Choose What You Need: Disentangled Representation Learning for Scene Text Recognition, Removal and Editing

Boqiang Zhang Hongtao Xie* Zuan Gao Yuxin Wang
University of Science and Technology of China
{cyril,zuangao}@mail.ustc.edu.cn {htxie,wangyx58}@ustc.edu.cn

Abstract

Scene text images contain not only style information (font, background) but also content information (character, texture). Different scene text tasks need different information, but previous representation learning methods use tightly coupled features for all tasks, resulting in sub-optimal performance. We propose a Disentangled Representation Learning framework (DARLING) aimed at disentangling these two types of features for improved adaptability in better addressing various downstream tasks (choose what you really need). Specifically, we synthesize a dataset of image pairs with identical style but different content. Based on the dataset, we decouple the two types of features by the supervision design. Clearly, we directly split the visual representation into style and content features, the content features are supervised by a text recognition loss, while an alignment loss aligns the style features in the image pairs. Then, style features are employed in reconstructing the counterpart image via an image decoder with a prompt that indicates the counterpart's content. Such an operation effectively decouples the features based on their distinctive properties. To the best of our knowledge, this is the first time in the field of scene text that disentangles the inherent properties of the text images. Our method achieves state-of-the-art performance in Scene Text Recognition, Removal, and Editing.

1. Introduction

As an important information carrier, language is widely found in natural scenes. Scene text is a significant topic in scene understanding and perception. There have been a number of researches on scene text including Scene Text Recognition (STR), Scene Text Editing (STE), Scene Text Removal (STRM), *etc*. These researches are widely used in human–computer interaction [4, 41], cross-modal understanding [13, 21, 22, 26], automatic pilots, *etc*.

*The corresponding author.

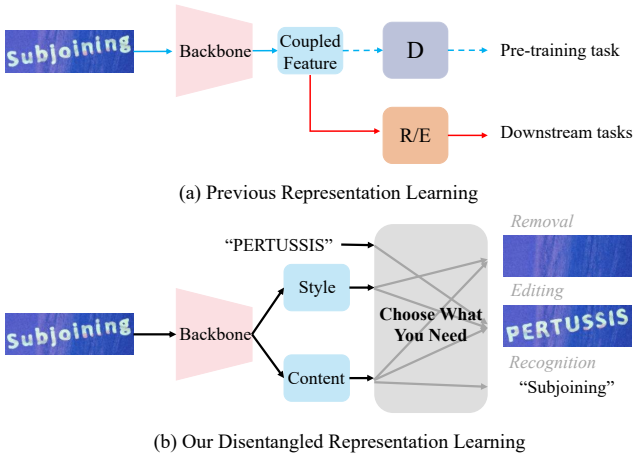

Figure 1. (a) The pipeline of previous representation learning methods that use a tightly coupled feature for all tasks. 'D' means decoder and 'R/E' represents the recognizer or eraser. (b) Our decoupled representation learning framework for multi-tasking.

In contemporary research, numerous studies try to leverage representation learning to enhance feature quality, thereby improving performance in downstream tasks. Within the domain of scene text, approaches employing mask image modeling (MIM) and feature contrastive learning (CL) have garnered notable success [16, 50]. As depicted in Fig. 1 (a), these works first use a decoder to pretrain the backbone, facilitating the completion of the pretasks. Then, the pre-trained backbone is used for fine-tuning with a task-specific decoder. Although achieve impressive performance, it is apparent that these pipelines face challenges. Using the same representation for different downstream tasks including discriminative and generative is sub-optimal, limiting the generalization of the methods.

To address the above issue, we investigate distinctive properties of scene text images that set them apart from general scene images. Specifically, cropped scene text images contain a focused region with high information density alongside a diverse background. We categorize these

distinctive attributes into style and content features. Style features encompass the background and text style elements (such as font, color, tilt, *etc*.), while content features encompass content and texture details. In STR, the essential information is content features. Style information is considered as noise that hinders accurate recognition. Nevertheless, in STE, the process can be divided into two stages: text removal and text rendering. The text removal stage resembles an image inpainting task that reconstructs the background pixels in the original text area. This process requires content features for stroke localization and style features for background reconstruction. The text rendering stage relies on content features to generate text strokes and style features to define the font. Therefore, different downstream tasks need different information (Fig. 1 (b)), and features irrelevant to a particular task may hinder task completion. The decomposition of these two types of features can be applied to various scene text tasks.

In this paper, we explore the representation learning from a novel perspective by considering the above unique properties of scene text. To decouple two types of features, we introduce a method as illustrated in Figure 2. To begin with, we synthesized a dataset containing pairs of images to facilitate the design of subsequent decoupled pre-training methods. The image pairs contain identical backgrounds and styles but differ in content. These pairs are simultaneously inputted into the network. Features of these image pairs are extracted using a ViT-based backbone along with a decoupling block. We directly divide the output tokens into two parts and subsequently achieve decoupling through different losses. Multi-task decoder utilizes these decoupled features to perform various tasks. To achieve the intended disentanglement of features and effectively train the model, we propose a training paradigm. This paradigm involves exclusively using the content features for recognition while aligning the style features of the image pairs. Additionally, style features are used to reconstruct the counterpart image with a text prompt. After pre-training, the multi-task decoder can effectively handle both generative and discriminative tasks, while also serving as a great starting for fine-tuning. Compared with previous methods, our method can accomplish both generative and discriminative tasks without the need for additional modules in the pre-training stage.

To summarize, our contributions are as follows:

- We propose to decouple the features for scene text tasks, leveraging the distinctive properties of scene text images. This endeavor may encourage the research community to reconsider the distinctiveness of textual images.
- We propose a training paradigm for feature disentanglement in scene text images.
- We propose a new synthetic dataset for the training and evaluating of scene text editing.

Compared with previous methods, our DARLING achieves state-of-the-art performance in scene text recognition, editing, and removal.

2. Related Works

2.1. Scene Text Tasks

We mainly review three widely researched scene text tasks: Scene Text Recognition, Editing, and Removal.

