This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Multi-scenario Overlapping Text Segmentation with Depth Awareness

Yang Liu[1*], Xudong Xie[1*], Yuliang Liu[1†], Xiang Bai[1]
[1]Huazhong University of Science and Technology
{yangliu1213,xdxie,ylliu,xbai}@hust.edu.cn

Abstract

Overlapping text poses significant challenges for text-related perception tasks, particularly in open scenes characterized by diverse fonts and visual effects. While existing research has primarily addressed the overlapping problem in documents, its applicability to other scenes remains limited. To bridge this gap, we propose a new task of multi-scenario overlapping text segmentation and introduce a corresponding real dataset in both English and Chinese, spanning various contexts such as printed text, bills, artistic designs, and house numbers. To further enhance the generalization of overlapping text segmentation models, we propose a hierarchical training data synthesis strategy that simulates diverse overlapping patterns across different scenarios. Furthermore, we found that depth maps can provide clear relative position relationships in three-dimensional space, assisting the model in capturing complex overlapping relationships between text instances. Building on this insight, we present a depth-guided decoder that seamlessly integrates image and depth features to capture overlapping interactions. Our proposed model achieves a 5.3% improvement in text mIoU and a 6.4% improvement in overall mIoU compared to existing SOTA methods on our benchmark and SignaTR6k datasets, respectively. Our code and dataset will be released at https://github.com/willpat1213/MOTS.

1. Introduction

In the field of OCR research, deep learning-based text recognition has made remarkable progress, offering effective solutions to challenges such as variations in lighting and shadow effects, blurriness in document images [9, 29], and text distortion caused by bending or tilting in complex scenes [19, 24, 27]. However, current state-of-the-art (SOTA) text recognition methods encounter significant challenges when dealing with overlapping text, where one piece of text partially or fully obscures another within an image. Overlapping text can be categorized into two types based on the degree of overlap: "text-in-text," where the texts significantly overlap, and "sticky text," where only the edges of the texts intersect. As shown in Fig. 1(b), these phenomena frequently occur in various everyday scenarios, such as invoices, receipts, educational documents, advertisements, slogans, artistic designs, house numbers, and other intricate outdoor settings.

(a) SignaTR6K

(b) MOT (ours)

Figure 1. The overlapping text examples of SignaTR6K and MOT (ours).

Previous studies [11, 14, 30] have recognized the critical need for tackling overlapping text in real-world applications. These works employ various segmentation models to deconstruct overlapping text into regular text. However, these studies have primarily concentrated on addressing overlapping patterns in document-specific scenarios, as illustrated in Fig. 1(a), without extending their focus to multi-scenario text overlap challenges. This paper aims to bridge this research gap by proposing a novel segmentation framework specifically designed for comprehensive overlapping text analysis across diverse real-world scenarios.

Unlike document scenario overlapping text segmentation, multi-scenario overlapping text segmentation presents two unique challenges, as outlined below: (1) Lack of overlapping text data in multiple scenarios. Previous overlapping text datasets primarily focus on specific document-based contexts, such as educational papers, historical archives and legal documents. These scenarios typically display similar overlapping patterns, which can be handled

* Co-first authors; † Corresponding author

by straightforward expert models. However, this limits the model's ability to generalize across diverse scenarios. (2) Complex and Diverse Challenges in multi-scenario. Multi-scenario overlapping text segmentation is especially difficult due to diverse backgrounds, dynamic text variations (such as changes in size, color, and font), and irregular occlusions from design elements or shifting perspectives. Furthermore, the arrangement and layout of text can vary significantly across scenes, further increasing the unpredictability of the segmentation task.

To address the above challenges, we propose the following solution:

(1) Real dataset MOT and synthetic strategy HSOT. We aim to develop a universal solution for overlapping text images applicable to multiple scenarios. To this end, we manually annotated a **M**ulti-scenario **O**verlapping **T**ext dataset **(MOT)** which contains 1,250 images from diverse and complex scenes, such as forms, bills, and open street scenes. The dataset includes various text types, including printed, handwritten, and artistic fonts. However, relying solely on real-world datasets does not fully enhance the model's generalization capabilities. To overcome this, we designed a **H**ierarchical **S**ynthetic **O**verlapping **T**ext strategy **(HSOT)** to simulate the sources and overlap patterns found in real datasets. Specifically, we classified the data into two categories: document-related and scene-related. For document scenarios, we extended SynthTiger [41], enabling it to generate overlapping text in a more flexible and realistic document style. For scene-related scenarios, we developed an end-to-end optimization model based on ControlNet [44], which generates overlapping text images that closely resemble real-world scene compositions. Experimental results show that pre-train with synthetic data from our proposed HSOT strategy achieves better performance than pre-train with data from a single pipeline.

(2) Depth-guided decoder. Depth maps represent the three-dimensional structure of the natural world, and we found that they can map overlapping text images into 3D space. In this space, the front-back and occlusion relationships between the overlapping texts become more distinct. This insight motivated us to investigate whether depth maps can serve as an effective auxiliary modality to help the model capture the features of overlapping texts. Based on this insight, we propose a depth-guided decoder, which consists of two core stage: a feature enhancement stage and a Depth-guided Cross-Attention stage. The feature enhancement refines the depth features to extract fine-grained depth information, while the Depth-guided Cross-Attention mechanism uses these depth features to guide the model in distinguishing between different texts. This approach enhances the model's ability to understand overlapping relationships, particularly for depth-in-occluded texts.

