

# MaskOCR: Text Recognition with Masked Encoder-Decoder Pretraining

Pengyuan Lyu, Chengquan Zhang, Shanshan Liu, Meina Qiao, Yangliu Xu,  
Liang Wu, Kun Yao, Junyu Han, Errui Ding, Jingdong Wang\*

Department of Computer Vision Technology (VIS), Baidu Inc.

{lvpengyuan, zhangchengquan, liushanshan07, qiaomeina, xuyangliu,  
wuliang11, yaokun01, hanjunyu, dingerrui, wangjingdong}@baidu.com

## Abstract

In this paper, we present a model pretraining technique, named MaskOCR, for text recognition. Our text recognition architecture is an encoder-decoder transformer: the encoder extracts the patch-level representations, and the decoder recognizes the text from the representations. Our approach pretrains both the encoder and the decoder in a sequential manner. (i) We pretrain the encoder in a self-supervised manner over a large set of unlabeled real text images. We adopt the masked image modeling approach, which shows the effectiveness for general images, expecting that the representations take on semantics. (ii) We pretrain the decoder over a large set of synthesized text images in a supervised manner and enhance the language modeling capability of the decoder by randomly masking some text image patches occupied by characters input to the encoder and accordingly the representations input to the decoder. Experiments show that the proposed MaskOCR approach achieves superior results on the benchmark datasets, including Chinese and English text images.

## 1 Introduction

Optical character recognition aims to recognize texts within a digital image, e.g., a scanned document, a photo of a document, a scene-photo, and so on. It has wide-range applications, such as visual search, document digitization, and so on. Optical character recognition generally consists of two tasks: text detection, localizing the text region, and text recognition, identifying the text from the localized region, which is the interest of this paper.

There are three main pipelines for text recognition. (i) The character-based pipeline [48; 23; 29] localizes each character, performs character recognition, and then groups them into words. (ii) The word-based pipeline [21] performs word classification directly. (iii) The sequence-based pipeline [42; 43; 50; 12] regards text recognition as a sequence labeling problem, and is adopted by most deep learning methods.

We follow the sequence-based pipeline and adopt an encoder-decoder transformer for text recognition. The encoder is a ViT architecture [11], a sequence of self-attention and FFN blocks, for text image patch representation extraction. The decoder is formed with the DETR-style [3] decoder, a sequence of self-attention, cross-attention and FFN blocks, mapping the patch representations to a text with a expected role of language modeling [27].

Our main work lies in exploring the pretraining technology for text recognition. Our approach, called MaskOCR, pretrains both the encoder and the decoder with the masking strategy in a sequential manner. We follow the self-supervised pretraining framework and adopt a masked image modeling

---

\*Corresponding Author.

approach to pretrain the encoder for semantic patch representation learning. We divide the text image into a set of vertical patches, and randomly mask some patches that may contain a part of some character, or some whole characters. We predict the representations of the masked patches from the visible patches in the representation space learned from the encoder, and map the predicted representations to the masked patch images.

We pretrain the decoder in a supervised manner with the masking strategy for language modeling over synthesized text images. We fix the pretrained encoder and only update the decoder, so that this pretraining task explores the language rule and the encoder is not affected by the synthesized text image style that might be different from the downstream tasks.

We validate the effectiveness of the proposed MaskOCR approach on the benchmark datasets, including Chinese and English text images. The experiments show that our approach achieves superior results over previous text recognition methods.

## 2 Related Work

**Scene text recognition** There are three main kinds of solutions to text recognition: character-based [48; 23; 29], word-based [21], and sequence-based [42; 43; 50; 12]. The character-based methods recognize text image by performing character localization, character classification and character grouping. The word-based methods directly treat each word as a category and recognition text by image classification. The sequence-based methods regard text recognition as a sequence labeling problem. Besides CTC [14] and attention mechanism [43; 34; 50; 12] are used to align the input image patch sequence and the output character sequence. The sequence-based solution has been extensively studied recently because ground-truth labeling is easier and the framework is more flexible. The architecture for the sequence-based solution consists of two main modules: encoder and decoder. The encoder aims to learn semantic representations for text images, and it can be formed using convolutional networks or transformers. The decoder aims to recognize the characters from the representations and map them to the texts.

Different from the previous sequence-based methods [42; 43; 12; 54; 1; 30; 50; 49], which consists of convolutional layers or recurrent layers, our proposed method is a fully transformer network which is simple but more effective.

**Pretraining** Representation pretraining, e.g., supervised or self-supervised pretraining on ImageNet in computer vision and self-supervised pretraining in natural language processing, has shown to be beneficial to downstream tasks. The self-supervised pretraining that is not limited by labeled data has attracted a lot of attention. In [17; 7; 8], the representation pretraining is learned with siamese networks by contrastive learning. In [16; 2; 10], the vision and nlp models are masked autoencoders, which enhance the representation by learning to predict the masked content.

**Encoder pretraining for text recognition** Recently, some pretraining methods were used to learn the representations for text recognition. TrOCR [27] pretraining the recognition model on synthesized data by supervised learning. In [1; 30; 54], self-supervised contrastive pretraining were introduced to learn representations from the input images.

