Toward Understanding WordArt: Corner-Guided Transformer for Scene Text Recognition

Xudong Xie[1], Ling Fu[1], Zhifei Zhang[2], Zhaowen Wang[2], and Xiang Bai[1](✉)

[1] Huazhong University of Science and Technology, China
{xdxie,ling_fu,xbai}@hust.edu.cn
[2] Adobe Research, USA
{zzhang,zhawang}@adobe.com

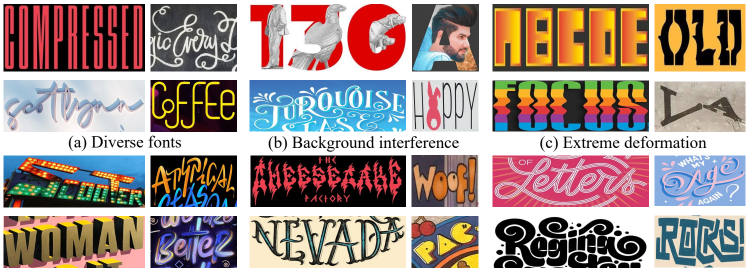
Abstract. Artistic text recognition is an extremely challenging task with a wide range of applications. However, current scene text recognition methods mainly focus on irregular text while have not explored artistic text specifically. The challenges of artistic text recognition include the various appearance with special-designed fonts and effects, the complex connections and overlaps between characters, and the severe interference from background patterns. To alleviate these problems, we propose to recognize the artistic text at three levels. Firstly, corner points are applied to guide the extraction of local features inside characters, considering the robustness of corner structures to appearance and shape. In this way, the discreteness of the corner points cuts off the connection between characters, and the sparsity of them improves the robustness for background interference. Secondly, we design a character contrastive loss to model the character-level feature, improving the feature representation for character classification. Thirdly, we utilize Transformer to learn the global feature on image-level and model the global relationship of the corner points, with the assistance of a corner-query cross-attention mechanism. Besides, we provide an artistic text dataset to benchmark the performance. Experimental results verify the significant superiority of our proposed method on artistic text recognition and also achieve state-of-the-art performance on several blurred and perspective datasets. The dataset and codes are available at: https://github.com/xdxie/WordArt.

Keywords: artistic text recognition · corner point · attention

1 Introduction

The artistic text is a kind of beautified text that is carefully designed by designers or artists. They use various complex fonts of different styles, combining word effects such as shadow, rotation, stereo transformation, deformation, and distortion. Meanwhile, the background patterns and text meaning are considered during the design. Artistic text is widely used in advertisements, slogans, exhibitions, decorations, magazines, and books. Fig. 1 shows some typical artistic text images with several unique properties.

✉ Corresponding author



(a) Diverse fonts (b) Background interference (c) Extreme deformation

(d) Word effect (e) Artistic design (f) Overlapping

Fig. 1. The artistic text examples of different types from the WordArt dataset

In view of this, artistic text recognition is an overlooked and extremely challenging task with importance and practicability in a wide range of applications. Unlike scene text recognition (STR) [39,41,8], artistic text recognition often has several difficulties and challenges: (1) As illustrated in Fig. 1 (a, c, d), the character appearance varies widely due to the different fonts, artistic design effects, and deformation. (2) In Fig. 1 (a, f), there are many complicated connections and overlaps between characters, which makes it difficult to focus on the center or the stroke of a character independently during the recognition process. (3) The design of the artistic text may use background elements to express characters or words and organically combine texts with patterns, causing serious background interference, as shown in Fig. 1 (b, e).

It is difficult for existing scene text recognition methods to be competent for this task. The approaches for regular scene text [12,39,15,5] only focus on horizontal text with standard printing fonts and cannot cope with instances with various shapes, artistic effects, and fonts. Other methods utilizing rectification [40,41,30] for irregular scene text recognition can rectify the text line but not the various character shapes. The existing methods based on the attention mechanism [54,27,26] cannot obtain accurate positions of artistic characters, as shown in Fig. 6. In a sense, irregular texts belong to a subset of artistic texts. In addition, handwriting contains a variety of fonts and ligatures, but the background of these instances is very simple without word effects and artistic designs. Therefore, the methods for handwriting recognition [9,2] fail to handle the artistic text with complex background. Recently, some researchers have introduced linguistic knowledge and corpora to help improve the performance of scene text recognition [36,8]. However, as shown in Fig. 5, the language model is also inefficient for the complex artistic text. Therefore, we need to learn more robust and representative visual features.

Considering these challenges, in order to obtain robust visual features to recognize the artistic text accurately, we propose to model image features at three levels. **(1) Local feature within character.** In the artistic text, the appearance and shape of characters vary widely from instance to instance. It is necessary to build an explicit invariant feature within characters to robustly represent the core key points or structures, suppressing the interference of appearance and deformation. Since the corner points [42,14] of the character strokes and the relative positions between these points are invariant, we use the corner point map as a robust representation of the input image. Moreover, the discreteness of the corner point map cuts off the connection and the overlap between characters, and the sparsity suppresses most of the background interference. In addition, we propose a corner-query cross-attention mechanism, treating the corner point as the query and the image as the key to make the corner seek the image features of interest. In this way, the corner guides the model to pay more accurate attention to the core strokes or character centers of the artistic text. **(2) Character-level feature.** Accurate character recognition is critical for text recognition. The huge visual differences between the same characters of artistic texts lead to the scattered distribution of their features in the feature space. To implicitly learn common representations for each class of characters, it is necessary to make the same class instances cluster together in the feature space and different classes away from each other. Therefore, we introduce a loss function based on the contrastive learning [4,23], significantly improving the clustering degree of their features (Fig. 7). For each character in a minibatch, its positive samples are characters of the same class, and other characters are negative samples. **(3) Global feature on image-level.** Global features of images can assist the overall text recognition because the characters can be reasoned through the visual and semantic information from context. Transformer [45] based on the self-attention mechanism has demonstrated its strong advantages and performance [26,8], benefiting from its global modeling ability. Therefore, to extract the global features of artistic text images with arbitrary shapes and model the global relationship of the corner points, we use Transformer [45] as our backbone and propose the CornerTransformer.

