2023 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-6891-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICME55011.2023.00107

LETTER EMBEDDING GUIDANCE DIFFUSION MODEL FOR SCENE TEXT EDITING

Changshuo Wang, Lei Wu, Xu Chen, Xiang Li, Lei Meng*, Xiangxu Meng*

School of Software, Shandong University, China
202115242@mail.sdu.edu.cn, i_lily@sdu.edu.cn,
listening@mail.sdu.edu.cn, 202035260@mail.sdu.edu.cn, lmeng@sdu.edu.cn, mxx@sdu.edu.cn

ABSTRACT

Scene text editing(STE) aims to modify the text in the scene image to the target text while retaining the original style. Existing models are based on GAN, where the source image and the target text are input only once during the generation process, and this approach could not fully obtain the style of the source image and content of the target text. In this paper, we propose an STE method based on the classifier-free guidance diffusion model. To our best knowledge, our model is the first work that developed diffusion models to handle the STE task. Specifically, we divide the STE task into multiple steps and extract style information and text content information in each step. In addition, we introduce the letter embedding method as guidance. We experimentally prove that our method outperforms other STE models in terms of overall realism and maintaining glyphs.

***Index Terms*—** Scene Text Editing, Diffusion Model, Text Synthesis

1. INTRODUCTION

Commonly found on billboards, store signs, and road signs, text in scenes contains rich information and plays a significant role in multimedia applications. Scene text editing (STE) is a technique to modify the text in scene images to the target text while preserving the original style and background, and is one of the cores of today's rapidly developing AR technology, which can be used for tasks such as image correction and AR translation. STE usually contains many subtasks, such as text style migration, background restoration, image fusion, etc. Therefore STE is a very challenging task.

STE was initially proposed by STEFANN [1], and existing STE models [2–4] are basically divided into two parts: text processing and background processing, i.e. Stylized text images and restored background images are generated separately and then fused. To the best of our knowledge, all previous works use the GAN model as the baseline, which has the advantage of fast training, but because GAN is a one-step generation method, it cannot fully extract the information of style image and text content, and often has the problem of poor text style conversion. Previous works mostly input the text content as a background-free text image in some fixed style, which will cause the text encoder to be disturbed by other unnecessary information when acquiring the text information and hinder the content generation. Although STEFANN [1] uses one-hot vectors for text input, it can only modify one letter at a time and cannot coordinate the overall style. In addition, previous models use some pre-trained models to assist in training, such as VGG, text recognizer and font classifier, which introduces supervised information outside of this task and makes the model more complex.

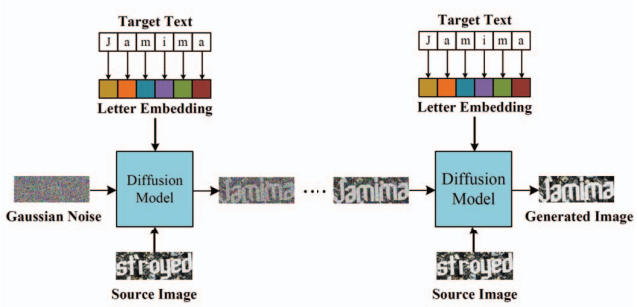

Fig. 1. **Overview of our approach.** Our method is based on the classifier-free guidance diffusion model, in each step of the generation, source image and target text are fed into the model in different ways.

We note that current image translation tasks based on diffusion models are generally built on datasets like ImageNet [5], which is more biased toward objects in nature, but few models focus on the text image, so the ability of diffusion models in generating text images has not been well explored. In this paper, we propose a new STE method based on the diffusion model. Compared with previous STE models, which contain various complex modules with different functions, our model consists of only one U-Net [6] and does not contain any pre-training modules. In our model, the target text is input as a string after being encoded to avoid the interference of other information in the text image. Compared to the previous STE model, this text input method does not require the

*Corresponding author

1945-788X/23/$31.00 ©2023 IEEE
DOI 10.1109/ICME55011.2023.00107

Authorized licensed use limited to: SHANDONG UNIVERSITY. Downloaded on September 27,2023 at 02:06:27 UTC from IEEE Xplore. Restrictions apply.

prior generation of standard-style text images, which further simplifies the model. We also use the scene text image with the background as the source image without separating the background from the text image. Using the multi-step generation feature of the diffusion model, we input the text and image information into the model at each step of the denoising process, as shown in Figure 1. Compared with the one-step generation based on GAN, our model can extract more comprehensive features and correct possible errors in the previous generation at each generation step to enhance the control of details. Our experiments on several public benchmarks demonstrate this model's advantages in generating quality.

