IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 May 2023, accepted 25 June 2023, date of publication 4 July 2023, date of current version 11 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3292264

APPLIED RESEARCH

Scene Text Segmentation via Multi-Task Cascade Transformer With Paired Data Synthesis

QUANG-VINH DANG AND GUEE-SANG LEE
Department of Artificial Intelligence Convergence, Chonnam National University, Buk-gu, Gwangju 61186, South Korea

Corresponding author: Guee-Sang Lee (gslee@jnu.ac.kr)

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2018R1D1A3B05049058.

ABSTRACT The scene text segmentation task provides a wide range of practical applications. However, the number of images in the available datasets for scene text segmentation is not large enough to effectively train deep learning-based models, leading to limited performance. To solve this problem, we employ paired data generation to secure sufficient data samples for text segmentation via Text Image-conditional GANs. Furthermore, existing models implicitly model text attributes such as size, layout, font, and structure, which hinders their performance. To remedy this, we propose a Multi-task Cascade Transformer network that explicitly learns these attributes using large volumes of generated synthetic data. The transformer-based network includes two auxiliary tasks and one main task for text segmentation. The auxiliary tasks help the network learn text regions to focus on, as well as the structure of the text through different words and fonts, to support the main task. To bridge the gap between different datasets, we train the proposed network on paired synthetic data before fine-tuning it on real data. Our experiments on publicly available scene text segmentation datasets show that our method outperforms existing methods.

INDEX TERMS Scene text segmentation, paired data synthesis, GANs, transformer, multi-task cascade.

I. INTRODUCTION

Scene text segmentation is a crucial task in computer vision that involves making precise predictions for the presence of text in a scene at the pixel level. This task is vital for various text-related applications, such as text recognition, font style transfer, text image editing, and scene text removal [3]. Effective text segmentation approaches are necessary to extract textual information accurately from natural images in such applications. However, despite significant progress in recent years, text segmentation in real-world scenarios remains a significant challenge due to the unconstrained nature of the scene environment. Such environments typically feature text in various sizes, colours, fonts, and spatial layouts, along with uncontrolled backgrounds. Additionally, the lack of annotated data in this task, as pointed out in [3] and [8], further exacerbates this problem. Table 1 demonstrates that the currently available human-annotated datasets suffer from limitations in terms of the volume of data for pixel-level text segmentation.

CNN-based methods have made significant strides in dealing with data scarcity. Several approaches have been proposed to address this issue, including the use of synthetic word datasets, which have been successful in various works, such as [9]. For example, in [10], Tang and Wu used a similar generation process to produce a more extensive collection of synthetic word images. They then employed a supervised-learning model to segment text in word images, rather than whole scene text images. Other techniques, such as those introduced in [11] and [12], have employed deep learning-based methods to generate pixel-level supervisions. The resulting annotations were then utilized to train deep convolutional neural networks for semantic segmentation. However, the quality of machine-generated ground truth data can be poor, as highlighted in [3] and [8], which can limit the final results. Our proposed method differs from the aforementioned approach in that it utilizes the generated ground truth data from the text image generator to generate a realistic scene text image that adheres to this generated ground truth.

The associate editor coordinating the review of this manuscript and approving it for publication was Zijian Zhang.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

TABLE 1. Statistical list of common human-annotated scene text datasets. The ground truth for pixel-level text segmentation occupies a very small ratio compared to the annotation of detection and recognition in available datasets. The "✓" and "✗" markers indicate that the corresponding annotation is present and absent in the dataset.

Dataset	#Images (train/val/test)	Word Detection	Word Recognition	Pixel-level Text Segmentation	Text Type
ICDAR13 FST [1]	229/0/233	✓	✓	✓	Scene
Total-Text [2]	1255/0/300	✓	✓	✓	Scene
TextSeg [3]	2646/340/1038	✓	✓	✓	Scene+Design
COCO-Text [4]	43686/10000/10000	✓	✓	✗	Scene
ICDAR MLT 2017 [5]	7200/1800/9000	✓	✓	✗	Scene
ICDAR MLT 2019 [6]	10000/0/10000	✓	✓	✗	Scene
ICDAR Art 2019 [7]	5603/0/4563	✓	✓	✗	Scene

This ensures that the quality of the ground truth data used for training is of high standard, as depicted in Figure 1. Previous research has explored the use of weak labels to improve text segmentation. For instance, Wang et al. [8] utilized polygon-level text masks that were extracted from text detection annotations as weak labels to aid in text segmentation. Meanwhile, Ren et al. [13] employed the annotation of text recognition to further support the text segmentation process. In a recent publication [3], a dataset containing human-annotated text segmentation ground truth was introduced. However, the dataset is limited by its small size and by the fact that only a portion of the images are scene text, while the rest are comprised of design text images.

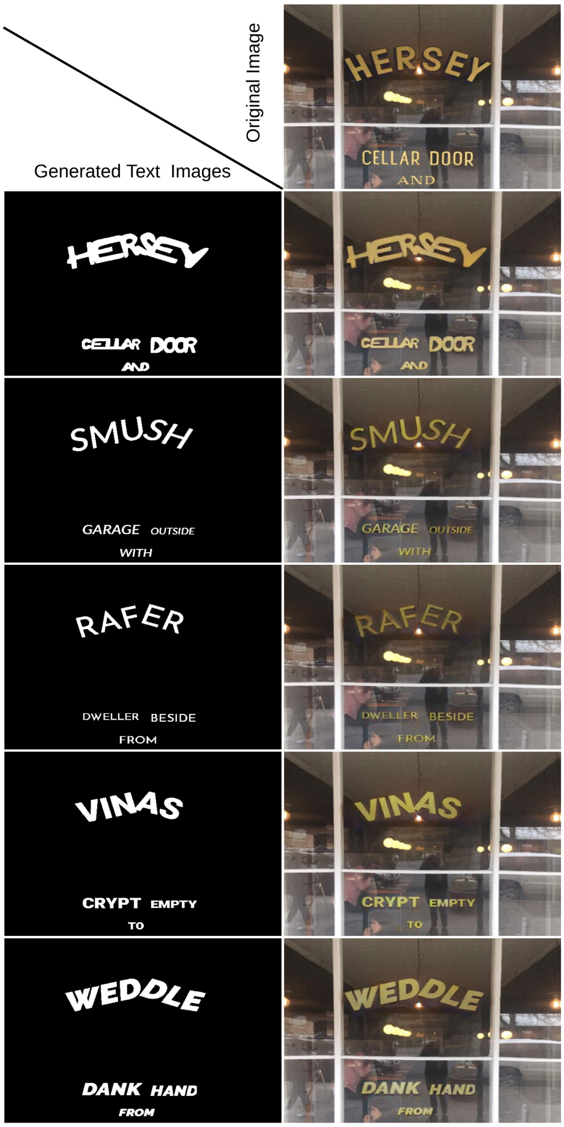
Based on the preliminary results in [14], we expand the paired data generation process to enhance diversity in terms of words and fonts. Our approach involves leveraging Text Image-conditional GANs to tackle the challenges associated with scene text segmentation by generating a wide range of realistic text images. The proposed GANs facilitate the generation of paired data by synthesizing scene text images based on diverse ground-truth images obtained from the Text Image Generator. By producing scene text images that follow different ground-truth images, our GANs effectively increase the diversity of generated texts, resulting in a naturally varied set of paired data. This means that with just one given scene text image and different text images, our GANs can generate various paired data for the text segmentation task, as demonstrated in Figure 1. Additionally, we propose the Multi-task Cascade Transformer to overcome the challenges in text segmentation and effectively learn the generated data. To this end, we utilize a transformer-based backbone [15] for the shared encoder, which benefits from being trained on large amounts of our synthetic data. The cascade decoder consists of three stages that are dedicated to tasks involving polygon-level text region, text skeleton, and pixel-level text. The input for the later stages comes from the earlier stage outputs. Predicting the polygon-level text region helps the main task know which parts of the image to pay attention to. The synthetic data used in the experiment comprise scene text images with varying fonts and word versions. To improve text segmentation, we employ the task of predicting text skeletons to encourage the network to learn the text structure explicitly. This approach is inspired by the fact that text skeletons can assist the network in better understanding the text structure, as mentioned in [16].

