

WAS: Dataset and Methods for Artistic Text Segmentation

Xudong Xie¹, Yuzhe Li¹, Yang Liu¹, Zhifei Zhang², Zhaowen Wang², Wei Xiong², and Xiang Bai¹^(✉)

¹ Huazhong University of Science and Technology, China
{xdxie,yzli12,yangliu1213,xbai}@hust.edu.cn

² Adobe, USA
{zzhang,zhawang}@adobe.com, wxiongur@gmail.com

Abstract. Accurate text segmentation results are crucial for text-related generative tasks, such as text image generation, text editing, text removal, and text style transfer. Recently, some scene text segmentation methods have made significant progress in segmenting regular text. However, these methods perform poorly in scenarios containing artistic text. Therefore, this paper focuses on the more challenging task of artistic text segmentation and constructs a real artistic text segmentation dataset. One challenge of the task is that the local stroke shapes of artistic text are changeable with diversity and complexity. We propose a decoder with the layer-wise momentum query to prevent the model from ignoring stroke regions of special shapes. Another challenge is the complexity of the global topological structure. We further design a skeleton-assisted head to guide the model to focus on the global structure. Additionally, to enhance the generalization performance of the text segmentation model, we propose a strategy for training data synthesis, based on the large multi-modal model and the diffusion model. Experimental results show that our proposed method and synthetic dataset can significantly enhance the performance of artistic text segmentation and achieve state-of-the-art results on other public datasets. The datasets and codes are available at: https://github.com/xdxie/WAS_WordArt-Segmentation.

Keywords: artistic text segmentation · momentum query · skeleton

1 Introduction

Text segmentation is dedicated to finely segmenting the strokes of text from complex scene images, discriminating whether each pixel belongs to the text foreground or the background. Accurate text segmentation results are the foundation for text-related generative tasks. For instance, tasks such as text image generation [5], text style transfer [17, 20], and text removal [33, 34] can produce excellent and practical generative outcomes based on text masks. However, although existing models have achieved outstanding performance in regular text

[✉] Corresponding author

segmentation tasks, they are difficult to accurately segment artistic text. Artistic text features complex appearances and shapes, making it difficult to distinguish from background patterns [39]. Therefore, artistic text segmentation is a challenging task that has not yet been studied by the academic community.

The first problem encountered in implementing this new task is the lack of datasets. The existing text segmentation datasets ICDAR13 FST [14], COCO_TS [2], MLT_S [3], and Total-Text [10] suffer from the problems of low annotation quality and insufficient quantity. Moreover, these data are scene images with regular text. Although TextSeg [40] provides high-quality annotated data and some artistic text images, the number of these images is still insufficient to train a high-performance artistic text segmentation model with strong generalizability. Therefore, we construct a real **WordArt Segmentation** dataset called **WAS-R**, consisting of 7100 artistic text images with word-level annotations of quadrilateral boxes, masks, and transcriptions. Additionally, to further enhance the accuracy and generalization ability of the text segmentation model, we also propose a synthetic dataset called **WAS-S**. Our designed synthetic pipeline utilizes the popular large multi-modal model and diffusion model to achieve realism, accuracy, and diversity in the generated images.

Besides, artistic text segmentation presents two unique challenges compared to general object segmentation and regular text segmentation. (1) The strokes of artistic text have flexible and changeable local shapes, such as slender tails or twisted ligatures. (2) The global topological structure of the artistic text is very complex, with many holes and intricate connections within the text. In contrast, the local stroke shapes and the global structure of regular text are almost invariant, and the topological structure of general objects is very simple. Therefore, the task we propose has clear academic value and practical significance.

There are currently few specialized models for text segmentation. Recent studies either require the aid of text detection modules [41, 43] or the assistance of character-level recognizers [40]. Moreover, these methods have not been specifically designed for artistic text. In view of this, we propose a WordArt segmentation model **WASNet**. To address the first challenge, we propose a Transformer decoder with the layer-wise momentum query. The input of the self-attention module is the momentum superposition of the masked queries from the current layer and the previous layers. This operation ensures that the model does not quickly fit a restricted regular mask area when updating attention, thereby ignoring some special-shaped stroke regions, which are precisely what the earlier layers are capable of capturing. To address the second challenge, we propose a skeleton-assisted head that enables the model to output both mask predictions and skeleton predictions simultaneously, guiding the model to capture the global topological structure. It enhances the decoder’s ability to perceive the overall structure of the text.

We conduct extensive experiments to verify the effectiveness of the proposed method and the synthetic dataset on the task of artistic text segmentation. We also verified the generalizability on other public datasets [2, 10, 40]. The

results achieved state-of-the-art (SOTA) performance. More importantly, the model trained on the WAS dataset can be directly tested on other datasets without the need for fine-tuning and still achieve competitive results. This opens up a new experimental paradigm for the task of text segmentation.

In summary, our contributions are four-fold:

- (1) We present a new challenging task: artistic text segmentation, and construct a real dataset to benchmark the performance of various models.
- (2) We design a training data synthesis strategy and generate a synthetic dataset consisting of 100k image-mask pairs.
- (3) We introduce the layer-wise momentum query to handle the changeable local strokes and skeleton-assisted head to capture the complex global structure.
- (4) We achieve new SOTA results in the tasks of artistic text segmentation and scene text segmentation, and simplify the experimental paradigm for text segmentation.

