

IPAD: Iterative, Parallel, and Diffusion-based Network for Scene Text Recognition

Xiaomeng Yang^{1,2}, Zhi Qiao³, Yu Zhou^{1*}

¹*VCIP & TMCC & DISec, College of Computer Science, Nankai University, Tianjin, 300350, China.

²Institute of Information Engineering, CAS, Beijing, 100089, China.

³Tomorrow Advancing Life, Beijing, 100080, China.

*Corresponding author(s). E-mail(s): yzhou@nankai.edu.cn;

Contributing authors: yang.xiaome@northeastern.edu; qiaozhi1@tal.com;

Abstract

Nowadays, scene text recognition has attracted more and more attention due to its diverse applications. Most state-of-the-art methods adopt an encoder-decoder framework with the attention mechanism, autoregressively generating text from left to right. Despite the convincing performance, this sequential decoding strategy constrains the inference speed. Conversely, non-autoregressive models provide faster, simultaneous predictions but often sacrifice accuracy. Although utilizing an explicit language model can improve performance, it burdens the computational load. Besides, separating linguistic knowledge from vision information may harm the final prediction. In this paper, we propose an alternative solution that uses a parallel and iterative decoder that adopts an easy-first decoding strategy. Furthermore, we regard text recognition as an image-based conditional text generation task and utilize the discrete diffusion strategy, ensuring exhaustive exploration of bidirectional contextual information. Extensive experiments demonstrate that the proposed approach achieves superior results on the benchmark datasets, including both Chinese and English text images.

Keywords: Scene Text Recognition, OCR, Discrete Diffusion, Non-autoregressive Decoding.

1 Introduction

Nowadays, the field of scene text detection [1] and recognition [2] is drawing substantial research attention because of its significant role in many downstream tasks, such as text spotting [3] and text-based visual question answering [4]. Scene text detection involves the localization of text instances in scene images, and scene text recognition (STR) transcribes the localized instances into an editable text format. STR, in particular,

is laden with challenges stemming from the diversity of characters and backgrounds. However, the advent and progression of deep learning have facilitated the achievement of convincing results in this domain.

Based on the decoding strategy, existing STR methods can be roughly categorized into three distinct types: Connectionist Temporal Classification (CTC) based [2, 5–8], attention mechanism based [9–13] and segmentation-based [14, 15] methods. From another perspective, these methods can also be classified into autoregressive

and non-autoregressive approaches. Specifically, autoregressive methods decode the text from left to right, and the number of regressions depends on the length of the text. Most attention mechanism-based methods adopt a left-to-right autoregressive decoding process. Recently, autoregressive methods have achieved great success in scene text recognition [11, 13, 16, 17].

Although the performance is satisfactory, the inference speed of autoregressive models is relatively slow, especially when dealing with long texts. On the contrary, non-autoregressive methods predict the text in parallel, with the CTC-based methods serving as typical exemplars. Such parallel inference improves the speed significantly but ignores the dependencies between characters. We argue that fully non-autoregressive methods lack the context information of characters, which is significant to the recognition of hard cases, and the assumption of independence increases the training difficulties of the hidden layers. Compared with autoregressive methods, the accuracy of non-autoregressive methods is relatively poor. An alternative approach involves supplementing non-autoregressive models with an explicit language module, refining predictions independently from visual data [12, 18]. However, this approach divides vision and language processing, which leads to a lack of visual information within the language model, so it can only adjust the results according to the pretrained language weight from the training text and may adjust the correct prediction from the vision model to a wrong one. Besides, this external language model-based solution poses substantial computational demands.

In this paper, we propose a parallel and iterative decoding model. In each iteration, the decoder still generates the text in parallel, and the context information is extracted depending on the previous predictions. Specifically, we adopt the easy-first [19] decoding strategy according to the iterative generative process. Easy-first strategy predicts the most confident and obvious characters in each iteration first and re-predicts other remaining characters in the next iteration. Unlike traditional left-to-right decoding, easy-first breaks the limitation of decoding order with higher flexibility. Inspired by Transformer [20], the parallel decoder consists of a masked self-attention module, a 2D

cross-attention module, and a feed-forward network (FFN).

Furthermore, although the iterative prediction brings the context information into the decoding, the pre-defined easy-first decoding training strategy can only allow the parallel decoder to exploit limited context according to the decoding order of the training dataset. Since the quantity of the training dataset is limited, the inter-character dependence may not be comprehensively contained in this scope. Therefore, it is more suitable during inference instead of training. To solve this problem, we contend that STR can be interpreted as a conditional text generation task depending on the input image. Therefore, we can employ the procedure of the discrete diffusion model for the STR task. Diffusion models [21–23] have demonstrated exceptional capabilities in tasks involving image and text generation, which is a potential solution. Their superior semantic learning capabilities and ability to integrate information across multiple stages of the decoding process could effectively address the complexities of STR. To achieve this, we adopt the discrete diffusion proposal into our parallel and iterative decoding framework. During training, the forward noising process is applied with a randomly chosen timestep, and the clean text is noised by replacing the characters with placeholders. In the denoising process, our decoder directly predicts the clean text. For inference, the denoising process is combined with our easy-first decoding process, which means the re-noising process is determined by the predicted confidence instead of the predefined noising process. In this paper, we call our proposed method an Iterative, Parallel, and Diffusion-based Network (IPAD), and the main framework of IPAD is shown in Fig. 1.

To be specific, in this paper, we propose an advanced framework that is different from our conference version PIMNet [24]¹ in the following aspects: (1) We deeply analyze the reason behind the performance gap between non-autoregressive and autoregressive models and propose to view the STR task as an image-based text-generation task and adopt a discrete diffusion training strategy for the STR task. With the training of the denoising process for numerous noise texts used at different timesteps, our parallel and iterative decoder could

¹A best paper candidate in ACM MM 2021.

learn sophisticated linguistic knowledge implicitly. (2) Different from English, the context of Chinese is more complex and the length of the Chinese text is longer, thus recognizing the Chinese text needs more advanced contextual information. Our method can learn the relationship between characters through easy-first decoding and diffusion training, which is more effective in Chinese text recognition. We implement more experiments and analyses, demonstrating that the proposed methods are effective and superior both on Chinese and English datasets.

The main contributions of our work are as follows:

- We propose a parallel and iterative decoding framework for text recognition. Different from previous works, our method decodes with constant iterations independent of the text length, which takes advantage of both fully autoregressive and fully non-autoregressive methods and achieves a good balance between accuracy and efficiency.
- To achieve parallel and iterative decoding, an easy-first strategy is designed for text recognition. Different from traditional left-to-right decoding, the easy-first strategy predicts the most confident characters in each iteration, which is more flexible.
- To enhance the ability of contextual learning and help the training of the parallel decoder, we employ the discrete diffusion training strategy, which views the text recognition process as an image-based conditional text generation.
- Extensive experiments are conducted to verify the effectiveness and efficiency of the proposed method, which achieves state-of-the-art or comparable accuracies on six popular benchmarks and three large datasets with a faster inference speed. Besides, it also achieves state-of-the-art performance on the Chinese benchmark datasets.

2 Related Work

2.1 Scene Text Recognition

Scene text recognition has been studied for many years, and existing methods can be classified into two primary categories: traditional methods and

deep learning-based methods. Traditional methods [25–32] usually employ a bottom-up framework, which initially detects and classifies characters, then subsequently group them utilizing a lexicon, a language model or heuristic rules. In recent years, deep learning-based methods have dominated this area because of the simple pipeline and compelling performance. Moreover, these methods can be further bifurcated, depending on their prediction strategies, into two groups: non-autoregressive and autoregressive methods as follows.

2.2 Autoregressive Text Recognition

Autoregressive methods typically adopt an encoder-decoder framework that predicts sequences from left to right. Most attention-based methods belong to the autoregressive model and can be divided into 1D attention-based and 2D attention-based. For 1D attention-based, Lee and Osindero [33] proposes a recursive CNN network to capture broader features and an attention-based decoder to transcribe sequence. Cheng et al. [34] introduces the problem of attention drift and proposes a focusing attention network to solve it. Fang et al. [35] proposes a fully CNN-based network to extract visual and language features separately. Some methods [9, 36, 37] rectify irregular text image first then recognize it with 1D attention-based decoder. ESIR [38] and ScRN [39] improve the quality of rectification with iteration and additional geometrical constraints, respectively. For 2D attention-based, Yang et al. [40] first introduces 2D attention into irregular text recognition, proposing an auxiliary segmentation task. Li et al. [11] suggests a tailored 2D attention operation.

Despite their robust performance resulting from the integration of preceding contextual information into the decoding process, autoregressive methods do face a set of challenges. Specifically, they encounter the issue of attention drift for long texts, and their one-way serial decoding process restricts the amount of contextual information that can be considered. Addressing the challenge of attention drift, DAN [41] offers a solution that decouples the prediction of attention weights from the decoding process, and Qiao et al. [42] introduces a Gaussian constrained

refinement module to refine the attention distribution. Further developments focus on employing advanced linguistic knowledge to improve the decoding process. SEED [43] introduces semantic global information to guide the decoding process. SCATTER [44] trains a deep BiLSTM encoder to extract broader contextual dependencies. Zheng et al. [45] adopts an external language model to incorporate useful context information into the decoding. PARSeq [13] learns an internal language model through permutations of autoregressive decoding. PTIE [46] trains the ViT encoder and transformer decoder with varying patch resolutions and AR decoding directions.

2.3 Non-Autoregressive Text Recognition

Non-autoregressive methods aim to generate the target text in a single iteration or a constant time independent of the text length. Jaderberg et al. [47] interprets word recognition as a classification task using CNN, which suffers from scalability issues due to a fixed vocabulary. ViT-STR [48] adapts the Vision Transformer (ViT) [49] for scene text recognition, which directly classifies each visual representation learned by the ViT encoder in parallel. Many methods treat text recognition as a sequence-to-sequence task, and non-autoregressive decoding generally relies on one of the three major technologies: CTC-based, segmentation-based, and parallel attention-based. A majority of the CTC-based methods [2, 5–7, 50–52] utilize a CNN to extract visual features and CTC to transcribe the final text with a short inference time. SVTR [8] further enhances this by designing local and global mixing blocks for a ViT-based encoder, capturing more refined visual features for the CTC-based decoder. Segmentation-based methods [14, 15, 53] regard text recognition as a task of semantic segmentation of characters, and they need additional character-level annotations.

Parallel attention-based methods have gained widespread attention in various fields, such as neural machine translation [54–56], automatic speech recognition [57–59] and image caption [60], thanks to their inference speed. Many recent non-autoregressive STR methods are grounded in this parallel attention decoder framework. However, one significant limitation of the complete one-time

parallel decoding methods is that they typically assume each character to be independent. This assumption hinders their ability to utilize contextual information during the decoding process, which is especially problematic when recognizing difficult cases. Therefore, parallel attention-based scene text recognition methods have introduced various strategies to incorporate linguistic knowledge into their models.