To benchmark the performance of different methods on the artistic text recognition task, we propose a dataset named WordArt. Experimental results show that our method outperforms the existing STR models on this challenging task. CornerTransformer performs well on many artistic texts containing complex fonts, ligatures, and overlaps. Furthermore, we achieve competitive or better results than other methods on common STR benchmarks. In particular, our model outperforms the SOTAs on Street View Text [48], SVT-Perspective [35] and ICDAR 2015 [21] benefits from the corner point map, as gradient-based corner point detection is robust to image resolution, noise and blur.

To summarize, the contributions of this paper are four-fold:

(1) We focus on a new challenging task: artistic text recognition, and propose the WordArt dataset to benchmark the performance.
(2) We notice the importance of the corner point on artistic text recognition and present a corner-query cross-attention mechanism, which allows the model to pay more accurate attention to the core strokes or character centers.
(3) We design a character contrastive loss to cluster the same class of character features, enabling the model to learn unified representations for characters.
(4) Our method significantly outperforms other models on artistic text recognition and also achieves new state-of-the-art results on scene text recognition.

2 Related Work

Scene text recognition. Scene text can be roughly divided into regular and irregular text. The sequence-to-sequence models based on CTC [12,39,15] and attention [5] for regular text recognition have made a great progress. However, these methods fail to cope with curved or rotated text, so irregular text recognition has recently attracted many research interests. The rectification-based methods [40,41,30,53] utilize the spatial transformer network [20] to transform the text image into a canonical shape, but the predefined transformation space limits the generalization of them. The segmentation-based methods [28,47] formulate the recognition task as a character segmentation problem, but character-level annotations are required. In addition, the recent methods with the 2D attention mechanism [54,27,26] also show strong performance on irregular text recognition, and we choose SATRN [26] as the baseline to build our model. Overall, it is difficult to directly apply these methods to artistic text recognition because of the limitations stated in Sec. 1.

Special text recognition. Beyond scene text recognition, other recognition tasks for special text are also significantly important. For example, handwriting recognition [17,60,2] has always been the focus of research given the changeable character shapes and varying writing styles. Another meaningful task that has emerged recently is handwritten mathematical expression recognition [51,25], which has wide applications in education. Manga text recognition [10] is also an interesting problem due to the unconstrained text in the manga. Moreover, Wang *et al.* [50] specifically explore font-independent features of scene texts via a glyph generative adversarial network [11]. Compared with the artistic text, the backgrounds of handwriting and manga images are very simple, and there is no character overlapping, artistic rendering, or word effects. To our knowledge, ours is the first work for artistic text recognition.

Text recognition with auxiliary information. Some segmentation-based methods [28,47] introduce character-level annotations to improve the recognition results. Other recent approaches [36,8] transfer linguistic knowledge to the vision model with a pre-trained language model. Through linguistic information, the model can predict characters according to the context. However, to utilize such information needs to pay the extra cost of data and computing. Besides the deep learning-based methods, other traditional methods explore robust text image presentations, such as SIFT descriptors [35], Strokelets [55], and HOG [43]. Access to such information is automatic and almost cost-free. In this paper, we use the corner point [42,14] to assist the Transformer-based [45,26] method for artistic text recognition.

Text recognition dataset. There exist several standard datasets for the task of scene text recognition. IIIT5k-Words (IIIT5k) [33], ICDAR 2013 (IC13) [22], and Street View Text (SVT) [48] only contain horizontal text with standard fonts. ICDAR 2015 (IC15) [21] contains many small, blurred, and irregular text. SVT-Perspective (SVTP) [35] is built based on the original SVT to evaluate perspective distorted text recognition. CUTE80 (CUTE) [37] and Total-Text [6] mainly focus on curved text. COCO-Text [46] is the first large-scale dataset for text in natural images. Besides, there are some multilingual datasets such as CTW [57], LSVT [44] and MLT [34]. However, most images in these datasets do not contain artistic text. Therefore, we construct a new dataset to benchmark the performance of artistic text recognition.

3 Methodology

3.1 Overview

The overall structure of CornerTransformer is shown in Fig. 2. Given an image $X \in \mathbb{R}^{H\times W\times 3}$, we first utilize a corner point detector to generate a corner point map $M \in \mathbb{R}^{H\times W\times 1}$. Then, X and M are fed into two convolutional layers respectively for producing features of $X^{'} \in \mathbb{R}^{\frac{H}{4}\times \frac{W}{4}\times C}$ and $M^{'} \in \mathbb{R}^{\frac{H}{4}\times \frac{W}{4}\times C}$, where C is the feature dimension. On the one hand, $X^{'}$ will learn the global features $X_g^{'}$ of the image through the multi-head self-attention mechanism. On the other hand, $M^{'}$ will combine with $X_g^{'}$ through the multi-head cross-attention mechanism. Then, the encoder output feature and the character sequence embedding will be fed into the Transformer decoder [45] to generate the feature sequence. Finally, we apply two linear branches to calculate the cross-entropy loss and the character contrastive loss separately.