2 Related Work

2.1 Text Segmentation Method

Some early text segmentation methods relied on thresholding [27], low-level features [31], or Markov Random Fields (MRF) [22] to segment foreground text from the background, but these methods could only achieve limited success in document processing. With the continuous advancement of deep learning, the corresponding text segmentation methods have shown great potential in complex scenes [3, 12, 28]. For instance, SMANet [3] employs the encoder-decoder architecture of PSPNet [48] and achieves a multi-scale attention module for text segmentation. PGTSNet [41] employs a pre-trained detector to ground out text regions before segmentation, further enhancing the accuracy of segmentation. TexRNet [40] incorporates character recognition and attention-based similarity checking to aid the model in segmenting text. Building on these methodologies, Yu *et al.* [43] developed a model featuring a lightweight detection head and a Text-Focused Module, elevating text segmentation performance in complex scenes to a new level. Nevertheless, recent high-performance text segmentation methods either utilize extra bounding box annotations and rely on text detection, or employ character-level supervision. Also, these methods lack specialized design for artistic text.

2.2 Text Segmentation Dataset

The construction of text segmentation datasets has not received enough attention in academia. ICDAR13 [14] and Total-Text [10] provide high-quality, pixel-level annotations for text segmentation, but their quantities are very limited with only 462 and 1,555 images, respectively. To address the issue of insufficient quantity, researchers have proposed a dataset COCO-TS [2] (14,690 images) based on COCO-Text [29] for text segmentation. Similarly, MLT_S [3]

(6,896 images) is also a large-scale text segmentation dataset based on ICDAR MLT [23]. Both of these datasets use automatic annotation strategies, resulting in low-quality dataset annotations. In view of these problems, Xu *et al.* [40] introduced a larger-scale and high-quality text segmentation dataset TextSeg (4,024 images), which includes character-level and word-level annotations of masks, bounding boxes, and transcriptions. Moreover, different from the datasets mentioned above, BTS [41] is a bilingual text segmentation dataset. These real datasets are all derived from natural images but lack a dedicated dataset for segmenting artistic text. Although TextSeg contains some artistic text images, the quantity is insufficient to train a robust model for artistic text segmentation.

2.3 Segmentation Dataset Generation

The construction of synthetic segmentation datasets plays an important role in studying visual perception problems. DatasetGAN [47] and BigDatasetGAN [15] only use a small number of manually labeled samples for each category to train the decoder and generate a large amount of new data. Diffumask [36] extends the text-driven image synthesis to semantic mask generation in Stable Diffusion [25] to create a high-resolution and class-discriminative pixel-wise mask. Dataset diffusion [24] leverages the pre-trained diffusion model and text prompts to generate segmentation maps corresponding to synthetic images. DatasetDM [35] is a generic dataset generation model that decode the latent code of the diffusion model as accurate perception annotations. MosaicFusion [38] is a diffusion-based data augmentation method that does not require training and does not rely on any label supervision, especially for rare and new categories. Then SegGen [42] integrates Text2Mask and Mask2Img synthesis to generate training data, improving the performance of state-of-the-art segmentation models in various segmentation tasks. These advanced methods often fail to align artistic text with their masks in images. In this paper, we avoid having the model generate both images and mask annotations. Instead, we pre-render the text masks and use ControlNet [45] to generate mask-conditioned images.

3 Dataset

As artistic text in the real world is incredibly diverse, we propose two new datasets: WAS-R composed of real-world text images, and WAS-S composed of synthetic text images. These multi-purpose artistic text datasets aim to bridge the gap between artistic text segmentation and real-world applications, accommodating the rapid advances in text vision research.

3.1 WAS-R Image Collection

The WAS-R dataset is composed of 7,100 images sourced from a variety of contexts, including posters, cards, covers, logos, goods, road signs, billboards, digital designs, and handwritten text. Among these, 4,100 images serve as the training

dataset, while the remaining 3,000 images constitute the test dataset. The artistic text can be categorized into two major types according to the way capturing images. A type of artistic text image is taken by cameras from various scenes, such as signboards. The other type is directly exported from design software, such as poster files. During data collection, we specifically balances these two types to create a diverse dataset for research and development.

3.2 WAS-R Image Annotation

The WAS-R dataset stands out due to its comprehensive annotations, surpassing existing datasets. Specifically, WAS-R provides minimum quadrilateral detection boxes with distinct segmentation mask labels for each word. It also provides text transcription for each word mask. Moreover, we annotates the word effects such as shadow, glow, 3D, which play a crucial role in distinguishing artistic text from conventional scene text and significantly impacts text segmentation. Fig. 1 shows examples of collected images and their annotations in WAS-R.



Fig. 1: Examples of images and annotations from the proposed WAS-R dataset.