For instance, SRN [12] and ABINet [18] design an independent semantic module following parallel visual attention. This arrangement allows the semantic module to learn contextual information and refine initial predictions made by the vision module. LevOCR [61] further explores the effective fusion of visual and linguistic features. However, the decoupled two-stage model structure makes context information limited by the predictions of the parallel visual attention module, and the context information may tend to accumulate errors due to the wrong predictions. To circumvent this, VisionLAN [62] introduces a visual reasoning module that randomly masks corresponding visual features of characters during training, enabling the implicit learning of linguistic knowledge within vision space. Furthermore, MGP-STR [63] includes subword representations to enable multi-granularity prediction.

In summary, while autoregressive models typically produce superior accuracy, their sequential generation approach constrains inference speed. Conversely, non-autoregressive models offer enhanced inference speed, but often at the expense of lower accuracy. Even though incorporating linguistic knowledge improves the performance of non-autoregressive methods, these usually require an additional semantic refinement module or a multi-stage training procedure. In contrast, our work seeks a balance between fully autoregressive and fully non-autoregressive models. Our objective is to explore the development of a straightforward yet powerful iterative decoding model that achieves an optimal balance between efficiency and accuracy. Furthermore, we intend to equip our model with internal linguistic knowledge to boost its performance further.

2.4 Diffusion Models

Diffusion models, characterized by a forward noising process and a reverse denoising operation, have

recently been classified as the current state-of-the-art in the family of deep generative models [64, 65]. They function by incrementally injecting noise into clean samples in the forward process and then employing the reverse process to counteract this corruption, recovering the original samples. These models can be categorized into continuous diffusion models and discrete diffusion models.

Continuous diffusion models [21, 66–68], which introduce noise into continuous-valued input or latent features, have been extensively applied in various tasks involving generation. They have achieved remarkable results in image generation [22, 69–71], video generation [72, 73], audio generation [74] and 3D generation [75]. Despite the inherent discreteness of text, researchers have made efforts to employ continuous diffusion models for text generation. These methods [76, 77] utilize designed mapping functions to bridge the gap between the continuous and discrete domains.

Regarding models of discrete diffusion [78], each element is treated as a discrete random variable with several categories, and the diffusion process happens within discrete states. ImageBART [79] introduces an innovative application of multinomial diffusion on the latent discrete code space of a VQ-VAE [80], learning a parametric model for image generation. VQ-Diffusion [81] alternatively substitutes Gaussian noise with a random walk in the discrete data space. For text generation, Hooeboom et al. [82] puts forward the use of multinomial diffusion for character-level text generation, employing the forward categorical noise through the Markov transition matrix. D3PM [83] broadens the scope of discrete text diffusion models by introducing the $\langle \text{MASK} \rangle$ token. Different from preceding methods, DiffusionBERT [23] integrates pretrained language models with absorbing-state discrete diffusion models for text. Additionally, within multi-modal tasks, DDCap [84] deploys the vector quantized discrete diffusion models for image captioning.

We propose viewing the text recognition task as a form of conditional text generation. In this paper, given the discrete nature of characters in texts, we employ the capabilities of the discrete diffusion model in scene text recognition. This approach trains a stochastic process and empowers the decoder to learn a suite of distributions. Consequently, it can progressively recognize and

refine results, informed by the internal linguistic knowledge acquired during training.

3 Methodology

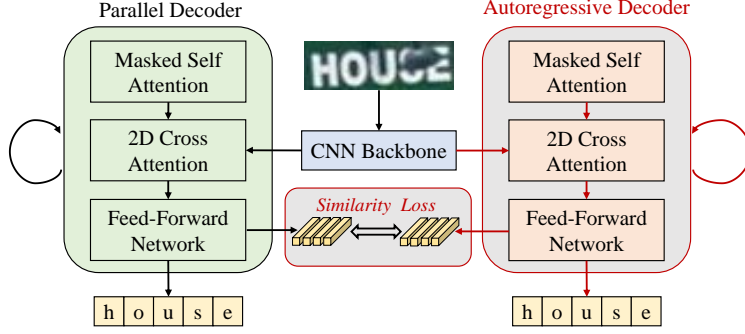
In this section, we will describe our proposed model in detail. As depicted in Figure 1, our model consists of two principal components: a ViT-based Encoder and a parallel decoder. To break the assumption of character independence in the fully parallel decoder, the easy-first decoding strategy is employed to predict the text iteratively during inference. Besides, to facilitate the internal contextual information learning of our parallel decoder, we incorporate a discrete diffusion procedure and make the parallel decoder denoise the text to clean text directly during training. In Section 3.1, we first introduce the encoder employed for visual feature extraction. Subsequent sections delineate the iterative decoding process and parallel decoder in Section 3.2 and Section 3.3, respectively. Section 3.4 describes the concept of deploying discrete diffusion models for STR decoding. Finally, in Section 3.5, we will outline the training and inference process, as well as the objective function.

3.1 ViT-based Encoder

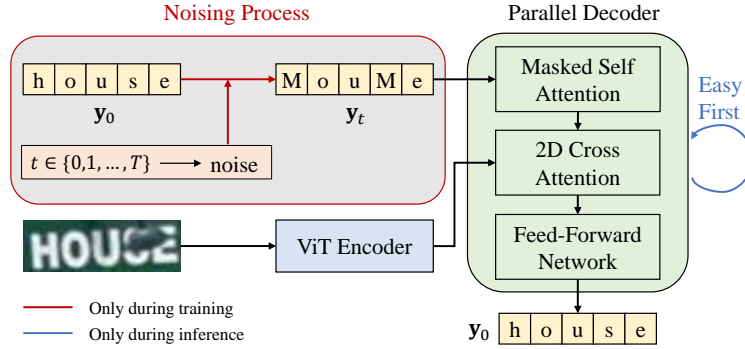
Contrasting with our conference version, which employs a ResNet50 equipped with FPN [85] as the encoder, we opt for a ViT-based [49] encoder in this work, considering its capability to provide sophisticated visual features. Same as PARSeq [13] and MGP-STR [63], we select the 12-layer ViT as our encoder. The classification head and the $\langle \text{CLS} \rangle$ token are ignored. When an image $\mathbf{x} \in \mathbb{R}^{\mathcal{H} \times \mathcal{W} \times \mathcal{C}}$ is input, the encoder partitions it into $\mathcal{N} = \mathcal{H}\mathcal{W}/(\mathcal{P}_w\mathcal{P}_h)$ patches of shape $\mathcal{P}_w \times \mathcal{P}_h$. These patches are then linearly projected into \mathcal{D} -dimensional tokens, together with the learned position embeddings, thus generating a sequence of visual features $\mathbf{z} \in \mathbb{R}^{\mathcal{N} \times \mathcal{D}}$ for the decoder.

3.2 Iterative Decoding Strategy

Easy first [19] is an iterative decoding strategy where the most confident predictions in each iteration are first predicted. Inspired by [86], we adopt a particular token $\langle \text{MASK} \rangle$ for the iterative decoding, which acts as a placeholder for the



(a) PIMNet [24] Architecture



(b) IPAD Architecture

Fig. 1: Comparison between the proposed IPAD and our conference version PIMNet [24]. For IPAD, the ViT-based encoder first extracts the visual feature, and then the parallel decoder adopts an iterative generation to extract the context information from the previous predictions. The discrete diffusion strategy is used during training, where the noising process adds noise by randomly sampling a timestep and then the parallel decoder works as the denoising network, which directly predicts the clean text without any iteration.

next iteration. As shown in Figure 2, all characters are unreachable in the first iteration, so the input to all of the decoder’s positions is $\langle \text{MASK} \rangle$ token. The next iterations can be separated into two main steps of prediction and update:

Prediction: In the current iteration, the parallel decoder predicts the corresponding probabilities of characters for each position which is still $\langle \text{MASK} \rangle$. On the contrary, the characters which have been already updated will not be re-predicted again:

$$\hat{y}_i^l = \begin{cases} \underset{c_n \in C}{\operatorname{argmax}}(P(c_n^l | y_{i-1})), & \text{if } y_{i-1}^l \text{ is } \langle \text{MASK} \rangle; \\ y_{i-1}^l, & \text{otherwise.} \end{cases} \quad (1)$$

where y_{i-1} is the character sequence in the previous iteration $i - 1$, \hat{y}_i^l indicates the predicted character in position $l \in [1, 2, \dots, L]$ in iteration i , $P(c_n^l)$ is the predicted probability for the character c_n in position l , and $\operatorname{argmax}(P)$ chooses the character with the largest probability. Here $C = \{c_1, c_2, \dots, c_{N+1}\}$ is the meaningful N -charset added with the $\langle \text{EOS} \rangle$ token. The conditional probability in P indicates that the context information has been fully utilized based on the predictions from the previous iteration in a bi-directional manner.

Update: After prediction, some positions of the target text are updated with the most confident predictions, and the others are abandoned

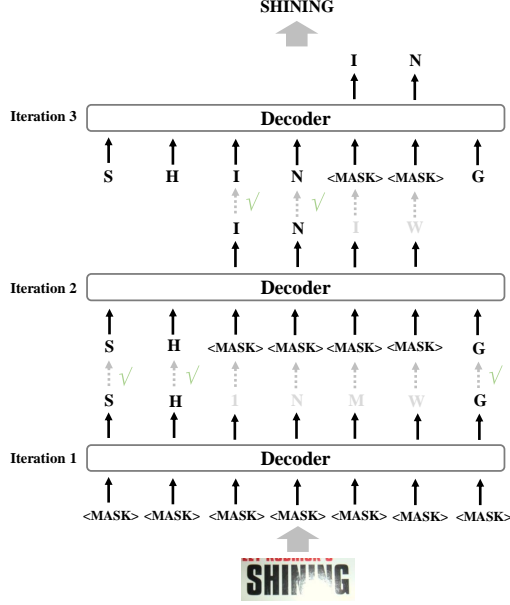


Fig. 2: An illustration of the easy-first decoding strategy. In each iteration, the characters in black represent the characters with high confidence and will be reserved in the next iteration. The characters with low confidence will be replaced with the $\langle \text{MASK} \rangle$ token and re-predicted again based on the other reachable predictions

and replaced by $\langle \text{MASK} \rangle$ again:

$$y_i^l = \begin{cases} \hat{y}_i^l, & \text{if } l \in \text{top}_k(\text{argmax}_{c_n \in C}(P(c_n^M | y_{i-1}))); \\ \langle \text{MASK} \rangle, & \text{otherwise.} \end{cases} \quad (2)$$

where y_i is the final prediction in this iteration and will act as the input in the next iteration, $\max(P)$ is the maximum probability for all candidate characters, M indicates all the $\langle \text{MASK} \rangle$ positions in the $(i-1)$ th iteration. top_k represents the most confident k predictions in the M positions which are picked to update the previous predictions. In our setting, k is equal to L/K where L is the max length of the target text, and K is the iteration number.

In summary, the predicted text is updated at the most top_k confident positions in each iteration, and the others are replaced by $\langle \text{MASK} \rangle$ again, which will be predicted in the following iterations. Once a character is reserved and updated, we will not update it again in the subsequent

iterations. We utilize the same post-processing method to deal with the length determination described in our conference version [24], where we let the decoder predict $\langle \text{EOS} \rangle$ token and choose the first $\langle \text{EOS} \rangle$ token from left to right to determine the text length.