The contributions we have made can be summarized as follows:

- To the best of our knowledge, we build the first scene text editing framework based on the classifier-free diffusion model with both source images and content characters as guides.
- We propose a guidance method based on letter embeddings, which uses semantic-level information and focuses more on the text content itself, without the interference of the image information of the text content.
- Experimental results show that our method exhibits superior performance on STE tasks compared to previous STE models and provides experience for future exploration of STE methods based on diffusion models.

2. RELATED WORK

2.1. Scene Text Editing

Scene Text Editing (STE) is a new task developed in recent years from Scene Text Recognition (STR), whose goal is to modify the text content in the scene text image while preserving its glyph style and background.

STE task was first proposed by STEFANN [1]. This model edits individual letters in a scene image. SRNet [2] generates and then fuses the background and foreground (text content) separately and uses a skeleton to constrain the text structure. SwapText [3] adds the attention module and Content Shape Transformation Network (CSTN) to SRNet to improve the generation capability for curved and skewed text. RewriteNet [7] finds that the model is vital for semantic understanding and adds a text recognition module to enhance network generation. TextStyleBrush [8] noticed the ability of StyleGAN [9] to control high-level semantic information in the hidden space and accomplished unsupervised STE tasks based on StyleGAN2 [10]. However, due to the one-step generation feature of GAN models, these models often do not perform the style transformation well.

2.2. Image-to-Image Translation

The rapid development of image generation[11, 12] models in previous years has given birth to many classical models, representative of which are GAN [13–19] and VAE [20]. Benefiting from the advancement of generation models, the image translation task is also developing rapidly. Isola et al. [21] proposes the pix2pix model based on conditional GAN, which implements the content of generated images through image control. VQ-VAE [22] discretizes the latent space of VAE to better fit some modalities in natural images.

In recent years, diffusion model [23] has become well-known for its excellent generation quality, but the training and sampling costs are much higher than those of GAN and VAE, so people are constantly exploring ways to optimize the diffusion model [24–27].

3. APPROACH

3.1. Overview

We propose a diffusion model-based approach for STE tasks, which follows the classifier-free guidance diffusion method [28], and the model structure is shown in Figure 2. In training phase, the input of the model can be represented by a triplet (y, s, x_0), where y is the source image with the background, s is the target text, and x_0 is the ground truth. The output is a Gaussian noise ϵ. In testing phase, the input is (y, s, x_T), where $x_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and the output is the target image x_0. For more details on the diffusion model, please see [23].

3.2. Diffusion Process for STE Task

Denoising Diffusion Probabilistic Models (DDPMs) learns the reverse process of a Markov chain that uses noise to pollute images gradually. Starting from a pure Gaussian noise x_T, DDPMs gradually remove the applied noise and obtain a noise-free image x_0.

The forward diffusion method and notations in [23] are used in our model, with adaptations for the STE task. Specifically, in the training phase, given a set of input data (y, s, x_0), along with a random integer t from $(0, T)$, where T denotes the maximum number of diffusion steps, x_0 is noise-added according to the following equation:

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1},\ \beta_t \mathrm{I}) \tag{1}$$

where β_t is the hyperparameter. Due to the additivity of the Gaussian distribution, it is possible to use only one step of the noise addition process:

$$q(x_t|x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0,\ (1-\bar{\alpha}_t)\mathrm{I}) \tag{2}$$

With the use of the reparameterization trick, Equation 2 can be written in a more comprehensible form:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t \tag{3}$$

where $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, $\bar{\alpha}_0 = 1$ and tends to 0 as t increases. Input (y, s, x_t) into the model,

Authorized licensed use limited to: SHANDONG UNIVERSITY. Downloaded on September 27,2023 at 02:06:27 UTC from IEEE Xplore. Restrictions apply.