Differently, the encoder pretraining of MaskOCR follows the masked image modeling framework, which is recently studied in computer vision and shows superior performance over contrastive pretraining [6; 16], to learn robust character representations from unlabeled text images.

**Decoder pretraining for text recognition** The attention decoder [43; 26; 50] in some sense is able to learn the text relation from the training data. Besides, the masked language modeling approach, such as BERT [10], is introduced to model the relation of the representations or the relation of the characters that are output from the CTC [14] or attention decoder. In [12], a masked language model is pretrained for correcting the predicted results from the visual model. In [27], the linguistic information are modeling by employing supervised learning on synthesized data.

We also attempt to pretrain the decoder from the synthesized data for enhancing the capability of solving the task: grouping the frame-wise predictions corresponding to the same character region and exploring the linguistic rule by masking some representations output from the encoder. This is different from [27]: the encoder is also pretrained, potentially degrading the encoding quality due to using the synthesized data for pretraining.

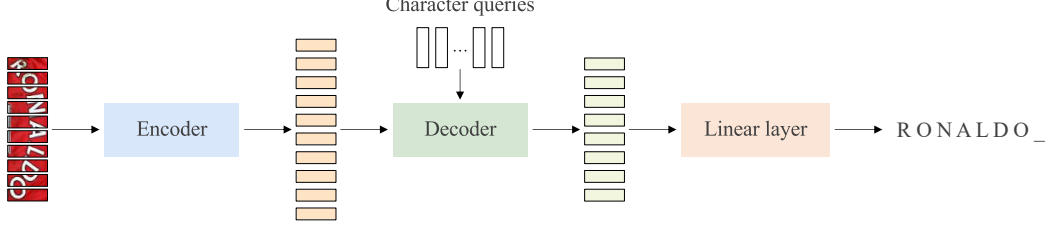


Figure 1: Encode-decoder transformer for text recognition. The encoder extracts a sequence of patch representations, and the decoder maps the patch representations to a sequence of representations, followed by a linear layer to recognize the sequence of characters.

### 3 Approach

#### 3.1 Encoder-Decoder Transformer for Text Recognition

We adopt the encoder-decoder transformer architecture for text recognition. The encoder extracts a sequence of patch representations, and the decoder maps the patch representations to a text. Figure 1 illustrates the encoder-decoder transformer architecture.

**Encoder.** The encoder receives an image,  $I \in \mathbb{R}^{3 \times H \times W}$ , as the input. We partition the image horizontally into a set of  $M$  vertical patches,  $[p_1, p_2, \dots, p_M]$ . The size of each patch is  $H \times W/M$ . We process the flattened patches using linear projection to get patch embeddings. We add the 1D positional embeddings, which is enough as we partition the images vertically. We use the ViT [11], consisting of a sequence of multi-head self-attention and FFN units, as the encoder and learn the patch-level representations,  $\mathbf{F} = [f_1, f_2, \dots, f_M]$ , for the text image.

**Decoder.** We form the text recognition decoder by following the decoder style of the DETR [3] that is original designed for object detection. A decoder is composed of a multi-head self-attention unit, a multi-head cross-attention unit, and an FFN unit.

The decoder transforms the  $N$  input embeddings,  $\mathbf{C} = [c_1, c_2, \dots, c_N]$ , called character queries, into output embeddings, which are then independently mapped into characters through a linear classifier. Multi-head self-attention is performed over the character queries for the first decoder layer, or the output embeddings of the previous decoder layer. Multi-head cross-attention, also called encoder-decoder attention, uses the encoder outputs  $\mathbf{F}$  as the keys and the values and the character queries for the first decoder layer, or the output embeddings of the previous decoder layer as the queries.

The multi-head self-attention, crucial for duplicate detection and non-maximum suppression removal in DETR, plays the role of merging the duplicate predictions from the patches that correspond to the same character region and promoting one prediction and suppressing other duplicate predictions.

Our decoder self-attention is different from masked attention that is widely used in machine translation [46] and also used for text recognition [27]. The decoder based on masked attention models the language in a unidirectional way [27], our decoders models the language in a bidirectional manner. During inference, the decoder based on masked attention predicts the letters one by one sequentially. In contrast, our decoder is more efficient and predicts the characters simultaneously.

**Loss.** We denote the character predictions by  $\mathbf{Y} = [y_1 \ y_2 \ \dots \ y_N]$ . Assuming  $N$  is larger than the number of characters in the text image. We consider the ground truth as  $\mathbf{Y}^* = [y_1^* \ y_2^* \ \dots \ y_N^*]$  padded with an end of sentence symbol [EOS]. The loss function is formulated as follows,

$$\ell(\mathbf{Y}, \mathbf{Y}^*) = \frac{1}{L+1} \sum_{l=1}^{L+1} \text{CE}(y_l, y_l^*), \quad (1)$$