3.3 WAS-S Synthetic Dataset Construction

Fig. 2 shows the pipeline of generating synthetic text images. The core idea is that we build a text image generation model which can generate aligned text images from text masks and input prompts. To this end, we construct the training pipeline as illustrated in Fig. 2 (a). Specifically, we first generate diverse and informative captions from the text images in the training set of WAS-R to obtain training triplets $\langle \text{caption}, \text{mask}, \text{image} \rangle$. Following that, we train a ControlNet [45] with these triplets for generating diverse images that are pixel-wisely aligned with the input text mask. During inference, as shown in Fig. 2 (b), we first construct diverse text masks using Mask Render, then use GPT-4 to extend the texts in the mask into a scene description caption. The constructed text mask and caption are send to our trained ControlNet to generate the synthetic text image. We describe the modeling details below.

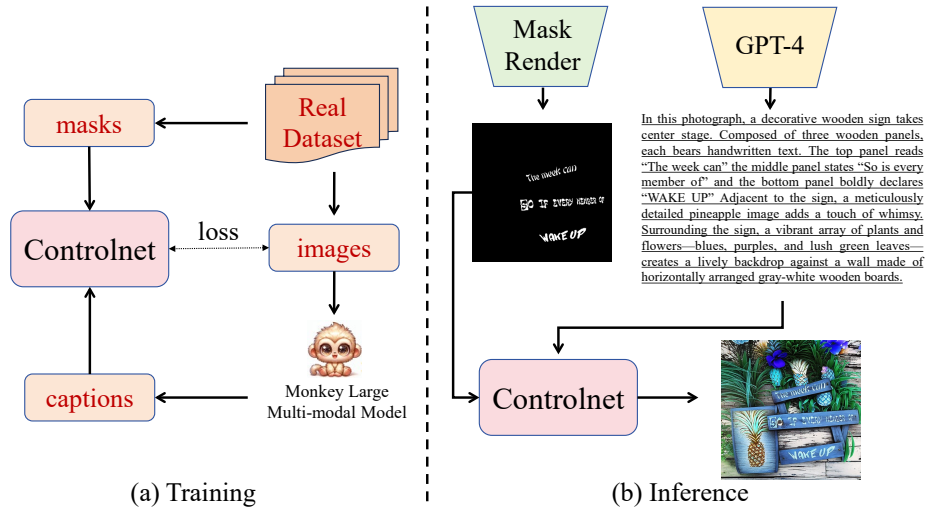


Fig. 2: (a) Training pipeline of ControlNet. (b) WAS-S data generation pipeline.

Training Pipeline. During training, we construct the dataset for training ControlNet [45] based on the training set of WAS-R. To this end, we obtain the image captions of existing training samples in WAS-R using the advanced multi-modal large language model called Monkey [18]. Formally, let I_t represent the real image in the training set of WAS-R, and C_t denote the prompt generated from I_t , we have $C_t = \text{Monkey}(I_t)$.

Having obtained the $\langle \text{caption, text mask, text image} \rangle$ training triplets, we train a ControlNet that maps the input prompt and text mask to a text image. Our goal is that the outline of the artistic text in the generated image should be well aligned with the input text mask. Moreover, the contents and styles of the generated images should be diverse enough as guided by the input prompts.

Inference Pipeline. During inference, as illustrated in Fig. 2(b), we first generate synthetic text masks denoted as M_{syn} using our proposed **Mask Render** technique. Specifically, for each mask, we randomly select 1-7 phrases from the 20 newsgroups dataset [1] based on the word distribution of each image in the real dataset WAS-R. These phrases are consist of 1-5 consecutive words. Additionally, we apply a random rotation in the range of $-30^\circ \leq \phi \leq 30^\circ$ to each phrase. The size of each phrase is limited to match the general width of the entire image, and we position them randomly within the image boundaries. Besides, we use 250 artistic fonts. Finally, an affine transformation is applied to each phrase to introduce skewness and distortion.

We use GPT-4 [4] to generate the prompts corresponding to the synthetic text masks. Specifically, we ask GPT-4 to mimic the style of captions we generate from the training set in WAS-R, and synthesize new prompts. Next, we incorporate the text information in the synthetic mask into the generated prompt to obtain the final prompt. Formally, we have: $C_{syn} = \text{GPT4}(C_t, C_M)$, where C_{syn} is the

caption we generate and C_M is the text in the synthetic mask. Fig. 3 shows examples of our synthetic prompts.

Following the construction of synthetic text mask and synthetic prompts, we use the trained ControlNet to generate the final text image. We have $I_{syn} = ControlNet_{\theta}(C_{syn}, M_{syn})$, where I_{syn} is the image we generate, θ denotes the trainable parameters in ControlNet. Fig. 3 shows examples of the final synthesized <text mask, prompt, image> triplet. More details can refer to the supplementary materials.







mask	prompt	image
	Amidst the serene forest, the figure stands—a canvas of mystery. Their white T-shirt, like a blank page, bears cryptic inscriptions: “In reality”:A whisper of secrets, etched in blue. What truth lies hidden behind this enigmatic phrase? Is it a riddle or a forgotten memory? “CbwCB”:Elegant cursive weaves letters into a dance. A code, perhaps? Each character a step, leading to an unknown destination. “ulkm1”:Bold and unapologetic, this word anchors the composition. Is it a name, a potion, or a spell? The forest holds its secret.	
	This image showcases a colorful billboard against a backdrop of lush green trees. The billboard has an abstract design in blue and orange colors, with partially visible text reading “In article”. Below this, there’s an orange section displaying the name “Charles M Kozierok” in white letters. The supporting structure appears aged and weathered, and a gravel path or road is visible at the bottom.	
	In the mall’ s mysterious corridors, cryptic signs beckon curious shoppers. A black placard ominously declares “Slashing His Wrist” hinting at an art installation or an edgy watch boutique. Meanwhile, a massive electronic screen scrolls news headlines, prominently featuring “IN ARTICLE” Is it breaking news or a promotional campaign? Near the sportswear section, an orange sign reads “AS for the Rangers Game” .	