In summary, this paper makes the following advantages:

- We present a challenging task: multi-scenario overlapping text segmentation and construct a comprehensive dataset (MOT) to benchmark the performance.
- We design a synthetic data generation strategy (HSOT) to improve the model's generalization by effectively simulating complex real-world overlapping scenes.
- We realize the value of depth information for overlapping text images and design a depth-guided decoder that uses depth maps to enhance the recognition of overlapping texts. This is achieved by refining depth features and employing a depth-guided cross-attention mechanism, significantly improving the model's understanding of occlusion relationships.
- Our model achieves (5.3% Text mIoU) and (6.4% mIoU) improvements on MOT and SignaTR6k datasets, respectively. We validate the effectiveness of the HSOT-generated dataset and depth-guided decoder.

2. Related Work

Since our method focuses on addressing the segmentation problem of multi-scenario overlapping text, we briefly review the recent advances in the following areas: overlapping text segmentation dataset and model.

2.1. Overlapping Text Segmentation Dataset

Based on the review of previous work on overlapping text in the preceding section, it is evident that past research has predominantly focused on document domains, with the proposed datasets reflecting this concentration. OverlapText-500 [14] is a dataset collected from financial documents and math exercise sheets, encompassing Arabic numerals, mathematical symbols, uppercase and lowercase English letters, and 3,000 commonly used Chinese characters. WGM-SYN [30] is a synthesized dataset derived from scanned images of historical archives, while SignaTR6K [11] originates from scanned images of legal documents. Both datasets focus on the overlap between handwritten and printed texts within document scenarios characterized by black text on a white background. In summary, previous work has focused on specific types of documents within a single scenario. These datasets often share similar backgrounds, fonts, colors, and overlapping patterns. However, there is a lack of a multi-scenario overlapping text dataset that reflects more general cases.

2.2. Overlapping Text Segmentation Model

Text Segmentation. Text segmentation involves predicting detailed masks for text in images with varying scene complexities. Traditional methods, such as thresholding techniques [22, 23, 26, 28] and low-level feature-based approaches [18, 28, 31], have faced challenges when applied to images with intricate colors and textures. With the advent of deep learning, models like SMANet [2] have leveraged

encoder-decoder structures and multi-scale attention mechanisms to improve segmentation performance. Further advancements, such as TextFormer [32] and TexRNet [36], have introduced hierarchical segmentation frameworks and fine-grained annotations to enhance text detail perception and segmentation accuracy. To address the lack of Chinese text in segmentation datasets, the BTS dataset [37] was introduced, while PGTSNet focuses on using pre-trained detection models to constrain segmentation to detected text regions. Addressing the challenge of variable stroke shapes in artistic text, WASNet [35] incorporates a transformer decoder with layer-wise momentum queries to preserve attention on special-shaped stroke regions. Similarly, EAFormer [42] integrates edge information through symmetric cross-attention submodules to guide the model's focus on text edges from the outset. Additionally, Hi-SAM [40] presents a unified framework for hierarchical text segmentation and layout analysis, based on the strong segmentation fundation model SAM [16].

Overlapping Text Segmentation. Early studies on overlapping text primarily focused on simple overlap scenarios within document scenes. These approaches commonly utilized Hidden Markov Models (HMMs) [12] and Support Vector Machines (SVMs) [10], following the binary classification paradigm typical of conventional text segmentation. Text was treated as instances, requiring models to learn abstract representations of overlaps. As a result, the separations were relatively coarse and often plagued by residual noise, posing challenges for downstream recognition tasks. RecycleNet [14] introduced a two-stage, end-to-end trainable instance segmentation network to help improve separation quality. WGM-MOD [30] and MFM [11] apply semantic segmentation to separate overlapping handwritten and printed texts; while WGM-MOD lacks explicit modeling of overlaps, MFM treats them as a separate category, assigning pixels to individual texts in post-processing. Both focus mainly on historical and legal documents, limiting their generalization to other scenarios. These works reflect ongoing research efforts to address specific application limitations in handling overlapping text.

3. MOT Dataset and HSOT Strategy

To address the issue of extensive overlapping text in both document and natural scenarios, we introduce a real dataset (MOT) and a synthetic strategy (HSOT). Relying on them can help the community achieve more valuable work in future overlapping text research.

3.1. MOT Dataset

3.1.1. Data Collection and Annotation

Previous overlapping text datasets primarily focus on document-based scenarios. To address this, our dataset includes a diverse range of scenarios with significant domain gaps, such as printed text, form tickets, art designs, door signs, and other forms of expression featuring overlapping text. To collect images of overlapping text in diverse scenes, we use search engines like Google and Baidu to search for relevant keywords related to scenarios where overlapping text is likely to appear, examples include: (1) Handwritten or printed text: Content written on a printed document; (2) Street view: Overlapping text created by the interaction of transparent foreground elements and background text on billboards; (3) Digital file processing: Overlapping text due to printing errors or anti-counterfeiting features; (4) Art design: In certain advertisements and art designs, multiple texts often overlap as part of the design elements such as some similar examples are in the Wordart dataset [34, 35].

To ensure image privacy, our data collection is sourced legally through public channels. Following this, multiple rounds of manual screening were conducted to remove poor-quality images that are indistinguishable to the human eye, as well as images that do not contain overlapping text. As a result, we collected 1,250 images containing 2,620 text instances. To facilitate model training and testing, we divided the dataset into two equal parts: 630 images for training and 620 images for testing.