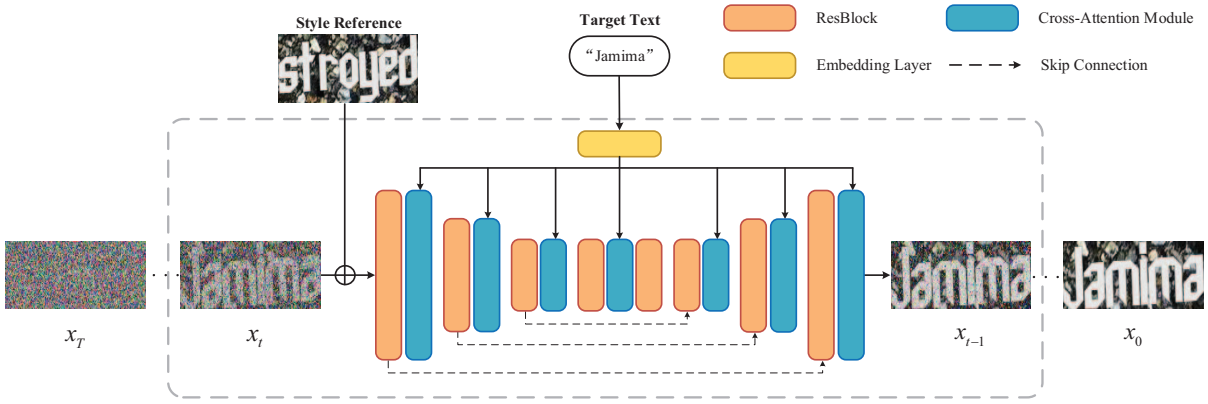

Fig. 2. **The overall structure of our model.** We use the classifier-free guidance diffusion model. In each step of the reverse diffusion, the source image is input to U-Net with the result of the previous step in the channel dimension splicing, and the target text enters the network as some letter embedding through the cross-attention module.

and the model should output the original x_0, or the noisy ϵ_t added to it, according to the style of y, the content of s and the ground truth x_t with noise. Here we make the model output noise ϵ_t:

$$\hat{\epsilon}_t = f_\theta(y, s, x_t) \tag{4}$$

Our model does not require any additional pre-trained model to induce losses and has only one loss function:

$$\mathcal{L} = \mathbb{E}_{y,s,x_0}\|\epsilon_t - \hat{\epsilon}_t(y, s, x_0)\|^2 \tag{5}$$

The training process is summarized in Algorithm 1.

In sampling phase, the input to the model is (y, s, x_T). Specifically, due to the setting of the hyperparameter α_t, x_T can be approximated as a random variable obeying $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ when $T = 1000$, which can be obtained by simple sampling. The sampling process has T steps. In each step, the model predicts the noise ϵ_t based on (y, s, x_t), and obtains $\hat{x}_0$ from the inverse form of Equation 3, and then adds noise using Equation 3 to obtain x_{t-1}, which is the input for the next step. The sampling process is summarized in Algorithm 2.

3.3. Guidance in Each Step

In order to make the diffusion model applicable to STE tasks, we adapted the reverse denoising step of the diffusion model. In each denoising step, the source image y is input to U-Net in the form of an image stitched with x_t in the channel dimension, and this input is used as the Query in the first cross-attention module of U-Net. Target text s is encoded and turned into letter embedding, which is used as Key and Value in the cross-attention module:

$$Q = W^Q f(x_t, y) \quad K = W^K Emb(s) \quad V = W^V Emb(s)$$

where $f(\cdot)$ represents the Resblock in the network and its inputs are x_t and y only if $f(\cdot)$ is the first Resblock, otherwise its input is the output of the previous attention module. $Emb(\cdot)$ represents the embedding layer. This fusion considers the different structures of y and s with the different information they provide and feeds them into the model in different ways to make the most of them.

4. EXPERIMENTS

In this section, we present the results in Figure 3 to demonstrate the excellent capability of our model for the STE task. We compare our model with other STE models to demonstrate the superiority of our approach. In addition, we also evaluate our model with a target text length test.

4.1. Datasets

Our experiments use a variety of datasets, including the synthetic dataset and the real dataset.

Synthetic data. We obtain synthetic data according to the method provided by SRNet [2], which generates pairs of data for supervised learning. We use 100 fonts and 1000 background images to generate a total of 110,000 training images and 1000 test images.

ICDAR 2013. [29] This set of images is part of the ICDAR 2013 Robust Reading Competition. Compared to other real datasets, the images of ICDAR 2013 have a higher resolution and prominent text. There are 848 and 1095 word images in the original training set and test set, respectively.

