72] Xiangyu Xu, Deqing Sun, Jinshan Pan, Yujin Zhang, Hanspeter Pfister, and Ming-Hsuan Yang. Learning to super-resolve blurry face and text images. In *ICCV*, 2017. 2

[73] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. GAN prior embedded network for blind face restoration in the wild. In *CVPR*, 2021. 3

[74] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 5

[75] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, 2021. 1, 2, 5, 7

[76] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[77] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *CVPR*, 2019. 2

[78] Cairong Zhao, Shuyang Feng, Brian Nlong Zhao, Zhijun Ding, Jun Wu, Fumin Shen, and Heng Tao Shen. Scene text image super-resolution via parallelly contextual attention network. In *ACM MM*, 2021. 3

[79] Minyi Zhao, Miao Wang, Fan Bai, Bingjia Li, Jie Wang, and Shuigeng Zhou. C3-STISR: Scene text image super-resolution with triple clues. In *IJCAI*, 2022. 1, 2, 3

[80] Barret Zoph, Ekin D Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V Le. Learning data augmentation strategies for object detection. In *ECCV*, 2020. 5