3.3 Parallel Decoder

We adopt a Transformer-based [20] decoder as our parallel decoder without any RNN structure for high parallelism. It contains three main components: a masked self-attention, a 2D cross-attention, and an FFN. The input of masked self-attention is the text embedding combined with position encoding, which is mixed with the $\langle \text{MASK} \rangle$ tokens. Different from the original Transformer, our parallel decoder is bidirectional, thus removing the original future-characters mask for the left-to-right generation. However, we add another mask to prevent the model from attending to the special $\langle \text{MASK} \rangle$ tokens. The masked self-attention can extract abundant context information ignored in the fully non-autoregressive models.

The 2D cross-attention is similar to the original Transformer, where the multi-head scale-dot attention operation is adopted. Specifically, the attention weights are calculated between the outputs of masked self-attention and the 2D visual feature map provided by the vision encoder. The calculation details are the same as the Transformer, so we don't describe it in detail. Finally, the FFN introduces the non-linear transformation into the outputs of 2D cross-attention, and new predictions are generated later with a simple linear function in parallel. Same as the original Transformer, the residual connections exist between each module.

3.4 Discrete Diffusion for STR

In our conference paper [24], we utilize the mimicking learning strategy to enhance the training of the parallel decoder, which makes the parallel decoder mimic the FFN outputs of the teacher model with an autoregressive decoder. However, the autoregressive decoder has its intrinsic defect. Its left-to-right serial language modeling strategy ignores the bidirectional contextual information and suffers from error accumulation. Therefore, we introduce another strategy, the discrete diffusion

model strategy, which could learn abundant bidirectional linguistic knowledge to enhance internal language model learning.

The easy first decoding could be viewed as a kind of iterative denoising process of a discrete diffusion model. Here, the corrupted text y_T is formed by all $\langle \text{MASK} \rangle$ tokens, and the denoising process needs to recover the correct text conditioned on the image visual information, which means that the scene text recognition task can be viewed as a conditional text-generation task, i.e., we need to generate (recognize) the text from the given image. Unlike the continuous diffusion models, which add continuous noises individually to the RGB values or latent features of each token, the discrete character token of a word determines it is direct to model the process as the vector quantized diffusion [81]. In the forward process, our model masks the character tokens randomly by replacing the original characters with $\langle \text{MASK} \rangle$ tokens according to the randomly chosen timestep and masking ratio. In the backward process, our model recovers the original texts from the corrupted masked texts conditioned on the input image.

3.4.1 Noising Process

The noising process can be thought of as the Markov process, where the forward noise process gradually corrupts the text \mathbf{y}_0 via a fixed Markov chain $q(\mathbf{y}_t|\mathbf{y}_{t-1})$. To be specific, the original text \mathbf{y}_0 is the word contained in the image, which is formed by the tokens in $C = \{c_1, c_2, \dots, c_{N+1}\}$ consisting of the meaningful N character categories and the special $\langle \text{EOS} \rangle$ token. At each timestep, we randomly replace some unmasked tokens of \mathbf{y}_{t-1} with the special token $\langle \text{MASK} \rangle$ with the probability β_t , leaving $1 - \beta_t$ to be unchanged. The tokens never change to other unmasked tokens, and the special $\langle \text{MASK} \rangle$ never changes to other tokens. For a character y_{t-1}^l of \mathbf{y}_{t-1} in position l (omit the superscripts l in the following description), if it is not $\langle \text{MASK} \rangle$, the transition is defined as:

$$q(y_t|y_{t-1}) = \begin{cases} 1 - \beta_t, & \text{if } y_t = y_{t-1}; \\ \beta_t, & \text{if } y_t = \langle \text{MASK} \rangle; \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

If y_{t-1} is a $\langle \text{MASK} \rangle$ token, the transition is defined as:

$$q(y_t|y_{t-1}) = \begin{cases} 1, & \text{if } y_{t-1} = \langle \text{MASK} \rangle; \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

We define $c_{N+2} = \langle \text{MASK} \rangle$. Denote $[\mathbf{Q}_t]_{ij} = q(y_t = c_i | y_{t-1} = c_j) \in \mathbb{R}^{(N+2) \times (N+2)}$ as the transition matrix at timestep t ,

$$\mathbf{Q}_t = \begin{bmatrix} 1 - \beta_t & 0 & 0 & \dots & 0 \\ 0 & 1 - \beta_t & 0 & \dots & 0 \\ 0 & 0 & 1 - \beta_t & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_t & \beta_t & \beta_t & \dots & 1 \end{bmatrix} \quad (5)$$

Then the forward Markov diffusion process for the whole token sequence can be written as:

$$q(\mathbf{y}_t|\mathbf{y}_{t-1}) = \mathbf{v}^\top(\mathbf{y}_t) \mathbf{Q}_t \mathbf{v}(\mathbf{y}_{t-1}) \quad (6)$$

$\mathbf{v}(\mathbf{y})$ is a one-hot column vector representation of \mathbf{y} with the length of $(N + 2)$. According to the property of the Markov chain rule, we can derive the corrupted text \mathbf{y}_t from \mathbf{y}_0 as:

$$q(\mathbf{y}_t|\mathbf{y}_0) = \mathbf{v}^\top(\mathbf{y}_t) \overline{\mathbf{Q}}_t \mathbf{v}(\mathbf{y}_{t-1}) \quad (7)$$

where $\overline{\mathbf{Q}}_t = \mathbf{Q}_t \dots \mathbf{Q}_1$. More specially, denote the $\alpha_t = 1 - \beta_t$ and $\overline{\alpha}_t = \prod_{i=1}^t \alpha_i$,

$$q(y_t|y_0) = \begin{cases} \overline{\alpha}_t, & \text{if } y_t = y_0; \\ 1 - \overline{\alpha}_t, & \text{if } y_t = \langle \text{MASK} \rangle; \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

After a fixed number of T timesteps, the probability of $\overline{\alpha}_t$ turns close to 0, and the noise process yields a sequence of pure noise tokens.

3.4.2 Conditional Denoising Process

The recognition process is the conditional denoising process from a sequence of all masked characters \mathbf{y}_T to the original text \mathbf{y}_0 based on the visual features \mathbf{z} provided by the encoder. To reverse the diffusion process, we train a denoising network $p_\theta(\mathbf{y}_{t-1}|\mathbf{y}_t, \mathbf{z})$ to estimate the posterior transition distribution. The denoising network used in our model is the Transformer-based parallel decoder described in Section 3.3. The conditional image

features are incorporated through the 2D cross-attention.

Since the noising and denoising process happens at the character level of texts in scene images, the timescale T in our task is significantly small compared with the popular image generation task. We adopt a scaling between our T and the popular used 8,000 timestep in the image generation task, i.e., $t_s = t * 8000/T$. To indicate current timestep t , we encode it as a sinusoidal positional embedding and inject it to the parallel decoder with the Adaptive Layer Normalization [87]:

$$\text{PE}(t_s) = \begin{cases} \sin(t_s/10000^{2i/d_{\text{model}}}), & \text{if } i < d_{\text{model}}/2; \\ \cos(t_s/10000^{2i/d_{\text{model}}}), & \text{if } i \geq d_{\text{model}}/2. \end{cases} \quad (9)$$

$$\text{AdaLN}(h, t_s) = W_1 \text{PE}(t_s) \text{LayerNorm}(h) + W_2 \text{PE}(t_s) \quad (10)$$

where h is the intermediate activations, W_1 and W_2 are learnable linear projection weights of the timestep embedding.

3.5 Training and Inference Strategies

In the training process, the decoder directly predicts the original noiseless text tokens \mathbf{y}_0 given the generated noise text \mathbf{y}_t according to the exploration of [83]. The objective function is just like the usual text recognition loss:

$$\mathcal{L}_{x_0} = -\log p_{\theta}(\mathbf{y}_0|\mathbf{y}_t, \mathbf{x}) \quad (11)$$

The standard diffusion inference process starts from the \mathbf{y}_T , and sequentially predict the $\mathbf{y}_{T-1}, \mathbf{y}_{T-2}, \dots, \mathbf{y}_0$. And the process of predicting \mathbf{y}_{t-1} from \mathbf{y}_t is computed as follows: First, estimate $\hat{\mathbf{y}}_0$ base on $p_{\theta}(\mathbf{y}_0|\mathbf{y}_t, \mathbf{x})$ from the trained denoising network. Then, the noise for the $t-1$ timestep is added on the predicted $\hat{\mathbf{y}}_0$ based on the predefined \mathbf{Q} to get the \mathbf{y}_{t-1} . Unlike the standard inference, we employ the easy-first strategy with the iterative denoising process. Since during training, we utilize the diffusion strategy and inject the information of the current timestep to the normalization, for the easy-first inference, we need to transcribe the iteration index of easy-first decoding to the corresponding timestep. For the iteration number K , if K is larger than T , we reduce the iteration to the maximum diffusion

steps T ; and if K is smaller than T , we map the current k to timestep t ,

$$t = \lfloor \frac{(K-k)*T}{K} \rfloor \quad (12)$$

The noise for the timestep is not determined by \mathbf{Q} , but by the easy-first strategy, i.e., mask the characters with low recognition confidence as described in Section 3.2.

4 Experiments

4.1 Datasets

4.1.1 English Datasets

To evaluate the effectiveness and efficiency of our method, we conduct extensive experiments on almost all open-source datasets. We utilize the popular synthetic training dataset MJSynth [47] (MJ) and SynthText [88] (ST) as our synthetic training data. Since the importance of real datasets has been proved, we validate the performance of our models on real datasets collected by [13]. Besides, we also train our models on the training set of the most recent benchmark dataset, Union14M-L [89].

For English text recognition, We first evaluate our methods on the challenging dataset Union14M-Benchmark [89]. Other representative challenging datasets are also involved in the evaluation. Three large datasets: ArT [90] contains 35.1k curved and rotated hard-case images; COCO-Text [91] comprises 9.8k occluded and distorted samples; Uber-Text [92] consists of 80.6k samples featuring vertical and rotated text. The occluded datasets [62] consists of images sheltered in weak (WOST) or heavy (HOST) degree. And the dataset WordArt [93] contains images with texts in artistic style.

Besides, we provide the six widely used public benchmarks for reference but the models' performance on this benchmark is nearly saturated with some mislabeled images in it [89]. IIIT5K-Words [25] (IIIT5K) contains 5000 images collected from the website. There are 3000 images for testing, most of which are horizontal with high quality. Street View Text [29] (SVT) consists of 647 cropped word images from 249 street view images, which targets regular text recognition. ICDAR2013 [94] (IC13) consists of 1095 regular-text images for testing. Two different versions are

used for evaluation: 1015 and 857 images, which discard images that contain non-alphanumeric characters and less than three characters, respectively. ICDAR2015 [95] (IC15) is a challenging dataset for recognition due to the degraded images collected without careful focusing, which contains 2077 cropped images. Some works used 1811 images for evaluation, discarding some distorted images. SVT-Perspective [96] (SVTP) comprises 645 images cropped from SVT, which is usually used for evaluating the performance of recognizing perspective text. CUTE80 [97] (CUTE) consists of 288 curved images without lexicon, which are used for irregular text recognition.