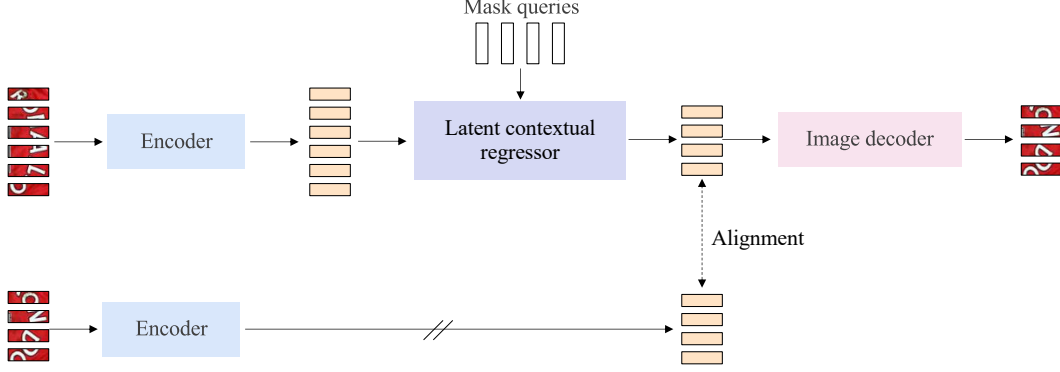


Figure 2: Encoder pretraining architecture. We adopt a masked image modeling approach, CAE [6] to pretrain the encoder for text image representation learning. Slightly different from the CAE, we use the image patches (RGB) as the prediction targets. In this example, six image patches (top) are visible patches, and the other four (bottom) to be predicted are masked patches.

where  $\text{CE}(\cdot, \cdot)$  is the cross-entropy loss.  $L$  is the number of characters in the text image. To balance the number of [EOS] and other characters, we only employ loss function on the characters as well as the first [EOS].

### 3.2 Masked Encoder Pretraining

We follow the masked image modeling framework, which is recently studied for general image representation pretraining, and adopt the context autoencoder-style method [6] to pretrain the encoder for text image representation learning.

The encoder pretraining process is given as follows. The text image is divided into a set of visible image patches and a set of masked image patches. The visible image bands are sent to the encoder, generating the representations of visible patches. Then, the representations of visible patches are fed into a latent contextual regressor with mask queries, predicting the representations for masked patches  $\mathbf{Z}_m$  which is expected to be close to the representations  $\mathbf{Z}_m^*$  of masked patches directly computed from the encoder. Last, the representations of masked patches go into the image decoder, predicting the targets  $\mathbf{T}_m$ . Figure 2 illustrates the encoder pretraining architecture.

We adopt the patch partitioning strategy, same with encoder-decoder transformer for text recognition, for encoder pretraining, and partition the text image into a sequence of  $M$  vertical patches with the size being  $H \times W/M$ . The number  $M$  of patches is usually larger than the number of characters contained in the text image, thus a patch is only a portion of a character. When many patches are masked, there are two cases: a part for some characters is masked, or some characters are completely masked.

Considering that the DALL-E [40] tokenizer used in CAE [6], that is learned over the general images, and we handle special images: text images, we use the patch RGB values, processed with layer normalization (Gaussian normalization), to form the targets. The loss function for encoder pretraining is a combination of representation alignment loss and prediction loss, and given as follows,

$$\ell_t(\mathbf{T}_m, \mathbf{T}_m^*) + \lambda \ell_z(\mathbf{Z}_m, \mathbf{Z}_m^*). \quad (2)$$

Here, both losses  $\ell_t(\cdot, \cdot)$  and  $\ell_z(\cdot, \cdot)$  are the MSE loss.  $\lambda$  is the tradeoff parameter, and set to be 0.05 in our implementation.

There are two main kinds of prediction tasks. One is to predict a part of a character from a visible part of the character and other characters. The other one is to predict some whole character from other characters. We assume that the predictions are more reliable if the representations for the patches take on semantics.

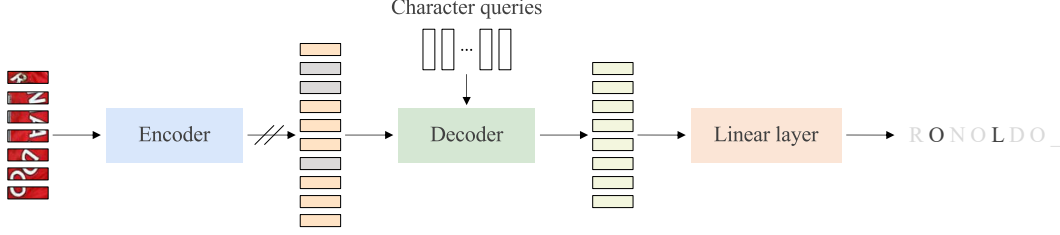


Figure 3: Decoder pretraining architecture. The whole pipeline is similar with the one in Figure 1. The difference is that the input to the encoder are the visible patches. The visible patches are formed by masking the image patches that correspond to the target characters (the patch number may be greater than the character number). The input representations to the decoder are a combination of encoded representations and zero representations added to the positions (gray boxes) where the masked patches are. The prediction targets are the characters that are masked.