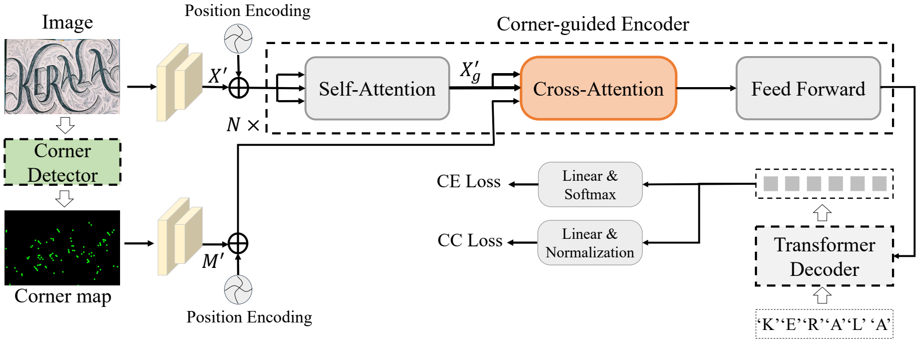


Fig. 2. The overall architecture of CornerTransformer consists of two inputs from different modalities, a corner-guided encoder and a Transformer decoder [45]. CE loss is the cross-entropy loss, and CC loss is our proposed character contrastive loss

3.2 Corner-guided Encoder

In the task of artistic text recognition, the deformation and distortion of characters are extremely diverse due to various fonts and artistic effects. Thus, it is necessary to transform the artistic text image into a more robust representation. As shown in Fig. 3, we observe that, for the great variance in the appearance

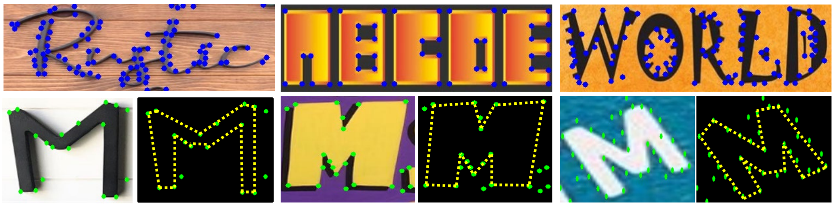


Fig. 3. Visualization of corner point detection. **Top:** The detected corner points of artistic text images. **Bottom:** Corner points of a character "M" with various appearances, whose structural relations are similar

of a specific character, the most critical corners of this character can almost always be detected. The structural relations formed by the connection of these key corners are similar. Moreover, these points are the positions that contain rich visual information of the image. Therefore, we utilize the corner map extracted from the image as an auxiliary input to provide an invariant visual representation. In addition, the connection and the position overlap between characters are extremely complicated, while the discrete corner map can naturally cut off the connection and suppress the overlapping effect of strokes. Furthermore, designers often use some background elements when designing the artistic text to perfectly integrate the artistic text with the background, which causes serious interference from the background during the recognition process. However, the corner map only retains the keypoints of the image, suppressing most background elements and making it easier for the model to focus on important text features.

Given an image, we use a classical corner point detector, Shi-Tomasi corner detector [42], to generate its corner map. This detector improves the stability of Harris detector [14], and can produce high-quality corner points. For each pixel (x, y) in the image, we first calculate the image structure tensor S, then the corner response function is defined as $R = \min(\lambda_1, \lambda_2)$, where λ_1 and λ_2 are the eigenvalues of S. If $R > threshold$, pixel (x, y) is a corner point and the value in position (x, y) of the corner map is 1, otherwise it is 0. Therefore, the corner map is a sparse matrix whose element of value 1 only represents the position information of corners.

After obtaining the corner map, considering that there are local correlations between corners instead of being independent of each other, we first use two convolutional layers to model local relations on the corner map and add 2D position encoding [26] to record the corner position information. A natural method to combine image and corner features is to concatenate them together and feed them into the Transformer encoder. However, this can not make full use of the auxiliary information of corners as shown in Tab. 3. Since the corner map is sparse, the model will still mainly focus on image features. Therefore, we design a corner-guided encoder to fuse corner features at each block. Specifically, we add a multi-head cross-attention layer after the self-attention layer. We utilize the image feature $X_g^{'}$ as the key and value, and the corner feature $M^{'}$ as the query. The corner-query cross-attention mechanism can be formulated as:

$$CA(Q, K, V) = CA(M^{'}, X_g^{'}, X_g^{'}) = softmax(\frac{M^{'} {X_g^{'}}^T}{\sigma})X_g^{'}, \tag{1}$$

where CA means Cross-Attention and σ is a scaling factor. Since corners represent keypoints inside characters, we use the corner map as a query to make the corner seek the image features of interest. Furthermore, the model can pay more accurate attention to the character positions of the artistic text in the image. For instance, for character "A" in a text image, its top corner point tends to focus on other positions of this character rather than other characters. Our ablation study and visualization analysis also prove the effectiveness of the corner-query cross-attention mechanism.

The corner-guided encoder is composed of a stack of N blocks, where each consists of a self-attention layer, a cross-attention layer, and a feed-forward layer. The query of each cross-attention layer is $M^{'}$.