We propose using multi-level features from the encoder to each branch, aiming to leverage multiple knowledge representations (MKR) for improved performance and robustness in text segmentation. The concept of MKR is supported by previous research [17], which highlights the benefits of combining multiple sources of information to enhance feature representation. By incorporating MKR into our model, we harness the power of diverse knowledge representations to capture better and understand the complexities of the text. This approach helps us achieve more comprehensive and accurate conceptual modelling, leading to superior performance in text segmentation tasks.

Our motivation for developing the Multi-task Cascade Transformer is to address the specific challenges of text segmentation, which differ from those of object segmentation. Text possesses unique properties, such as font, shape, size, layout, and location, that require a specialized model for accurate segmentation. We observe limitations in the baseline model, originally designed for object segmentation when applied to text segmentation.

To overcome these limitations, we extend the baseline model to focus on text properties, resulting in superior performance. By leveraging a transformer-based model, we take advantage of its effectiveness in handling large volumes of synthesized data, surpassing the performance of previous CNN-based approaches in scene text segmentation tasks. Our innovative decoder module plays a crucial role in integrating different tasks and knowledge representations, prioritizing text region location and capturing the unique structure of the text. These contributions establish our paper as pioneering work in transformer-based methods for text segmentation, addressing the unique challenges of the task and achieving improved performance compared to existing approaches.

In summary, this paper's key contributions are divided into three categories: (1) A synthetic scene text segmentation dataset having different font and word versions is constructed with the help of our proposed Text Image Generator and Text Image-conditional GANs. The data generation process does not incur labour-intensive costs, yet compared to the existing



(a) Generated ground-truth images

(b) Synthetic Scene Text Images

FIGURE 1. Results of paired synthetic data on the training set. The generated text images (a) from the proposed text image generator have different words and fonts. The original image is from the public scene text image dataset. With one given original image and the different text images, the proposed GANs can generate different scene text images naturally (b). As a result, we have new paired data (scene text image and text segmentation ground truth) for the scene text segmentation task.

datasets, the dataset provides high-quality and high-diversity segmentation ground truth; (2) We propose a scene text segmentation network, called Multi-task Cascade Transformer, to explicitly learn distinctive text attributes. We design two auxiliary tasks and one main task for text segmentation. The functions of the two auxiliary tasks are to learn the text region to pay attention to and the structure of the text through various words and their fonts, and then they support the main task; (3) We perform extensive experiments on three text segmentation benchmarks, and show the superior performance of the proposed method compared to current models.

II. RELATED WORKS

Traditional approaches to scene text segmentation. Prior to the advent of deep learning, scene text segmentation was extensively studied using conventional techniques. In [18], Yang et al. proposed a modified K-means clustering algorithm to generate initial text region candidates. These candidates were subsequently verified utilizing a Markov Random Field model that incorporated collinearity weight, enabling better alignment of the detected text regions. Building upon character-level properties, Zhang and Kasturi [19] introduced a text extraction approach that employed stroke edge similarity and link energy to group individual characters into coherent text objects. In [20], Mishra et al. focused on addressing the binarization problem, formulating it within a Markov Random Field framework. They devised an iterative graph cut scheme to minimize an energy function, thus enhancing the robustness of the segmentation process to variations in colour. Additionally, Lafferty et al. [21] presented a framework utilizing Conditional Random Fields (CRFs) for segmenting and labelling sequence data. This approach offered advantages over hidden Markov models by relaxing independence assumptions and achieving more accurate text segmentation results. These works demonstrate a comprehensive understanding of the conventional techniques employed in scene text segmentation.

Deep learning-based scene text segmentation. CNNs have recently shown promising results in a variety of applications that include object classification [22], binarization [23], and detection [24]. Furthermore, pixel-level scene text segmentation has attracted the interest of the scientific community. However, in the early stage, only two public datasets, i.e., ICDAR13-FST [1] and Total-Text [2], contain less than 3k pixel-level annotated images that do not meet large-scale standards for the deep learning-based model. Therefore, the researchers employed the extra annotation of text detection to support pixel-level scene text segmentation. A CNN-based approach was proposed in [10], with three stages: detecting, segmenting, and filtering candidate text regions. In [11] and [12], Bonechi et al. proposed a weakly supervised method by generating the pixel-level annotations for COCO-Text [4] and MLT dataset [5], resulting in two new datasets, COCO-TS and MLT-S. Then, these two datasets were employed to train a text segmentation network. Because the ground truth of the proposed dataset is machine-generated, its quality is very low compared to that of human annotation. In [8], Wang et al. proposed a dual-task mutually guided network, which comprises a common encoder, but two decoders for two tasks: the pixel-level and polygon-level masks. However, the method did not take advantage of the structural information of the text, but only relied on the text area, leading to missing parts of the strokes of the scene text segmentation results when the strokes are ambiguous. In [3], Xu et al. proposed a new text segmentation dataset that includes only part of the scene text images and the remaining part of the design text source. Additionally, they also presented a Text Refinement Network that could be fully supervised-trained with available datasets for text segmentation. Recently, the ARM-Net method proposed in [13] improved the accuracy of scene text segmentation by using both implicit low-level text appearance information and higher-level text semantic information.