4.1.2 Chinese Datasets

To validate our models’ performance on Chinese recognition, we train and evaluate our models on the Chinese benchmark dataset BCTR [98], which consists of four subsets: scene, web, document, and handwriting, and each subset contains datasets for training, validation, and test. The scene dataset includes several public datasets, such as RCTW [99], ReCTS [100], LSVT [90], ArT [101] and CTW [102], with 112,471 samples for training, 14,059 samples for validation, and 14,059 samples for testing. The web dataset is collected from the MTWI [103] dataset and contains 112,471, 14,059, and 14,059 samples for training, validation, and testing, respectively. The document subset is a synthetic dataset generated using Text Render in document style with 400,000 training samples, 50,000 validation samples, and 50,000 test samples. The handwriting subset is collected from a handwriting dataset SCUT-HCCDoc [104], which consists of 74,603, 18,651, and 23,389 training, validation, and testing samples, respectively.

4.2 Implementation Details

4.2.1 Model Settings

For English recognition on most datasets, there are 37 symbols covered for recognition, including numbers, case-insensitive characters, and $\langle \text{EOS} \rangle$ token. For English recognition on Union14M, we employ the same charset with [89] for fair comparison. We utilize the same charset with [98] for Chinese recognition. We set the maximum length L to 25 for English and 40 for Chinese recognition, respectively. The input image size is 32×128 ,

Table 1: Configuration for small, base and large models. d_{model} indicates the dimensions of the feature maps. d_{MLP} represents the dimension of the intermediate features in the MLP layer. h refers to the number of attention heads. n represents the number of layers used in encoder and decoder.

Models	d_{model}	d_{MLP}	Encoder		Decoder	
			h	n	h	n
Small	384	1536	6	12	12	1
Base	768	3072	12	12	24	1
Large	1024	4096	16	24	32	1

and the patch size is 4×8 . The details of the encoder and decoder transformer units settings are illustrated in Table 1, where n is the number of Transformer units, h is the number of attention heads, and d_{model} is the dimension of the hidden layer in FFN. For English recognition, we utilize the DeiT-small configuration following [13], while for Chinese recognition with all three configurations. The number of easy-first iterations of the parallel decoder is set to 5.

4.2.2 Model Training

For English recognition, we train models on the synthetic (S), real (R), or Union14M-L datasets. For Chinese recognition, we combine the four subsets for model pre-training and then finetune the model on the specific subset. The input images are resized to 32×128 directly. We utilize Adam [105] as our optimizer with the 1cycle [106] learning rate scheduler. After 85% of the total iterations, we replace it with Stochastic Weight Averaging (SWA) [107]. A weight decay of $1e - 4$ and a warmup ratio of 0.05 are employed. Learning rates vary per model.

We set the training epochs as 500 for Chinese recognition pre-training and 100 for specific subset finetuning. We train IPAD 50, 150, and 50 epochs for English recognition on synthetic, real, and Union14M-L datasets, respectively. The batch size is 384, 192, and 96 for small, base, and large models, respectively. Data augmentations like Gaussian blur, Poisson noise, and rotation are randomly performed the same with PARSeq [13].

4.2.3 Model Evaluation

We evaluate the English recognition for 36-charset word accuracy, including case-insensitive alphabets and digits. The 7248-sample average accuracy of six usual benchmarks is calculated with IC13-857 and IC15-1811, and the 7672-sample average accuracy is calculated with IC13-1015 and IC15-2077. For Chinese recognition evaluation, we post-process the predictions of the models following [98]: (i) convert the full-width characters to half-width characters; (ii) convert all traditional Chinese characters to simplified characters; (iii) convert all English characters to lowercase; (iv) remove all spaces.

4.3 Comparisons with State-of-the-Art Methods

4.3.1 English Recognition

Given recent studies highlighting performance saturation on six usual benchmarks, we first conduct experiments on the Union14M-Benchmark to demonstrate the performance of our model. As shown in Table 2, trained with the Union14M-L, our IPAD improves the accuracy by an average of 4.9%. Notably, on the curved, multi-oriented, and salient datasets, the IPAD surpasses the previous best model by 4.7%, 15.1%, and 2.0%, respectively, which demonstrates the assistance of abundant contextual information. However, leveraging contextual information during recognition also has drawbacks, such as the unsatisfactory performance on the incomplete dataset. As shown in Table 2, the best model on the incomplete dataset is CRNN [2] and SATRN [41], which do not leverage linguistic knowledge. This occurs because models that leverage contextual information are trained on large volumes of complete English words. Incomplete words fall outside their learned linguistic patterns, causing these context-dependent models to fill in missing characters in an attempt to align the results with their internal linguistic knowledge. For example, when faced with an incomplete word “convenienc”, the model is likely to add an “e” to form “convenience”, as this fits the contextual information they get during training.

Trained with synthetic or real datasets, we evaluate our model on large and intricate datasets.

These include the three extensive datasets presented in Table 3 and the occluded and artistic datasets featured in Table 4. As evidenced in Table 3, when trained on synthetic datasets, our IPAD exhibits superior performance to the autoregressive (AR) version PARSeq_A. While models trained on real datasets may not surpass the AR model in terms of performance, they excel in reduced inference time. Compared with other non-autoregressive (NAR) models, our models manifest enhanced performance, with IPAD showing improvements of 1.2%. The results in Table 4 suggest that our models excel in recognizing text on occluded images due to their advanced context learning capabilities. Additionally, they surpass the CornerTransformer [93] in artistic text recognition.

Following previous works, We also compare our model IPAD with other published state-of-the-art methods on the six widely used benchmarks. As delineated in Table 5, it is evident that AR methods typically outperform their NAR counterparts. Even though an external language module can function as a recognition checker and refine the preliminary recognition results, the performance of LM methods consistently lags behind that of the AR models. Furthermore, by leveraging the diffusion model, IPAD elevates performance across nearly all datasets, marking a notable impact on irregular ones. This paper tries to balance the accuracy and the efficiency. Our IPAD demonstrates accuracy comparable to most autoregressive methods while ensuring accelerated inference speed. Compared to the AR-based method PARSeq 1 refinement [13], our methodology produces similar accuracy on average but needs 2.5ms less time. Aligning with PARSeq [13], we also train our models utilizing real data. As delineated in Table 5, trained on real datasets, our IPAD can also align in accuracy with the refined autoregressive model, PARSeq_A.

4.3.2 Chinese Recognition

To elucidate the capabilities of our model in more complex, non-Latin character recognition scenarios, we compare IPAD against the state-of-the-art Chinese recognition models on the BCTR dataset. As shown in Table 6, our IPAD without pre-training has surpassed the preceding methods on three subsets. Specifically, it has introduced

Table 2: Recognition results on Union14M-Benchmark. The average is calculated by directly averaging the seven results. MAERec-S* represents the model without pre-train. PIMNet is reproduced with the ViT small encoder.

Method	Curve	Multi-Oriented	Artistic	Contextless	Salient	Multi-Words	General	Avg	Incomplete (↓)
CRNN [2]	19.4	4.5	34.2	44.0	16.7	35.7	60.4	30.7	0.9
SAR [11]	68.9	56.9	60.6	73.3	60.1	<u>74.6</u>	76.0	67.2	2.1
SATRN [41]	74.8	64.7	67.1	76.1	72.2	74.1	75.8	72.1	0.9
SRN [12]	49.7	20.0	50.7	61.0	43.9	51.5	62.7	48.5	2.2
ABINet [18]	75.0	61.5	65.3	71.1	72.9	59.1	79.4	69.2	2.6
VisionLAN [62]	70.7	57.2	56.7	63.8	67.6	47.3	74.2	62.5	<u>1.3</u>
SVTR [8]	72.4	68.2	54.1	68.0	71.4	67.7	77.0	68.4	2.0
MATRNet [108]	<u>80.5</u>	64.7	<u>71.1</u>	74.8	<u>79.4</u>	67.6	77.9	74.6	1.7
MAERec-S* [89]	75.4	66.5	<u>76.1</u>	<u>76.1</u>	72.6	77.0	<u>80.8</u>	73.5	3.5
PIMNet [24]	80.3	<u>79.8</u>	68.4	75.9	77.8	68.3	<u>80.8</u>	<u>75.9</u>	<u>1.3</u>
IPAD (Ours)	85.2	83.3	72.1	78.4	81.4	73.7	82.2	79.5	<u>1.3</u>

Table 3: Text recognition results on three large datasets. S and R represent the synthetic and real training dataset. ‡ means that the model is reproduced by [13]. The subscripts _A and _N represent autoregressive and non-autoregressive version. PIMNet is reproduced with the ViT small encoder.

Method	Train Data	ArT	COCO	Uber	Avg
TRBA [‡] [10]	S	68.2	61.4	38.0	48.3
PARSeq _A [13]	S	70.7	<u>64.0</u>	42.0	<u>51.8</u>
ABINet [‡] [18]	S	65.4	57.1	34.9	45.2
ViTSTR-S [‡] [48]	S	66.1	56.4	37.6	47.0
PARSeq _N [13]	S	69.1	60.2	39.9	49.7
PIMNet [24]	S	70.3	63.8	42.3	<u>51.8</u>
IPAD (Ours)	S	<u>70.5</u>	64.2	42.6	52.1
TRBA [‡] [10]	R	82.5	77.5	81.2	81.3
PARSeq _A [13]	R	84.5	79.8	84.5	84.1
ABINet [‡] [18]	R	81.2	76.4	71.5	74.6
ViTSTR-S [‡] [48]	R	81.1	74.1	78.2	78.7
PARSeq _N [13]	R	83.0	77.0	82.4	82.1
PIMNet [24]	R	83.7	78.6	83.2	83.0
IPAD (Ours)	R	<u>83.8</u>	<u>78.8</u>	<u>83.6</u>	<u>83.3</u>

improvements of 12.1% on the scene subset, 7.1% on the web subset, 0.9% on the document subset, and an average enhancement of 4.8%.

Moreover, we initially pre-train the IPAD on the combined BCTR training dataset before fine-tuning it on specific subsets since the dataset for Chinese recognition is relatively small, and the STR models still demand large datasets. This approach produces notable improvements, and

Table 4: Recognition results on occluded and artist images. PIMNet is reproduced with the ViT small encoder. All models are trained on synthetic datasets.

Model	HOST	WOST	WordArt
VisionLAN [62]	50.3	70.3	-
CornerTransformer [93]	-	-	70.8
PIMNet [24]	<u>70.8</u>	<u>81.5</u>	<u>71.2</u>
IPAD (Ours)	71.7	82.2	71.4

our pretrained models demonstrate their superior performance. Notably, IPAD_{Small} manifests substantial advancements in web and handwriting text recognition with improvements of 11.1% and 18.0%, respectively. We attribute part of MaskOCR [113]’s superior performance in handwriting recognition to its reliance on extensive pre-training datasets, potentially including a wealth of handwriting samples and contextually analogous texts. However, the marked proficiency of our model in scene and web subsets underscores its ability to perceive linguistic information. Besides, the pronounced efficacy of our model in a smaller configuration is particularly notable (10.6% and 8.4% improvements compared with MaskOCR_{ViT-S} on scene and web subsets, respectively), demonstrating its practical utility and efficiency.

Table 5: Recognition results on six benchmark datasets. S¹ LM pretrained on WikiText-103 [109]. MGP-STR[‡] represents the small model for fair comparison. ‡ means the model is reproduced by [13]. The subscripts _A and _N represent autoregressive and Non-autoregressive version. PIMNet is reproduced with the ViT small encoder. The inference time is averaged on 7,672 images.