Fig. 3: The generated <mask, prompt, image> triplet. The left column is the generated masks. The middle column shows the prompt generated by GPT-4, imitating styles of the prompt in the training set. The right column is the final generated images.

4 Methodology

In this section, we introduce our artistic text segmentation model WASNet. We first present the overall architecture, followed by detailed descriptions of the local and global designs.

4.1 Overall Architecture

The overall framework of WASNet is shown in Fig. 4. We take an excellent semantic segmentation model Mask2Former [8] as the meta-architecture. It is a mask classification architecture that directly predicts multiple binary masks and corresponding category labels, instead of performing per-pixel classification. We add a skeleton-assisted head and improve the Transformer decoder with a mechanism

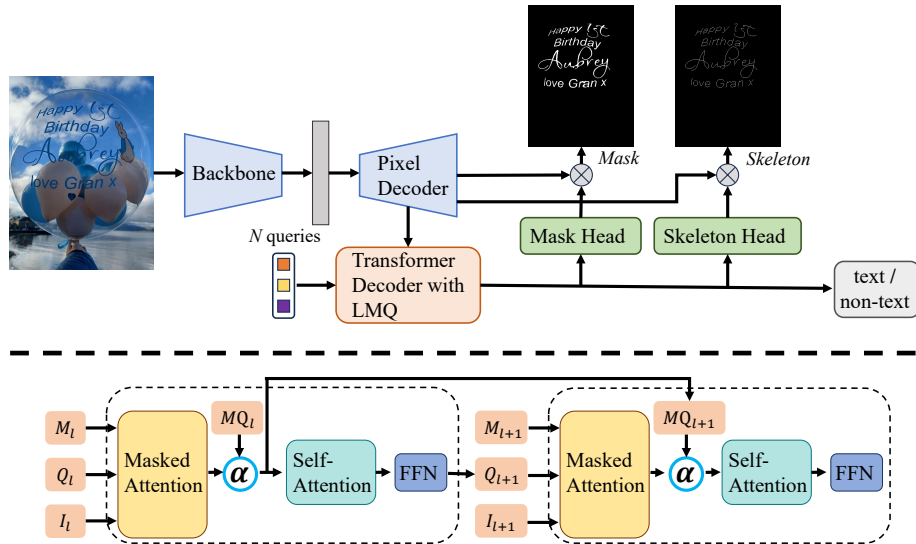


Fig. 4: Up: The overall architecture of our WASNet. **Down:** The Transformer decoder with layer-wise momentum query (LMQ).

of layer-wise momentum query. The backbone extracts low-resolution features from an image. The pixel decoder upsamples the image features and generates a feature pyramid. The multi-scale features are fed into the Transformer decoder, with each resolution corresponding to each layer’s input of the decoder. Besides, each layer of the Transformer decoder also receives the mask prediction and query generated from the previous layer as input. Finally, the mask head and the skeleton head generate binary mask and skeleton predictions respectively, by decoding the per-pixel embeddings from the pixel decoder and object queries from the Transformer decoder. The ground truth of the skeleton is obtained by thinning the binary mask labels through a skeleton extraction algorithm [46].

4.2 Transformer Decoder with Layer-wise Momentum Query

Artistic text segmentation faces the challenge of the local stroke shapes being flexible and changeable. Due to designers using hundreds of different artistic fonts and applying various text effects, the local strokes of the same character can differ significantly. This results in some slender strokes spanning across other areas, as well as twisted ligatures leading to complex text edges. In contrast, normal scene text typically utilizes regular printed fonts without special designs, and the stroke shapes are almost invariant. Therefore, it is necessary for the decoder to pay attention to these special local strokes.

First, we use the masked attention mechanism [8], constraining cross-attention to within the local text mask region for each query, instead of attending to the

full feature map. This mechanism can be expressed as:

$$\mathbf{MA}_l = \text{softmax}(\mathbf{M}_l + \mathbf{Q}_l \mathbf{I}_l^T) \mathbf{I}_l, \quad (1)$$

where l is the layer index, and \mathbf{Q}_l is the input queries. $\mathbf{I}_l \in \mathbb{R}^{H_l W_l \times C}$ is the image feature input to the l -th layer, which comes from the feature pyramid of the pixel decoder. H_l and W_l indicate the spatial resolution of the image feature and C is the feature dimension. \mathbf{M}_l is transformed from the binary mask output of the previous layer, with the value of the text region being 0 and the value of the non-text region being $-\infty$ [8]. \mathbf{MA}_l is the output of the masked attention module. We have omitted the residual connection and normalization here.