Generative adversarial networks (GANs). GANs aims to generate realistic-looking images by training a generator to produce synthetic samples that resemble real images. The synthetic samples are evaluated by a discriminator network that assesses their similarity to real images. GANs enable a wide range of applications, such as image-to-image translation [25], [26], de-raining [27], [28], inpainting [29], [30], and editing text [16]. The image-to-image translation is the mapping task from image to image. Under the given conditions, the model in [31] generated the required images. The modified model can generate new images with the constraints of the given text image. However, to produce an image with a complex scene background, the generated synthetic image was far from real-world scene text images. De-raining aims to recover clean image content by eliminating rain components in scene images without artefacts. Inpainting is the task of filling in missing pixels in an image such that the completed image looks realistic and preserves the original context. The missing region is empty, without any prior condition inside. Editing text in natural images replaces words in the original image with different ones that maintain a natural-looking appearance. In the existing papers [16], the style of generated text followed the original text, while its content followed the given text. Therefore, the generated scene text image and the given text image were not paired. Motivated by previous works, we introduce the Text Image-conditional GANs that generate paired data, realistic scene text image and corresponding text segmentation ground truth for the scene text segmentation task. Paired data generation in the scene text image has never been tried before, and to the best of our knowledge, this is the first attempt in the literature. In [32], Zhan et al. introduced a multi-modal spatial learning technique that transforms a source-domain image into multiple images with different spatial views, resembling the target domain. While this model produces realistic scene text images, it cannot provide ground truth for pixel-level text segmentation. The papers [33] and [34] presented image synthesis techniques for generating annotated scene text images to train robust text detection and recognition models. These methods employed a geometry synthesizer to learn the contextual geometries of background images and randomly placed foreground objects, including text, within them. However, this random placement may lead to text appearing in unreasonable positions within the background. Additionally, these papers used background images that already contained original text, lacking ground truth for text segmentation. Consequently, although these methods can offer ground truth for synthetic text segmentation, they cannot provide ground truth for real text segmentation, limiting their suitability for generating paired synthetic data for scene text segmentation. In contrast, our proposed method realistically replaces the original text within background images with synthetic text, ensuring the availability of corresponding ground truth for text segmentation. Notably, our GAN is the first to specifically generate synthetic data for pixel-level text segmentation, setting it apart from previous works that mainly focus on data generation for text detection and recognition tasks. By emphasizing these distinctions, we emphasize the unique contributions of our proposed synthesis method, addressing the limitations of prior studies and presenting a novel approach for generating synthetic data for pixel-level text segmentation.

Transformers for vision. The Transformer concept is originally designed for natural language sequence processing. Transformers are now state-of-the-art in many Natural Language Processing (NLP) tasks. Self-attention mechanisms are used in these models, which capture long-range dependencies between tokens (words) in a sentence. Transformers are also well-suited to parallelization, making training on large datasets easier. The success of transformers in NLP has inspired various approaches in computer vision, including handwriting text recognition [35], image captioning [36], and human action recognition [37]. The Vision Transformer (ViT) [38] presented a convolution-free transformer for image classification that processes input images as token patch sequences. SegFormer [15] has a unique transformer encoder that can output features at different scales. It does not require positional encoding, avoiding the requirement for interpolation of positional codes, which results in worse performance when the testing resolution differs from the training resolution. Here, our approach adapts the SegFormer backbone to model long-range dependencies.

III. PAIRED DATA SYNTHESIS AND SCENE TEXT SEGMENTATION NETWORK

The overview of the proposed framework is composed of the text image generator and two deep learning-based networks: Text Image Generator, Text Image-conditional GANs, and Multi-task Cascade Transformer. Given the output of the Text Image Generator, polygon-level masks of text and original scene text image, Text Image-conditional GANs attempts to create a new version of the scene text image by incorporating the original scene background with various text foregrounds naturally. Then, the Multi-task Cascade Transformer performs text segmentation from the newly generated scene text image, fully supervised by the new corresponding ground truth. The overview framework is depicted in Figure 2.

A. PAIRED DATA SYNTHESIS

1) TEXT IMAGE GENERATOR

Text Image Generator, as in Figure 3, is based on a non-parametric algorithm. This takes text string, bounding polygon coordinates of words, and randomly selected fonts as input. We generate text images from text strings by rendering text based on the library Matplotlib [39]. Their locations are aligned along the centre line of the bounding polygon.

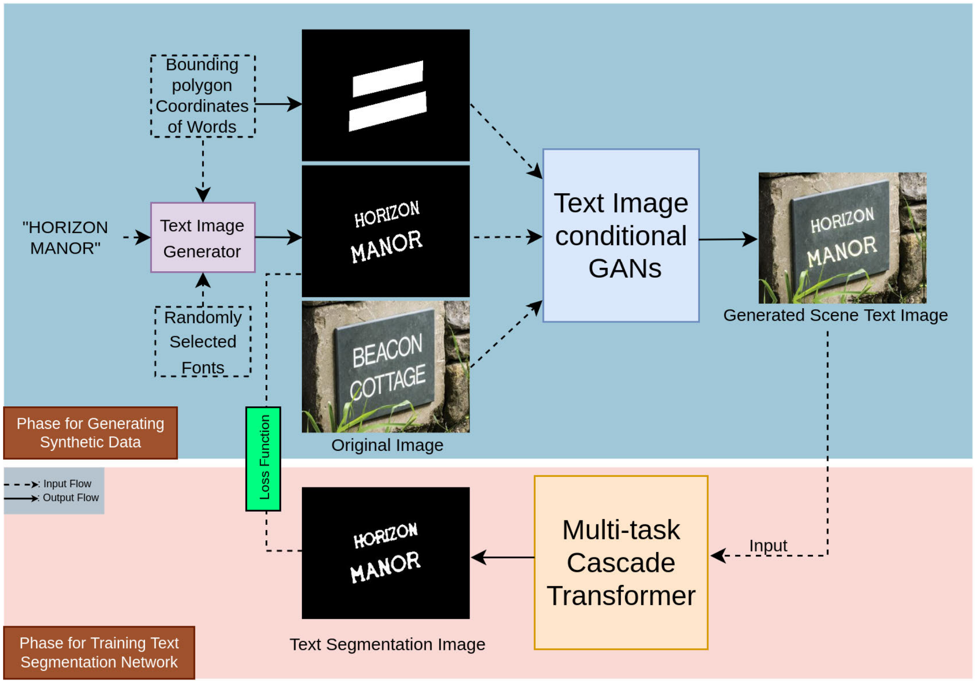

Bounding polygon Coordinates of Words
"HORIZON MANOR"
Text Image Generator
Randomly Selected Fonts
HORIZON MANOR
BEACON COTTAGE
Original Image
Text Image conditional GANs
Generated Scene Text Image
Phase for Generating Synthetic Data
Input Flow
Output Flow
Loss Function
Multi-task Cascade Transformer
Input
Text Segmentation Image
Phase for Training Text Segmentation Network