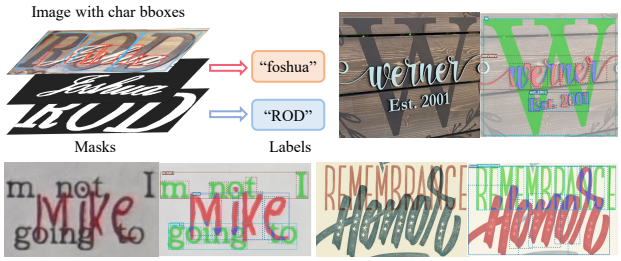


Figure 2. Example annotations from the proposed MOT dataset. The top-left image displays the original image with character-level bounding boxes. The pixel-level character masks for each instance are organized into overlapping layers, ordered from foreground to background, with each instance corresponding to a unique label.

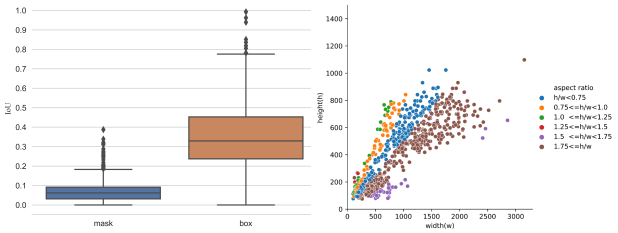
Our dataset includes detailed, multi-level annotations to support future research in related fields. After defining the labeling format, we labeled different datasets and conducted cross-validation after annotation to minimize the impact of human factors on label quality. Labeling experts identified the text in overlapping areas based on their prior knowledge of the characters. Due to the richness of the information in our dataset, the labeling process was complex and took six months, including significant time spent on data cleaning. The annotation details for the dataset are shown in Fig. 2. Our annotations include three levels: pixel-level text masks, character-level bounding boxes, and text-level labels, along

https://www.google.com/
https://www.baidu.com/

with other relevant information. To represent the overlapping relationships between text instances, we annotate the text masks from foreground to background, with each occlusion layer identified by its corresponding file name. This implicit annotation provides valuable information for future researchers, aiding in the separation of overlapping text based on these relationships.

3.1.2. Data Statistics

We aim to ensure that the degree of overlap between text instances is not overly extreme—that is, neither completely non-overlapping (as in standard multi-line text) nor entirely overlapping (making it impossible for humans to predict or annotate). To verify the rationality and predictability of the MOT dataset, we analyzed the overlap patterns among text instances in each image. As shown in Fig. 3(a), we present the bbox plots for mask IoU and bbox IoU within the dataset. Mask IoU represents the overlap of strokes between different text instances. A higher mask IoU indicates more severe occlusion of text strokes, making them harder to discern. The results show that mask IoU is primarily distributed around 0.2, with some hard cases not exceeding 0.5, demonstrating that there are no instances in the dataset that are impossible to predict or annotate. Additionally, bbox IoU reflects the overlap between bounding boxes. The bbox IoU distribution reveals a range of overlapping patterns, including cases with slight overlaps (IoU near 0) and more challenging text-in-text overlaps (IoU approaching 0.8). Fig. 3(b) further illustrates the distribution of image sizes in the dataset, which includes images with varying aspect ratios and resolutions.



(a) Distribution of Mask IoU and Bbox IoU (b) Distribution of Image size

Figure 3. Statistical analysis on MOT dataset.

3.2. HSOT Strategy

As our overlapping text algorithm is applicable to all scenarios and existing methods center around processing document-type situations, we propose a Hierachical Synthetic Overlapping Text (HSOT) construction pipeline, as illustrated in the Fig. 4.

3.2.1. Document-related

For document-related scenarios, as shown in Fig. 4(a), we synthesize overlapping text images using the non-deep

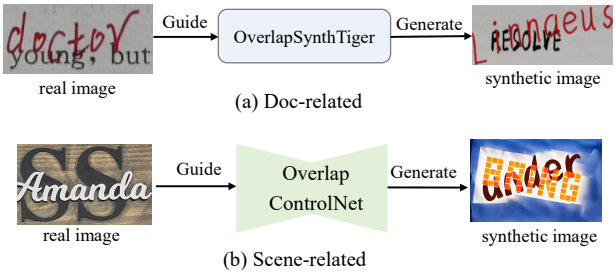


Figure 4. The proposed HSOT construction pipeline.

learning data synthesizer SynthTiger. The process involves the following steps: first, we update SynthTiger's synthesis pipeline to enable the superposition of multiple texts. Then, we sample parameters such as text color, font, and perspective from a real dataset to ensure they follow the same distribution as in the real world. The background images are directly extracted from the real dataset to maintain a high degree of similarity with real data. This approach ensures that the distribution and characteristics of the synthesized data closely match those of real-world documents.

3.2.2. Scene-related

For scene-related scenarios, as shown in Fig. 5. we design a text-to-image generation model that creates aligned overlapping text images by using masks and input prompts.

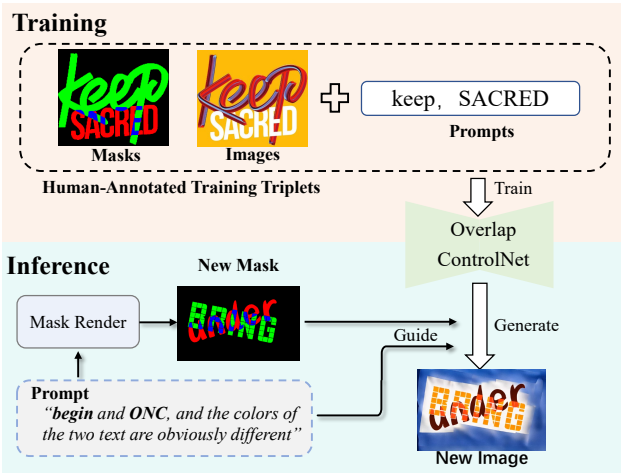


Figure 5. The proposed OverlapControlNet pipeline.