Scene Text Recognition (STR) has been a significant research term in computer vision. There are three kinds of methods: CTC-based, attention-based, and segmentation-based. (1) CTC-based methods [6, 30, 39] use a Connectionist Temporal Classification (CTC) [9] decoder to directly transform the image features into text sequences. Such an operation has a fast inference speed but a limited performance. (2) Attention-based methods [1, 7, 38, 40, 44, 47, 55, 57] use a learnable query and a cross-attention operation to decode the recognition result. Such methods have high recognition accuracy but usually not as fast as CTC-based methods. (3) Segmentation-based methods [23, 32] treat the scene text images as a combination of characters. Recently, some works [10, 11] use pre-processing algorithms to obtain the segmentation map which can further direct the model to get better performance. Our method uses the attention-based decoder because of its high performance and elegant consistency with the transformer structure.

Scene Text Editing (STE) aims to replace text in a scene image with new text while maintaining the original background and styles. Due to the great development of Generating Adversarial Networks (GAN) [8], GAN-based scene text editing methods attract increasing research interest. SRNet [48] first proposes to divide the editing task into three sub-process: background inpainting, text conversion and fusion, which inspires many subsequent works [35, 37, 51]. Nowadays, with the advances in diffusion models, some works [2–4, 41] use the diffusion process to achieve excellent editing results. However, due to the complexity of the diffusion model, these methods are less efficient. From a unified perspective, we abandon the use of diffusion and GAN, in favor of an efficient attention structure that achieves high-quality generation through decoupled features.

Scene Text Removal (STRM) can be seen as the first step in STE. Recently a variety of approaches [5, 24, 25, 28, 33, 42, 46, 56] tried to accomplish text removal on a whole scene image. However, as texts are sparse in the scene images, text erasing directly on the scene image tends to affect non-text areas and seems uncontrollable. For uniformity, our approach uses cropped images.

2.2. Representation Learning for Scene Text Tasks

Several studies employ representation learning to enhance features for scene text tasks. MaskOCR [29] employs

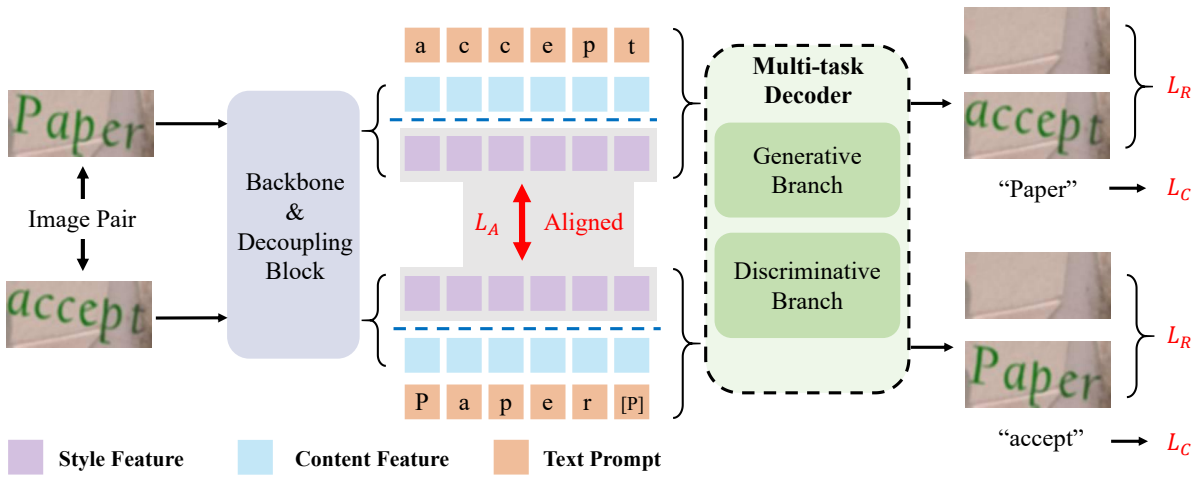

Figure 2. The pipeline and training paradigm of our DARLING. The Decoupling Block divides features from the backbone into style and content features. The multi-task decoder processes these features to perform both discriminative and generative tasks. '[p]' is the padding symbol. Image pairs with the same style but different content are input. The style features are aligned and recognition loss supervises the content features to eliminate the style from content features.

masked image modeling (MIM) to glean implicit knowledge from substantial unlabeled data. DiG [50] proposes the combined use of MIM and contrastive learning to bolster backbone representation for accurate text recognition. [16] introduces a high-quality dataset for unsupervised pre-training, leading to significant improvements in STR. In the field of STRM, recent efforts [5, 33, 46] leverage synthetic datasets for pre-training, enhancing performance and yielding more robust features. However, these methods focus on using a tightly coupled feature to accomplish downstream tasks, which limits the generalization of the method. SimAN [27] introduces a Similarity-Aware Normalization module, implicitly decomposing features and achieving notable performance across some tasks. Despite this, the decomposition using instance normalization is not thorough. We first propose to address multiple tasks by feature disentanglement and introduce a disentangled training paradigm.

3. Method

Our DARLING is a pre-training method for scene text tasks including STR, STE, and STRM. In this section, we detail the pipeline of the proposed method in Sec. 3.1. Then, we propose the disentanglement training paradigm and multi-task decoder in Sec. 3.2 and Sec. 3.3. Finally, we describe the training objective in Sec. 3.4.

3.1. Pipeline

The pipeline of our DARLING is shown in Fig. 2. Given a scene text image $\mathbf{I}$, the feature tokens $\mathbf{F} \in \mathbb{R}^{L \times D}$ are first extracted by a transformer-based backbone. Then, the features are fed into a decoupling block. We directly separate the output of the decoupling block $\mathbf{F}_D$ into two components denoted as $\mathbf{F}_D = [\mathbf{F}_S, \mathbf{F}_C]$, where $\mathbf{F}_S \in \mathbb{R}^{\frac{L}{2} \times D}$ indicates the style features and $\mathbf{F}_C \in \mathbb{R}^{\frac{L}{2} \times D}$ represents the content features. The decoupling block comprises multiple self-attention layers designed to capture long-range information essential for feature separation. Subsequently, the two kinds of features are inputted into the multi-task decoder with a text prompt indicating the desired text for rendering. This decoder is capable of handling both discriminative tasks like recognition and generative tasks such as editing and removal. Meanwhile, multi-task supervision aids in constraining features, fostering feature disentanglement, and acquiring diverse features.