### 3.3 Masked Decoder Pretraining

We pretrain the decoder on synthesized text images using the supervision loss. Considering the style of synthesized text images might be different from the real text images, we keep the encoder (learned from encoder pretraining) not updated and only optimize the decoder, so that this pretraining stage does not influence the representation quality.

In addition to removing duplicate predictions, the decoder is capable of modeling the language for better text recognition. To further enhance the capability of language modeling, we adopt the idea of masked language modeling and introduce a masked decoder pretraining scheme. We randomly mask some characters and accordingly the patches, and send the remaining visible patches to the encoder, obtain the representations of visible patches,  $\mathbf{F}_v$ . Then, we insert the zero representations  $\mathbf{F}_m$  to the positions corresponding to the masked patches, and feed the combined representations  $\mathbf{F} = [\mathbf{F}_v \mathbf{F}_m]$  with the corresponding positional embeddings added into the decoder, predicting the text:  $\tilde{\mathbf{Y}} = [\tilde{y}_1 \tilde{y}_2 \dots \tilde{y}_N]$ . The loss function is similar to BERT and is merely about masked characters,

$$\ell(\tilde{\mathbf{Y}}_l, \mathbf{Y}_l^*) = \frac{1}{L} \sum_{l=1}^L \text{CE}(\mathbf{y}_{n_l}, \mathbf{y}_{n_l}^*), \quad (3)$$

where  $L$  is the number of masked characters, and  $\{n_1, n_2, \dots, n_L\}$  are the positions of the masked characters.

## 4 Experiments

### 4.1 Datasets

**Chinese text line images.** The pretraining set consists of 100 million unlabeled text line images collected from practical scenarios for encoder pretraining, and 100 million synthetic text line images for decoder pretraining. The real images are collected from document and street view, and the text in them are almost in Chinese. We collect text corpus from Chinese corpus<sup>2</sup>, and generate 100 million images with 64 commonly used fonts using Text Render<sup>3</sup>. Specifically, for each synthetic sample, the text transcription as well as the character bounding boxes are given.

We first pretrain the encoder and decoder serially on the collected real images and the synthetic images, and then evaluate our model on a large-scale Chinese text image benchmark BCTR [5]. BCTR consists of four subsets (scene, web, document, and handwriting) and provides 1.4 million fully labeled images in total. The scene subset is derived from some scene text datasets, including RCTW [44], ReCTS [53], LSVT [45], ArT [9], and CTW [51], resulting in 636,455 images. The web subset is constructed based on the MTWI [18] dataset and contains 140589 text images. The

<sup>2</sup><https://github.com/crownpku/awesome-chinese-nlp>

<sup>3</sup>[https://github.com/oh-my-ocr/text\\_renderer](https://github.com/oh-my-ocr/text_renderer)

document subset is composed of 500000 synthetic text images generated by Text Render in document style. The handwriting subset is collected from a handwriting dataset SCUT-HCCDoc [52], and 116643 text images are included.

**English text word images.** We collect about 4.2 million unlabeled English word images from COCO-Text [47] and Book32 [19] for encoder pretraining. To be specific, we obtain about 0.5 million cropped images from COCO-Text with the official annotations. Besides, about 3.7 million cropped images are collected from Book32 with the predicted bounding boxes yielded by DB<sup>4</sup> [28]. In addition, we also synthesize 100 million English word images for decoder pretraining. Similarly, we collect corpus from WikiText103 [35] and generate synthetic images with Text Render and 10 commonly used English fonts.

Following [43; 50; 12; 49; 54], two synthetic datasets MJSynth [20] and SynthText [15] are used for the training of downstream recognition task. We evaluate our model on six public scene text datasets: ICDAR 2013 (IC13) [25], Street View Text (SVT) [48], IIIT5K-Words (IIIT5K) [36]), ICDAR 2015 (IC15) [24], Street View Text-Perspective (SVTP) [37], and CUTE80 (CUTE) [41]). The samples in the first three datasets are all regular text images and the remaining datasets may contain perspective or curved text images.

## 4.2 Implementation Details

**Encoder-decoder transformer.** The image patches are fed into a linear projection layer, and then sent to the ViT. Three ViT structures are studied: ViT-S (12 transformer blocks with dimension 384), ViT-B (12 transformer blocks with dimension 768), and ViT-L (24 transformer blocks with dimension 1024). The decoder consists of four decoder layers, each of which includes a self-attention unit, a cross-attention unit, and an FFN unit. Each attention module is a 12-head attention with dimension 384.

We train the encoder-decoder transformer with AdamW optimizer [32], cosine learning rate decay [31], a weight decay of 0.05, a drop path ratio of 0.1 and a batch size of 512. When the model is trained from scratch, the learning rate is set to  $1e - 3$ . Otherwise, the model is optimized with a initial learning rate of  $1e - 4$ . We set the training epochs as 120 and 20 for the Chinese text line recognition model and the English word recognition model with a warm-up of 5 epochs and 0.5 epoch respectively.

**Masked encoder pretraining.** The CAE regressor consists of four regressor layers. Each layer includes a cross-attention unit, where the queries are learnable mask queries for the first layer and the output embeddings of the previous layer for other layers, and an FFN unit. The CAE decoder consists of four layers, and each layer includes a self-attention unit and an FFN unit. Each attention module is also a 12-head attention with dimension 384. Following [16], we use the normalized pixel values of each masked patch as task.