ICDAR 2015. [30] This set of images is part of the ICDAR 2015 Robust Reading Competition, which was designed to be more challenging than ICDAR 2013. Most of the images in this set have low resolution and viewpoint irregularities.

Street View Text. [31] (SVT) is collected from Google Street View and contains 647 images in the test set. Many of the images are heavily corrupted by noise and blur, or have very low resolution.

Authorized licensed use limited to: SHANDONG UNIVERSITY. Downloaded on September 27,2023 at 02:06:27 UTC from IEEE Xplore. Restrictions apply.

Fig. 3. **Scene Text Editing results.** We paste the output image directly back to the location where the source image image is located, without using any fusion algorithm.

Algorithm 1 Training Process

Input: Dataset $D\{(\boldsymbol{x}_0^i, \boldsymbol{y}^i, \boldsymbol{s}^i)_{i=1}^N\}$
Output: Noise $\boldsymbol{\epsilon}$ added to $\boldsymbol{x}_0$

repeat
  $(\boldsymbol{x}_0, \boldsymbol{y}, \boldsymbol{s}) \sim D$
  $t \sim Uniform(\{1, ..., T\})$
  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  $\boldsymbol{x}_t \leftarrow \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$
  Take gradient descent step on
    $\nabla_\theta \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{y}, \boldsymbol{s}, t)\|^2$
until converged

Algorithm 2 Sampling Process

Input: y: scene text image that provides style; s: target text;
Output: x_0: scene text image that has the style of y with the content of s;

$\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
for $t = T$ to 1 **do**
  $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ if $t > 1$, else $\boldsymbol{z} = \mathbf{0}$
  $\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{y}, \boldsymbol{s}, t)\right) + \boldsymbol{\beta}_t \boldsymbol{z}$
end for
return $\boldsymbol{x}_0$

4.2. Implementation Details

In the training phase of the diffusion model, we set $T = 1000$ and the forward process variances to constants increasing linearly from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$. The letter embedding length is 20. In the sampling phase, we follow the sampling method of DDPM [23]. In our generated dataset, the image size is 64×128, and the batch size is 32. The Adam optimizer with β_1 = 0.9, β_2 = 0.999, and lr gradually decays from 2^{-4} to 2^{-6} after 200 epochs are used to optimize the whole framework. We used an Nvidia Geforce RTX 3090 to train the model for seven days, accumulating 600 epochs.

Table 1. Recognition accuracy on real world datasets.

Model	ICDAR13	ICDAR15	SVT
SRNet[2]	37.13	16.21	29.83
STEFANN[1]	18.62	7.84	21.81
DSE-NET[32]	73.08	62.97	69.24
Ours	**92.34**	**89.21**	**84.15**

Table 2. Quantitative results on synthetic test dataset.

Model	Synthetic dataset			
	Ac ↑	FID ↓	SSIM ↑	PSNR ↑
SRNet[2]	17.40	103.26	**0.32**	16.46
STEFANN[1]	12.25	188.84	0.10	10.02
DSE-NET[32]	57.21	128.47	0.06	8.30
Ours	**82.62**	**89.40**	0.23	**17.81**

4.3. Performance Comparison

We compare our model with three models: SRNet [2], STEFANN [1] and DSE-NET [32]. Although DSE-NET was not designed for STE tasks, we trained it in the same way as other STE tasks. During the comparison, the text content was entered only in character encoding in our model and in a standard format with Arial font and gray background images in the other models due to the different ways of data input.

Visual Comparison. The comparison results are shown in Figure 4. We show the results of comparing our model with the rest of the models under several different source images. Due to the specificity that STEFANN can only edit a single letter, we generate many results in each comparison and select the best visually appealing image from them for display. From the comparison results, we can see that our model not only performs well on source images with standard fonts but also performs significantly better than other models on source images with more fancy fonts and even source images in different languages.

Quantitative Comparison. We further conduct quantitative comparisons in terms of MSE, FID, PSNR, SSIM, and

Authorized licensed use limited to: SHANDONG UNIVERSITY. Downloaded on September 27,2023 at 02:06:27 UTC from IEEE Xplore. Restrictions apply.