3.2. Disentangled Training Paradigm

To achieve the expected disentanglement of the two feature types, we introduce a pre-training paradigm. The concept is depicted in Fig. 2. First, we employ a synthesis engine to create pairs of images that share identical backgrounds and fonts but differ in content. Subsequently, pairs of images are simultaneously fed into our proposed network. Due to the similarity in background and font, which we categorize as style features, we employ an alignment loss $\mathcal{L}_A$ to align the style features of image pairs. The formulation is as follows:

$$\mathcal{L}_A = \frac{1}{2} \sum (\mathbf{F}_{S1} - \mathbf{F}_{S2})^2, \tag{1}$$

where $\mathbf{F}_{S1}$ and $\mathbf{F}_{S2}$ represents the style features of two images, respectively.

Furthermore, we exchange the content of two images by using each other's content as the text prompt $\mathbf{C}_T$. The decoder is tasked to reconstruct the other image and supervised using a reconstruction loss $\mathcal{L}_R$. In the multi-task de-

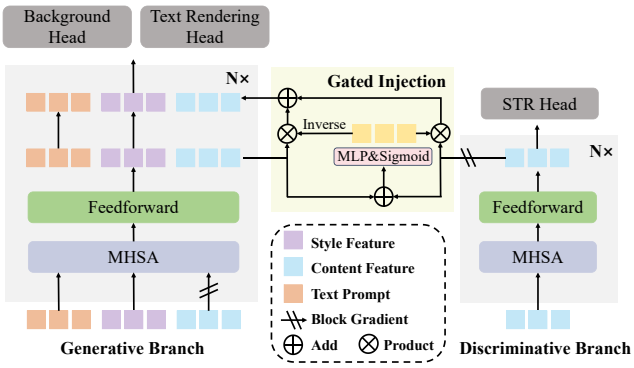

Figure 3. The structure of Multi-task Decoder. It comprises the Generative Branch (GEB) and the Discriminative Branch (DIB), each dedicated to specific tasks. Gated Injection strategy is proposed to convey fine-grained details from DIB to GEB.

coder, the text recognition task only uses the content feature. Therefore, the supervision of recognition loss will eliminate the style information from content features, which further ensures the decoupling.

Discussion. Through the utilization of the proposed disentangled pre-training, the model is constrained to align the common style features shared between a pair of images and use it for text editing. Meanwhile, since scene text recognition is inherently a fine-grained task that solely requires content features while treating style features as noise, we employ the recognition loss to guide $\mathbf{F}_C$ toward content features. As a result of the pre-training and the set of loss functions, the intended decoupling of the two feature types is achieved. Additionally, in contrast to previous works that solely pre-train the backbone, our method includes pre-training of the multi-task decoder, which proves beneficial for subsequent fine-tuning processes.

3.3. Multi-task Decoder

The Multi-task Decoder (MTD) is designed to handle both generative tasks and discriminative tasks. As depicted in Fig. 3, due to the distinct feature requirements of these tasks, the MTD structure comprises two branches: generative and discriminative. Additionally, we propose a gated injection strategy to integrate the information from the discriminative branch into the generative branch.

The discriminative branch (DIB) aims to accomplish STR task. This task demands fine-grained details but suffers from interference with style features. The input of the discriminative branch is $\mathbf{F}_C$ and the branch consists of N self-attention layers to further extract the fine-grained details. Subsequently, the output features $\mathbf{F}_C^N$ are employed for text prediction by an STR head, utilizing a cross-attention operation and a linear projection. The prediction process is formalized in Eq. (2), we have projection weights $\mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{D\times D}$ and $\mathbf{W}_R \in \mathbb{R}^{D\times C}$. $\mathbf{R} \in \mathbb{R}^{T\times C}$ represents the recognition result and $\mathbf{Q} \in \mathbb{R}^{T\times D}$ is a learnable query for decoding. T denotes the maximum text length, while C represents the number of character classes.

$$\mathbf{R} = \mathbf{Q}(\mathbf{F}_C^N\mathbf{W}_K)^\top(\mathbf{F}_C^N\mathbf{W}_V)^\top\mathbf{W}_R. \tag{2}$$

The generative branch (GEB) is designed to fulfill generative tasks such as STE and STRM. This branch is followed by two heads: the background head and the text rendering head, responsible for obtaining the background reconstruction result and the edited text image, respectively. Similar to DIB, GEB shares the same structure but receives different input. Considering the necessity for both content features in generating or removing text areas and style features in background reconstruction, we configure GEB's input as a three-part concatenation, as illustrated in Eq. (3). $\mathbf{C}_T$ serves as the text prompt indicating the target text we intend to render on the original image.

$$input = concat(\mathbf{C}_T, \mathbf{F}_S, \mathbf{F}_C). \tag{3}$$

In response to the absence of fine-grained details in $\mathbf{F}_C$, which are subsequently enhanced by DIB, we introduce a Gated Injection strategy to provide well-managed fine-grained information to GEB. To be specific, different layers of features in DIB represent information at different levels of detail, all under the supervision of recognition loss. We introduce an adaptive fusion mechanism that operates between each layer of DIB and GEB. As formalized in Eq. (4), $\odot$ represents dot production and $\mathcal{T}$ stands for the self-attention operation. The superscript indicates the layer number. $\mathbf{W}_G \in \mathbb{R}^{D\times 1}$ transforms the features into a gating weight $\mathbf{G}$, which is further used to control the fusion of $\hat{\mathbf{F}}_C$ and $\mathbf{F}_C$, where $\hat{\mathbf{F}}_C$ denotes the content features in each layer of GEB.

$$\begin{aligned}
\mathbf{G}^i &= Sigmoid((\hat{\mathbf{F}}_C^i + \mathbf{F}_C^i)\mathbf{W}_G), \\
\mathbf{F}_G^i &= \hat{\mathbf{F}}_C^i \odot (1 - \mathbf{G}^i) + \mathbf{F}_C^i \odot \mathbf{G}^i, \\
[\mathbf{C}_T^i, \mathbf{F}_S^i, \hat{\mathbf{F}}_C^i] &= \mathcal{T}([\mathbf{C}_T^{i-1}, \mathbf{F}_S^{i-1}, \mathbf{F}_G^{i-1}]).
\end{aligned} \tag{4}$$

Finally, the background head and the text rendering head leverage a cross-attention operation to generate the residual of the background and the text expected, similar to Eq. (3).

Overall, in our MTD, the decoupled learning approach guides the model in acquiring more discriminative features, while diverse task-guided training enhances feature diversity. Clearly, The recognition loss facilitates the decoupling of features, and the gated injection strategy combines diverse fine-grained details supervised by recognition loss to better accomplish generative tasks.