We optimize the model with AdamW optimizer and set learning rate with the linear learning rate scaling rule [13]:  $lr = base\_lr \times batchsize/256$ . By default, the  $base\_lr$  is set to  $1.5e - 4$  with cosine learning rate decay and a 0.5 epoch warm-up. We train the encoder for 10 epochs with the batch size being 4096 for ViT-S and ViT-B, and 2048 for ViT-L due to the memory limitation of GPU.

**Masked decoder pretraining.** We mask some characters and accordingly mask the patches that contain the characters. This might lead to that a different number of patches are masked for different text images as one character may correspond to a different number of patches. We adopt masked attention to replace the original attention in the encoder with the parameters unchanged.

We pretrain the decoder for 5 epochs with a batch size of 512, a initial learning rate of  $1e - 4$ , a 0.5 epoch warmup and a cosine learning rate decay.

**Data preprocessing.** Since the Chinese text line images vary greatly in width, we resize the height of the input image to 32 with the aspect ratio kept and pad the width of the input images to 400. For the English word samples, we directly resize all input images to  $32 \times 128$ . We set the width of the split vertical patch to 4 for all datasets by default. During the training of downstream recognition, some data augmentations like rotation, distortion, and colorJitter are also used.

---

<sup>4</sup><https://github.com/MhLiao/DB>

Table 1: Ablation about encoder pretraining. We report the text recognition results for three cases: randomly initialized encoder, pretrained encoder fixed during downstream task training, and pretrained encoder fine-tuned during downstream task training. One can see that the score is also good when the pretrained encoder is fixed, indicating that the pretrained encoder is already able to learn the semantics. Fine-tuning pretrained encoder outperforms fixing pretrained encoder, implying that the adaptation to downstream tasks, exploiting the domain knowledge, is helpful. The superiority over random encoder demonstrates the effectiveness of encoder pretraining.

	Scene	Web	Document	Handwriting	Avg
Training from random encoder	68.8	70.7	98.6	49.4	75.8
Fixing pretrained encoder	62.8	67.5	98.1	54.3	73.6
Fine-tuning pretrained encoder	<b>72.3</b>	<b>73.7</b>	<b>99.2</b>	<b>62.5</b>	<b>79.8</b>

Table 2: Ablation about masking ratio in encoder pretraining. In our experiments, we choose 0.45 as the masking ratio.

Masking ratio	Scene	Web	Document	Handwriting	Avg
0.30	71.5	73.1	99.1	61.8	79.3
0.45	<b>72.3</b>	<b>73.7</b>	<b>99.2</b>	<b>62.5</b>	<b>79.8</b>
0.60	72.0	73.6	99.1	60.7	79.4

### 4.3 Evaluation

We evaluate BCTR by first processing the predictions and ground truth with the following rules as [5]: (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. After that, we compute the accuracy in sentence level over each subsets and the whole dataset (avg).

To evaluate the six scene English text datasets, we follow [43; 50; 12; 49; 54] and evaluate the recognition performance of our model with the case-insensitive word accuracy. We also report the average accuracy (avg) over both datasets.

### 4.4 Ablation Studies

In this section, we conduct ablation studies on BCTR dataset to verify the effectiveness of pretraining. All experiments are conducted on 8 A100 GPUs with the ViT-B as encoder.

**Encoder pretraining.** We compare the results of downstream tasks for three cases: (i) randomly initialized encoder, (ii) pretrained encoder fixed during downstream task training, and (iii) pretrained encoder fine-tuned during downstream task training. The results are given in Table 1. We use linear probing (ii) to evaluate the pretraining quality of representation learning by fixing the pretrained encoder and only update the rest modules. The acceptable result in Table 1 shows that the pretrained encoder can learn meaningful representations from the input image. We also fine-tune the pretrained encoder during the downstream tasking, and the accuracy is improved by 4.0% over training from scratch, also showing the effectiveness of the encoder pretraining.

We also explore the effect of different masking ratios: 0.30, 0.45, and 0.60. The results are shown in Table 2. The masking ratio of 0.45 works better in downstream recognition task. We observe that our optimal masking ratio is not so high as [16] (0.75), perhaps because the text images have higher information density.

**Decoder pretraining.** We report the results for: (i) randomly initialized decoder with fixing pretrained encoder, (ii) pretrained encoder and decoder fixed during downstream task training (only fine-tuning the linear classifiers), (iii) pretrained decoder fine-tuned only during downstream task training, and (iv) fine-tuning both the encoder and the decoder. The results are given in Table 3. When conducting linear probing (ii), our method achieves the accuracy of 47.8%, showing that our pretraining can learn useful representations for the downstream recognition task. We also finetune the decoder with the encoder fixed and fine-tuned, and 0.8% (73.6% v.s. 74.4%) and 1.0% (79.8% v.s.