3.3 Character Contrastive loss

Corner-based representation mainly focuses on the local modeling within the character, while Transformer tends to the global modeling of the whole image. To bridge these two representation levels, we introduce a middle-level (character-level) representation learning method. For the artistic text, different instances of the same character show a variety of appearances, including font, shadow, rotation, and other effects. Therefore, in the training process, it is necessary to learn an implicit and unified character-level representation for each character class, so that instances of the same character class are clustered together in the feature space, and features of different classes are far away from each other.

Inspired by the popular thought of contrastive learning [4,23,59], we propose a Character Contrastive loss (CC loss) to achieve our motivation. In short, for a character in a minibatch, the characters of the same class are positive samples, and other characters are negative samples. Specifically, given a minibatch of N images, each image contains variable-length text. We unify the length of text labels to $m = 25$, and there are $N \times m$ characters in a minibatch. For the ith character, x_i is the feature vector and y_i is the class label, where $i \in I \equiv \{1, 2, ..., N \times m\}$. When the ith character is an anchor, its positive set is $P(i) \equiv \{p \in I : y_p = y_i, p \neq i\}$, and the negative set is $N(i) \equiv \{n \in I : y_n \neq y_i, n \neq i\}$. The character contrastive loss can be formulated as:

$$\mathcal{L}_{\mathrm{CC}} = \sum_{i \in I} \frac{-1}{N_p} \sum_{p \in P(i)} log \frac{exp(x_i \cdot x_p/\tau)}{\sum_{s \in P(i)} exp(x_i \cdot x_s/\tau) + \sum_{t \in N(i)} exp(x_i \cdot x_t/\tau)}, \tag{2}$$

where N_p is the number of positive samples, and τ a scaling factor.

Finally, the full optimization objective is defined as:

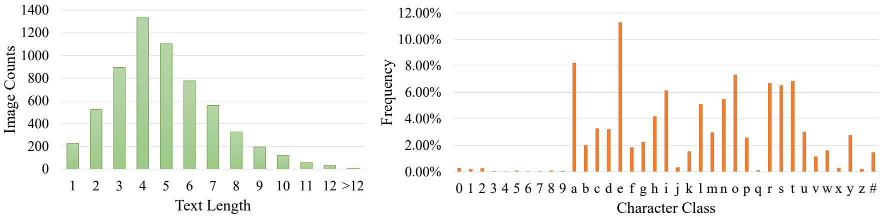
$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda \mathcal{L}_{\mathrm{CC}}, \tag{3}$$

where $\mathcal{L}_{\mathrm{CE}}$ is the cross entropy loss. We set $\lambda = 0.1$ by default.

4 Experiments

4.1 WordArt Dataset

To benchmark the performance of different models on the artistic text recognition task, we collect a dataset of artistic text named WordArt. Thanks to the TextSeg [52] dataset, which contains images of posters, greeting cards, covers, billboards, handwriting, etc. There exist many artistic texts in these images. In view of this, we first crop the word images with the word bounding box annotations and then carefully pick over the artistic text following the definition of the artistic text as stated in Sec. 1. Finally, our WordArt dataset consists of 6316 artistic text images. Following the splitting rule of TextSeg, the training set contains 4805 images, and the testing set contains 1511 images. The statistical analysis is presented in Fig. 4. The distributions of text length and character frequency roughly align with the English corpus. The qualitative presentation of the WordArt dataset is shown in Fig. 1.



(a) Text length distribution (b) Character frequency

Fig. 4. Statistical analysis for the WordArt dataset. (a) The number of images with different text lengths. (b) Frequency distribution of all characters in the whole dataset

4.2 Implementation Details

In our CornerTransformer, the feature dimension of all the attention layers is set to 512, with 8 heads for each layer. We set $N = 12$ for the corner-guided encoder. To calculate the character contrastive loss, we add two linear layers with 2048 hidden dimension and an $L2$ normalization to transform the decoder output features into a normalized feature space. By default, we jointly use CE loss and CC loss to train our model. τ in CC loss is set to 0.1. As common practice [56,1], we use two synthetic datasets MJSynth (MJ) [18,19] and SynthText [13] as training datasets and directly evaluate the performance on the WordArt dataset and STR datasets after training. The input images are resized to 32×128 for both training and testing with batch size 256. The model is trained with Adam optimizer [24] with the initial learning rate $3e^{-4}$. The total epoch is 6, and the learning rate will decay to $3e^{-5}$ after 4 epochs. We adopt several data augmentation strategies such as rotation, Gaussian noise, etc.

4.3 Ablation study

For artistic text recognition, our main contributions to the method are introducing the corner map with designing the corner-query cross-attention and proposing the character contrastive loss. We will verify the effectiveness of each design in detail. Since we need to use the global modeling capabilities of Transformer [45], we choose SATRN [26] as our baseline and reproduce its model by replacing the dimensions the query and key from 128 to 512. To comprehensively evaluate different designs, besides the word accuracy, we present character recall and character precision to assist the evaluation. All the results for the ablation study are evaluated on our WordArt dataset.

The effectiveness of the corner map. Since we add an attention module to fuse the corner map, extra parameters are introduced, increasing the capacity of the model. In order to verify that the performance improvement comes from the role of the corner map rather than the extra parameters, we replace the input of the corner branch in Fig. 2 with the same image as the main branch. The results shown in the third row of Tab. 1 are lower than baseline. We attribute this to the large amount of noise and redundant information contained in the image, which adds false guidance to the model when directly applying it as a query. Besides, we also remove the extra input branch but retain the added attention module. In this case, the cross-attention turns into self-attention. As shown in the fourth row of Tab. 1, this only gives a slight boost to the baseline results. Therefore, the role of the corner map is crucial for artistic text recognition, indicating keypoints and filtering out noise.