3.4. Training Objective

As shown in Eq. (5), the final objective function of the proposed method contains three parts: reconstruction loss $\mathcal{L}_R$, scene text recognition loss $\mathcal{L}_C$, and feature align loss $\mathcal{L}_A$.

$$\mathcal{L} = \mathcal{L}_R + \lambda_C \mathcal{L}_C + \lambda_A \mathcal{L}_A. \quad (5)$$

The scene text recognition loss is a cross-entropy loss which is formulated in Eq. (6), where G_t is the ground truth, T is the max length of the character sequence.

$$\mathcal{L}_C = -\frac{1}{T}\sum_{i=0}^{T} log(P(\mathbf{R}|G_t)). \quad (6)$$

4. Experiment

4.1. Datasets

For pre-training, we generate a dataset using the publicly available synthesis engine [52][1]. We use the background images provided by SynthText [12] and the lexicon provided by MJSynth [14, 15] and SynthText [12]. The dataset contains 4M image pairs for training (TSE-4M) and 10K for evaluation (TSE-10k). Some image samples are shown in Fig. 4. This dataset features more complex text styles and backgrounds, a broader array of fonts, and includes low-quality images affected by blurring, noise, *etc*. The evaluation set serves as a more comprehensive benchmark for assessing the performance of scene text editing methods. The dataset will be made publicly available.

For the evaluation of STE, we utilize the synthetic (Tamper-Syn2k) and real (Tamper-Scene) datasets introduced by MOSTEL [35] along with our STE-10k. For STR, we conduct fine-tuning on Union-L [16] for fair comparison. Then, the performance is evaluated on 7 commonly used benchmarks including Union-benchmark[16], IIIT 5K-Words (IIIT5K) [31], ICDAR2013 (IC13) [17], ICDAR2015 (IC15) [18], Street View Text (SVT) [43], Street View Text-Perspective (SVTP) [34], and CUTE80 (CUTE) [36]. Comprehensive details regarding these datasets can be found in prior works [7, 44]. As for STRM, we fine-tune and evaluate the removal part of our model on SCUT-EnsText [24].

4.2. Implementation Details

In our implementation, The layer number of the decoupling block is set to 3 and the layer number of the multi-task decoder is set to 6. Following most previous methods, the image size is fixed at 128 × 32. We conduct the experiments on 4 NVIDIA 4090 GPUs with a batch size of 384. For scene text recognition, the vocabulary size C is set to 98, including 0 - 9, a - z, A-Z, special characters, [PAD] for padding symbol and [EOS] for ending symbol. λ_A and λ_C are both set to 0.5.

The network is pre-trained end-to-end using Adam [19] optimizer with an initial learning rate set at 1e-4. The learning rate is then decayed to 1e-5 after seven epochs. The pre-training phase encompasses 200K iterations. For fine-tuning in the STR task, an additional 400K iterations are executed, maintaining the same learning rate schedule as in pre-training. For STRM, we fine-tuned 100K iterations.

[1] https://github.com/clovaai/synthtiger

Figure 4. Some sample images from our generated datasets: TSE-4M and TSE-10k. The datasets comprise more diverse images with a variety of fonts and backgrounds, including low-quality images. TSE-10k can facilitate a more comprehensive evaluation of the model's performance.

4.3. Comparisons with State-of-the-Arts

Our multi-task decoder is adept at handling both discriminative (STR) and generative (STE, STRM) tasks. In this section, we conduct a series of experiments to prove the effectiveness of the MTD and our disentangled representation learning framework.

4.3.1 Scene Text Recognition

For the task of scene text recognition, we evaluated our approach on six widely used benchmarks (common benchmarks) as well as the more diverse Union14M-Benchmark [16]. The results are shown in Tab. 1, the Baseline is the result of our model trained on STE-4M and Union-L without disentangled pre-train. Compared with it, our pre-trained model obtains a remarkable performance enhancement across all datasets. Compared with other methods, our approach surpasses the 0.5% average accuracy of the state-of-the-art model while employing fewer parameters. This outcome substantiates that our approach attains a high-quality representation adept at effectively handling diverse text images in real-world scenes. The Union14M-Benchmark encompasses a variety of challenging images, such as curved, multi-oriented, artistic, and salient text. Our significant performance improvement is evident across these datasets. Notably, our approach outperforms the SOTA performance by 4.4%, 2.6%, and 0.5% on curved, salient, and multi-oriented datasets, respectively. Our approach demonstrates a slightly diminished performance on multi-word images which contain several words within a single image. This limitation can be mitigated by employing a robust text detector.

Compared with other pre-training methods, our approach still exhibits high performance. Note that we just pre-train

Table 1. Comparison with state-of-the-art STR methods on Common benchmarks and Union-14M Benchmarks. †stands for reproducing by ourselves. ‡means the method has a pre-training stage. All methods are trained or fine-tuned on Union-L [16].