Furthermore, since the masks predicted by each layer are different, the previous layers yield coarse masks that may include special-shaped stroke regions. However, the subsequent layers are inclined to predict more precise regions of regular strokes, overlooking those local special regions. Therefore, in order to prevent the model’s attention from being quickly confined to regular regions, we design a mechanism of Layer-wise Momentum Query (LMQ). The momentum superposition of the masked queries from the current and previous layers is input to the self-attention module before the module gathers contextual information. Eq. (2) illustrates this mechanism.

$$\mathbf{MQ}_{l+1} = \alpha \mathbf{MQ}_l + (1 - \alpha) \mathbf{MA}_l, \quad (2)$$

where $\alpha \in [0, 1)$ is a momentum coefficient. \mathbf{MQ} is the momentum query that is input to the self-attention module. We ultimately use this decoder in WASNet with layer-wise momentum query.

4.3 Skeleton-Assisted Head

Different from regular text and general objects, the global topological structure of artistic text is very complex, and there are many holes and intricate connections inside. This presents new challenges for the segmentation task. The model needs to capture the global structure of the text object rather than just a region. Inspired by DeepSkeleton [26] and DeepFlux [32], we found that the skeleton is an effective representation to describe the shape and topology of text because it can extract the central axis of the object. Therefore, we use skeletons to assist text segmentation.

As shown in Fig. 4, we add a skeleton-assisted head to WASNet, enabling the model to simultaneously predict the mask and the skeleton, thus endowing it with the capability to perceive the global topological structure. Since the binary mask is a finely annotated label for semantic segmentation, the ground truth for the skeleton can be obtained by processing the mask with the classic Zhang-Suen [46] skeleton extraction algorithm. The algorithm progressively removes pixels that satisfy certain template structural conditions through an iterative process, until no more pixels meeting the conditions are deleted.

We use the binary cross-entropy loss and the dice loss [21] for our skeleton loss and mask loss:

$$\mathcal{L}_{skeleton} = \mathcal{L}_{mask} = \lambda_{ce}\mathcal{L}_{ce} + \lambda_{dice}\mathcal{L}_{dice}. \quad (3)$$

We set $\lambda_{ce} = \lambda_{dice} = 5$. The final loss is the combination of skeleton loss, mask loss, and classification loss:

$$\mathcal{L}_{final} = \mathcal{L}_{skeleton} + \mathcal{L}_{mask} + \lambda_{cls}\mathcal{L}_{cls}, \quad (4)$$

where $\lambda_{cls} = 2$ for predictions matched with labels and 0.1 for predictions that have not been matched with any labels.

During the inference phase, it is unnecessary to output the predictions of the skeleton. Therefore, we follow the post-processing method in [9] to obtain the final output of text semantic segmentation.

Table 1: Performance comparison with other methods on WAS dataset. * TextFormer trains a text detection module using additional bounding box labels. “pre-train” indicates that the model was firstly trained on WAS-S and then fine-tuned on WAS-R.

Methods	Venue	WAS	
		fgIoU	F-score
PSPNet [48]	CVPR’17	71.15	0.831
DeepLabV3+ [7]	CVPR’18	79.65	0.887
OCRNet [44]	ECCV’20	79.06	0.883
SegFormer [37]	NeurIPS’21	79.46	0.886
TexRNet [40]	CVPR’21	77.19	0.850
DDP [13]	ICCV’23	81.07	0.896
TextFormer* [43]	ACM MM’23	80.12	0.889
Mask2Former [9]	NeurIPS’21	80.21	0.890
WASNet (ours)	-	82.11	0.901
Mask2Former (pre-train)	-	82.42	0.902
WASNet (pre-train)	-	84.18	0.913

5 Experiments

5.1 Implementation Details

Our experiments are mainly based on the MMSegmentation [11] toolbox. The overall hyperparameter configuration is the same as [8]. The pixel decoder is a multi-scale deformable attention Transformer [49] with 6 layers. The Transformer decoder consists of 9 layers, each with an auxiliary loss. We use the AdamW [19] optimizer and the poly [6] learning rate schedule with an initial learning rate

of 10^{-4} and a weight decay of 0.05. The data augmentation strategies include random scale jittering, random color jittering, random cropping as well as random horizontal flipping. We use a crop size of 512×512 and a batch size of 16. The models are trained with 8 RTX4090 GPUs. If the model is only trained on the real dataset, we set the number of iterations to 100k. If the model needs to be pre-trained on the synthetic dataset WAS-S, we first pre-train the model for 50k iterations and then fine-tune it on the real dataset for 50k iterations. For the momentum coefficient α in Eq. (2), we set $\alpha = 0.8$ by default. Following the previous text segmentation methods [40, 43], we use foreground (text) Intersection-over-Union (fgIoU) as the major metric and F-score measurement on foreground pixels as the auxiliary metric.

5.2 Results of Artistic Text Segmentation

To verify the superiority of our method in the task of artistic text segmentation, we trained several representative models on our WAS-R dataset, including six semantic segmentation models and two text segmentation models. We use the officially released code for TexRNet [40], DDP [13], and TextFormer [43], and the code reproduced by MMSegmentation [11] for other models. For a fair comparison, we did not apply the character-level glyph discriminator for TexRNet.