Table 1. Ablation study about the effectiveness of the corner map

Input	word acc	char recall	char precision
Baseline (Self-attn)	67.0	84.6	84.2
Corner+Image	**69.1**	**85.7**	**84.8**
Image+Image	66.0	83.8	83.3
Self-attn ×2	67.6	85.2	83.3

Different corner detectors. In view of the importance of the corner map for model performance, it is necessary to choose a suitable corner detector to obtain high-quality corner maps. The detector used in our model is the Shi-Tomasi corner detector [42]. We also experiment with the Harris detector [14] but found it often produces more extra noise corners, which has slight damage to performance. In addition, we use a deep learning-based corner detector SuperPoint [7]. We load its pre-trained model to produce corner maps, and the results are presented in Tab. 2. Although SuperPoint can generate high-quality corner maps, it uses an additional neural network model that increases the feed-forward time.

It is worth noting that no matter which detector we use, they can all capture the most critical corner locations and the structure of the text. Therefore, the

Table 2. Results of different corner detectors

Corner Detector	word acc	char recall	char precision
Shi-Tomasi [42]	**69.1**	**85.7**	**84.8**
Harris [14]	68.4	85.1	84.6
SuperPoint [7]	69.0	85.3	84.7

Table 3. Results of different fusion strategies

Fusion strategy	word acc	char recall	char precision
Baseline	67.0	84.6	84.2
Corner-query	**69.1**	**85.7**	**84.8**
Corner-key/value	66.9	84.1	84.2
Concat	67.0	84.7	84.4
Add	66.6	84.9	84.3
Multiply	67.4	84.7	84.4

results in Tab. 2 using corner maps are better than the other results in Tab. 1.

Fusion strategy. It is crucial to efficiently fuse the features of the corner map and the image, which determines whether the model can make full use of the important information carried by the corners. Given these two features obtained from convolutional layers, we can fuse them into one by Concat, Add and Multiply operations and straightly feed the fused feature to Transformer. As shown in Tab. 3, there is no significant improvement in these results. Add operation introduces additive noise to image features. Multiply operation makes the image filter out valuable features based on the corners, bringing a slight improvement. Moreover, for the cross-attention module, we swap the roles of corner and image, so that corner features are used as the key and value. But the results are not improved, although this operation introduces extra parameters compared to the baseline. The reason is that a lot of information is lost when the corner map is used as the value. Therefore, our corner-query cross-attention mechanism is an efficient fusion strategy.

Table 4. Ablation study on character contrastive loss

Hyperparameters	word acc	char recall	char precision
$\lambda = 0$ (without CC loss)	67.0	84.6	84.2
$\lambda = 0.1$, $\tau = 0.05$, $d = 512$	66.5	84.0	83.6
$\lambda = 0.1$, $\tau = 0.1$, $d = 512$	68.1	85.5	85.3
$\lambda = 0.1$, $\tau = 0.15$, $d = 512$	67.7	84.9	84.6
$\lambda = 0.1$, $\tau = 0.1$, $d = 2048$	**68.6**	**85.8**	**85.9**
$\lambda = 0.01$, $\tau = 0.1$, $d = 2048$	67.2	84.4	84.3
$\lambda = 1$, $\tau = 0.1$, $d = 2048$	66.6	83.9	83.7

Character contrastive loss. According to the previous work of contrastive learning [4,23], the scaling factor τ of the loss function in formula (2) plays an important role in final performance. Relatively low values of τ make hard negatives have more weight but the feature space will be less smooth when τ is extremely low. We conduct an ablation study on τ as shown in Tab. 4, and found $\tau = 0.1$ is optimal. Besides, the dimension of the final output feature vector x_i also affects performance. Generally, higher dimension brings better results because the feature vector represents more information. If the weight of the CC loss is small ($\lambda = 0.01$), it will not bring a significant performance improvement. In contrast, if $\lambda = 1$, it will interfere the joint optimization, resulting in performance degradation. As a result, we adopt $\lambda = 0.1$, $\tau = 0.1$, $d = 2048$ in our model. The results of character recall and character precision show that CC loss actually improves the performance of character recognition.

4.4 Performance for Artistic Text Recognition

In order to demonstrate the superiority of our CornerTransformer on the artistic text recognition task, we compare it with several state-of-the-art scene text recognition methods in Tab. 5. All the results of these methods are obtained by directly loading their released checkpoints to be evaluated on WordArt. Our CornerTransformer shows a significant superiority, thanks to the corner-query cross-attention and the character contrastive loss. Fig. 5 presents some hard examples that are successfully recognized by CornerTransformer. Our model can cope with artistic texts containing complex fonts, ligatures, overlaps, and many extremely curved and deformed texts.