Methods	Common Benchmarks							Union14M-Benchmark								#Params
	IIIT5K 3000	IC13 1015	SVT 647	IC15 2077	SVTP 645	CUTE 288	AVG	Curve	Multi-Oriented	Artistic	Context-less	Salient	Multi-Words	General	AVG	(M)
CRNN [39]	91.8	91.8	83.8	71.8	70.4	80.9	81.6	19.4	4.5	34.2	44.0	16.7	35.7	60.4	30.7	8.3
ASTER [40]	94.3	92.6	88.9	77.7	80.5	86.5	86.7	38.4	13.0	41.8	52.9	31.9	49.8	66.7	42.1	27.2
NRTR [38]	96.2	96.9	94.0	80.9	84.8	92.0	90.8	49.3	40.6	54.3	69.6	42.9	75.5	75.2	58.2	-
SATRN [20]	97.0	97.9	95.2	87.1	91.0	96.2	93.9	74.8	64.7	67.1	76.1	72.2	74.1	75.8	72.1	-
RobustScanner [54]	96.8	95.7	92.4	86.4	83.9	93.8	91.2	66.2	54.2	61.4	72.7	60.1	74.2	75.7	66.4	-
SVTR-S [6]	95.9	95.5	92.4	83.9	85.7	93.1	91.1	72.4	68.2	54.1	68.0	71.4	67.7	77.0	68.4	10.3
SRN [53]	95.5	94.7	89.5	79.1	83.9	91.3	89.0	49.7	20.0	50.7	61.0	43.9	51.5	62.7	48.5	55
VisonLAN [45]	96.3	95.1	91.3	83.6	85.4	92.4	91.3	70.7	57.2	56.7	63.8	67.6	47.3	74.2	62.5	33
PARSeq [1]†	98.0	96.8	95.2	85.2	90.5	96.5	93.5	79.8	79.2	67.4	77.4	77.0	76.9	80.6	76.9	23.8
ABINet [7]	97.2	97.2	95.7	87.6	92.1	94.4	94.0	75.0	61.5	65.3	71.1	72.9	59.1	79.4	69.2	37
MAERec-S [16]‡	98.0	97.6	96.8	87.1	93.2	**97.9**	95.1	81.4	71.4	72.0	**82.0**	78.5	**82.4**	82.5	78.6	35.8
CCD [11]†‡	98.3	97.6	97.1	88.3	92.3	97.2	95.1	79.1	76.8	72.2	80.0	78.0	80.2	81.5	78.3	36
Baseline	98.1	98.2	96.9	87.2	91.9	96.2	94.8	79.4	71.1	68.4	77.5	75.9	77.5	81.0	75.8	18.7
DARLING (Ours)‡	**98.5**	**98.7**	**97.8**	**88.5**	**93.3**	96.5	**95.6**	**85.8**	**79.7**	**72.3**	80.4	**81.1**	78.7	**83.3**	**80.2**	18.7

Table 2. Comparison with state-of-the-art STR methods on some more complicated datasets including occlusion and wordart.

Methods	WOST	HOST	WordArt	# Params(M)
CRNN [39]	-	-	47.5	8.3
ASTER [40]	-	-	57.9	27.2
RobustScanner [54]	-	-	61.3	-
VisonLAN [45]	70.8	49.8	69.1	33
ABINet [7]	75.3	57.9	67.4	37
PARSeq [1]	73.6	55.4	79.2	23.8
MGP-Small [44]	76.0	62.8	69.0	52.6
SVTR-S [6]	74.6	58.9	65.9	10.3
CCD [11]†‡	80.6	67.3	79.3	36
Baseline	78.3	62.8	78.7	18.7
DARLING (Ours)‡	**82.5**	**70.8**	**81.2**	18.7

our model on 4M synthetic image pairs we proposed. In contrast, other pre-training methods rely on approximately 10M **real** images for pre-training. This dataset is considerably more difficult to acquire than our synthetic data.

Furthermore, we conduct experiments on more challenging datasets in Tab. 2. These datasets encompass scenes with occlusions (WOST, HOST [45]) and art characters (WortArt [49]). A more discriminative representation is required in these scenarios, and our approach achieves the best performance (1.9%, 3.5%, 1.9% performance gain) compared to previous methods. Despite being a language-independent framework, our approach yields high-quality features that offer dependable information in scenarios involving occlusion and art words.

4.3.2 Scene Text Editing

Our method uses a disentangled training paradigm, allowing for the direct acquisition of a scene text editing network without the need for fine-tuning. We assess its performance on the datasets introduced by MOSTEL [35] and our STE-10k. To comprehensively evaluate the edited images of our method on synthetic datasets, we adopt the following commonly used metrics: 1) MSE, the $L2$ distances; 2) PSNR, the ratio of peak signal to noise; 3) SSIM, the mean structural similarity; 4) FID, the distances between features extracted by Inception V3. In experiments with real-scene datasets, the absence of ground truth poses a challenge. We employ a metric named SeqAcc, which measures recognition accuracy using a widely utilized OCR engine[2]. Furthermore, we found in our experiments that the above metrics being good does not mean that the edited images are of high quality. In order to better assess the authenticity of edited images, we propose a new metric called ClassAcc. Specifically, we employ a simple convolutional network trained on the original and edited images to distinguish the authenticity of the images. The accuracy of this network's classification acts as the metric for edit quality, where lower accuracy represents more authentic generated images, thus indicating superior editing quality. For fair comparison, we resize the images to 128×32. The results are shown in Tab. 3.

It's noteworthy that other methods often necessitate more supervision and pre-training. For instance, SRNet demands supervision for text masks, text skeleton images, and standard text images, along with the use of a discriminator for adversarial training. MOSTEL, on the other hand, requires supervision for text masks and pre-training for both text removal and text recognition models. In contrast, our approach solely relies on image pairs and does not require a discriminator or additional pre-training.

Compared with other approaches, our method achieves significant performance on both synthetic and real datasets. On synthetic datasets, we have the best performance except FID score. This is due to the constraints in the synthesizer's synthesis quality, causing the ground truth image to not always be the most realistic and unique. Due to the limits of synthetic images, real-world scenarios can better reflect the

[2]https://github.com/PaddlePaddle/PaddleOCR

Table 3. Comparison with state-of-the-art STE methods on synthetic and real datasets. †stands for the methods we reproduce. ClassAcc is the metric we proposed to evaluate the realism of the generated images. The SSIM, SeqAcc, and ClassAcc are presented in percent (%).

Methods	Tamper-Syn2k					Tamper-Scene		STE-10k				
	MSE↓	PSNR↑	SSIM↑	FID↓	SeqAcc↑	SeqAcc↑	ClassAcc↓	MSE↓	PSNR↑	SSIM↑	FID↓	SeqAcc↑
pix2pix†	0.1450	9.18	34.15	127.21	6.1	13.26	94.74	0.1291	9.78	31.69	132.52	12.2
SRNet [48]†	0.0216	18.66	49.97	64.37	-	30.36	-	-	-	-	-	-
SwapText [51]†	0.0194	19.43	52.43	53.23	-	54.83	-	-	-	-	-	-
MOSTEL [35]	0.0135	20.27	56.94	**33.79**	35.9	66.54	69.81	0.0169	19.24	43.56	**18.46**	50.71
DARLING (Ours)	**0.0120**	**20.80**	**60.07**	44.48	**38.3**	**70.85**	**66.46**	**0.0100**	**21.77**	**59.95**	37.75	**61.00**

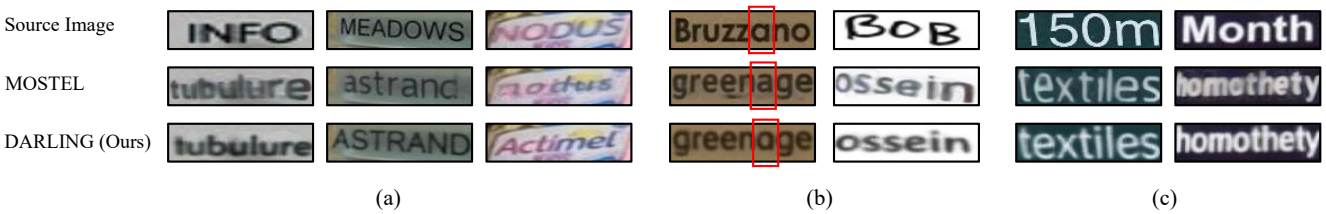

Figure 5. Qualitative examples of text editing in real scenes. (a) Comparison of the generation quality. (b) Comparison of the ability to maintain style. (c) Comparison of the realism when the generated images are clear and readable.