Table 3: Ablation about decoder pretraining. We consider four cases: randomly initialized decoder, pretrained decoder fixed during downstream task training (only fine-tuning the linear classifier), pretrained decoder fine-tuned during downstream task training, and fine-tuning both the encoder and the decoder. The first three cases do not fine-tune the pretrained encoder. For the case of fixing pretrained decoder, we only fine-tune the linear classifier. The result (47.8%) shows that the pretrained decoder has certain capability about decoding. The reason that the result is not high is that the dataset for pretraining the decoder is synthesized and there is a domain difference from the downstream dataset, BCTR.

	Scene	Web	Document	Handwriting	Avg
Training from random decoder	62.8	67.5	98.1	54.3	73.6
Fixing pretrained decoder	39.1	54.7	66.3	28.1	47.8
Fine-tuning pretrained decoder	63.9	67.8	98.4	55.4	74.4
Fine-tuning both pretrained encoder and decoder	<b>73.9</b>	<b>74.8</b>	<b>99.3</b>	<b>63.7</b>	<b>80.8</b>

Table 4: Ablation about the masking strategy in decoder pretraining. It can be seen that masking helps the performance, especially for scene.

	Scene	Web	Document	Handwriting	Avg
w/o masking	73.1	74.6	<b>99.3</b>	63.6	80.5
w/ masking	<b>73.9</b>	<b>74.8</b>	<b>99.3</b>	<b>63.7</b>	<b>80.8</b>

Table 5: Studying if the pretrained encoder is simultaneously retrained during decoder pretraining. The performance for encoder retraining is inferior. The reason is that the style of synthesized text image is different from the downstream text images.

	Scene	Web	Document	Handwriting	Avg
Retraining encoder	69.0	71.4	99.0	53.3	76.7
Fixing encoder	<b>73.9</b>	<b>74.8</b>	<b>99.3</b>	<b>63.7</b>	<b>80.8</b>

Table 6: Ablation about the vertical patch size.

Patch size	Scene	Web	Document	Handwriting	Avg
$32 \times 4$	<b>68.8</b>	<b>70.7</b>	<b>98.6</b>	<b>49.4</b>	<b>75.8</b>
$32 \times 8$	64.0	67.3	97.5	43.3	72.2

80.8%) improvements are obtained respectively, further validating the effectiveness of our decoder pretraining.

We report the results of decoder pretraining for two cases in Table 4: masking patches and not masking patches. Compared to the model of without decoder pretraining, the two cases of decoder pretraining contribute an improvement of 0.7% and 1.0% respectively. Besides, the better results from masking patches suggest that the ability of language modeling is enhanced.

We also study the performance for the case of retraining the encoder from the pretrained weights when pretraining the decoder. The results are shown in Table 5. The accuracy of 76.7% is achieved, which is better than the model from scratch (76.7% *v.s.* 75.8%). However, the accuracy drops by 4.1% when compared with our serially pretraining (76.7% *v.s.* 80.8%), which indicates that the pretrained encoder is affected by the synthesized text images.

**Vertical patch size.** We consider two kinds of patch sizes and evaluate the performance over the BCTR dataset without pretraining. The results are shown in Table 6. The larger patch size achieves worse performance and drops the accuracy by 3.6%, probably due to the embedding token of larger patch size has higher information density, which is more difficult to learn.



Table 7: Text recognition results on the BCTR dataset.

Methods	Pretraining	Scene	Web	Document	Handwriting	Avg	#Params
CRNN [42]	×	53.4	54.5	97.5	46.4	67.0	-
ASTER [43]	×	54.5	52.3	93.1	38.9	64.7	-
MORAN [33]	×	51.8	49.9	95.8	39.7	64.3	-
SAR [26]	×	62.5	54.3	93.8	31.4	67.3	-
SRN [50]	×	60.1	52.3	96.7	18.0	65.0	-
SEED [39]	×	49.6	46.3	93.7	32.1	61.2	-
TransOCR[4]	×	63.3	62.3	96.9	53.4	72.8	84M
MaskOCR (ours, ViT-S)	✓	71.4	72.5	98.8	55.6	78.1	36M
MaskOCR (ours, ViT-B)	✓	73.9	74.8	99.3	63.7	80.8	100M
MaskOCR (ours, ViT-L)	✓	<b>76.2</b>	<b>76.8</b>	<b>99.4</b>	<b>67.9</b>	<b>82.6</b>	318M

Table 8: Text recognition results on six English scene text datasets.

Methods	Pretraining	IC13	SVT	IIIT5K	IC15	SVTP	CUTE	Avg	#Params
ASTER [43]	×	91.8	89.5	93.4	76.1	78.5	79.5	86.7	-
SEED [39]	×	92.8	89.6	93.8	80.0	81.4	83.6	88.3	-
PIMNet [38]	×	95.2	91.2	95.2	83.5	84.3	84.4	90.5	-
SRN [50]	×	95.5	91.5	94.8	82.7	85.1	87.8	90.4	55M
VisionLan [49]	×	95.7	91.7	95.8	83.7	86.0	88.5	91.2	33M
ABINet [12]	✓	97.4	93.5	96.2	86.0	89.3	89.2	92.7	37M
ConCLR [54]	✓	97.7	94.3	<b>96.5</b>	85.4	89.3	91.3	92.8	37M
PerSec [30]	✓	97.2	94.6	96.3	84.4	89.5	90.2	92.4	-
MaskOCR (ours, ViT-S)	✓	97.1	<b>95.7</b>	95.5	87.0	90.1	90.3	92.9	31M
MaskOCR (ours, ViT-B)	✓	<b>98.1</b>	94.7	95.8	87.3	89.9	89.2	93.1	97M
MaskOCR (ours, ViT-L)	✓	97.8	94.1	<b>96.5</b>	<b>88.7</b>	<b>90.2</b>	<b>92.7</b>	<b>93.8</b>	315M