Training Pipeline. During training, we first extract word-level information from MOT dataset images and use it as prompts to create training triples in the form of ¡prompt, mask, image¿. The masks are colorized to help the model better distinguish different text regions. We then train a ControlNet model to map the prompt and mask to a generated image, ensuring the outline of overlapping text aligns precisely with the mask. This approach enables the model

to generate diverse, pixel-aligned images.

Inference Pipeline. During inference, we generate diverse synthetic text masks using our Mask Render. For each mask, we randomly select two words (4 to 10 letters) based on the letter distribution in the MOT dataset, apply a random rotation between -30° and 30°, and position the words within the image boundaries. We use 250 scene fonts and apply an affine transformation to introduce skew and distortion. Finally, we generate the prompt by incorporating the text information from the mask into a template, which is used to create the overlapping text image. After constructing the synthetic text mask and prompt, we use the trained ControlNet to generate the final text image. Specifically, the final image is produced by feeding the prompt and mask into the ControlNet, which generates the corresponding overlapping text image.

4. Methodology

In this section, we first give an overview of our proposed model, and then provide a detailed introduction to our proposed Depth-guided decoder.

4.1. Overall

As shown in the Fig. 7, we propose a multimodal overlapping text segmentation method using Mask2Former [7] as the meta-architecture. The core component is the depth-guided decoder, which mainly consists of two stages: deep feature fusion and depth-guided cross-attention. To begin, we provide a brief overview of the segmentation process.

Extract Features. First, Given the input overlapping text image $\mathbf{X} \in \mathbb{R}^{3\times H\times W}$, a pre-trained monocular depth estimation model [38, 39] is employed to compute the corresponding depth map. Then, we employ the same encoder for depth and image, but without weight sharing, to extract depth features $\mathbf{D} = \{\mathbf{D_1}, \mathbf{D_2}, \mathbf{D_3}, \mathbf{D_4}\}$ and image features $\mathbf{I} = \{\mathbf{I_1}, \mathbf{I_2}, \mathbf{I_3}, \mathbf{I_4}\}$ separately, where $\mathbf{D}_i$ and $\mathbf{I}_i$ represents the features at the i-th layer of the ResNet-like [13] backbone.

Multi-Modal Feature Fusion. For image features $\mathbf{I}$, we use a pixel decoder [6] to process the low-resolution features from the backbone, enhancing their expressive capability. This results in a high-resolution mask feature $\mathbf{M}_I$ and a multi-scale image feature $\hat{\mathbf{I}} = \{\hat{\mathbf{I}}_1, \hat{\mathbf{I}}_2, \hat{\mathbf{I}}_3\}$ after fusion. For depth features, we design a lightweight and fully convolutional process (see Sec. 4.2) that generates a high-resolution depth mask feature $\mathbf{M}_D$ and multi-scale depth. After enhancement, the features from both modalities are fed into the Depth-guided Cross Attention module (see Algorithm 1) for further fusion. This process yields the final enhanced depth aware mask feature M_F and depth aware multi-scale features $\hat{\mathbf{F}} = \{\hat{\mathbf{F}}_1, \hat{\mathbf{F}}_2, \hat{\mathbf{F}}_3\}$ for segmentation.

Decoding. Following the Deformable DETR Decoder [3, 45] in Mask2Former, the enhanced multi-scale fusion features are sparsely sampled using the deformable attention mechanism. The mask head generates a binary mask by decoding the per-pixel embeddings from the pixel decoder and the learnable queries from the Deformable DETR Decoder.

Figure 6. Visualization of depth map.

4.2. Depth-guided Decoder

Overlapping text segmentation differs from conventional text segmentation due to challenges such as feature similarity between overlapping texts and occlusion between different text instances. Standard feature representations are insufficient to address these issues. As shown in Fig. 6, we observed that the depth map provides valuable three-dimensional spatial information, enhancing feature representation and improving the handling of occlusion relationships. To leverage this, we introduced depth modality information and designed a Depth-guided Decoder (shown in the gray area in Fig. 7) to integrate depth information into the segmentation process.

To leverage text-level depth relationships for segmentation guidance, we implement a feature enhancement stage within our depth-guided decoder. This sub-process employs a lightweight and fully convolutional method to extract hierarchical features: First, multi-scale depth features $\mathbf{D}$ are transformed to a unified scale using a 1x1 convolution, resulting in $\hat{\mathbf{D}}$. The largest-scale feature $\mathbf{D}_4$ is then processed with a 3x3 convolution and added to $\hat{\mathbf{D}}_3$ to generate the high-resolution depth mask feature $\mathbf{M}_D$, forming a structure similar to FPN [17]. This can be expressed as:

$$\hat{\mathbf{D}}_k = \mathrm{Conv}_{1\mathrm{x}1}(\mathbf{D}_k), \quad k \in [1, 3], \tag{1}$$

$$\mathbf{M}_D = \mathrm{Conv}_{3\mathrm{x}3}(\mathbf{D}_4) + \hat{\mathbf{D}}_3, \tag{2}$$

4.2.1. Depth-guided Cross Attention

As seen the Algorithm 1, **Stage 1:** the module first concatenates the mask features $\mathbf{M}_I, \mathbf{M}_D$ from two modalities into a global mask feature $\mathbf{M}^G$ Simultaneously, it fuses multi-scale features $\{\hat{\mathbf{I}}_k, \hat{\mathbf{D}}_k\}_{k=1}^3$ across scales by concatenation to form global multi-scale features $\hat{\mathbf{F}}^G$. This step ensures the integration of both local mask information and multi-scale context from different modalities.