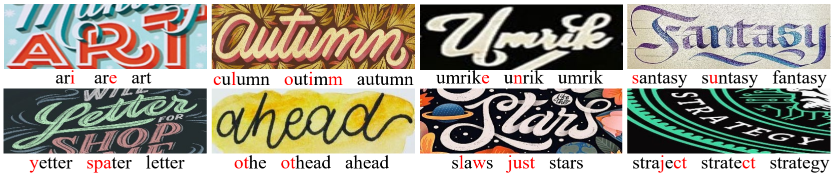


Fig. 5. Qualitative recognition results on WordArt dataset. Each example is along with the results from ABINet-LV [8], our baseline and the proposed CornerTransformer, separately. Hard examples successfully recognized by CornerTransformer

4.5 Evaluation on STR Benchmarks

To further verify the generalization of CornerTransformer, we also conduct evaluations on six STR benchmarks: IIIT5k [33], IC13 [22], SVT [48], IC15 [21], SVTP [35] and CUTE [37]. The results compared with other state-of-the-art methods are shown in Tab. 6. We can achieve state-of-the-art results on SVT and IC15 because most images are severely corrupted by noise and blur, while

Table 5. Performance comparison with other methods on WordArt dataset. * indicates the baseline of SATRN[26] reimplemented by ourselves, replacing the dimensions of the query and key from 128 to 512. Inference time is estimated using an NVIDIA TITAN Xp by averaging 3 trials, based on Pytorch implementation. † indicates the inference time is estimated based on the TensorFlow implementation. "WiKi" indicates using a language model trained with WiKiText-103 [32].

Methods	Training Data	WordArt	Params (M)	Time (ms)
CRNN [39]	ST+MJ	47.5	8.3	9.9
ASTER [41]	ST+MJ	57.9	21	247.9
TRBA [1]	ST+MJ	55.8	49.6	28.8
DAN [49]	ST+MJ	52.4	18.2	41.7
NRTR [38]	ST+MJ	58.5	66.7	350.8
RobustScanner [58]	ST+MJ+SA+R	61.3	48.0	71.0
SAR [27]	ST+MJ+SA+R	63.8	57.5	109.2
SEED [36]	ST+MJ	60.1	25.0	158.8
SCATTER [29]	ST+MJ+SA	64.0	119.7	142.7
SATRN† [26]	ST+MJ	65.7	55.5	494.1
ABINet-LV [8]	ST+MJ+WiKi	67.4	36.7	42.4
Baseline*	ST+MJ	67.0	65.6	274.7
Baseline + Corner	ST+MJ	69.1	80.5	294.9
Baseline + CC loss	ST+MJ	68.6	70.9	274.7
CornerTransformer	ST+MJ	**70.8**	85.7	294.9

gradient-based corner detection is robust to image resolution, noise and blur. Besides, we also obtain a competitive result on CUTE and the best result on SVTP. The texts in these datasets are perspective and curved, while the relative position between corner points is invariant.

4.6 Further Visualization and Analysis

Corner directs more accurate attention. To intuitively verify the effectiveness of our corner-query cross-attention, exploring the essential mechanism why the corner map can improve the model performance, we visualize the feature map of the final output from our corner-guided encoder, as shown in Fig 6. Evidently, for various text images with deformation, ligature, art design, and curve, our encoder can accurately focus on the position of each character, and there are apparent margins between characters. More importantly, our encoder can sometimes even focus on fine-grained features like character strokes, despite not providing any character-level or stroke-level annotations. All these good properties benefit from the corner-query cross-attention. The corner map contains the keypoints of the character strokes, and the corner-query attention enables the corner to seek the image features of interest (that is to seek other positions of the current character but not another character). Therefore, a corner point can gradually focus on the stroke feature up to the whole character feature. Besides, the corner map is very sparse and naturally separates each character.

Table 6. Accuracy comparison with other STR methods on six standard benchmarks

Methods	Training Data	Regular				Irregular			
		IIIT5k	SVT	IC13	Avg	SVTP	IC15	CUTE	Avg
CRNN [39]	ST+MJ	78.2	80.9	89.4	81.0	-	-	-	-
ASTER [41]	ST+MJ	93.4	89.5	91.8	92.5	78.5	76.1	79.5	76.9
TRBA [1]	ST+MJ	87.9	87.5	92.3	88.8	79.2	77.6	74.0	77.6
DAN [49]	ST+MJ	94.3	89.2	93.9	93.5	80.0	74.5	84.4	76.6
NRTR [38]	ST+MJ	90.1	91.5	95.8	91.5	86.6	79.4	80.9	81.1
RobustScanner [58]	ST+MJ	95.3	88.1	94.8	94.2	79.5	77.1	90.3	78.9
SAR [27]	ST+MJ	91.5	84.5	91.0	90.4	76.4	69.2	83.3	72.1
SEED [36]	ST+MJ	93.8	89.6	92.8	93.0	81.4	80.0	83.6	80.6
SCATTER [29]	ST+MJ	93.2	90.9	94.1	93.1	86.2	82.0	84.8	83.2
SATRN [26]	ST+MJ	92.8	91.3	94.1	92.9	86.5	79.0	87.8	81.5
Text is Text [3]	ST+MJ	92.3	89.9	93.3	92.2	84.4	76.9	86.3	79.4
ABINet-LV [8]	ST+MJ+WiKi	**96.2**	93.5	**97.4**	**96.1**	89.3	86.0	89.2	87.0
S-GTR [16]	ST+MJ	95.8	94.1	96.8	95.8	87.9	84.6	**92.3**	86.0
Baseline*	ST+MJ	94.7	92.3	95.5	94.5	87.1	83.3	89.6	84.7
Baseline + Corner	ST+MJ	95.1	94.1	95.7	95.1	90.1	84.9	90.3	86.5
Baseline + CC loss	ST+MJ	95.4	92.0	96.1	95.1	88.2	83.9	89.8	85.4
CornerTransformer	ST+MJ	95.9	**94.6**	96.4	95.8	**91.5**	**86.3**	92.0	**88.0**