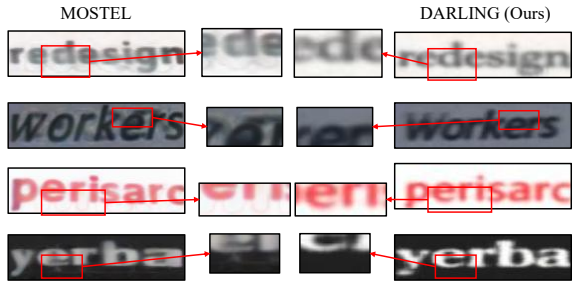

Figure 6. Comparison of generating details in real scenes. Details like artifacts, textures, and sharp edges lead to the fake appearance.

generative capability. Our model attains 10.94% SeqAcc improvement on the real dataset (Tamper-Scene), showcasing the readability of our outputs. As shown in Fig. 5 (a), the generated images exhibit more accurate text pixels. Additionally, our method effectively preserves the style of the source image, as demonstrated in Fig. 5 (b). Furthermore, we argue that achieving readability and correct stylization in generated images is insufficient for measuring generation authenticity. As shown in Fig. 5 (c), in certain instances, the generated images are easily recognizable as fake. Hence, we introduce the ClassAcc metric, and our method surpasses the state-of-the-art by 3%. We further magnify the local details of the generated images in Fig. 6. In these instances, details like artifacts, textures, and sharp edges lead to fake appearance, but our results remain high quality.

4.3.3 Scene Text Removal

Our model incorporates a background head in the multi-task decoder, enabling it to perform the scene text removal task. We feed the image into the model with the text prompt $\mathbf{C}_T$ set as '[B][E][P][P]...' to complete the STRM task, where '[B]', '[E]', and '[P]' represent the beginning, end, and padding symbols, respectively. The evaluation was conducted on the widely used real scene dataset SCUT-EnsText [24]. The results are detailed in Tab. 4. In comparison to our baseline without disentangled pre-training, our pre-training approach demonstrates a noteworthy improvement. This underscores the advantageous impact of disentangled pre-training on representation learning. When compared to state-of-the-art methods, our approach achieves the best performance across all metrics. Although our method of erasing a cropped image is naturally better than the previous method of erasing over the entire image, we can still see the superior erasing power of our method from the table comparison. Moreover, some qualitative examples are shown in Fig. 7, our method can obtain a more effective removal outcome while keeping the unrelated areas unaffected.

4.4. Ablation Study

The disentanglement of our model: We propose a disentangled representation learning framework for scene text tasks. Leveraging the disentangled training paradigm, we align style features across style-consistent image pairs. The recognition loss then guides content features to eliminate style information, achieving effective decoupling. The superior performance across various tasks substantiates the effectiveness of our disentangled representation learning paradigm. We further visualize the features in a scatter-plot with the help of PCA algorithm to validate the decoupling capability of our model. Some simple images are depicted in Fig. 8. Along the axis of the style feature, different positions roughly correspond to different backgrounds and

Table 4. Comparison with state-of-the-art STRM methods on the real scene dataset SCUT-EnsText [24]. ‡means the methods with a pre-training stage. The SSIM is presented in percent (%).

Methods	MSE↓	PSNR↑	SSIM↑	FID↓	# Params
pix2pix	0.0037	26.70	88.56	46.88	54.42
EnsNet [56]	0.0024	29.54	92.74	32.71	12.40
MTRNet++ [42]	0.0028	29.63	93.71	35.68	18.67
EraseNet [24]	0.0015	32.30	95.42	19.27	17.82
PERT [46]	0.0014	33.25	96.95	-	14.00
CTRNet [25]	0.0009	35.20	97.36	13.99	159.81
FETNet [28]	0.0013	34.53	97.01	-	8.53
PEN [5]‡	**0.0005**	35.72	96.68	-	-
ViTEraser-Tiny [33]‡	**0.0005**	36.80	97.55	10.79	65.39
Baseline	0.0011	33.11	95.90	11.03	23.70
DARLING (Ours)	**0.0005**	**38.85**	**98.25**	**10.11**	23.70

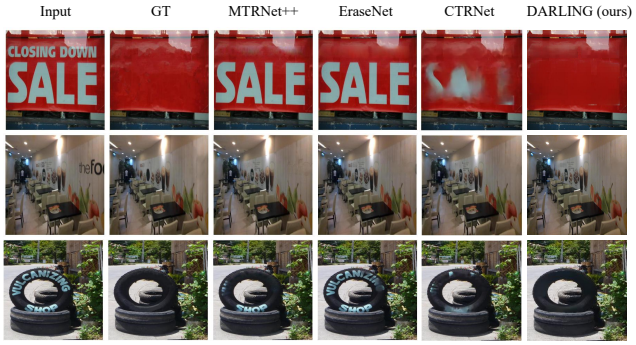

Figure 7. Some qualitative examples in STRM task.

fonts, and along the axis of the content features, different positions roughly correspond to different contents. This visualization affirms that the representation of two types of features differs significantly. When handling diverse tasks, we selectively employ different features to achieve notable performance improvements.

Effectiveness of Gated Injection: The proposed gated injection can provide well-managed fine-grained details of different levels to assist in the generation of text pixels. To substantiate this, we conducted an ablation study (see Tab. 5) and presented examples in Fig. 9. It is evident that the inclusion of the gated injection strategy significantly enhances text editing performance, resulting in a clearer and more realistic generation of text details.