Stage 2: Depth features $\hat{\mathbf{D}}$ are normalized using LayerNorm [1] and treated as queries, while the global multi-scale features $\hat{\mathbf{F}}^G$ serve as both keys and values. Multi-head attention is then computed to dynamically weight feature interactions, prioritizing depth-informed regions for precise

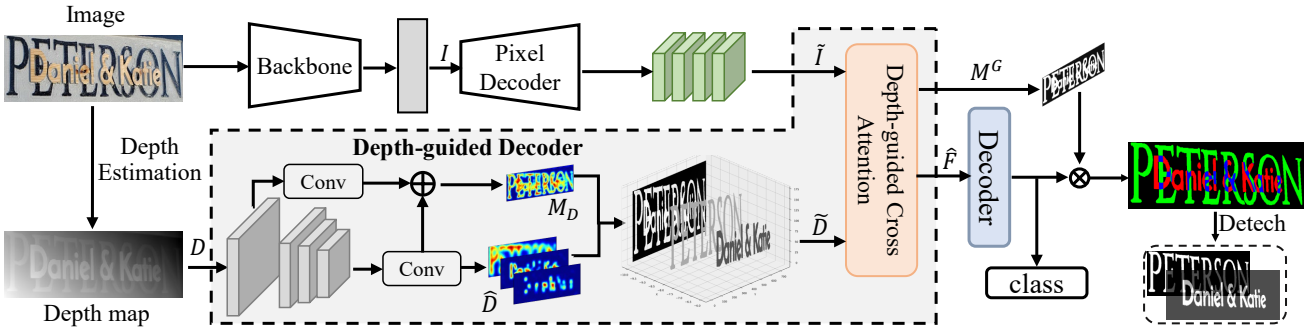


Figure 7. The overall architecture of our model.

Algorithm 1 Depth-guided Cross Attention

Input: Multi-scale feat. $\{\hat{\mathbf{I}}_k, \hat{\mathbf{D}}_k\}_{k=1}^{3}$, mask feat. $\mathbf{M}_I$, $\mathbf{M}_D$
Output: Fused feat. $\{\hat{\mathbf{F}}_i\}_{i=1}^{3}$, text mask $\mathbf{M}^G$

Stage 1: Feature Fusion
  $\mathbf{M}^G \leftarrow \text{Concat}(\mathbf{M}_I, \mathbf{M}_D)$
  $\hat{\mathbf{F}}_i^G \leftarrow \text{Concat}(\hat{\mathbf{I}}_i, \hat{\mathbf{D}}_i),\ \forall i \in [1, 3]$
Stage 2: Cross Attention
for $i = 1$ **to** 3 **do**
    $\mathbf{Q}_i \leftarrow \text{LN}(\hat{\mathbf{D}}_i)$ ▷ Layer normalization
    $(\mathbf{K}_i, \mathbf{V}_i) \leftarrow (\hat{\mathbf{F}}_i^G, \hat{\mathbf{F}}_i^G)$ ▷ Shared features
    $\hat{\mathcal{V}}_i \leftarrow \text{CA}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i)$
end for
Stage 3: Feature Aggregation
for $i = 1$ **to** 3 **do**
    $\hat{\mathcal{V}}_i^D \leftarrow \hat{\mathbf{D}}_i + \hat{\mathcal{V}}_i$ ▷ Residual addition
    $\hat{\mathbf{F}}_i \leftarrow \text{FFN}(\text{LN}(\hat{\mathcal{V}}_i^D)) + \hat{\mathcal{V}}_i^D$
end for
return $\mathbf{M}^G$, $\{\hat{\mathbf{F}}_i\}_{i=1}^{3}$

fusion. This stage emphasizes depth-driven feature alignment between modalities.

Stage 3: The output of the cross-attention is combined with original depth features through a residual connection. This aggregated result undergoes two consecutive transformations: (1) a feed-forward network (FFN) followed by LayerNorm, and (2) another residual addition. This dual-path design preserves feature integrity while refining fused representations through non-linear modeling. The module ultimately outputs the global mask feature $\mathbf{M}^G$ and the refined fused features$\{\hat{\mathbf{F}}_i\}_{i=1}^{3}$, which capture rich multi-modal correlations guided by depth information.

4.3. Loss Functions

We utilize cross-entropy loss for the classification of different semantics and dice loss [21] for mask prediction. Given that overlaps are relatively rare, we further incorporate focal loss to address the long-tail distribution issue:

$$\mathcal{L} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}, \tag{3}$$

Following the baseline's settings, we set $\lambda_{\text{cls}} = 2$, $\lambda_{\text{ce}} = \lambda_{\text{dice}} = 5$ to increase the model's focus on overlapping areas (hard class). The higher weights on λ_{ce} and λ_{dice} specifically amplify the contribution of pixel-wise classification and region overlap accuracy in these challenging regions.