Fig. 4. **Visual comparison with other methods.** Except for our model, the text content of all models is input as images, and only the text content is shown. Since the real-world image does not have the ground truth, it is not shown.

Recognition Accuracy. MSE, FID, PSNR, and SSIM are used to measure the image quality and style similarity of the output image, and recognition accuracy is used to measure how well the output image text content matches the target text content. To obtain recognition accuracy, here we use a pre-trained scene text recognition model[1]. It is worth mentioning that since all the OCR datasets we used did not have ground truth for the STE task, we calculated only on the synthetic dataset for metrics other than recognition accuracy. The quantitative comparison results on the OCR datasets and the synthetic dataset are shown in Table 1 and Table 2, respectively. The quantitative comparison results show that our model achieves the best results in recognition accuracy, indicating that the ability to extract information about text content is far superior to other models. Regarding image quality metrics, SRNet is slightly better at restoring background details than our model because it uses an additional module to separate the background from the foreground. Therefore, experiments conducted on the synthetic dataset show that SRNet's SSIM is better than our model but slightly inferior to ours in terms of FID and PSNR.

4.4. Effectiveness of Letter Embedding Guidance

In this section, we compare the two approaches: the letter embedding + cross-attention approach proposed in this paper with the textual content-as-image input approach widely used by other STE frameworks. Here we introduce variance as a metric to measure the two methods' ability to control the diffusion model's randomness. This is done by generating ten results with the same target text for each test image, calculating the variance of these ten results, and taking the mean value for all test images. The test results are shown in Table 3. To facilitate the measurement of variance, we scaled the pixel values of the images to between 0 and 10. The experimental results show that the letter embedding + cross-attention approach outperforms the content-as-image approach in all metrics. In particular, variance metric indicates that our method can provide better constraints on the generated results.

[1]https://github.com/clovaai/deep-text-recognition-benchmark

Table 3. Effectiveness of letter embedding guidance.

Method	Ac ↑	FID ↓	SSIM ↑	PSNR ↑	Var ↓
content as image	56.71	125.69	0.16	11.58	3.24
letter embedding	**82.60**	**89.40**	**0.23**	**17.81**	**0.93**

4.5. Robustness for Different Text Lengths.

To demonstrate that our model is robust to the length of the target text, we conducted multiple sets of experiments, as shown in Figure 5. Multiple sets of results with the text content of different lengths are generated for the same stylized reference image. The results show that our model has excellent generation results for text contents of different lengths.

5. CONCLUSION

This paper proposes a novel STE method based on the diffusion model. Compared to most previous methods that generate and fuse foreground and background separately, our model contains only a U-Net, which requires less supervised information and does not contain any pre-trained model to guide the generation. We also propose a method to input textual information into the U-Net in a letter embedding manner based on a classifier-free guided diffusion model, which

Authorized licensed use limited to: SHANDONG UNIVERSITY. Downloaded on September 27,2023 at 02:06:27 UTC from IEEE Xplore. Restrictions apply.

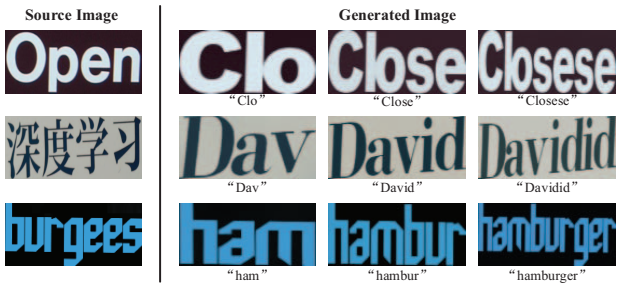

Fig. 5. **Robustness for Different Text Lengths.** Generated images from our model when the lengths of target texts are different. When the text length changes, the generation quality is not affected.

can more significantly reduce the input of distracting information than the text image stitching method used in previous STE models, and can better control the diversity that diffusion models have. Our visual and quantitative experiments on multiple datasets demonstrate the effectiveness of our method.

Acknowledgements

This work is supported in part by the National Key R&D Program of China (Grant no. 2021YFC3300203), the "20 Regulations for New Universities" funding program of Jinan (Grant no. 2021GXRC073), and the Excellent Youth Scholars Program of Shandong Province (Grant no. 2022HWYQ-048).

References

[1] Prasun Roy, Saumik Bhattacharya, et al., "Stefann: scene text editor using font adaptive neural network," in *CVPR*, 2020, pp. 13225–13234.