FIGURE 2. **The overview of the proposed framework. This consists of two deep learning networks: Text Image-conditional GANs, and Multi-task Cascade Transformers. In the phase for generating synthetic data, the GANs takes the output of the Text Image Generator, together with polygon-level text masks and original image as inputs to generate realistic scene text image. As a result, the generated paired data includes text images as segmentation ground truths and synthetic scene text images. This is extra training data for the Multi-task Cascade Transformer in the scene text segmentation task. In the testing phase, the trained segmentation network applies to only the original images of testing data, independent of training and validation data, to generate the model's final output.**

Their sizes are limited by text regions. We can choose text strings randomly. However, for the content of the text to be contextual, the text string is taken from the text recognition annotations of the dataset or their semantically similar words. We search similar words for an input word based on cosine similarities that are computed between their corresponding word vectors. The word vectors used in this process are obtained from the pre-trained Glove model [40]. For example, similar words to “hotel” are “hostel”, “motel”, “lodging”, “house”, etc. Therefore, the different versions of text images are based on the original word and semantically similar words, along with their randomly selected fonts. If the language is not English, the word is randomly chosen in the Glove library as long as the length of the word is equal to the original one.

2) TEXT IMAGE-CONDITIONAL GANS

In the phase for generating synthetic images, our goal is to generate realistic paired data, including scene text image, and corresponding text segmentation ground truth.

In the training process, we pre-train the proposed GANs following the inpainting-based self-supervision approach to the collection of large-scale public datasets that have the annotation for text detection but do not have the ground truth for pixel-level text segmentation. Then, we employ public datasets having scene text image R and original ground truth for text segmentation image S, along with the polygon-level text mask M (1 for the text region, 0 for the background) extracted from text detection annotation to feed into GANs. The masked image containing text in the polygon-level mask is denoted $R \odot (1-M) + S$. The mask M is stacked with the masked image containing text, leading to the input for GANs $R' = stack(R \odot (1-M) + S, M)$. We employ a feed-forward network $G_\theta(.)$. The training is performed on the masked image containing original text in polygon-level mask $R' = stack(R \odot (1-M) + S, M)$ and real image R. The predicted image is $\hat{R} = G_\theta(R')$. Figure 4 shows the proposed Text Image-conditional GANs in the training phase.

In the testing process, we employ the synthetic text image produced by the text image generator as the new segmentation ground truth S'. We replace the ground truth S with synthetic ground truth S'. Therefore, the testing is performed on the masked image containing synthetic text in polygon-level mask $R'' = stack(R \odot (1-M)+S', M)$, and produces realistic image $\check{R} = G_\theta(R'')$.

We employ a Fast Fourier convolution-based network [42] for the generator network $G_\theta(.)$. Fast Fourier convolutions (FFCs) is based on a channel-wise fast Fourier transform [43], and features an image-wide receptive field. Because of the image-wide receptive field, which is critical for high-resolution images, FFCs enable the generator to account for the global context from the early layers. Furthermore, FFCs

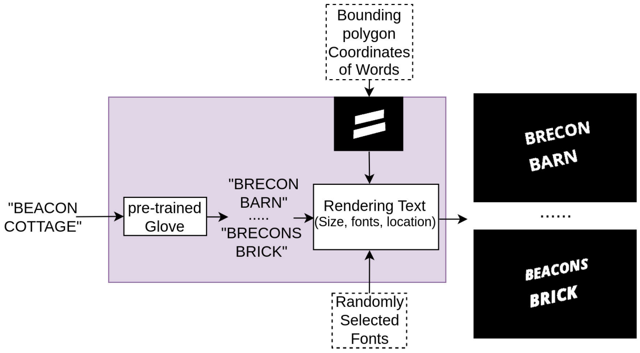


FIGURE 3. The overview of the proposed text image generator.

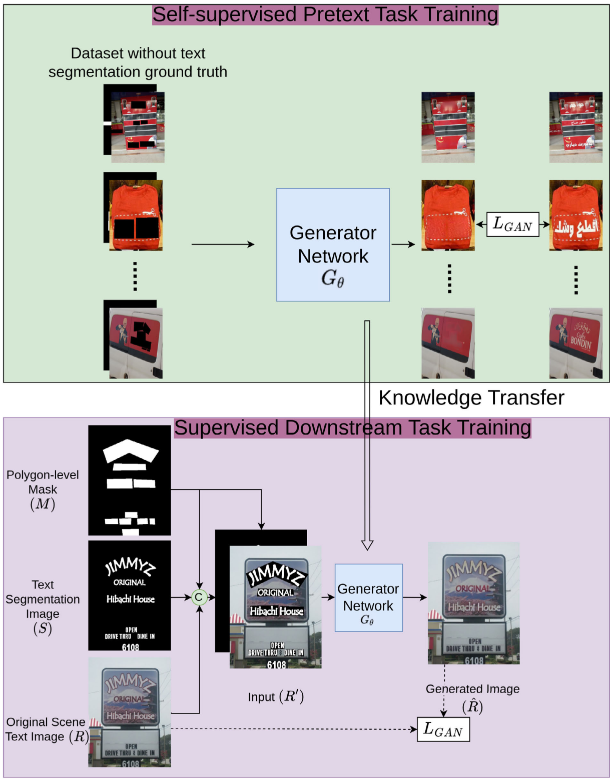

Self-supervised Pretext Task Training
Dataset without text segmentation ground truth
Generator Network G_θ
L_{GAN}
Knowledge Transfer
Supervised Downstream Task Training
Polygon-level Mask (M)
Text Segmentation Image (S)
Original Scene Text Image (R)
C
Input (R')
Generator Network G_θ
Generated Image ($\hat{R}$)
L_{GAN}

FIGURE 4. The training scheme of text image-conditional GANs.

are highly suited to capturing periodic structures [42] found in artificial environments, such as signs, doors, walls, etc. There could be many reasonable fillings to the polygon-level mask containing text.

Adversarial loss. We use adversarial objective L_{adv} [44] to generate natural-looking text region details. A discriminator $D_\eta(.)$ is defined to distinguish between real and fake patches, as in Figure 5. Patches that overlap with text regions are labelled as fake.

$$\begin{aligned} L_D = &-E_R[logD_\eta(R)] - E_{R,M}[logD_\eta(\hat{R}) \odot (1-M)] \\ &- E_{R,M}[log(1-D_\eta(\hat{R})) \odot M] \qquad (1) \\ L_G = &-E_{R,M}[logD_\eta(\hat{R})] \qquad (2) \end{aligned}$$