The experimental results in Tab. 1 indicate that our WASNet outperforms all of these advanced models. Moreover, when we train the baseline models and WASNet with the synthetic dataset WAS-S, their performance can be further improved. Our final results have achieved a significant SOTA performance.

5.3 Results of Scene Text Segmentation

To further verify the generalizability of WASNet, we also conducted experiments on three publicly available scene text segmentation datasets [2, 10, 40], as shown in Tab. 2. We can draw the same conclusion as in Sec. 5.2 regarding the effectiveness of WASNet and our synthetic dataset. It is worth mentioning that character-level annotations were used to train TexRNet on TextSeg. Extra bounding box labels were used to train the text detection module of TextFormer on all three datasets. However, we only use binary mask labels of the full images. Despite this, we still achieved competitive or state-of-the-art results. Due to the highly inaccurate annotation quality of COCO_TS [2] and the fact that Total-Text [10] contains only 300 test images, the conclusions drawn from the evaluation results of the models on these two datasets may be inconsistent.

Furthermore, we directly evaluate the performance of WASNet on the three datasets using the model trained on the synthetic and real WAS datasets, as shown in the last row of Tab. 2. Note that the results in this row have not been fine-tuned on specific datasets but are still competitive. Therefore, to simplify the experimental paradigm and evaluation process of text segmentation models, we encourage researchers to train on WAS and test directly on other datasets.

Table 2: Performance comparison with other methods on three publicly available scene text segmentation datasets. * TextFormer trains a text detection module using additional bounding box labels. “pre-train” indicates that the model was firstly trained on WAS-S and subsequently fine-tuned on the specific datasets.

Methods	COCO_TS [2]		Total-Text [10]		TextSeg [40]	
	fgIoU	F-score	fgIoU	F-score	fgIoU	F-score
PSPNet [48]	-	-	-	0.740	-	-
SMANet [2]	-	-	-	0.770	-	-
DeepLabV3+ [7]	72.07	0.641	74.44	0.824	84.07	0.914
HRNetV2-W48 [30]	72.07	0.641	74.44	0.824	85.03	0.914
OCRNet [44]	69.54	0.627	76.23	0.832	85.98	0.918
SegFormer [37]	63.17	0.774	73.31	0.846	84.59	0.916
TexRNet [40]	72.39	0.720	78.47	0.848	86.84	0.924
DDP [13]	70.04	0.824	72.55	0.841	84.37	0.915
TextFormer* [43]	73.40	0.847	82.10	0.902	<u>87.11</u>	<u>0.931</u>
Mask2Former [9] (baseline)	70.03	0.823	75.54	0.832	84.95	0.911
WASNet (ours)	71.10	0.830	77.26	0.840	86.56	0.921
Mask2Former (pre-train)	70.89	0.830	78.05	0.851	86.00	0.919
WASNet (pre-train)	<u>73.28</u>	0.848	<u>79.30</u>	<u>0.863</u>	87.42	0.932
WASNet (WAS dataset)	69.22	0.817	75.39	0.836	84.26	0.906

Table 3: Ablation study on our proposed modules and datasets. **Table 4:** Ablation study on datasets.

Methods	WAS		Dataset	WAS	
	fgIoU	F-score		fgIoU	F-score
Baseline [9]	80.21	0.890	WAS-S	84.18	0.913
+ LMQ	80.95	0.898	5w images	83.25	0.906
+ Skeleton	82.11	0.901	20w images	84.01	0.912
+ WAS-S	84.18	0.913	BLIP2 [16]	83.07	0.906
			1000 fonts	82.35	0.902

Table 5: Ablation study on the momentum coefficient.

α	0.9	0.8	0.6	0.5	0.4	0.2	0.1
fgIoU	82.03	82.11	82.06	81.87	81.92	81.36	81.31

5.4 Ablation Study

In this section, we conduct the ablation study on the artistic text segmentation dataset WAS. We first validate the effectiveness of our proposed modules and

the synthetic dataset. As shown in Tab. 3, as we gradually apply the design of LMQ and the skeleton to the baseline, the performance of WASNet is incrementally improved. Pre-training WASNet on WAS-S can continue to enhance the performance of artistic text segmentation.

Therefore, the synthetic dataset is an important contribution of this paper. We conduct ablation experiments regarding some synthetic details of the dataset in Tab. 4. It is crucial to control the number of synthesized mask-image pairs. Less data will weaken performance, but more data will lead to a performance plateau. The reasons will be analyzed in Sec. 5.6. We also use other large multi-modal model BLIP2 [9] to generate the image captions, but the performance is limited. This is because the overall performance of BLIP2 is inferior to Monkey [18] we used. Besides, we applied more fonts to generate masks, but the performance actually decreased. The dataset of 1000 fonts includes a large number of regular fonts, which reduces the learning difficulty of the dataset.

Furthermore, we explored the impact of different momentum coefficient values α on the performance of WASNet in Tab. 5 and found that the approximate optimal value is 0.8. A coefficient that is too large can cause the model to be overly influenced by the coarse predictions from earlier layers. A coefficient that is too small diminishes the positive effect of momentum queries.