[2] Liang Wu, Chengquan Zhang, et al., "Editing text in the wild," in *27th ACM MM*, 2019, pp. 1500–1508.

[3] Qiangpeng Yang et al., "Swaptext: Image based texts transfer in scenes," in *CVPR*, 2020, pp. 14688–14697.

[4] Lin Zhao, Changsheng Chen, et al., "Deep learning-based forgery attack on document images," *IEEE Transactions on Image Processing*, vol. 30, pp. 7964–7979, 2021.

[5] Jia Deng, Wei Dong, et al., "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[7] Junyeop Lee, Yoonsik Kim, et al., "Rewritenet: Realistic scene text image generation via editing text in real-world image," *arXiv preprint arXiv:2107.11041*, 2021.

[8] Praveen Krishnan, Rama Kovvuri, et al., "Textstylebrush: transfer of text aesthetics from a single example," *arXiv preprint arXiv:2106.08385*, 2021.

[9] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019, pp. 4401–4410.

[10] Tero Karras, Samuli Laine, et al., "Analyzing and improving the image quality of stylegan," in *CVPR*, 2020, pp. 8110–8119.

[11] Chuang Lin, Sicheng Zhao, Lei Meng, and Tat-Seng Chua, "Multi-source domain adaptation for visual sentiment classification," *AAAI*, vol. 34, no. 03, pp. 2661–2668, Apr. 2020.

[12] Lei Meng, Long Chen, et al., "Learning using privileged information for food recognition," in *ACM MM*, 2019, p. 557–565.

[13] Ian Goodfellow, Jean Pouget-Abadie, et al., "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[14] Pei Dong, Lei Wu, et al., "Disentangled representations and hierarchical refinement of multi-granularity features for text-to-image synthesis," in *ICMR*, 2022, p. 268–276.

[15] Xiangxian Li, Haokai Ma, Lei Meng, and Xiangxu Meng, "Comparative study of adversarial training methods for long-tailed classification," in *ACM MM Workshop*, 2021, p. 1–7.

[16] Jingyu Li, Haokai Ma, Xiangxian Li, Zhuang Qi, Lei Meng, and Xiangxu Meng, "Unsupervised contrastive masking for visual haze classification," in *ICMR*, 2022, p. 426–434.

[17] Xu Chen, Lei Wu, et al., "Mlfont: Few-shot chinese font generation via deep meta-learning," in *ICMR*, 2021, p. 37–45.

[18] Pei Dong, Lei Wu, et al., "Hr-prgan: High-resolution story visualization with progressive generative adversarial networks," *Information Sciences*, vol. 614, pp. 548–562, 2022.

[19] Lei Wu, Xi Chen, et al., "Multitask adversarial learning for chinese font style transfer," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8.

[20] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.

[22] Aaron Van Den Oord, Oriol Vinyals, et al., "Neural discrete representation learning," *NeurIPS*, vol. 30, 2017.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[24] Jiaming Song, Chenlin Meng, et al., "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[25] Alexander Quinn Nichol and Prafulla Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8162–8171.

[26] Cheng Lu, Yuhao Zhou, et al., "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," *arXiv preprint arXiv:2206.00927*, 2022.

[27] Cheng Lu, Yuhao Zhou, et al., "Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models," *arXiv preprint arXiv:2211.01095*, 2022.

[28] Jonathan Ho and Tim Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[29] Dimosthenis Karatzas, Faisal Shafait, et al., "Icdar 2013 robust reading competition," in *ICDAR*, 2013, pp. 1484–1493.

[30] Dimosthenis Karatzas, Faisal Shafait, et al., "Icdar 2015 competition on robust reading," in *ICDAR*, 2015, pp. 1156–1160.

[31] Kai Wang, Boris Babenko, and Serge Belongie, "End-to-end scene text recognition," in *ICCV*, 2011, pp. 1457–1464.

[32] Xiang Li, Lei Wu, Xu Chen, Lei Meng, and Xiangxu Meng, "Dse-net: Artistic font image synthesis via disentangled style encoding," in *ICME*, 2022, pp. 1–6.

Authorized licensed use limited to: SHANDONG UNIVERSITY. Downloaded on September 27,2023 at 02:06:27 UTC from IEEE Xplore. Restrictions apply.