	Method	Train Data	IIIT5K 3000	SVT 647	IC13 857 1015	IC15 1811 2077	SVTP 645	CUTE 288	Average 7248 7672	Params. (M)	Time (ms/img.)
AR	ASTER [9]	S	93.4	89.5	-	91.8	76.1	-	78.5	79.5	-
	TRBA [10]	S	87.9	87.5	93.6	92.3	77.6	71.8	79.2	74.0	84.6 82.8 49.6 16.7
	SEED [43]	S	93.8	89.6	-	92.8	80.0	-	81.2	83.6	-
	RobustScanner [110]	S	95.3	88.1	-	94.8	-	77.1	79.5	90.3	-
	PTIE [46]	S	96.3	94.9	-	97.2	87.8	84.3	90.1	91.7	- 92.4 45.9 52.0
	LevOCR [61]	S ¹	96.6	92.9	96.9	-	86.4	-	88.1	91.7	92.8 - 92.6 60.5
	PARSeq _A [13]	S	97.0	93.6	97.0	<u>96.2</u>	86.5	82.9	88.9	<u>92.2</u>	<u>93.2</u> <u>91.9</u> 23.8 17.5
NAR	ViTSTR-S [48]	S	86.6	87.3	92.1	91.2	77.9	71.7	81.4	77.9	84.3 82.5 85.5 7.92
	SRN [12]	S	94.8	91.5	95.5	-	82.7	-	85.1	87.8	90.4 - 54.7 12.1
	VisionLAN [62]	S	95.8	91.7	95.7	-	83.7	-	86.0	88.5	91.2 - 32.8 14.8
	ABINet [18]	S ¹	96.2	93.5	97.4	95.7	86.0	85.1	89.3	89.2	92.6 - 36.7 23.4
	PARSeq _N [13]	S	95.7	92.6	96.3	95.5	85.1	81.4	87.9	91.4	92.0 90.7 23.8 11.7
	MGP-STR [‡] [63]	S	95.3	93.5	96.4	-	86.1	-	87.3	87.9	92.0 - 52.6 9.37
	PIMNet [24]	S	96.6	93.2	96.6	95.7	86.0	82.3	87.8	91.0	92.6 91.4 24.2 14.6
	IPAD (Ours)	S	<u>96.8</u>	<u>94.3</u>	<u>97.0</u>	95.6	<u>86.7</u>	<u>83.0</u>	<u>89.3</u>	93.4	93.3 <u>91.9</u> 24.7 15.0
AR	TRBA [‡] [10]	R	98.6	97.0	97.6	97.6	89.8	88.7	93.7	97.7	95.7 95.2 49.8 21.5
	PARSeq _A [13]	R	99.1	97.9	<u>98.3</u>	98.4	<u>90.7</u>	<u>89.6</u>	95.7	98.3	96.4 96.0 23.8 17.5
NAR	ViTSTR-S [‡] [48]	R	98.1	95.8	97.6	97.7	88.4	87.1	91.4	96.1	94.7 94.3 21.4 6.94
	ABINet [‡] [18]	R	98.6	97.8	98.0	97.8	90.2	88.5	93.9	97.7	95.9 95.2 36.9 22.8
	PARSeq _N [13]	R	98.3	97.5	98.0	<u>98.1</u>	89.6	88.4	94.6	97.7	95.7 95.2 23.8 11.7
	PIMNet [24]	R	98.9	<u>97.8</u>	98.4	<u>98.1</u>	89.6	88.4	94.3	<u>97.9</u>	95.9 95.4 24.2 14.6
	IPAD (Ours)	R	<u>99.0</u>	97.7	98.1	<u>98.1</u>	90.8	89.8	<u>95.5</u>	<u>97.9</u>	96.4 96.0 24.7 15.0

Table 6: Chinese text recognition results on the BCTR datasets. The ‘Combined’ dataset contains about 1.1M data.

Method	(Pre)Training Data	Finetuning Data	Scene	Web	Doc.	Handw.	Avg	Params. (M)
CRNN [2]	Specific	-	54.9	56.2	97.4	48.0	68.0	12
ASTER [9]	Specific	-	59.4	57.8	97.6	45.9	69.8	27
MORAN [36]	Specific	-	54.7	49.6	91.7	30.2	62.7	29
SAR [11]	Specific	-	53.8	50.5	96.2	31.0	64.0	28
SEED [43]	Specific	-	45.4	31.4	96.1	21.1	57.1	36
MASTER [111]	Specific	-	62.1	53.4	82.7	18.5	61.4	63
ABINet [18]	Specific	-	60.9	51.1	91.7	13.8	62.9	53
TransOCR [112]	Specific	-	67.8	62.7	97.9	51.7	74.8	84
PIMNet _{Small} [24]	Specific	-	78.8	67.8	98.4	37.6	77.9	30
IPAD _{Small} (Ours)	Specific	-	79.9	69.8	98.8	45.4	79.9	31
MaskOCR _{ViT-S} [113]	100M Unlabeled Real + 100M Synthetic	Specific	71.4	72.5	98.8	55.6	78.1	36
MaskOCR _{ViT-B} [113]			73.9	74.8	99.3	63.7	80.8	100
MaskOCR _{ViT-L} [113]			76.2	76.8	99.4	67.9	82.6	318
PIMNet _{Small} [24]	Combined	Specific	81.0	80.0	98.9	59.0	83.4	30
IPAD _{Small} (Ours)			82.0	80.9	99.1	61.7	84.4	31
IPAD _{Base} (Ours)			83.5	81.9	99.2	63.0	85.4	110
IPAD _{Large} (Ours)			84.2	82.6	99.2	64.2	85.9	342

4.4 Ablation Studies

In this subsection, we conduct ablation studies to compare IPAD with our conference version PIMNet [24] and show the effectiveness of each proposed module and strategy within both English

and Chinese recognition contexts. All experiments within this section utilize small models to maintain consistency and focus on component-specific impacts. In PIMNet, we utilize the mimicking learning strategy to enhance the training of the parallel decoder, where the FFN outputs of the

parallel decoder are aligned to the FFN feature of an autoregressive decoder with the same encoder. Additionally, to maintain consistency in the encoder, we replace the CNN encoder from the conference version with the same small ViT encoder used in IPAD.

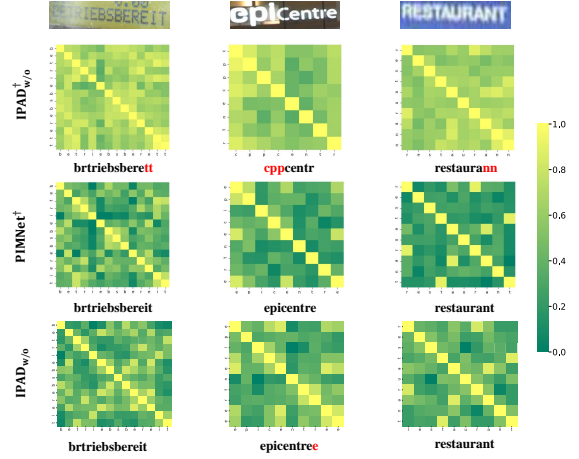
4.4.1 Analysis of Different Components

We conduct ablation studies on different components to dissect their contribution towards performance improvement, as detailed in Table 7. The foundational baseline consists of a purely non-autoregressive model with our parallel decoder. Subsequently, for $\text{IPAD}_{w/o}$ (without any mimicking or diffusion step), we implement the easy-first decoding strategy, utilizing linguistic information throughout the iterative decoding process and formulating predictions for the masked characters based on their relationships with the unmasked ones. As shown in Table 7, the deployment of easy-first decoding exhibits discernible improvements across both English and Chinese datasets, highlighting a notable 2.22% improvement on the Chinese datasets.

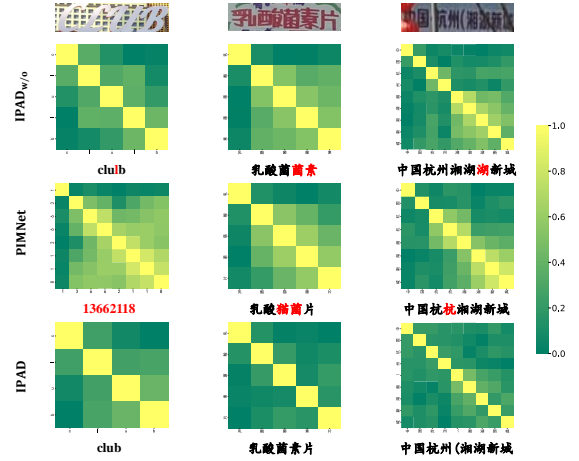
The diffusion strategy empowers the IPAD with a profound implicit understanding of contextual relationships within words by incorporating a time-dependent noising and denoising generation process during the training phase. It could explore more possible combinations of context. Therefore, the IPAD performs better compared with $\text{IPAD}_{w/o}$, leading to advancements of 0.78% and 0.54% on English trained with synthetic and real datasets, and 1.30% on Chinese, respectively. The diffusion strategy’s effectiveness demonstrates its pivotal role in enabling the model to navigate and comprehend the intricate linguistic landscapes and relationships inherent in text recognition tasks. In contrast, incorporating the mimicking learning strategy used in our conference version only enables PIMNet marginally superior performance relative to $\text{IPAD}_{w/o}$. The increments are almost negligible, especially for models trained with real datasets for English and Chinese recognition.

4.4.2 Comparison between Mimicking Learning and Discrete Diffusion

To analyze and compare the efficacy of the mimicking learning and discrete diffusion strategies



(a) Normalized cosine similarities of FFN outputs for relatively simple English test images.



(b) Normalized cosine similarities of FFN outputs for more challenging English and Chinese test images.

Fig. 3: Comparison of normalized cosine similarities of FFN outputs across different encoders. \dagger indicates models using a CNN encoder. w/o denotes models without mimicking or diffusion steps.

further, we provide visual representations of the cosine similarities of FFN outputs. As depicted in Fig. 3b, the parallel decoder in the early iterations tends to predict similar outputs due to the fully parallel decoding, thus the similarities among neighboring positions tend to be large. The similar outputs of FFN may mislead the final predictions, which tend to be duplicated and wrong. In PIMNet, erroneous predictions of preceding characters can influence the recognition of subsequent

Table 7: Experimental results using different components. Recognition accuracy on English is the average accuracy over the 7248 samples. The Chinese recognition trains on the combined BCTR dataset without finetuning. w/o means without any mimicking or diffusion step.

Model	Base	Easy-first	Mimicking	Diffusion	English		Chinese
					Synthetic	Real	
Baseline	✓	✗	✗	✗	92.00	95.35	80.56
IPAD $_{w/o}$	✓	✓	✗	✗	92.47	95.85	82.78
PIMNet	✓	✓	✓	✗	92.63	95.94	82.86
IPAD	✓	✓	✗	✓	93.25	96.39	84.08

characters and introduce noise into the FFN outputs. This occurs as it mimics the FFN feature of the autoregressive decoder, which is inherently incapable of learning bidirectional context and is prone to significant error accumulation. Thus, the similarities across positions in the left part are relatively large, which may mislead the final predictions.