5. Limitation

Compared to other self-supervised works using real-world data, our method, utilizing synthetic data, offers certain advantages. Firstly, synthetic data is more easily accessible. Secondly, unlabeled real data cannot be employed to pre-train the decoder, a distinctive feature of our method. However, our approach has its limitations. Synthetic data exhibits a domain gap with real data, which may impact performance. Additionally, a substantial amount of unlabeled real data already exists. The question arises of how

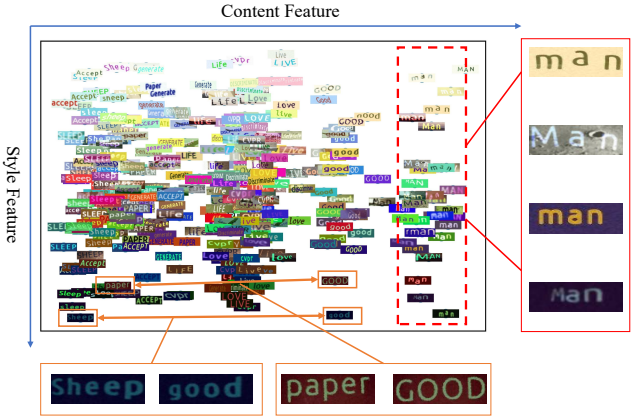

Figure 8. Feature visualization about the disentangled capabilities of our pre-trained model.

Figure 9. Some samples about the gated injection strategy. "w/o GI" is without gated injection. "w/ GI" means using gated injection.

Methods	Tamper-Syn2k STE-10k		
	MSE↓	PSNR↑	SSIM↑
w/o GI	0.0136	19.56	54.02
	0.0148	20.46	57.69
w/ GI	**0.0120**	**20.80**	**60.07**
	0.0100	**21.77**	**59.95**

Table 5. Ablation study of gated injection strategy. "w/o GI" denotes the absence of gated injection, while "w/ GI" means the utilization of gated injection.

to integrate them into our method or directly design self-supervised decoupled representation learning.

6. Conclusion

we explore the distinctions between scene text and general scene images, proposing to decouple the two distinctive features (style and content) within scene text images. Employing our disentangled representation learning framework, the model acquires more discriminative features. When addressing various downstream tasks, distinct features are utilized. Specifically, for STR, only content features are employed to eliminate style interference. For generative tasks such as STE and STRM, style features coupled with well-managed content features are utilized to generate more realistic images. Our approach achieves state-of-the-art performance in STR, STE, and STRM. We believe this work offers valuable insights into differentiating scene text images from general images, fostering inspiration for future research in the field of scene text.

Acknowledgments: This work is supported by the National Key Research and Development Program of China (2022YFB3104700), the National Nature Science Foundation of China (U23B2028, 62121002, 62102384).

References

[1] Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. In *European Conference on Computer Vision*, pages 178–196. Springer, 2022. 2, 6

[2] Haoxing Chen, Zhuoer Xu, Zhangxuan Gu, Jun Lan, Xing Zheng, Yaohui Li, Changhua Meng, Huijia Zhu, and Weiqiang Wang. Diffute: Universal text editing diffusion model. 2

[3] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *arXiv preprint arXiv:2305.10855*, 2023.

[4] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. *arXiv preprint arXiv:2311.16465*, 2023. 1, 2

[5] Xiangcheng Du, Zhao Zhou, Yingbin Zheng, Xingjiao Wu, Tianlong Ma, and Cheng Jin. Progressive scene text erasing with self-supervision. *Computer Vision and Image Understanding*, 233:103712, 2023. 2, 3, 8

[6] Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoting Yin, Tianlun Zheng, Chenxia Li, Yuning Du, and Yu-Gang Jiang. Svtr: Scene text recognition with a single visual model. *arXiv preprint arXiv:2205.00159*, 2022. 2, 6

[7] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7098–7107, 2021. 2, 5, 6

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[9] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006. 2

[10] Tongkun Guan, Chaochen Gu, Jingzheng Tu, Xue Yang, Qi Feng, Yudi Zhao, and Wei Shen. Self-supervised implicit glyph attention for text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15285–15294, 2023. 2

[11] Tongkun Guan, Wei Shen, Xue Yang, Qi Feng, Zekun Jiang, and Xiaokang Yang. Self-supervised character-to-character distillation for text recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19473–19484, 2023. 2, 6

[12] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2315–2324, 2016. 5

[13] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091, 2022. 1

[14] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, 2014. 5

[15] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *International journal of computer vision*, 116(1):1–20, 2016. 5

[16] Qing Jiang, Jiapeng Wang, Dezhi Peng, Chongyu Liu, and Lianwen Jin. Revisiting scene text recognition: A data perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20543–20554, 2023. 1, 3, 5, 6

[17] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*, pages 1484–1493. IEEE, 2013. 5

[18] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1156–1160. IEEE, 2015. 5

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[20] Junyeop Lee, Sungrae Park, Jeonghun Baek, Seong Joon Oh, Seonghyeon Kim, and Hwalsuk Lee. On recognizing texts of arbitrary shapes with 2d self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 546–547, 2020. 6

[21] Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. Momentdiff: Generative video moment retrieval from random to real. In *Advances in neural information processing systems*, 2023. 1

[22] Pandeng Li, Chen-Wei Xie, Liming Zhao, Hongtao Xie, Jiannan Ge, Yun Zheng, Deli Zhao, and Yongdong Zhang. Progressive spatio-temporal prototype matching for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4100–4110, 2023. 1

[23] Minghui Liao, Jian Zhang, Zhaoyi Wan, Fengming Xie, Jiajun Liang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Scene text recognition from two-dimensional perspective. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8714–8721, 2019. 2

[24] Chongyu Liu, Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and Yongpan Wang. Erasenet: End-to-end text removal in the wild. *IEEE Transactions on Image Processing*, 29:8760–8775, 2020. 2, 5, 7, 8

[25] Chongyu Liu, Lianwen Jin, Yuliang Liu, Canjie Luo, Bangdong Chen, Fengjun Guo, and Kai Ding. Don't forget me: accurate background recovery for text removal via modeling local-global context. In *European Conference on Computer Vision*, pages 409–426. Springer, 2022. 2, 8

[26] Zhihang Liu, Jun Li, Hongtao Xie, Pandeng Li, Jiannan Ge, Sun-Ao Liu, and Guoqing Jin. Towards balanced alignment: Modal-enhanced semantic modeling for video moment retrieval. *arXiv preprint arXiv:2312.12155*, 2023. 1