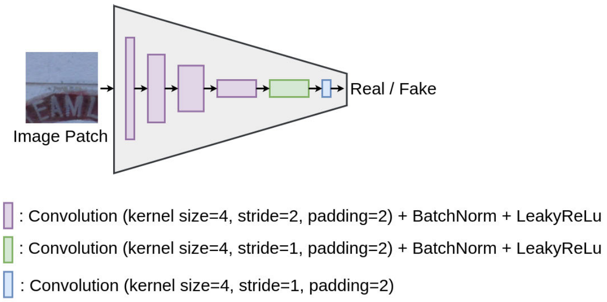


Image Patch
Real / Fake
: Convolution (kernel size=4, stride=2, padding=2) + BatchNorm + LeakyReLu
: Convolution (kernel size=4, stride=1, padding=2) + BatchNorm + LeakyReLu
: Convolution (kernel size=4, stride=1, padding=2)

FIGURE 5. The architecture of discriminator.

$$L_{adv} = L_D + L_G \rightarrow \min_{\theta,\eta} \qquad (3)$$

Feature matching loss L_{fm} [45] is proven to stabilize training, while in some circumstances, also improving performance. High receptive field perceptual loss L_{hrfpl} [42] is in charge of the supervised signal and global structural consistency. Therefore, we combine the loss functions for the loss of the proposed GANs:

$$L_{GAN} = \alpha_1 L_{adv} + \alpha_2 L_{hrfpl} + \alpha_3 L_{fm} \qquad (4)$$

where α_1, α_2 and α_3 are weight parameters

B. SCENE TEXT SEGMENTATION NETWORK

With the generated paired data in the previous phase, we propose the Multi-task Cascade Transformer network for text image segmentation. Compared with the traditional CNN-based methods, transformers have a better ability to understand shape and geometry [46]. In addition, the local nature of convolutional filters restricts access to global information [47], which is critical for segmentation because the labelling of local patches is frequently dependent on the global image context. Furthermore, Transformer-based models benefit from training on large data [38] that is available from our synthetic data. Therefore, we employ a transformer-based backbone for the shared encoder. We choose the backbone based on [15] to avoid the interpolation of positional codes leading to reduced performance when the testing image resolution varies from the training one. The cascade decoder has three stages that address tasks: polygon-level text region, text skeleton, and pixel-level text. However, in our method, unlike many multi-task learning applications, a later stage is conditional on the outputs of an earlier stage, resulting in a causal cascade, as in Figure 6. The polygon-level masks guide pixel-level text segmentation, providing prior knowledge to help the network better localize text regions and where to pay more attention. Because the synthetic data comprises many scene text images with different text versions, we use the text diversity to force the network to learn the text structure. Specifically, we employ text skeletons to help the network better understand the text structure explicitly [16], leading to improved results.

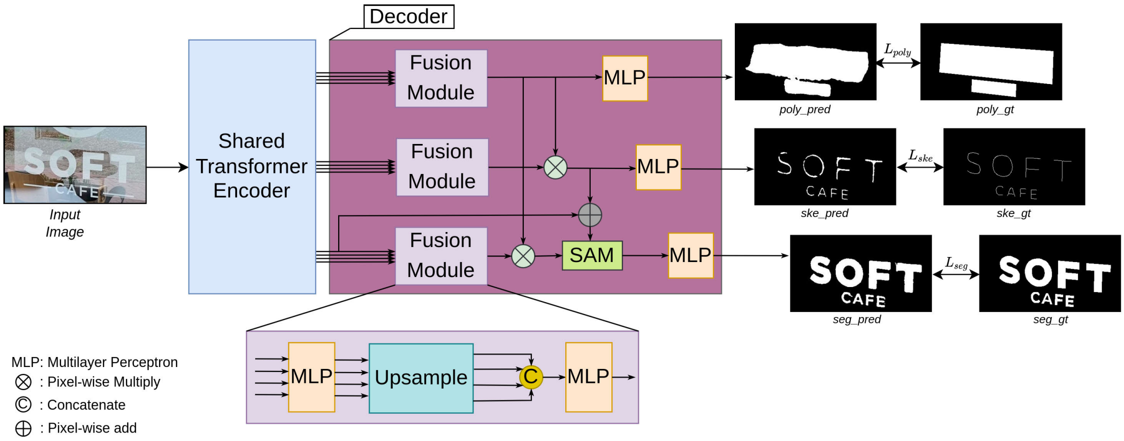


FIGURE 6. The proposed Multi-task Cascade Transformer for scene text segmentation. The shared transformer encoder extracts coarse and fine features. The Fusion Module is to upsample and fuse features. The similarity aggregation module (SAM) [41] injects text detailed appearance and structure features into the fused feature output. The ground truth of the polygon-level text mask is generated from provided word detection annotation. The ground truth of the text skeleton is extracted from the ground truth of pixel-level text segmentation.

1) SHARED TRANSFORMER ENCODER

Firstly, we develop a model based on a transformer. The encoder uses image patches as inputs and takes the transformer to propagate global contexts among all patches, modelling long-range dependencies to understand the whole image generally. Specifically, we utilize a transformer-based backbone [15] for the shared encoder that benefits from training on large data of our synthetic data. Unlike ViT [38], which can only build a single-resolution feature map, the purpose of this module is to construct CNN-like multi-level features from an input image. These features give high-resolution coarse features as well as low-resolution fine-grained features, which help semantic segmentation perform better. Furthermore, it does not require positional encoding, avoiding the need for positional code interpolation, which leads to poor performance when the image resolutions in testing and training differ.

2) MULTI-TASK CASCADE DECODER

To adapt the transformer-based encoder to the text segmentation task, we design a cascade decoder for text segmentation on top of four multi-scale feature maps generated by different four stages. Among these feature maps, high-resolution coarse features in the first layer give detailed appearance information of text, and low-resolution fine-grained features in the later layers provide high-level features.

The multi-task cascade transformer uses a lightweight decoder comprised of MLP layers, eliminating the compute-complex components found in other CNN-based methods. Our hierarchical Transformer encoder has a bigger effective receptive field than typical CNN encoders. It is a crucial factor in enabling such a straightforward decoder. CNN's limited receptive field necessitates the decoder to resort to context modules, which expand the receptive field but are inevitably heavy. Our decoder design takes advantage of Transformers' non-local attention and results in a larger receptive field while remaining simple. However, due to the small receptive field, the same decoder design does not operate well on CNN backbones. Furthermore, Using the non-local attention solely from the final layer is insufficient for good results. Our decoder design takes advantage of features extracted from the transformer encoder that yields both highly local and non-local attention simultaneously. Our MLP decoder can unify them to provide complementary and effective representations with fewer parameters. It is another important factor that influenced our decoder design.

The cascade decoder has three stages that address tasks: polygon-level text mask, text skeleton, and pixel-level text. The input of the later stage is from the outputs of an earlier stage. The task for predicting the polygon-level text mask provides information on where to pay more attention to the main task. The synthetic data comprises scene text images with different text fonts and word versions. We use the task to predict text skeletons to force the network to learn the text structure, because text skeletons can help the network better understand the text structure explicitly [16], resulting in improved text segmentation. Specifically, multi-level features from the encoder go through each branch, including Fusion Module. Multi-scale features are fed to the Fusion Module that unifies the channel dimension by MLP layer before up-sampling and concatenating together, then fusing the concatenated features by MLP layer. The fused features in the task for predicting polygon-level text mask provide other tasks for attention to text region. The fused features in the task for predicting text skeleton provide the main task to get text structure information. The similarity aggregation module (SAM) [41] injects detailed text appearance and structural features into the fused feature in the main task. Finally, another MLP layer in each branch takes the fused feature to predict the corresponding segmentation.