Conversely, the application of the diffusion strategy allows IPAD to yield more distinguishable FFN outputs in challenging cases, for instance, against complex backgrounds, and in cases of occluded and irregular text, as demonstrated in Fig. 3b. This substantiates the diffusion strategy’s role in enhancing the model’s adaptability in deciphering text under intricate conditions.

4.4.3 Analysis of Encoder

We analyze the influence of different encoders to gain insights into their impacts on performance. In IPAD $_{w/o}^{\dagger}$ and PIMNet † , we employ a Feature Pyramid Network (FPN) coupled with a ResNet-50, supplemented by two additional transformer units as the encoder. In IPAD $_{w/o}$ and PIMNet, a Vision Transformer (ViT) with 4×8 image patches is deployed.

As revealed in Table 8, leveraging ViT improves the performance of PIMNet on the six standard English recognition benchmarks markedly, introducing improvements exceeding 2%. Furthermore, because of the refined quality of image features encoded by ViT, the model exhibits diminished reliance on the decoder. Consequently, the mimicking learning strategy utilized in our conference paper yields marginal gains of 0.1% with the ViT encoder, contrasting the 0.4% gains observed with the previous CNN encoder. This phenomenon indicates that the mimicking

learning strategy in our conference paper is somewhat inadequate to enhance the parallel decoder’s contextual information learning capabilities in the context of a ViT encoder.

In Fig. 3a, the CNN encoder shows limited capacity in capturing relationships among detailed image features, which is why mimicking learning helps PIMNet † produce more distinguishable FFN outputs and improve prediction accuracy. However, when switching to the ViT encoder, which synthesizes higher-quality visual features and better captures visual relationships, the need for mimicking learning decreases, and its limitations become more apparent. The ViT encoder enables IPAD $_{w/o}$ to generate clearer and more discriminative FFN features on the same test images as IPAD $_{w/o}^{\dagger}$. Notably, IPAD $_{w/o}$ is able to recognize these relatively simple test images in Fig. 3a almost correctly without any specific decoding design. To further illustrate the advantages of IPAD, Fig. 3b presents more challenging English and Chinese test images. As shown in Fig. 3b, the FFN pattern for IPAD $_{w/o}$ is also distinct enough. The second row of Fig. 3b shows the FFN outputs of PIMNet with the ViT encoder. By comparing this with the first row, it becomes clear that early recognition errors (such as those seen with the “club” figure in the mimicking learning PIMNet) negatively impact the discriminability of subsequent predictions. This occurs because mimicking learning attempts to replicate the autoregressive decoding process, which accumulates errors from earlier stages of decoding.

In summary, Table 8 demonstrates the numerical improvements gained from the ViT encoder. Meanwhile, Fig. 3a compares the FFN patterns of both CNN-based and ViT-based models. The CNN-based IPAD $_{w/o}^{\dagger}$ exhibits less clarity, whereas the ViT-based IPAD $_{w/o}$ generates more distinct

Table 8: Ablation study on different encoder. IC13 here is IC13-857 and IC15 is IC1-1811. [†] indicates the model uses a CNN encoder. _{w/o} means without any mimicking or diffusion step.

Model	encoder	IIT5K	SVT	IC13	IC15	SVTP	CUTE	Avg.
IPAD _{w/o} [†]	CNN	94.9	90.1	94.7	82.9	83.4	86.1	90.1
PIMNet [†]	CNN	95.2	91.2	95.2	83.5	84.3	84.4	90.5
IPAD _{w/o}	ViT	96.7	94.0	96.4	85.0	87.6	91.0	92.5
PIMNet	ViT	96.6	93.2	96.6	86.0	87.8	91.0	92.6

and discriminative FFN features, even without specific decoding strategies. Fig.3a further highlights the performance on more challenging English and Chinese test images, showing that the ViT encoder effectively handles complex visual features and improves feature representation, reinforcing the advantages of our proposed method.

4.4.4 Effect of Timestep Setting in Discrete Diffusion

The diffusion model’s performance is intricately linked to how noise samples are generated, and this process is significantly influenced by the specific timestep schedule. How to set the timestep embedding plays a pivotal role in the model’s denoising capabilities. In the context of the presented results in Table 9, it becomes evident that if the model ignores the timestep information during noising and denoising, its performance will suffer greatly. This decline in accuracy is more obvious for Chinese recognition since the linguistic context in Chinese is more complex. The phenomenon underscores the significance of timestep information in equipping the model to select the most pertinent contextual cues effectively. When Adaptive Layer Normalization is utilized with a learnable timestep embedding, the accuracy is also less than satisfactory compared with the sinusoidal positional embedding’s results. The model demonstrates commendable results when provided with a little timestep than the length of the text. For example, although the prescribed text length for Chinese is 40, setting the timestep to $T = 25$ seems optimal for Chinese recognition. This preference can be explained by the fact that most Chinese text images contain fewer than 25 characters.

Table 9: The influences of different ways to set the timestep embedding. W/o Timestep means that we do not indicate the timestep during the denoising process.

Method	English		Chinese
	Synth	Real	
W/o Timestep	92.89	96.03	81.97
Learnable Embedding	92.95	96.21	82.65
Sinusoidal Embedding (T=10)	93.16	96.32	83.58
Sinusoidal Embedding (T=25)	93.25	96.39	84.08
Sinusoidal Embedding (T=40)	93.02	96.14	83.99
Sinusoidal Embedding (T=80)	93.10	95.93	83.96

4.4.5 Analysis of Iteration Number

The number of iterations is an important hyperparameter in our method, so we conduct experiments to analyze the effect caused by the iterations number. As shown in Table 10 and Table 11, the fully parallel IPAD model with only one iteration works poorly on both English and Chinese recognition. As we discussed previously, the fully parallel decoding lacks useful context information, which impacts the optimization procedure. When an additional iteration is adopted, the performance is improved significantly, especially on Scene (from 77.17% to 80.53%), Web (from 77.61% to 79.72%), and Handwriting (from 50.82% to 57.52%) subsets of BCTR. With the increment of iterations, the accuracy and inference time increase accordingly. Five is selected as the number of iterations in the implementation to achieve a balance. Note that the timestep T of diffusion in our implementation is 25, so the model with 25 iterations is the best. By utilizing different numbers of iterations, we can achieve various goals, i.e., achieve high performance or speed, which also shows the flexibility of our model.

Table 10: The comparison of accuracy and inference time with different iterations.

Iter	1	2	3	5	10	15	25
S	92.15	92.98	92.96	93.25	93.41	93.46	93.60
R	94.90	95.94	96.27	96.39	96.37	96.36	96.27
Time	9.39	11.3	12.9	15.0	23.9	32.4	47.8

Table 11: The comparison of accuracy and inference time with different iterations for Chinese recognition.

Iter	Scene	Web	Doc	Hand	Avg.	Time (ms/img.)
1	77.17	77.61	98.66	50.82	80.23	9.45
2	80.56	79.68	98.94	57.51	82.98	11.6
3	81.30	80.15	99.01	59.24	83.63	13.9
5	81.85	80.38	99.00	60.30	84.04	16.8
10	82.18	80.49	99.04	60.84	84.28	26.0
15	82.25	80.56	99.06	60.99	84.35	34.7
25	82.23	80.50	99.06	61.07	84.35	51.6

Table 12: Text recognition results on the BCTR datasets. w/o means without any mimicking or diffusion step. ‘Spec.’ means that we utilize the corresponding subset to train a specific model for the four datasets, while ‘Comb.’ means that we combine the four subsets. FT indicates whether the finetuning is employed. For finetuning, the model is finetuned on specific dataset.

Model	Data	FT	Scene	Web	Doc.	Handw.
IPAD w/o	Spec.	✗	78.62	66.44	98.07	33.90
	Comb.	✗	80.71	79.06	98.27	57.14
	Comb.	✓	81.10	79.81	98.63	58.00
IPAD	Spec.	✗	79.94	69.75	98.79	43.74
	Comb.	✗	81.85	80.38	99.00	60.30
	Comb.	✓	82.02	80.89	99.11	61.70

4.4.6 Training Procedure for Chinese Recognition

As described in section 4.1.2, the BCTR dataset consists of four subsets. Previous methods usually train and test on each specific subset. And some methods utilize synthetic datasets to pre-train their models and then finetune them on different subsets. However, we notice that combining the four subsets for training and directly

Table 13: The comparison of inference time among different decoding strategies. w/o means without any mimicking or diffusion step. We evaluated the model, trained on synthetic datasets, using the six English benchmarks to maintain consistency with previous studies.

Methods	Time(ms/image)	Acc.
CTC based	8.14	90.3
Attention based	43.9	92.0
IPAD w/o (1 iteration)	9.39	90.8
IPAD w/o (5 iterations)	14.6	92.5
IPAD (1 iteration)	9.39	92.2
IPAD (5 iteration)	15.0	93.3

testing the model on the particular task can outperform training on the specific dataset, especially on the Web and Handwriting dataset, as shown in Table 12. And the validation set is also the combination of the four sub-datasets. We argue that the reason is the small amount of data because combining all four datasets can only result in 1,096,238 samples for training, whose amount is one fourteenth of the English synthetic dataset and about a third of the English real dataset. Besides, the Chinese charset is much larger than the English, which demands more samples to perform similarly. If we further finetune the model trained on the whole dataset on a particular sub-dataset, we could improve the model’s performance further.

4.4.7 Analysis of Inference Speed

To further verify our advantage in efficiency, we compare our method with CTC and AR attention-based decoders. To remove the influence of the encoder, we evaluate all methods with the same encoder as our IPAD. As shown in Tab 13, our method with 5 iterations is almost 3 times faster than AR attention-based decoders while achieving higher recognition accuracy. When the iteration number is 1, the inference time becomes comparable to the CTC-based decoder since they are both fully parallel decoding. In other words, our method is flexible between efficiency and accuracy. The iteration number can be adjusted based on the efficiency and accuracy requirements of different real-world applications.

					
CRNN:	嵊州_糕	安全警钟时刻条鸡	北京启点百室	曙光街道社区店品	商头卷购
ASTER:	嵊州宝盖	安全警钟时刻尚房	北京启点百室	曙光街道社区店民	蜀烤购
MORAN:	嵊州空港	安全警钟时刻染店	北京启点百室	曙光街道社区店品	柳华购片
SAR:	嵊州全盖	安全警钟时刻贺鸣	北京启点百室	曙光街道社区店_	顺投购购
SEED:	嵊州红糕	安全警钟时刻章	北京点南堂室	曙光街道社区店_	商头_
TransOCR:	嵊州鱼糕	安全警钟时刻盘鸡	北京启点百堂	曙光街道社区活品	酒类牛肉
IPAD:	嵊州年糕	安全警钟时刻长鸣	北京启点画室	曙光街道社区居民	源头采购

					
CRNN:	和辣鱼	华库红包	祛烧八味	满轮增鸭	百世界币
ASTER:	美都鱼	华等红包	龙状八味	演轮增强	百世界浴
MORAN:	声深典	华丰丝包	花城八味	汤馆增强	百世界_
SAR:	老琳鱼	坐军红包	龙坎八味	渴松增鸭	百世界完
SEED:	手小煲	华辉烟包	坛北公味	网轮胎压	百世界_
TransOCR:	天津泉	丝等红包	姑烧八味	潮输增鸭	百世界万
IPAD:	江湖鱼	坐等红包	坛烧八味	涡轮增鸭	百世泉酒

Fig. 4: Comparison of recognition results on the scene subset of the BCTR dataset. The characters in red are wrongly recognized. The red “_” means missing characters.