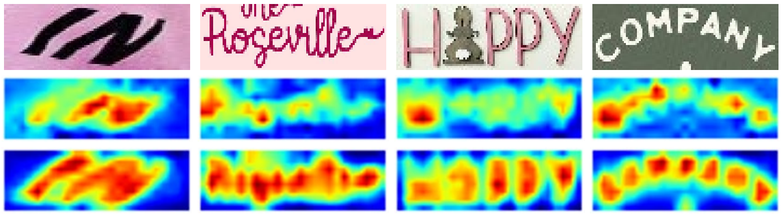
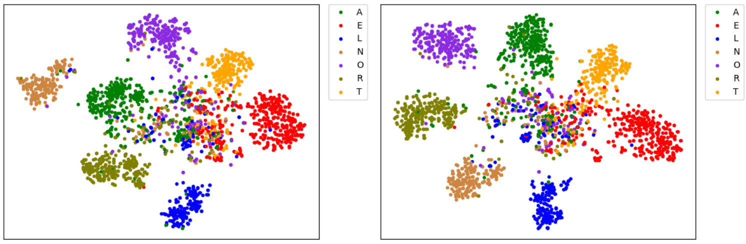


Fig. 6. Visualization for the feature map of the encoder output. First row: input images; Second row: feature maps of the baseline; Third row: feature maps of the baseline equipped with the corner-query cross-attention



(a) CE loss (b) CE loss + CC loss

Fig. 7. Visualization for the character feature distribution of the decoder output

Character contrastive loss improves class representation. In order to verify the effectiveness of our character contrastive loss and justify the motivation for designing this loss, we perform dimension reduction on the final output features of the CornerTransformer decoder and use t-SNE [31] to visualize the distribution of character features. Fig. 7 demonstrates the feature distributions of randomly selected 7 characters. Obviously, compared with the baseline using only the cross-entropy loss, when adding the character contrastive loss, the features of each character class are clustered together, and the features of different classes are far away from each other. This phenomenon is in line with our design that characters of the same category are positive samples and those of different categories are negative samples.

4.7 Limitations

For some extremely difficult artistic texts, CornerTransformer may fail to achieve correct results. A few failure examples are shown in Fig. 8. When decorative patterns from the background have exactly the same appearance and similar shape as the texts, it is difficult to distinguish whether these patterns belong to texts or not. These are indeed challenging examples for any text recognizer.

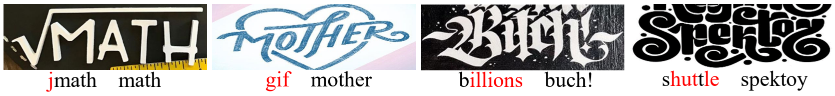


Fig. 8. Failure examples for artistic text recognition. Each image is along with our result and the ground truth

5 Conclusion

In this paper, we focus on a new challenging task of artistic text recognition. To tackle the difficulties of this task, we introduce the corner point map as a robust representation for the artistic text image and present the corner-query cross-attention mechanism to make the model achieve more accurate attention. We also design a character contrastive loss to learn the invariant features of characters, leading to tight clustering of features. In order to benchmark the performance of different models, we provide the WordArt dataset. Experimental results demonstrate the remarkable superiority of our CornerTransformer on artistic text recognition. Interestingly, we achieve state-of-the-art results on several scene text datasets with small and blurred images. We hope the proposed WordArt dataset can encourage more advanced text recognition models, and the corner-based design can offer insights to other challenging recognition tasks.

Acknowledgements This work was supported by the National Natural Science Foundation of China 61733007.

References

1. 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) 8, 12, 13
2. Bhunia, A.K., Ghose, S., Kumar, A., Chowdhury, P.N., Sain, A., Song, Y.Z.: Metahtr: Towards writer-adaptive handwritten text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15830–15839 (2021) 2, 4
3. Bhunia, A.K., Sain, A., Chowdhury, P.N., Song, Y.Z.: Text is text, no matter what: Unifying text recognition using knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 983–992 (2021) 13
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020) 3, 7, 11
5. 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 international conference on computer vision. pp. 5076–5084 (2017) 2, 4
6. Ch'ng, C.K., Chan, C.S., Liu, C.L.: Total-text: toward orientation robustness in scene text detection. International Journal on Document Analysis and Recognition (IJDAR) **23**(1), 31–52 (2020) 4
7. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 224–236 (2018) 9, 10
8. 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) 2, 3, 4, 11, 12, 13
9. Frinken, V., Uchida, S.: Deep blstm neural networks for unconstrained continuous handwritten text recognition. In: 2015 13th international conference on document analysis and recognition (ICDAR). pp. 911–915. IEEE (2015) 2
10. Gobbo, J.D., Matuk Herrera, R.: Unconstrained text detection in manga: A new dataset and baseline. In: European Conference on Computer Vision. pp. 629–646. Springer (2020) 4
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014) 4
12. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on Machine learning. pp. 369–376 (2006) 2, 4
13. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2315–2324 (2016) 8
14. Harris, C., Stephens, M., et al.: A combined corner and edge detector. In: Alvey vision conference. vol. 15, pp. 10–5244. Citeseer (1988) 3, 4, 6, 9, 10
15. He, P., Huang, W., Qiao, Y., Loy, C.C., Tang, X.: Reading scene text in deep convolutional sequences. In: Thirtieth AAAI conference on artificial intelligence (2016) 2, 4