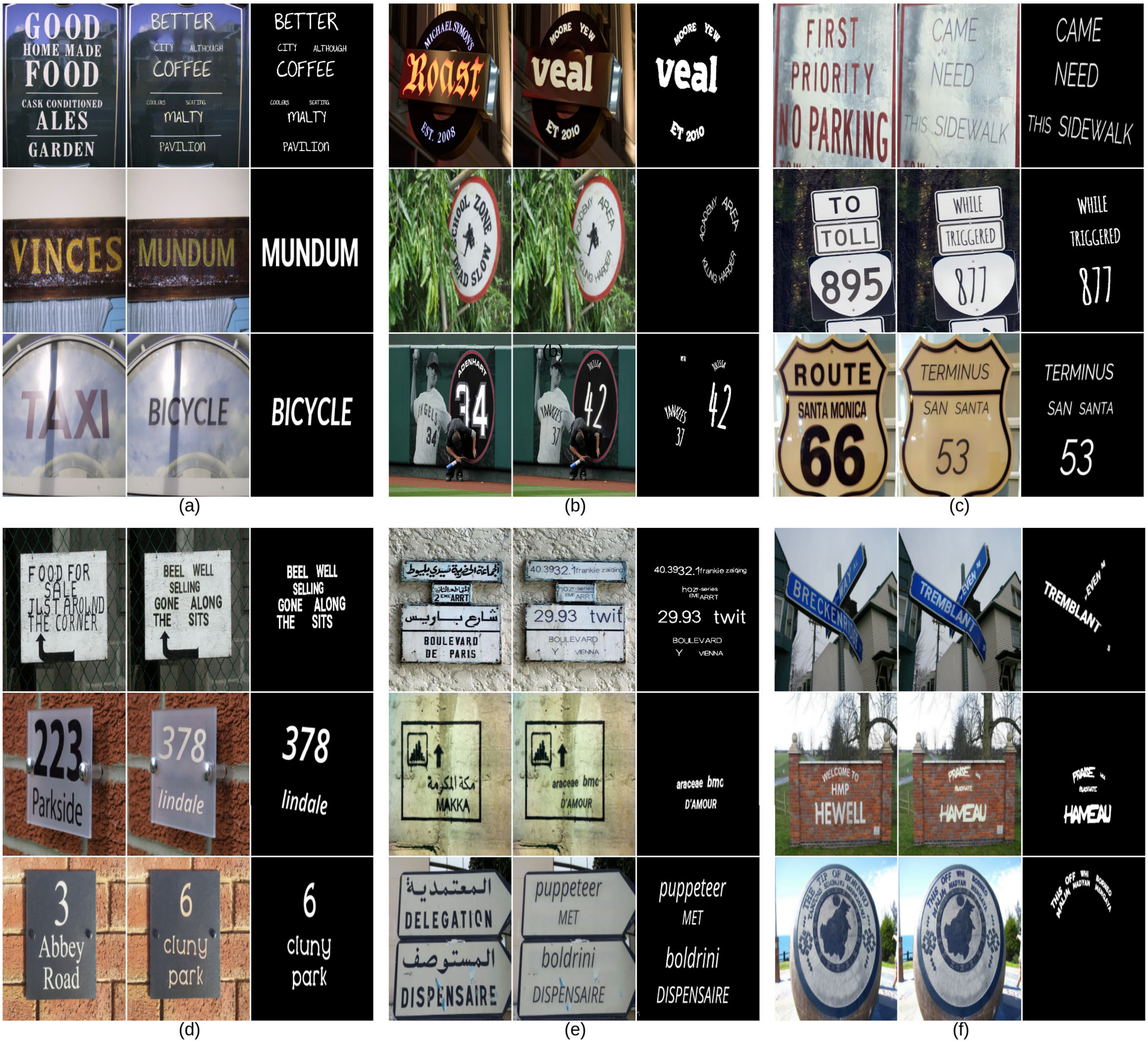


FIGURE 7. Generated paired image examples from the proposed Text Image-conditional GANs. (a), (b), (c), (d), (e) and (f) show the generated paired images based on scene background from ICDAR13 FST, Total-Text, TextSeg, ICDAR MLT 2017, ICDAR MLT 2019, and ICDAR Art 2019 datasets, respectively. The first, second, and third columns of each are original images, synthetic images, and generated ground truths, respectively.

3) LOSS FUNCTION

Ground truth of polygon-level text mask ($poly\_gt$) is generated from provided word detection annotation by the library OpenCV [50].

The ground truth of the text skeleton (ske_gt) is extracted from the ground truth of pixel-level text segmentation (seg_gt) by the library OpenCV.

$poly\_pred$, ske_pred and seg_pred indicate the result of 3 tasks, polygon-level text mask, text skeleton, and pixel-level text segmentation, respectively.

Each task involves a loss term. Because a later task's loss relies on the output of an earlier task, the loss terms are not independent. We train the proposed Multi-task Cascade Transformer with a unified loss function. This can be formulated:

$$L = \beta_1 L_{poly} + \beta_2 L_{ske} + \beta_3 L_{seg} \quad (5)$$

$$L_{poly} = BCE(poly\_pred, poly\_gt) \quad (6)$$

$$L_{ske} = BCE(ske\_pred, ske\_gt) \quad (7)$$

$$L_{seg} = BCE(seg\_pred, seg\_gt) \quad (8)$$

where, L_{poly}, L_{ske}, and L_{seg} are loss functions for polygon-level segmentation, text skeleton prediction, and pixel-level segmentation, respectively. All of them are only based on modified binary cross-entropy (BCE) loss. Unlike the regular BCE loss function [51], which considers all pixels identically,

TABLE 2. Results on the F-score(Black) and IoU(Blue) to ICDAR13 FST, Total-Text and TextSeg. The "-" marker indicates that the result is not reported in the corresponding paper.