#### 4.5 Comparison with State-of-the-art Methods

**Chinese Text Line Recognition.** We evaluate the ability of our model to recognize Chinese text line on the BCTR dataset. We set the number of input embeddings  $N$  to 40, since most of the samples in BCTR have less than 40 characters. We show the results of our method and representative existing methods on the BCTR dataset in Table 7. When training from scratch, our method with ViT-B as encoder outperforms all the previous methods which with the similar model size. Specifically, our method is better than the previous best method TransOCR [4] by 3.0% (72.8% *v.s.* 75.8%). When training with pretrained encoder and decoder, our models outperforms all the previous methods and surpass the previous best results by large margins. In detail, our method shows steady improvement with the increase of the model size, and improve over the state-of-the-art by 5.3%, 8.0%, and 9.8% respectively.

**English scene text recognition.** Following [42; 43], we set the number of input embeddings  $N$  to 25 which exceeds the lengths of most English words. Since scene text appeared in natural scenes always with distortions or irregular layout, we employ a spatial transformer network [22] which is adopted in [43] to rectify the input image and train it with our recognizer jointly. We report the results of our method and representative existing methods on English Scene text datasets in Table 8. Our small model achieves the best performance over the previous counterparts with the similar model size. Specifically, our method achieves better accuracy than PerSec [30] which pretrained with 100 million real data, while we only use 4.2 million real data for pretraining. ABINet [12] and its extension ConCLR [54], performing similarly to our approach with small ViT, use an explicit pretrained masked language model to correct the predicted result iteratively, which is complementary to our approach and may benefit our approach. Our method benefit from bigger models: ours large model pushes the SOTA result to a new level: 93.8% .

## 5 Conclusion

The core of the proposed approach, MaskOCR, for text recognition lies in that we pretrain the recognition model, including both the encoder and the decoder. The encoder pretraining is able to benefit from large scale real text images that are easily available without the need of text annotation. The decoder pretraining is able to benefit from the synthetic text images that are also easily available with the character-level annotation easily obtained. Experiments verify the effectiveness of encoder-decoder pretraining.

## References

- [1] Aviad Aberdam, Ron Litman, Shahar Tsiper, Oron Anschel, Ron Slossberg, Shai Mazor, R. Manmatha, and Pietro Perona. Sequence-to-sequence contrastive learning for text recognition. In *CVPR*, pages 15302–15312, 2021.
- [2] Hangbo Bao, Li Dong, and Furu Wei. Beit: BERT pre-training of image transformers. *CoRR*, abs/2106.08254, 2021.
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020.
- [4] Jingye Chen, Bin Li, and Xiangyang Xue. Scene text telescope: Text-focused scene image super-resolution. In *CVPR*, pages 12026–12035, 2021.
- [5] Jingye Chen, Haiyang Yu, Jianqi Ma, Mengnan Guan, Xixi Xu, Xiaocong Wang, Shaobo Qu, Bin Li, and Xiangyang Xue. Benchmarking chinese text recognition: Datasets, baselines, and an empirical study. *CoRR*, abs/2112.15093, 2021.
- [6] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *CoRR*, abs/2202.03026, 2022.
- [7] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020.
- [8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pages 9620–9629, 2021.
- [9] Chee Kheng Chng, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, Chee Seng Chan, Lianwen Jin, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, and Junyu Han. ICDAR2019 robust reading challenge on arbitrary-shaped text - rrc-art. In *ICDAR*, pages 1571–1576, 2019.
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019.
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.
- [12] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *CVPR*, pages 7098–7107, 2021.
- [13] Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: training imagenet in 1 hour. *CoRR*, abs/1706.02677, 2017.