5. Experiments

5.1. Implementation Details

The proposed model is implemented in PyTorch and evaluated on a server with 8 RTX 3090 GPUs. We use the AdamW [20] optimizer with a weight decay of 0.05 and apply a poly [4] learning rate schedule, starting with an initial learning rate of 10^{-4}. The batch size is set to 4, and the number of iterations is 60,000. For each experiment, we calculate the IoU for three classes: Occlusion (Occ), Occluded (Occd), and Overlap (Ov). Additionally, we compute the mean IoU as the average of the IoUs for these three classes.

5.2. Experimental Results

5.2.1. Results on the MOT dataset

To evaluate the effectiveness of our overlapping text segmentation method, we conducted extensive experiments on the MOT dataset comparing against SOTA semantic/text segmentation models. Experimental results in Tab. 1 demonstrate our method achieves superior performance across all metrics, outperforming existing approaches. Notably, our pre-trained version (Ours (pre-train)) shows significant improvement with 75.53% mIoUText - a 4.97% gain over the non-pretrained variant, confirming the value of synthetic and multi-source data for domain adaptation. The most substantial improvement appears in IoUov (60.66%) improved by 7.84%, indicating enhanced capability in processing occluded and low-visibility text instances.

5.2.2. Results on the SignaTR6K dataset

Additionally, we validated the effectiveness of our proposed model on another overlapping text dataset, SignaTR6K. As shown in Tab. 2, our model, without requiring pre-training,

Model	IoU_{Occ}	IoU_{Occd}	IoU_{Ov}	$mIoU_{Text}$
Unet [25]	80.23	65.68	40.73	62.21
Deeplab v3+ [5]	83.16	71.20	49.25	67.87
OCRNet [43]	81.04	68.47	47.75	65.75
Segformer [33]	83.59	74.11	49.27	68.99
Maskformer [6]	83.47	70.26	51.40	68.38
TexRNet [36]	84.22	73.16	49.25	68.88
EAFormer [42]	83.78	74.23	50.47	69.06
WASNet [35]	84.81	74.35	53.12	70.76
Mask2former [7]	84.72	73.29	52.82	70.28
Ours	85.17	77.54	54.93	72.55
Ours (pre-train)	**86.77**	**79.17**	**60.66**	**75.53**

Table 1. Performance comparison with other segmentation methods on MOT dataset. "pre-train" indicates that the model was firstly trained on HSOT-generated data and MOT datasets and subsequently fine-tuned on the MOT dataset.

outperforms the MFM model in its best configuration on SignaTR6K across all IoU categories, including printed text (PT), handwritten text (HT), and background (BG). The mean IoU improved by 6.35%. Notably, the performance gain in the occluded category (PT) reaches 17.41%, demonstrating that our model deeply understands overlapping phenomena and can effectively detect occluded text.

Model	Training & Testing Data	PT (%)
MFM [11]	SignaTR6K	73.05
ours	SignaTR6K	**90.46**

Table 2. Performance comparison with other segmentation methods on SignaTR6K dataset.

5.3. Ablation Study

As shown in Tab. 3, we evaluate the effectiveness of our proposed dataset synthesis strategy and model design. First, the Hierarchical Synthetic Overlapping Text (HSOT) strategy explicitly synthesizes multi-scale overlapping text patterns during pre-training, enabling the model to learn the inherent spatial relationships between overlapping characters. This data-centric approach significantly improves the separation of overlapping regions, boosting IoU_{OV} by 6.74% and $mIoU_{Text}$ by 3.90%.

Furthermore, the Depth-guided Decoder (D-Decoder) leverages depth information to explicitly guide the segmentation of occluded text regions. By dynamically integrating depth-aware features, it refines boundary localization for partially obscured text, leading to a 3.19% improvement in IoU_{OD}. The synergy between HSOT and D-Decoder achieves a final $mIoU_{Text}$ of 75.53%, with an additional 1.35% gain over the pre-trained baseline. These results underscore HSOT's unique ability to resolve overlapping text ambiguity through data synthesis and D-Decoder's strength in utilizing depth cues for occlusion reasoning, together advancing SOTA performance in multi-scenario overlapping text segmentation tasks.

Methods	IoU_{Occ}	IoU_{Occd}	IoU_{Ov}	$mIoU_{Text}$
baseline [7]	84.72	73.29	52.82	70.28
+ HSOT	86.00	76.98	59.56	74.18
+ D-decoder	**86.77**	**79.17**	**60.66**	**75.53**

Table 3. Ablation study on our proposed datasets and modules.

To further explore how the data synthesis strategy HSOT enhances model performance, we present a detailed ablation study on HSOT, including: the scale of synthetic data and the scenario of synthetic data.

Scaling Up of Synthetic Data. To ensure sufficient pre-training data for effective model training, we investigated model performance with varying amounts of pre-training data. To maintain consistency in data distribution, we kept the proportion of different scenes fixed while only adjusting the volume of synthetic data. As shown in Tab. 4, we observed that increasing the amount of pre-training synthetic data led to a steady improvement in model performance, which also demonstrates the effectiveness of our HSOT strategy. Additionally, when the data scale exceeded 280k, model performance approached saturation. Therefore, we selected 280k synthetic data as the pre-training data for other experiments.

Scaling up	IoU_{Occ}	IoU_{Occd}	IoU_{Ov}	$mIoU_{Text}$
70k	86.00	78.16	60.14	74.77
140k	86.16	77.77	60.66	74.86
210k	86.50	78.98	59.67	75.05
280k	86.77	**79.17**	60.66	75.53
350k	**86.81**	79.12	**60.71**	**75.59**

Table 4. Ablation Study on Synthesis Data Scale.