Method	ICDAR13 FST	Total Text	TextSeg	Params
PSPNet [12]	0.804/-	0.753/-	-	65.1M
SMANet [11]	0.858/-	0.781/-	-	-
DeeplabV3+ [48]	0.806/0.693	0.815/0.739	0.908/0.841	62.7M
Wang et al. [8]	0.745/-	0.805/-	-	-
HRNetV2-W48 + OCR [3], [49]	0.830/0.725	0.832/0.762	0.918/0.860	70.5M
ARMNet [13]	0.851/-	0.854/-	0.927/-	-
TexRNet [3]	0.850/0.734	0.848/0.785	0.924/0.868	67.1M
TexRNet [3] (S.D.)	0.856/0.741	0.855/**0.788**	0.929/0.869	67.1M
Ours	0.846/0.732	0.851/0.779	0.916/0.863	66.9M
Ours (S.D.)	**0.870/0.745**	**0.862**/0.786	**0.932/0.877**	66.9M

"S.D." denotes using our synthetic data in the model. For instance, referring to "Ours (S.D.)" signifies that our proposed model incorporates our synthetic data.

the modified BCE takes each pixel's importance into account and gives higher weights to hard pixels. β_1, β_2 and β_3 are weight parameters.

IV. EXPERIMENTAL RESULTS

This section introduces the experimental datasets, implementation details, and experimental results. Then, we compare the proposed text segmentation model to state-of-the-art methods.

A. DATASETS

We collect publicly available datasets. They are:

ICDAR13 FST [1]: This includes scene text images. It only contains 229 training and 233 testing images with ground truth for text segmentation, recognition, and detection. The bounding box for each word is a rectangle.

Total-Text [2]: This contains 1255 training and 300 testing scene text images, and has multi-oriented and curved texts. The word-bounding polygons are available, along with ground truth for text segmentation and recognition.

TextSeg [3]: TextSeg consists of 4024 text images, including scene text and design text. It is split into training, validation, and testing sets with (2646, 340, and 1038) images, respectively. The ground truth for text segmentation, recognition and detection is available.

We utilize a self-supervised inpainting technique to pre-train GANs on a large-scale dataset that has annotations for text detection but lacks ground truth for pixel-level text segmentation. To achieve this, we combine four widely used datasets in the field, namely Coco-text, ICDAR MLTS 17, ICDAR MLT 19, and ICDAR Art 19, which provide a diverse range of text data for the pre-training process.

Coco-text [4]: This is the largest scene text dataset, with 63686 images. It is split into training, validation, and testing sets with (43686, 10000, and 10000) images, respectively. Although annotation for text recognition and detection in the training and validation set is available, it does not have the ground truth for text segmentation.

ICDAR MLT 2017 [5]: There are 18000 images in total, which are divided into three sets: training, validation, and testing, each comprising 7200, 1800, and 9000 images. The annotation for text recognition and detection in the training and validation set is available.

ICDAR MLT 2019 [6]: This has a total of 20000 images. There are 10000 and 10000 images in each training and testing set, respectively. The annotation for text recognition and detection in the training set is provided.

ICDAR Art 2019 [7]: This consists of 10166 images. It is split into training and testing sets with 5603 and 4563 images, respectively. In the training set, there is an annotation for text recognition and detection.

B. TRAINING

We pre-train the proposed GANs following the inpainting-based self-supervision approach on collected common large-scale datasets that have the annotation for text detection but do not have the ground truth for pixel-level text segmentation, such as Coco-text, ICDAR MLTS 17, ICDAR MLT 19, and ICDAR Art 19. We use a polygon-level text mask extracted from detection annotation. This is trained on low-resolution 256 × 256 patches (1M iterations). The patches are cropped as long as they overlap the text regions. Then, we train the GANs on only the training set of ICDAR13 FST, Total-Text, and TextSeg datasets to ensure fairness. We employ the Adam optimizer to train our network with the initial learning rate of 0.001 and 0.0001 for the generator and discriminator of GANs, respectively (600K iterations). We take $\alpha_1 = 10$, $\alpha_2 = 30$ and $\alpha_3 = 80$ for GANs throughout the experiments. The model is trained with a batch size of 24. As a result, synthetic paired data for scene text segmentation is generated by the trained Text Image-conditional GANs. The scene backgrounds of original images from the

TABLE 3. Ablation study of the proposed model on TextSeg.

Model	F-score	IoU
1.Baseline	0.905	0.843
2.Basline+Polygon-level Segmentation	0.912	0.856
3.Basline+Polygon-level Segmentation+Text Skeleton Prediction	0.916	0.863
4.Basline+Polygon-level Segmentation+Text Skeleton Prediction+Synthetic data	**0.932**	**0.877**

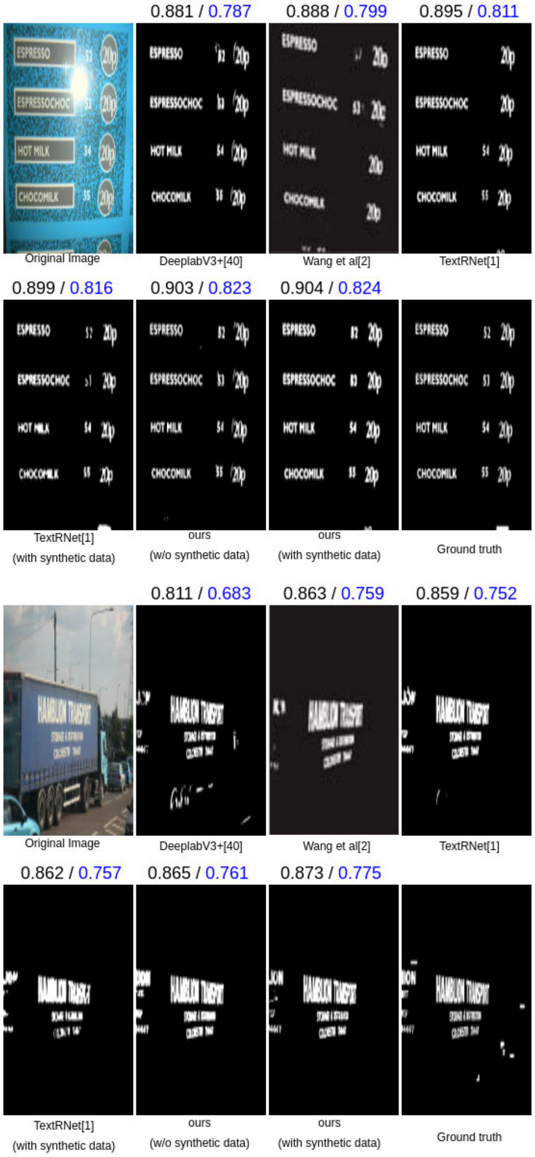


FIGURE 8. Text segmentation results on ICDAR13 FST. We provide F1-score(Black) and IoU(Blue) for individual results.