5.5 Further Analysis

To further verify the effectiveness of our proposed WASNet, we visualize the inference outputs of our baseline model Mask2Former [9] and WASNet in Fig. 5. According to Fig. 5 (a), it is evident that WASNet can capture special-shaped stroke regions such as slender tails or twisted ligatures. This is attributed to our Transformer decoder with the layer-wise momentum query. Additionally, according to Fig. 5 (b), WASNet exhibits good scale adaptability. It can achieve fine results for both large-scale and small-scale text with complex structures. This is because the skeleton-assisted head can obtain the global topological structure of the text through the thinning operation, guiding fine segmentation.

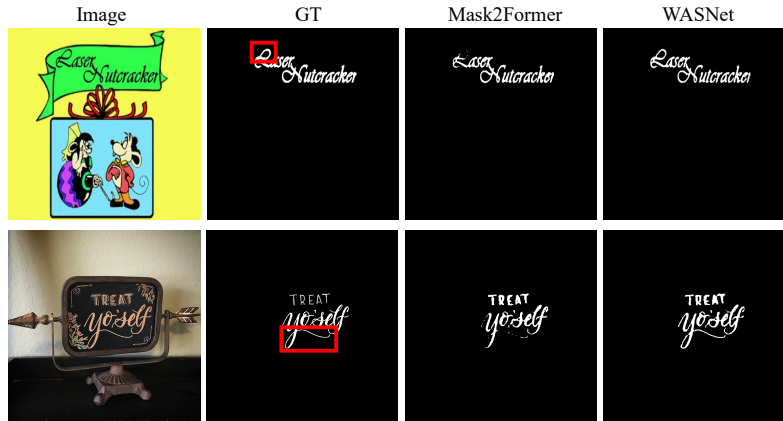
Once accurate text stroke masks are obtained, downstream text-related generative tasks can demonstrate excellent results. The application effects of text removal, text background replacement, and text style transfer are shown in the supplementary materials.

5.6 Limitation

Although the proposed synthetic dataset can improve the performance of text segmentation models, the enhancement is limited and does not *significantly* increase. Even when we further increased the amount of synthetic data, the performance remained unchanged. This could be caused by the bottleneck encountered in the diversity and realism of the synthetic images. In the future, we are considering designing more advanced generative models.

6 Conclusion

The paper focuses on a new challenging task of artistic text segmentation. We propose a real dataset for this task to train the models and benchmark the performance. We also construct a synthetic dataset to further enhance the accuracy and generalization ability. In order to meet the challenges of this task, we introduce the layer-wise momentum query to handle the changeable local strokes and the skeleton-assisted head to capture the complex global structure. Experimental results have demonstrated the effectiveness and superiority of our method in the tasks of artistic text segmentation and scene text segmentation. We hope that more researchers can focus on this task in the future and that the dataset we propose can change the experimental paradigm of text segmentation.



(a) WASNet captures local special-shaped stroke regions



(b) WASNet captures global topological structures

Fig. 5: Qualitative comparison between the baseline model Mask2Former [9] and our WASNet. The two innovations of our method alleviate the two main problems of artistic text segmentation respectively.

Acknowledgements

This work was supported by the National Science Fund for Distinguished Young Scholars of China (Grant No.62225603).

References

1. Albishre, K., Albathan, M., Li, Y.: Effective 20 newsgroups dataset cleaning. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015, Singapore, December 6-9, 2015 - Volume III. pp. 98–101. IEEE Computer Society (2015)
2. Bonechi, S., Andreini, P., Bianchini, M., Scarselli, F.: Coco_ts dataset: Pixel-level annotations based on weak supervision for scene text segmentation. In: International Conference on Artificial Neural Networks. pp. 238–250. Springer (2019)
3. Bonechi, S., Bianchini, M., Scarselli, F., Andreini, P.: Weak supervision for generating pixel-level annotations in scene text segmentation. *Pattern Recognition Letters* **138**, 1–7 (2020)
4. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T., Zhang, Y.: Sparks of artificial general intelligence: Early experiments with gpt-4 (2023)
5. Chen, J., Huang, Y., Lv, T., Cui, L., Chen, Q., Wei, F.: Textdiffuser: Diffusion models as text painters. In: Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Advances in Neural Information Processing Systems*. vol. 36, pp. 9353–9387 (2023)
6. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(4), 834–848 (2018)
7. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *ECCV* (2018)
8. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 1290–1299 (2022)
9. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems* **34**, 17864–17875 (2021)
10. Ch'ng, C.K., Chan, C.S.: Total-text: A comprehensive dataset for scene text detection and recognition. In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. vol. 1, pp. 935–942. IEEE (2017)
11. Contributors, M.: MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. <https://github.com/open-mmlab/mms Segmentation> (2020)
12. Diaz-Escobar, J., Kober, V.: Natural scene text detection and segmentation using phase-based regions and character retrieval. *Mathematical Problems in Engineering* **2020** (2020)
13. Ji, Y., Chen, Z., Xie, E., Hong, L., Liu, X., Liu, Z., Lu, T., Li, Z., Luo, P.: Ddp: Diffusion model for dense visual prediction. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 21741–21752 (2023)