Scenario of Synthetic Data. A key feature of the HSOT synthesis strategy is the use of different synthesis methods tailored to specific scenarios—doc-related and scene-related. This approach enhances the model's generalization ability across diverse scenarios and better supports multi-scenario situations in the MOT dataset. To verify this, we conducted separate evaluations based on the scenario. As shown in the Tab. 5, r1 was fine-tuned using only real data, r2 and r3 were pre-trained with scene-related and doc-related data, respectively, and r4 was pre-trained using both datas. The results show that using either scene-related or

document-related synthetic data yields varying degrees of performance improvement, while combining both datasets simultaneously leads to a further enhancement in model performance.

	Dataset			IoU_{Occ}	IoU_{Occd}	IoU_{Ov}	$mIoU_{Text}$
	Real	Doc*	Scene*				
r1	✓			85.17	77.54	54.93	72.55
r2	✓		✓	85.60	77.72	56.41	73.24
r3	✓	✓		85.86	78.90	60.64	75.13
r4	✓	✓	✓	**86.77**	**79.17**	**60.66**	**75.53**

Table 5. Ablation Study of the HSOT Strategy Across Various Scenarios. * denotes document-related and scene-related data generated using the HSOT strategy, respectively.

5.4. Discussions

Downstream Applications. To evaluate the effectiveness of our approach, we conducted a text recognition experiment on the MOT test set and compared our method with SOTA recognizers. Notably, the MOT dataset includes images with overlapping multi-line text, which can negatively impact recognition performance. To isolate the influence of multi-line text, we excluded such images and curated a subset of 556 images from the test set for this experiment. The results (Tab. 6) reveal distinct outcomes across the three categories of overlapping text handling strategies. Unlike one-stage recognition methods, our approach employs a "segment-then-recognize" pipeline: overlapping text is first decoupled and binarized into conventional single-line text images, enabling standard recognizers to operate with high accuracy. Our method outperforms the one-stage baseline significantly in both accuracy and edit distance metrics. Furthermore, attention-based recognizers achieved superior results over CTC-based models on images with minimal overlap. This disparity arises because CTC-based models are inherently disadvantaged when processing overlapping text sequences, whereas attention-based mechanisms exhibit greater robustness to mild overlap.

Category	Model	Acc.↑	Edit Dis.↑
Recognizer	SVTRv2 [8]	4.04	37.01
	MAERec [15]	8.72	48.65
Segm-based	Ours	**78.15**	**86.95**

Table 6. Performance of ours in enhancing the text recognition task.The recognizers process the overlapping image directly. SVTRv2 is used as the recognizer for the segm-based methods.

Advantages of Depth-Guided Model in Training. The training loss curves (Fig. 8) reveal that ours achieves both faster convergence and a lower final loss plateau compared to baseline. This stems from the depth decoder's ability to disentangle occluded regions through explicit 3D spatial priors. While conventional models rely on ambiguous RGB cues to infer overlaps, depth maps explicitly model layer ordering (e.g., foreground vs. background), sharpening the model's focus on physically plausible occlusion boundaries.

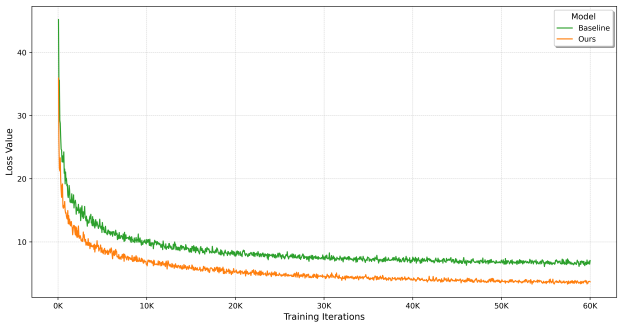


Figure 8. Training loss comparison with baseline.

Limitations. In cases where overlapping texts share nearly identical visual features (e.g., same font, color, and size), the depth-guided decoder may fail to prioritize subtle boundary cues. This is particularly evident in "text-in-text" overlaps with high mask IoU (>0.5), where even human annotators face ambiguity. We may propose some solutions to this problem in future work.

6. Conclusion

The paper addresses the critical challenge of overlapping text segmentation across diverse real-world scenarios, a problem underexplored in prior research focused primarily on document contexts. We propose a novel task, multi-scenario overlapping text segmentation, and introduce two key contributions to advance the field. First, we collected and annotated a real multi-scenario text dataset MOT for benchmarking the performance, complemented by our HSOT synthesis strategy that enhances model generalization through scenario-specific text overlap simulation. Second, we pioneer the use of depth-guided 3D spatial reasoning for overlapping text analysis. Our Depth-guided Decoder effectively leverages depth maps to resolve occlusion relationships, enabling precise segmentation of intertwined texts across variable layouts and backgrounds. Extensive experiments validate the superiority of our model, highlighting the significance of synthesizing multi-scenario training data and integrating depth-aware feature fusion. These advancements are crucial for a wide range of applications, from invoice parsing to outdoor signage recognition.

Acknowledgements

This work was supported by the NSFC (62206104, 62206103 and 62225603).