			
Iteration 1	人口防量·高人口素	浙江新兴房化产及有限公司	瑞桥委丝
Iteration 2	人口防量提高人口素	浙江新兴房化产开发有限公司	瑞得委丝
Iteration 3	人口数量提高人口素质	浙江新兴房地产开发有限公司	瑞得委丝
Iteration 4	人口数量提高人口素质	浙江新兴房地产开发有限公司	瑞得委丝

Fig. 5: Some examples to illustrate the iterative generation of IPAD with easy first.

4.5 Qualitative Analysis

4.5.1 Easy First Decoding

As shown in Fig. 5 We also provide a depiction of the easy-first decoding process employed by IPAD for Chinese recognition. Given the intricate nature of Chinese texts, which often operate at the sentence level without definitive word boundaries, understanding context becomes paramount. The combined prowess of the easy-first and discrete diffusion strategies enables our model to harness the inherent contextual ties between characters, refining initial, less accurate predictions. A case in point is the first image: an optical illusion initially led the model astray, resulting in the prediction “人口防量·高人口素”. However, subsequent iterations capitalized on the contextual cues, leading to improved accuracy. The third character “数” eventually gets recognized, aided by the surrounding characters.

4.5.2 Chinese Recognition

In light of the outstanding performance exhibited by our IPAD compared to other state-of-the-art methods, especially on the Scene subset of BCTR, we undertook qualitative comparisons against previous methodologies. Figure 4 offers illustrative examples, underscoring IPAD’s robustness in recognizing occluded texts, texts in non-standard fonts, and texts embedded in intricate backgrounds. As illustrated, most previous methods recognize the final image in the first row wrongly because it is influenced by significant disturbances. They also do not perform well on the inaugural image in the second row with artistic text formations. However, these two images are interpreted by our IPAD correctly.

5 Conclusion

In this paper, we propose a Parallel, Iterative, and Diffusion-based network for STR, which utilizes the easy-first strategy to refine the coarse prediction iteratively and successfully balance the accuracy and efficiency of scene text recognition. To further improve the contextual learning ability of the parallel decoder, the IPAD enables the parallel decoder to learn more abundant bidirectional contextual information with the discrete diffusion strategy by viewing STR as an image-based conditional text generation process. Extensive experiments demonstrate that our model achieves

comparable results on most English benchmarks and a faster inference speed compared with other state-of-the-art autoregressive methods. Besides, it boosts the performance of BCTR datasets for Chinese recognition, especially the Scene subset. In the future, we will explore the potential of introducing the concept of large language models in NLP to our model.

Acknowledgements. Supported by the National Natural Science Foundation of China (Grant NO 62376266), and by the Key Research Program of Frontier Sciences, CAS (Grant NO ZDBS-LY-7024).

Data Availability Statement. The data that support the findings of this study are openly available from the corresponding reference papers.

References

- [1] Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., Liang, J.: EAST: An efficient and accurate scene text detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5551–5560 (2017)
- [2] Shi, B., Bai, X., Yao, C.: 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)
- [3] Wang, W., Zhou, Y., Lv, J., Wu, D., Zhao, G., Jiang, N., Wang, W.: TPSNet: Reverse thinking of thin plate splines for arbitrary shape scene text representation. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 5014–5025 (2022)
- [4] Zeng, G., Zhang, Y., Zhou, Y., Yang, X., Jiang, N., Zhao, G., Wang, W., Yin, X.-C.: Beyond OCR+VQA: Towards end-to-end reading and reasoning for robust and accurate textvqa. *Pattern Recognition* **138**, 109337 (2023)
- [5] He, P., Huang, W., Qiao, Y., Loy, C., Tang, X.: Reading scene text in deep convolutional sequences. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30 (2016)
- [6] Su, B., Lu, S.: Accurate recognition of words in scenes without character segmentation using recurrent neural network. *Pattern Recognition* **63**, 397–405 (2017)
- [7] Wang, J., Hu, X.: Gated recurrent convolution neural network for ocr. *Advances in Neural Information Processing Systems* **30** (2017)
- [8] Du, Y., Chen, Z., Jia, C., Yin, X., Zheng, T., Li, C., Du, Y., Jiang, Y.-G.: SVTR: Scene text recognition with a single visual model. In: Proceedings of the 31st International Joint Conference on Artificial Intelligence, pp. 884–890 (2022)
- [9] Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: ASTER: An attentional scene text recognizer with flexible rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(9), 2035–2048 (2018)
- [10] Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S.J., Lee, H.: What is wrong with scene text recognition model comparisons? Dataset and model analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4715–4723 (2019)
- [11] Li, H., Wang, P., Shen, C., Zhang, G.: Show, attend and read: A simple and strong baseline for irregular text recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8610–8617 (2019)
- [12] Yu, D., Li, X., Zhang, C., Liu, T., Han, J., Liu, J., Ding, E.: Towards accurate scene text recognition with semantic reasoning networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12113–12122 (2020)
- [13] Bautista, D., Atienza, R.: Scene text recognition with permuted autoregressive sequence models. In: Proceedings of the

- European Conference on Computer Vision, pp. 178–196 (2022)
- [14] Liao, M., Zhang, J., Wan, Z., Xie, F., Liang, J., Lyu, P., Yao, C., Bai, X.: Scene text recognition from two-dimensional perspective. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8714–8721 (2019)
 - [15] Wan, Z., He, M., Chen, H., Bai, X., Yao, C.: Textscanner: Reading characters in order for robust scene text recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12120–12127 (2020)
 - [16] Zhong, D., Lyu, S., Shivakumara, P., Yin, B., Wu, J., Pal, U., Lu, Y.: SGBANet: Semantic GAN and balanced attention network for arbitrarily oriented scene text recognition. In: Proceedings of the European Conference on Computer Vision, pp. 464–480 (2022)
 - [17] Yang, X., Qiao, Z., Wei, J., Yang, D., Zhou, Y.: Masked and permuted implicit context learning for scene text recognition. *IEEE Signal Processing Letters* (2024)
 - [18] Fang, S., Xie, H., Wang, Y., Mao, Z., Zhang, Y.: 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, pp. 7098–7107 (2021)
 - [19] Goldberg, Y., Elhadad, M.: An efficient algorithm for easy-first non-directional dependency parsing. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 742–750 (2010)
 - [20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in Neural Information Processing Systems* **30** (2017)
 - [21] Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems* **34**, 8780–8794 (2021)
 - [22] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
 - [23] He, Z., Sun, T., Tang, Q., Wang, K., Huang, X., Qiu, X.: DiffusionBERT: Improving generative masked language models with diffusion models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, pp. 4521–4534 (2023)
 - [24] Qiao, Z., Zhou, Y., Wei, J., Wang, W., Zhang, Y., Jiang, N., Wang, H., Wang, W.: PIMNet: A parallel, iterative and mimicking network for scene text recognition. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 2046–2055 (2021)
 - [25] Mishra, A., Alahari, K., Jawahar, C.: Scene text recognition using higher order language priors. In: British Machine Vision Conference (2012)
 - [26] Mishra, A., Alahari, K., Jawahar, C.: Top-down and bottom-up cues for scene text recognition. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2687–2694 (2012). IEEE
 - [27] Neumann, L., Matas, J.: Real-time scene text localization and recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3538–3545 (2012)
 - [28] Novikova, T., Barinova, O., Kohli, P., Lempitsky, V.: Large-lexicon attribute-consistent text recognition in natural images. In: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part VI 12, pp. 752–765 (2012). Springer

- [29] Wang, K., Babenko, B., Belongie, S.: End-to-end scene text recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1457–1464 (2011)
- [30] Wang, K., Belongie, S.: Word spotting in the wild. In: Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5–11, 2010, Proceedings, Part I 11, pp. 591–604 (2010). Springer
- [31] Wang, T., Wu, D.J., Coates, A., Ng, A.Y.: End-to-end text recognition with convolutional neural networks. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp. 3304–3308 (2012). IEEE
- [32] Yao, C., Bai, X., Shi, B., Liu, W.: Strokelets: A learned multi-scale representation for scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4042–4049 (2014)
- [33] Lee, C.-Y., Osindero, S.: Recursive recurrent nets with attention modeling for OCR in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2231–2239 (2016)
- [34] Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., Zhou, S.: Focusing attention: Towards accurate text recognition in natural images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5076–5084 (2017)
- [35] Fang, S., Xie, H., Zha, Z.-J., Sun, N., Tan, J., Zhang, Y.: Attention and language ensemble for scene text recognition with convolutional sequence modeling. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 248–256 (2018)
- [36] Luo, C., Jin, L., Sun, Z.: MORAN: A multi-object rectified attention network for scene text recognition. *Pattern Recognition* **90**, 109–118 (2019)
- [37] Shi, B., Wang, X., Lyu, P., Yao, C., Bai, X.: Robust scene text recognition with automatic rectification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4168–4176 (2016)
- [38] Zhan, F., Lu, S.: ESIR: End-to-end scene text recognition via iterative image rectification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2059–2068 (2019)
- [39] Yang, M., Guan, Y., Liao, M., He, X., Bian, K., Bai, S., Yao, C., Bai, X.: Symmetry-constrained rectification network for scene text recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9147–9156 (2019)
- [40] Yang, X., He, D., Zhou, Z., Kifer, D., Giles, C.L.: Learning to read irregular text with attention mechanisms. In: Proceedings of the 27th International Joint Conferences on Artificial Intelligence, vol. 1, p. 3 (2017)
- [41] Wang, T., Zhu, Y., Jin, L., Luo, C., Chen, X., Wu, Y., Wang, Q., Cai, M.: Decoupled attention network for text recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 12216–12224 (2020)
- [42] Qiao, Z., Qin, X., Zhou, Y., Yang, F., Wang, W.: Gaussian constrained attention network for scene text recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 3328–3335 (2021). IEEE
- [43] Qiao, Z., Zhou, Y., Yang, D., Zhou, Y., Wang, W.: SEED: Semantics enhanced encoder-decoder framework for scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13528–13537 (2020)
- [44] Litman, R., Anshel, O., Tsiper, S., Litman, R., Mazor, S., Manmatha, R.: SCATTER: Selective context attentional scene text recognizer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