14. 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)
15. Li, D., Ling, H., Kim, S.W., Kreis, K., Fidler, S., Torralla, A.: Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 21298–21308. IEEE (2022)
16. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th International Conference on Machine Learning. vol. 202, pp. 19730–19742. PMLR (2023)
17. Li, W., He, Y., Qi, Y., Li, Z., Tang, Y.: Fet-gan: Font and effect transfer via k-shot adaptive instance normalization. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 1717–1724 (2020)
18. Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., Bai, X.: Monkey: Image resolution and text label are important things for large multi-modal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26763–26773 (2024)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019)
20. Lyu, P., Bai, X., Yao, C., Zhu, Z., Huang, T., Liu, W.: Auto-encoder guided gan for chinese calligraphy synthesis. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 1095–1100. IEEE (2017)
21. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
22. Mishra, A., Alahari, K., Jawahar, C.: An mrf model for binarization of natural scene text. In: 2011 International Conference on Document Analysis and Recognition. pp. 11–16. IEEE (2011)
23. Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., Luo, Z., Pal, U., Rigaud, C., Chazalon, J., Khlif, W., Luqman, M.M., Burie, J.C., Liu, C.I., Ogier, J.M.: Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification - rrc-mlt. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 1454–1459 (2017)
24. Nguyen, Q., Vu, T., Tran, A., Nguyen, K.: Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. In: Advances in Neural Information Processing Systems. pp. 76872–76892 (2023)
25. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 10674–10685. IEEE (2022)
26. Shen, W., Zhao, K., Jiang, Y., Wang, Y., Bai, X., Yuille, A.: Deepskeleton: Learning multi-task scale-associated deep side outputs for object skeleton extraction in natural images. *IEEE Transactions on Image Processing* **26**(11), 5298–5311 (2017)
27. Su, B., Lu, S., Tan, C.L.: Binarization of historical document images using the local maximum and minimum. In: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems. pp. 159–166 (2010)

28. Tang, Y., Wu, X.: Scene text detection and segmentation based on cascaded convolution neural networks. *IEEE Transactions on Image Processing* **26**(3), 1509–1520 (2017)
29. 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)
30. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* **43**(10), 3349–3364 (2020)
31. Wang, X., Huang, L., Liu, C.: A novel method for embedded text segmentation based on stroke and color. In: *2011 International Conference on Document Analysis and Recognition*. pp. 151–155. IEEE (2011)
32. Wang, Y., Xu, Y., Tsogkas, S., Bai, X., Dickinson, S., Siddiqi, K.: Deepflux for skeletons in the wild. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 5287–5296 (2019)
33. Wang, Y., Xie, H., Wang, Z., Qu, Y., Zhang, Y.: What is the real need for scene text removal? exploring the background integrity and erasure exhaustivity properties. *IEEE Transactions on Image Processing* (2023)
34. Wu, L., Zhang, C., Liu, J., Han, J., Liu, J., Ding, E., Bai, X.: Editing text in the wild. In: *Proceedings of the 27th ACM International Conference on Multimedia*. p. 1500–1508. MM '19, Association for Computing Machinery, New York, NY, USA (2019)
35. Wu, W., Zhao, Y., Chen, H., Gu, Y., Zhao, R., He, Y., Zhou, H., Shou, M.Z., Shen, C.: Datasetdm: Synthesizing data with perception annotations using diffusion models. In: *Advances in Neural Information Processing Systems*. vol. 36, pp. 54683–54695 (2023)
36. Wu, W., Zhao, Y., Shou, M.Z., Zhou, H., Shen, C.: Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In: *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. pp. 1206–1217. IEEE (2023)
37. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: *Advances in Neural Information Processing Systems*. vol. 34, pp. 12077–12090. Curran Associates, Inc. (2021)
38. Xie, J., Li, W., Li, X., Liu, Z., Ong, Y.S., Loy, C.C.: Mosaicfusion: Diffusion models as data augmenters for large vocabulary instance segmentation. *CoRR abs/2309.13042* (2023)
39. Xie, X., Fu, L., Zhang, Z., Wang, Z., Bai, X.: Toward understanding wordart: Corner-guided transformer for scene text recognition. In: *European Conference on Computer Vision*. pp. 303–321. Springer (2022)
40. 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)
41. Xu, X., Qi, Z., Ma, J., Zhang, H., Shan, Y., Qie, X.: Bts: a bi-lingual benchmark for text segmentation in the wild. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 19152–19162 (2022)
42. Ye, H., Kuen, J., Liu, Q., Lin, Z.L., Price, B., Xu, D.: Seggen: Supercharging segmentation models with text2mask and mask2img synthesis. *CoRR abs/2311.03355* (2023)

43. Yu, H., Wang, X., Niu, K., Li, B., Xue, X.: Scene text segmentation with text-focused transformers. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 2898–2907 (2023)
44. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: Computer Vision – ECCV 2020. pp. 173–190. Springer International Publishing (2020)
45. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 3813–3824. IEEE (2023)
46. Zhang, T.Y., Suen, C.Y.: A fast parallel algorithm for thinning digital patterns. *Communications of the ACM* **27**(3), 236–239 (1984)
47. Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J., Barriuso, A., Torralba, A., Fidler, S.: Datasetgan: Efficient labeled data factory with minimal human effort. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 10145–10155. Computer Vision Foundation / IEEE (2021)
48. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. pp. 2881–2890 (2017)
49. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021)