[27] Canjie Luo, Lianwen Jin, and Jingdong Chen. Siman: exploring self-supervised representation learning of scene text via similarity-aware normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1039–1048, 2022. 3

[28] Guangtao Lyu, Kun Liu, Anna Zhu, Seiichi Uchida, and Brian Kenji Iwana. Fetnet: Feature erasing and transferring network for scene text removal. *Pattern Recognition*, 140: 109531, 2023. 2, 8

[29] Pengyuan Lyu, Chengquan Zhang, Shanshan Liu, Meina Qiao, Yangliu Xu, Liang Wu, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. Maskocr: text recognition with masked encoder-decoder pretraining. *arXiv preprint arXiv:2206.00311*, 2022. 2

[30] Weihong Ma, Hesuo Zhang, Lianwen Jin, Sihang Wu, Jiapeng Wang, and Yongpan Wang. Joint layout analysis, character detection and recognition for historical document digitization. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 31–36. IEEE, 2020. 2

[31] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *BMVC-British machine vision conference*. BMVA, 2012. 5

[32] Dezhi Peng, Lianwen Jin, Weihong Ma, Canyu Xie, Hesuo Zhang, Shenggao Zhu, and Jing Li. Recognition of handwritten chinese text by segmentation: A segment-annotation-free approach. *IEEE Transactions on Multimedia*, 2022. 2

[33] Dezhi Peng, Chongyu Liu, Yuliang Liu, and Lianwen Jin. Viteraser: Harnessing the power of vision transformers for scene text removal with segmim pretraining. *arXiv preprint arXiv:2306.12106*, 2023. 2, 3, 8

[34] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 569–576, 2013. 5

[35] Yadong Qu, Qingfeng Tan, Hongtao Xie, Jianjun Xu, Yuxin Wang, and Yongdong Zhang. Exploring stroke-level modifications for scene text editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2119–2127, 2023. 2, 5, 6, 7

[36] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, 41(18):8027–8048, 2014. 5

[37] Prasun Roy, Saumik Bhattacharya, Subhankar Ghosh, and Umapada Pal. Stefann: scene text editor using font adaptive neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13228–13237, 2020. 2

[38] Fenfen Sheng, Zhineng Chen, and Bo Xu. Nrtr: A no-recurrence sequence-to-sequence model for scene text recognition. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 781–786. IEEE, 2019. 2, 6

[39] 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 transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2016. 2, 6

[40] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2035–2048, 2018. 2, 6

[41] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. *arXiv preprint arXiv:2311.03054*, 2023. 1, 2

[42] Osman Tursun, Rui Zeng, Simon Denman, Sabesan Sivapalan, Sridha Sridharan, and Clinton Fookes. Mtrnet: A generic scene text eraser. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 39–44. IEEE, 2019. 2, 8

[43] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *2011 International conference on computer vision*, pages 1457–1464. IEEE, 2011. 5

[44] Peng Wang, Cheng Da, and Cong Yao. Multi-granularity prediction for scene text recognition. In *European Conference on Computer Vision*, pages 339–355. Springer, 2022. 2, 5, 6

[45] Yuxin Wang, Hongtao Xie, Shancheng Fang, Jing Wang, Shenggao Zhu, and Yongdong Zhang. From two to one: A new scene text recognizer with visual language modeling network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14194–14203, 2021. 6

[46] Yuxin Wang, Hongtao Xie, Zixiao Wang, Yadong Qu, and Yongdong Zhang. What is the real need for scene text removal? exploring the background integrity and erasure exhaustivity properties. *IEEE Transactions on Image Processing*, 2023. 2, 3, 8

[47] Zixiao Wang, Hongtao Xie, Yuxin Wang, Jianjun Xu, Boqiang Zhang, and Yongdong Zhang. Symmetrical linguistic feature distillation with clip for scene text recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 509–518, 2023. 2

[48] Liang Wu, Chengquan Zhang, Jiaming Liu, Junyu Han, Jingtuo Liu, Errui Ding, and Xiang Bai. Editing text in the wild. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1500–1508, 2019. 2, 7

[49] Xudong Xie, Ling Fu, Zhifei Zhang, Zhaowen Wang, and Xiang Bai. Toward understanding wordart: Corner-guided transformer for scene text recognition. In *European Conference on Computer Vision*, pages 303–321. Springer, 2022. 6

[50] Mingkun Yang, Minghui Liao, Pu Lu, Jing Wang, Shenggao Zhu, Hualin Luo, Qi Tian, and Xiang Bai. Reading and writing: Discriminative and generative modeling for self-supervised text recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4214–4223, 2022. 1, 3

[51] Qiangpeng Yang, Jun Huang, and Wei Lin. Swaptext: Image based texts transfer in scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14700–14709, 2020. 2, 7

[52] Moonbin Yim, Yoonsik Kim, Han-Cheol Cho, and Sungrae Park. Synthtiger: Synthetic text image generator towards better text recognition models. In *International Conference on Document Analysis and Recognition*, pages 109–124. Springer, 2021. 5

[53] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Towards accurate scene text recognition with semantic reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12113–12122, 2020. 6

[54] Xiaoyu Yue, Zhanghui Kuang, Chenhao Lin, Hongbin Sun, and Wayne Zhang. Robustscanner: Dynamically enhancing positional clues for robust text recognition. In *European Conference on Computer Vision*, pages 135–151. Springer, 2020. 6

[55] Boqiang Zhang, Hongtao Xie, Yuxin Wang, Jianjun Xu, and Yongdong Zhang. Linguistic more: Taking a further step toward efficient and accurate scene text recognition. *arXiv preprint arXiv:2305.05140*, 2023. 2

[56] Shuaitao Zhang, Yuliang Liu, Lianwen Jin, Yaoxiong Huang, and Songxuan Lai. Ensnet: Ensconce text in the wild. In *Proceedings of the AAAI conference on artificial intelligence*, pages 801–808, 2019. 2, 8

[57] Tianlun Zheng, Zhineng Chen, Shancheng Fang, Hongtao Xie, and Yu-Gang Jiang. Cdistnet: Perceiving multi-domain character distance for robust text recognition. *International Journal of Computer Vision*, pages 1–19, 2023. 2