16. He, Y., Chen, C., Zhang, J., Liu, J., He, F., Wang, C., Du, B.: Visual semantics allow for textual reasoning better in scene text recognition. arXiv preprint arXiv:2112.12916 (2021) 13
17. Hu, J., Brown, M.K., Turin, W.: Hmm based online handwriting recognition. IEEE Transactions on pattern analysis and machine intelligence **18**(10), 1039–1045 (1996) 4
18. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and artificial neural networks for natural scene text recognition. Eprint Arxiv (2014) 8
19. 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) 8
20. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Advances in neural information processing systems **28** (2015) 4
21. 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: 2015 13th international conference on document analysis and recognition (ICDAR). pp. 1156–1160. IEEE (2015) 3, 4, 11
22. Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i 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: 2013 12th International Conference on Document Analysis and Recognition. pp. 1484–1493. IEEE (2013) 4, 11
23. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in Neural Information Processing Systems **33**, 18661–18673 (2020) 3, 7, 11
24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 8
25. Le, A.D.: Recognizing handwritten mathematical expressions via paired dual loss attention network and printed mathematical expressions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 566–567 (2020) 4
26. Lee, J., Park, S., Baek, J., Oh, S.J., Kim, S., Lee, H.: On recognizing texts of arbitrary shapes with 2d self-attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 546–547 (2020) 2, 3, 4, 6, 9, 12, 13
27. 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. vol. 33, pp. 8610–8617 (2019) 2, 4, 12, 13
28. 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) 4
29. Litman, R., Anschel, 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) 12, 13
30. Luo, C., Jin, L., Sun, Z.: Moran: A multi-object rectified attention network for scene text recognition. Pattern Recognition **90**, 109–118 (2019) 2, 4
31. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008) 14
32. Merity, S., Xiong, C., Bradbury, J., Socher, R.: Pointer sentinel mixture models. arXiv preprint arXiv:1609.07843 (2016) 12

33. Mishra, A., Alahari, K., Jawahar, C.: Scene text recognition using higher order language priors. In: BMVC-British machine vision conference. BMVA (2012) 4, 11
34. Nayef, N., Patel, Y., Busta, M., Chowdhury, P.N., Karatzas, D., Khlif, W., Matas, J., Pal, U., Burie, J.C., Liu, C.l., et al.: Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019. In: 2019 International conference on document analysis and recognition (ICDAR). pp. 1582–1587. IEEE (2019) 5
35. 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) 3, 4, 11
36. 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) 2, 4, 12, 13
37. 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) 4, 11
38. Sheng, F., Chen, Z., Xu, B.: Nrtr: A no-recurrence sequence-to-sequence model for scene text recognition. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 781–786. IEEE (2019) 12, 13
39. 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) 2, 4, 12, 13
40. Shi, B., Wang, X., Lyu, P., Yao, C., Bai, X.: Robust scene text recognition with automatic rectification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4168–4176 (2016) 2, 4
41. 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) 2, 4, 12, 13
42. Shi, J., et al.: Good features to track. In: 1994 Proceedings of IEEE conference on computer vision and pattern recognition. pp. 593–600. IEEE (1994) 3, 4, 6, 9, 10
43. Su, B., Lu, S.: Accurate scene text recognition based on recurrent neural network. In: Asian Conference on Computer Vision. pp. 35–48. Springer (2014) 4
44. 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: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1557–1562. IEEE (2019) 5
45. 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) 3, 4, 5, 9
46. 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) 4
47. 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) 4
48. Wang, K., Babenko, B., Belongie, S.: End-to-end scene text recognition. In: 2011 International conference on computer vision. pp. 1457–1464. IEEE (2011) 3, 4, 11

49. 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. vol. 34, pp. 12216–12224 (2020) 12, 13
50. Wang, Y., Lian, Z.: Exploring font-independent features for scene text recognition. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 1900–1920 (2020) 4
51. Wu, J.W., Yin, F., Zhang, Y.M., Zhang, X.Y., Liu, C.L.: Handwritten mathematical expression recognition via paired adversarial learning. International Journal of Computer Vision **128**(10), 2386–2401 (2020) 4
52. Xu, X., Zhang, Z., Wang, Z., Price, B., Wang, Z., Shi, H.: Rethinking text segmentation: A novel dataset and a text-specific refinement approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12045–12055 (2021) 8
53. 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) 4
54. Yang, X., He, D., Zhou, Z., Kifer, D., Giles, C.L.: Learning to read irregular text with attention mechanisms. In: IJCAI. vol. 1, p. 3 (2017) 2, 4
55. Yao, C., Bai, X., Shi, B., Liu, W.: Strokelets: A learned multi-scale representation for scene text recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4042–4049 (2014) 4
56. 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) 8
57. 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**(3), 509–521 (2019) 5
58. Yue, X., Kuang, Z., Lin, C., Sun, H., Zhang, W.: Robustscanner: Dynamically enhancing positional clues for robust text recognition. In: European Conference on Computer Vision. pp. 135–151. Springer (2020) 12, 13
59. Zhang, X., Zhu, B., Yao, X., Sun, Q., Li, R., Yu, B.: Context-based contrastive learning for scene text recognition. AAAI (2022) 7
60. Zhang, X.Y., Yin, F., Zhang, Y.M., Liu, C.L., Bengio, Y.: Drawing and recognizing chinese characters with recurrent neural network. IEEE transactions on pattern analysis and machine intelligence **40**(4), 849–862 (2017) 4