- [14] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, pages 369–376, 2006.
- [15] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *CVPR*, pages 2315–2324, 2016.
- [16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021.
- [17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735, 2020.
- [18] Mengchao He, Yuliang Liu, Zhibo Yang, Sheng Zhang, Canjie Luo, Feiyu Gao, Qi Zheng, Yongpan Wang, Xin Zhang, and Lianwen Jin. ICPR2018 contest on robust reading for multi-type web images. In *ICPR*, pages 7–12, 2018.
- [19] Brian Kenji Iwana and Seiichi Uchida. Judging a book by its cover. *CoRR*, abs/1610.09204, 2016.
- [20] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *CoRR*, abs/1406.2227, 2014.
- [21] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *IJCV*, pages 1–20, 2016.
- [22] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NeurIPS*, pages 2017–2025, 2015.
- [23] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Deep features for text spotting. In *ECCV*, pages 512–528, 2014.
- [24] Dimosthenis Karatzas, Lluís Gomez-Bigorda, Angelos Nicolaou, Suman K. Ghosh, Andrew D. Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. ICDAR 2015 competition on robust reading. In *ICDAR*, pages 1156–1160, 2015.
- [25] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluís Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernández Mota, Jon Almazán, and Lluís-Pere de las Heras. *Icdar*. pages 1484–1493, 2013.
- [26] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. Show, attend and read: A simple and strong baseline for irregular text recognition. In *AAAI*, pages 8610–8617, 2019.
- [27] Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. *CoRR*, abs/2109.10282, 2021.
- [28] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *AAAI*, pages 11474–11481, 2020.
- [29] Minghui Liao, Jian Zhang, Zhaoyi Wan, Fengming Xie, Jiajun Liang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Scene text recognition from two-dimensional perspective. In *AAAI*, pages 8714–8721, 2019.
- [30] Hao Liu, Bin Wang, Zhimin Bao, Mobai Xue, Sheng Kang, Deqiang Jiang, Yinsong Liu, and Bo Ren. Perceiving stroke-semantic context: Hierarchical contrastive learning for robust scene text recognition. In *AAAI*, 2022.
- [31] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR*, 2017.
- [32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

- [33] Canjie Luo, Lianwen Jin, and Zenghui Sun. MORAN: A multi-object rectified attention network for scene text recognition. *Pattern Recognit.*, pages 109–118, 2019.
- [34] Pengyuan Lyu, Zhicheng Yang, Xinhang Leng, Xiaojun Wu, Ruiyu Li, and Xiaoyong Shen. 2d attentional irregular scene text recognizer. *CoRR*, abs/1906.05708, 2019.
- [35] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *ICLR*, 2017.
- [36] Anand Mishra, Karteek Alahari, and C. V. Jawahar. Scene text recognition using higher order language priors. In Richard Bowden, John P. Collomosse, and Krystian Mikolajczyk, editors, *BMVC*, pages 1–11, 2012.
- [37] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *ICCV*, pages 569–576, 2013.
- [38] Zhi Qiao, Yu Zhou, Jin Wei, Wei Wang, Yuan Zhang, Ning Jiang, Hongbin Wang, and Weiping Wang. Pimnet: A parallel, iterative and mimicking network for scene text recognition. In *ACM MM*, pages 2046–2055, 2021.
- [39] Zhi Qiao, Yu Zhou, Dongbao Yang, Yucan Zhou, and Weiping Wang. SEED: semantics enhanced encoder-decoder framework for scene text recognition. In *CVPR*, pages 13525–13534, 2020.
- [40] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831, 2021.
- [41] Anhar Risnumawan, Palaiahnakote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. *Expert Syst. Appl.*, pages 8027–8048, 2014.
- [42] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *TPAMI*, pages 2298–2304, 2017.
- [43] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. ASTER: an attentional scene text recognizer with flexible rectification. *TPAMI*, pages 2035–2048, 2019.
- [44] Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge J. Belongie, Shijian Lu, and Xiang Bai. ICDAR2017 competition on reading chinese text in the wild (RCTW-17). In *ICDAR*, pages 1429–1434, 2017.
- [45] Yipeng Sun, Dimosthenis Karatzas, Chee Seng Chan, Lianwen Jin, Zihan Ni, Chee Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, and Jingtuo Liu. ICDAR 2019 competition on large-scale street view text with partial labeling - RRC-LSVT. In *ICDAR*, pages 1557–1562, 2019.
- [46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.
- [47] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge J. Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *CoRR*, abs/1601.07140, 2016.
- [48] Kai Wang, Boris Babenko, and Serge J. Belongie. End-to-end scene text recognition. In *ICCV*, pages 1457–1464, 2011.
- [49] Yuxin Wang, Hongtao Xie, Shancheng Fang, Jing Wang, Shenggao Zhu, and Yongdong Zhang. From two to one: A new scene text recognizer with visual language modeling network. In *ICCV*, pages 14174–14183, 2021.

- [50] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Towards accurate scene text recognition with semantic reasoning networks. In CVPR, pages 12110–12119, 2020.
- [51] Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, and Shi-Min Hu. A large chinese text dataset in the wild. JCST, pages 509–521, 2019.
- [52] Hesuo Zhang, Lingyu Liang, and Lianwen Jin. Scut-hccdoc: A new benchmark dataset of handwritten chinese text in unconstrained camera-captured documents. Pattern Recognit., page 107559, 2020.
- [53] Rui Zhang, Mingkun Yang, Xiang Bai, Baoguang Shi, Dimosthenis Karatzas, Shijian Lu, C. V. Jawahar, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, and Minghui Liao. ICDAR 2019 robust reading challenge on reading chinese text on signboard. In ICDAR, pages 1577–1581, 2019.
- [54] Xinyun Zhang, Binwu Zhu, Xufeng Yao, Qi Sun, Ruiyu Li, and Bei Yu. Context-based contrastive learning for scene text recognition. In AAAI, 2022.