- Recognition, pp. 11962–11972 (2020)
- [45] Zheng, Y., Qin, W., Wijaya, D., Betke, M.: Lal: Linguistically aware learning for scene text recognition. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 4051–4059 (2020)
 - [46] Tan, Y.L., Kong, A.W.-K., Kim, J.-J.: Pure transformer with integrated experts for scene text recognition. In: European Conference on Computer Vision, pp. 481–497 (2022)
 - [47] Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision* **116**(1), 1–20 (2016)
 - [48] Atienza, R.: Vision transformer for fast and efficient scene text recognition. In: International Conference on Document Analysis and Recognition, pp. 319–334 (2021)
 - [49] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., *et al.*: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
 - [50] Chao, L., Chen, J., Chu, W.: Variational connectionist temporal classification. In: Proceedings of the European Conference on Computer Vision, pp. 460–476 (2020)
 - [51] Feng, W., He, W., Yin, F., Zhang, X.-Y., Liu, C.-L.: Textdragon: An end-to-end framework for arbitrary shaped text spotting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9076–9085 (2019)
 - [52] Hu, W., Cai, X., Hou, J., Yi, S., Lin, Z.: GTC: Guided training of CTC towards efficient and accurate scene text recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11005–11012 (2020)
 - [53] Guan, T., Gu, C., Tu, J., Yang, X., Feng, Q.: A glyph-driven topology enhancement network for scene text recognition. *arXiv preprint arXiv:2203.03382* (2022)
 - [54] Ghazvininejad, M., Levy, O., Liu, Y., Zettlemoyer, L.: Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324* (2019)
 - [55] Gu, J., Bradbury, J., Xiong, C., Li, V.O., Socher, R.: Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281* (2017)
 - [56] Qian, L., Zhou, H., Bao, Y., Wang, M., Qiu, L., Zhang, W., Yu, Y., Li, L.: Glancing transformer for non-autoregressive neural machine translation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, pp. 1993–2003 (2021)
 - [57] Chan, W., Saharia, C., Hinton, G., Norouzi, M., Jaitly, N.: Imputer: Sequence modelling via imputation and dynamic programming. In: International Conference on Machine Learning, pp. 1403–1413 (2020). PMLR
 - [58] Tian, Z., Yi, J., Tao, J., Bai, Y., Zhang, S., Wen, Z.: Spike-triggered non-autoregressive transformer for end-to-end speech recognition. *arXiv preprint arXiv:2005.07903* (2020)
 - [59] Chi, E.A., Salazar, J., Kirchhoff, K.: Align-refine: Non-autoregressive speech recognition via iterative realignment. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1920–1927 (2021)
 - [60] Guo, L., Liu, J., Zhu, X., He, X., Jiang, J., Lu, H.: Non-autoregressive image captioning with counterfactuals-critical multi-agent learning. *arXiv preprint arXiv:2005.04690* (2020)
 - [61] Da, C., Wang, P., Yao, C.: Levenshtein OCR. In: Proceedings of the European Conference on Computer Vision, pp. 322–338

- (2022)
- [62] Wang, Y., Xie, H., Fang, S., Wang, J., Zhu, S., Zhang, Y.: 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, pp. 14194–14203 (2021)
 - [63] Wang, P., Da, C., Yao, C.: Multi-granularity prediction for scene text recognition. In: Proceedings of the European Conference on Computer Vision, pp. 339–355 (2022)
 - [64] Croitoru, F.-A., Hondru, V., Ionescu, R.T., Shah, M.: Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(9), 10850–10869 (2023)
 - [65] Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Shao, Y., Zhang, W., Cui, B., Yang, M.-H.: Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796* (2022)
 - [66] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* **33**, 6840–6851 (2020)
 - [67] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations, pp. 1–20 (2021)
 - [68] Zhang, Q., Tao, M., Chen, Y.: gDDIM: Generalized denoising diffusion implicit models. In: International Conference on Learning Representations, pp. 1–31 (2023)
 - [69] Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In: International Conference on Machine Learning, vol. 162, pp. 16784–16804 (2022)
 - [70] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022)
 - [71] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., *et al.*: Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* **35**, 36479–36494 (2022)
 - [72] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. *arXiv preprint arXiv:2204.03458* (2022)
 - [73] Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., *et al.*: Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022)
 - [74] Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B.: Diffwave: A versatile diffusion model for audio synthesis. In: International Conference on Learning Representations, pp. 1–17 (2021)
 - [75] Xu, J., Wang, X., Cheng, W., Cao, Y.-P., Shan, Y., Qie, X., Gao, S.: Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20908–20918 (2023)
 - [76] Li, X., Thickstun, J., Gulrajani, I., Liang, P.S., Hashimoto, T.B.: Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems* **35**, 4328–4343 (2022)
 - [77] Gong, S., Li, M., Feng, J., Wu, Z., Kong, L.: DiffuSeq: Sequence to sequence text generation with diffusion models. In: International Conference on Learning Representations, pp. 1–20 (2023)
 - [78] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium

- thermodynamics. In: International Conference on Machine Learning, pp. 2256–2265 (2015)
- [79] Esser, P., Rombach, R., Blattmann, A., Ommer, B.: ImageBART: Bidirectional context with multinomial diffusion for autoregressive image synthesis. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W. (eds.) *Advances in Neural Information Processing Systems*, vol. 34, pp. 3518–3532 (2021)
- [80] Razavi, A., Oord, A., Vinyals, O.: Generating diverse high-fidelity images with VQ-VAE-2. In: *Neural Information Processing Systems*, pp. 14837–14847 (2019)
- [81] Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10696–10706 (2022)
- [82] Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., Welling, M.: Argmax flows and multinomial diffusion: Learning categorical distributions. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W. (eds.) *Advances in Neural Information Processing Systems*, vol. 34, pp. 12454–12465 (2021)
- [83] Austin, J., Johnson, D.D., Ho, J., Tarlow, D., Van Den Berg, R.: Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems* **34**, 17981–17993 (2021)
- [84] Zhu, Z., Wei, Y., Wang, J., Gan, Z., Zhang, Z., Wang, L., Hua, G., Wang, L., Liu, Z., Hu, H.: Exploring discrete diffusion models for image captioning. *arXiv preprint arXiv:2211.11694* (2022)
- [85] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)
- [86] Kenton, J.D.M.-W.C., Toutanova, L.K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186 (2019)
- [87] Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. *arXiv preprint arXiv:1607.06450* (2016)
- [88] Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2315–2324 (2016)
- [89] Jiang, Q., Wang, J., Peng, D., Liu, C., Jin, L.: Revisiting scene text recognition: A data perspective. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20543–20554 (2023)
- [90] Sun, Y., Ni, Z., Chng, C.-K., Liu, Y., Luo, C., Ng, C.C., Han, J., Ding, E., Liu, J., Karatzas, D., *et al.*: ICDAR 2019 competition on large-scale street view text with partial labeling – RRC-LSVT. In: *International Conference on Document Analysis and Recognition*, pp. 1557–1562 (2019)
- [91] Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S.: COCO-Text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140* (2016)
- [92] Zhang, Y., Gueguen, L., Zharkov, I., Zhang, P., Seifert, K., Kadlec, B.: Uber-Text: A large-scale dataset for optical character recognition from street-level imagery. In: *SUNw: Scene Understanding Workshop-CVPR*, p. 5 (2017)
- [93] Xie, X., Fu, L., Zhang, Z., Wang, Z., Bai, X.: Toward understanding WordArt: Corner-guided transformer for scene text recognition. In: *Proceedings of the European Conference on Computer Vision*, pp. 303–321 (2022)

- [94] Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., Bigorda, L.G., Mestre, S.R., Mas, J., Mota, D.F., Almazan, J.A., De Las Heras, L.P.: ICDAR 2013 robust reading competition. In: International Conference on Document Analysis and Recognition, pp. 1484–1493 (2013)
- [95] Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., *et al.*: ICDAR 2015 competition on robust reading. In: International Conference on Document Analysis and Recognition, pp. 1156–1160 (2015)
- [96] Phan, T.Q., Shivakumara, P., Tian, S., Tan, C.L.: Recognizing text with perspective distortion in natural scenes. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 569–576 (2013)
- [97] Risnumawan, A., Shivakumara, P., Chan, C.S., Tan, C.L.: A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications* **41**(18), 8027–8048 (2014)
- [98] Chen, J., Yu, H., Ma, J., Guan, M., Xu, X., Wang, X., Qu, S., Li, B., Xue, X.: Benchmarking chinese text recognition: Datasets, baselines, and an empirical study. *arXiv preprint arXiv:2112.15093* (2021)
- [99] Shi, B., Yao, C., Liao, M., Yang, M., Xu, P., Cui, L., Belongie, S., Lu, S., Bai, X.: ICDAR2017 competition on reading chinese text in the wild (RCTW-17). In: International Conference on Document Analysis and Recognition, vol. 1, pp. 1429–1434 (2017)
- [100] Zhang, R., Zhou, Y., Jiang, Q., Song, Q., Li, N., Zhou, K., Wang, L., Wang, D., Liao, M., Yang, M., *et al.*: ICDAR 2019 robust reading challenge on reading chinese text on signboard. In: International Conference on Document Analysis and Recognition, pp. 1577–1581 (2019)
- [101] Chng, C.K., Liu, Y., Sun, Y., Ng, C.C., Luo, C., Ni, Z., Fang, C., Zhang, S., Han, J., Ding, E., *et al.*: ICDAR2019 robust reading challenge on arbitrary-shaped text (RRC-ArT). In: International Conference on Document Analysis and Recognition, pp. 1571–1576 (2019)
- [102] Yuan, T.-L., Zhu, Z., Xu, K., Li, C.-J., Mu, T.-J., Hu, S.-M.: A large chinese text dataset in the wild. *Journal of Computer Science and Technology* **34**, 509–521 (2019)
- [103] He, M., Liu, Y., Yang, Z., Zhang, S., Luo, C., Gao, F., Zheng, Q., Wang, Y., Zhang, X., Jin, L.: ICPR2018 contest on robust reading for multi-type web images. In: International Conference on Pattern Recognition, pp. 7–12 (2018)
- [104] Zhang, H., Liang, L., Jin, L.: SCUT-HCCDoc: A new benchmark dataset of handwritten chinese text in unconstrained camera-captured documents. *Pattern Recognition* **108**, 107559 (2020)
- [105] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations, pp. 1–15 (2015)
- [106] Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. In: *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006, pp. 369–386 (2019)
- [107] Izmailov, P., Podoprikin, D., Garipov, T., Vetrov, D.P., Wilson, A.G.: Averaging weights leads to wider optima and better generalization. In: *Conference on Uncertainty in Artificial Intelligence* (2018)
- [108] Na, B., Kim, Y., Park, S.: Multi-modal text recognition networks: Interactive enhancements between visual and semantic features. In: *Proceedings of the European Conference on Computer Vision*, pp. 446–463 (2022)
- [109] Merity, S., Xiong, C., Bradbury, J., Socher, R.: Pointer sentinel mixture models. In: *International Conference on Learning Representations* (2017)

- [110] Yue, X., Kuang, Z., Lin, C., Sun, H., Zhang, W.: RobustScanner: Dynamically enhancing positional clues for robust text recognition. In: Proceedings of the European Conference on Computer Vision, pp. 135–151 (2020)
- [111] Lu, N., Yu, W., Qi, X., Chen, Y., Gong, P., Xiao, R., Bai, X.: MASTER: Multi-aspect non-local network for scene text recognition. *Pattern Recognition* **117**, 107980 (2021)
- [112] Chen, J., Li, B., Xue, X.: Scene text telescope: Text-focused scene image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12026–12035 (2021)
- [113] Lyu, P., Zhang, C., Liu, S., Qiao, M., Xu, Y., Wu, L., Yao, K., Han, J., Ding, E., Wang, J.: MaskOCR: Text recognition with masked encoder-decoder pretraining. *arXiv preprint arXiv:2206.00311* (2022)