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[2] Simone Bonechi, Monica Bianchini, Franco Scarselli, and Paolo Andreini. Weak supervision for generating pixel–level annotations in scene text segmentation. *Pattern Recognition Letters*, 138:1–7, 2020. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 5

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 6

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 7

[6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 5, 7

[7] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 5, 7

[8] Yongkun Du, Zhineng Chen, Hongtao Xie, Caiyan Jia, and Yu-Gang Jiang. Svtrv2: Ctc beats encoder-decoder models in scene text recognition. *arXiv preprint arXiv:2411.15858*, 2024. 8

[9] Nicolas Dutly, Fouad Slimane, and Rolf Ingold. Phti-ws: a printed and handwritten text identification web service based on fcn and crf post-processing. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, pages 20–25. IEEE, 2019. 1

[10] Bala Mallikarjunarao Garlapati and Srinivasa Rao Chalamala. A system for handwritten and printed text classification. In *2017 UKSim-AMSS 19th International Conference on Computer Modelling & Simulation (UKSim)*, pages 50–54. IEEE, 2017. 3

[11] S Gholamian and A Vahdat. Handwritten and printed text segmentation: a signature case study (2023), 2023. 1, 2, 3, 7

[12] Jinhong Katherine Guo and Matthew Y Ma. Separating handwritten material from machine printed text using hidden markov models. In *Proceedings of sixth international conference on document analysis and recognition*, pages 439–443. IEEE, 2001. 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[14] Yiqing Hu, Yan Zheng, Xinghua Jiang, Hao Liu, Deqiang Jiang, Yinsong Liu, Bo Ren, and Rongrong Ji. Recyclenet: An overlapped text instance recovery approach. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1102–1110, 2021. 1, 2, 3

[15] Qing Jiang, Jiapeng Wang, Dezhi Peng, Chongyu Liu, and Lianwen Jin. Revisiting scene text recognition: A data perspective. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20543–20554, 2023. 8

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3

[17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 5

[18] Xiaoqing Liu and Jagath Samarabandu. Multiscale edge-based text extraction from complex images. In *2006 IEEE International Conference on Multimedia and Expo*, pages 1721–1724. IEEE, 2006. 2

[19] Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9809–9818, 2020. 1

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[21] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 6

[22] Wan Azani Mustafa and Mohamed Mydin M Abdul Kader. Binarization of document image using optimum threshold modification. In *Journal of Physics: Conference Series*, page 012022. IOP Publishing, 2018. 2

[23] Nobuyuki Otsu et al. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975. 2

[24] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 569–576, 2013. 1

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 7

[26] Jaakko Sauvola and Matti Pietikäinen. Adaptive document image binarization. *Pattern recognition*, 33(2):225–236, 2000. 2

[27] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2035–2048, 2018. 1

[28] Bolan Su, Shijian Lu, and Chew Lim Tan. Binarization of historical document images using the local maximum and minimum. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 159–166, 2010. 2

[29] Nishant Subramani, Alexandre Matton, Malcolm Greaves, and Adrian Lam. A survey of deep learning approaches for ocr and document understanding. *arXiv preprint arXiv:2011.13534*, 2020. 1

[30] Mahsa Vafaie, Oleksandra Bruns, Nastasja Pilz, Jörg Waitelonis, and Harald Sack. Handwritten and printed text identification in historical archival documents. In *Archiving Conference*, pages 15–20. Society for Imaging Science and Technology, 2022. 1, 2, 3

[31] Quang Nhat Vo, Soo Hyung Kim, Hyung Jeong Yang, and Gueesang Lee. Binarization of degraded document images based on hierarchical deep supervised network. *Pattern Recognition*, 74:568–586, 2018. 2

[32] Xiaocong Wang, Chaoyue Wu, Haiyang Yu, Bin Li, and Xiangyang Xue. Textformer: component-aware text segmentation with transformer. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1877–1882. IEEE, 2023. 3

[33] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. 7

[34] Xudong Xie, Ling Fu, Zhifei Zhang, Zhaowen Wang, and Xiang Bai. Toward understanding wordart: Corner-guided transformer for scene text recognition. In *European Conference on Computer Vision*, pages 303–321. Springer, 2022. 3

[35] Xudong Xie, Yuzhe Li, Yang Liu, Zhifei Zhang, Zhaowen Wang, Wei Xiong, and Xiang Bai. Was: Dataset and methods for artistic text segmentation. In *European Conference on Computer Vision*, pages 237–254. Springer, 2025. 3, 7

[36] Xingqian Xu, Zhifei Zhang, Zhaowen Wang, Brian Price, Zhonghao Wang, and Humphrey Shi. 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*, pages 12045–12055, 2021. 3, 7

[37] Xixi Xu, Zhongang Qi, Jianqi Ma, Honglun Zhang, Ying Shan, and Xiaohu Qie. Bts: a bi-lingual benchmark for text segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19152–19162, 2022. 3

[38] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 5

[39] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2025. 5

[40] Maoyuan Ye, Jing Zhang, Juhua Liu, Chenyu Liu, Baocai Yin, Cong Liu, Bo Du, and Dacheng Tao. Hi-sam: Marrying segment anything model for hierarchical text segmentation. *arXiv preprint arXiv:2401.17904*, 2024. 3

[41] Moonbin Yim, Yoonsik Kim, Han-Cheol Cho, and Sungrae Park. Synthtiger: Synthetic text image generator towards better text recognition models. In *International conference on document analysis and recognition*, pages 109–124. Springer, 2021. 2

[42] Haiyang Yu, Teng Fu, Bin Li, and Xiangyang Xue. Eaformer: scene text segmentation with edge-aware transformers. In *European Conference on Computer Vision*, pages 410–427. Springer, 2025. 3, 7

[43] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020. 7

[44] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2

[45] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable {detr}: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. 5