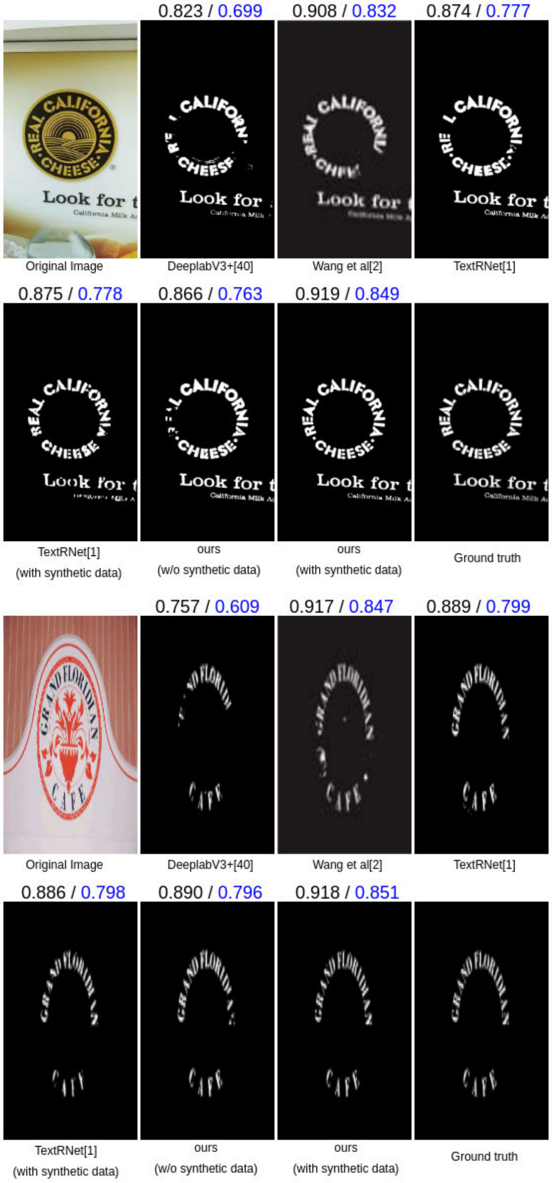


FIGURE 9. Text segmentation results on Total-Text. We provide F1-score(Black) and IoU(Blue) for individual results.

TABLE 4. Improvement by the predicted mask for text spotting.

Method	ICDAR13	ICDAR15	Total Text
DeepSolo [52]	90.1	76.9	86.2
DeepSolo(predicted Mask)	91.4 (+1.3)	78.1 (+1.2)	87.8 (+1.6)

training sets of datasets (ICDAR13 FST, Total-Text, TextSeg, ICDAR MLT 2017, ICDAR MLT 2019, and ICDAR Art 2019) are employed as in Figure 7. Because we use the testing set of three datasets (ICDAR13 FST, Total-Text, and TextSeg) for evaluation in scene text segmentation, we generate five different pairs of images, as in Figure 1, that include the scene backgrounds of the training set of ICDAR13 FST, Total-Text, and TextSeg to reduce imbalanced data and domain gaps

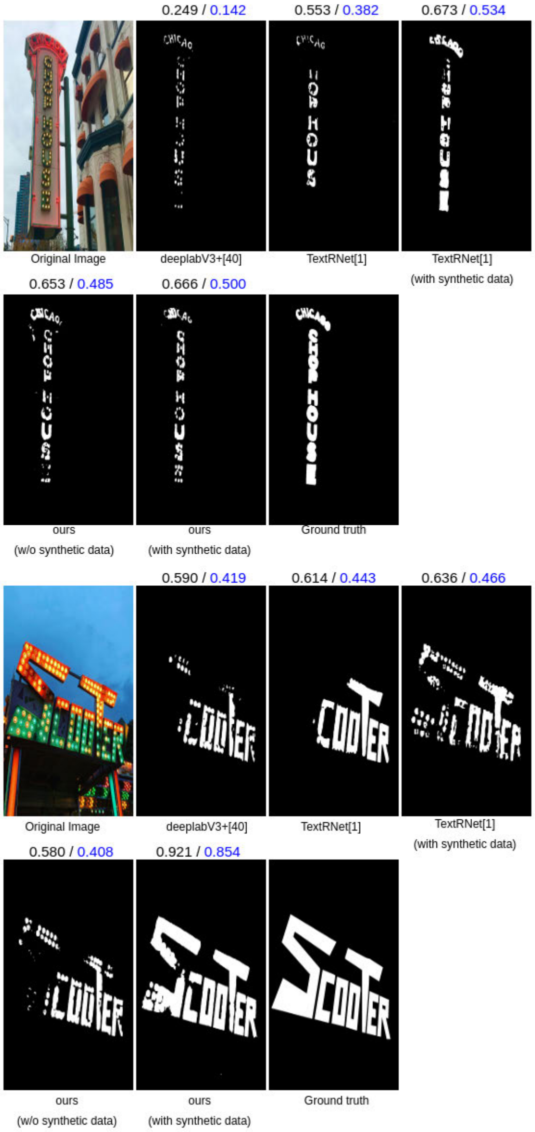


FIGURE 10. Text segmentation results on TextSeg. We provide F1-score(Black) and IoU(Blue) for individual results.

between different datasets. Therefore, our generated data has a total of 43453 images.

In the next phase, the Multi-task Cascade Transformer initializes weight on ImageNet-1K [53]. Then, Multi-task Cascade Transformer trains on the synthetic paired data before fine-tuning alternatively on the training set of ICDAR13 FST, Total-Text and TextSeg that have human-annotated ground truth for pixel-level text segmentation. We use AdamW optimizer with an initial learning rate of 5.0×10^{-5} for training on synthetic data (100K iterations) and 5.0×10^{-6} for fine-tuning alternatively on ICDAR13 FST, Total-Text and TextSeg (20K iterations). The polynomial learning rate scheduler is applied. $\beta_1 = 0.5$, $\beta_2 = 0.5$, and $\beta_3 = 1$ are chosen throughout the experiments. We utilize different data augmentation for training. The images are subjected to several transformations prior to being input into the neural network. These transformations include random resizing within a ratio range of 0.5 to 2.0, random horizontal flipping, and random cropping to a resolution of 512 × 512 pixels.

Our network is implemented based on the public framework PyTorch 1.7.1 and runs on a computer equipped with an Intel(R) Core(TM) i9-10900 CPU @ 2.80GHz. We utilize 2 x GPU RTX 3090 (with 24GB memory) for our experiments. We employ F-score and IOU measurement on foreground pixels in the same fashion to [3] for our evaluation metrics.

C. RESULTS AND DISCUSSION

Table 2 compares the results between our approach and state-of-the-art methods. Figures 8, 9, & 10 show some qualitative results on ICDAR13 FST, Total-Text and TextSeg, respectively. When we apply synthetic data to the previous method and our multi-task cascade transformer, the results improve. We apply our synthetic data to TextRNet. It increases by 0.6 % for ICDAR13 FST, 0.7 % for Total Text and 0.5 % for Texseg. This increase is moderate and reasonable. Therefore, the impact of our synthetic data is effective but not limited to TextRNet based on CNNs. However, our synthetic data is fully exploited when applying it to transformer-based models because transformers work effectively when training with large-scale data [38] that is available with our synthetic data. More precisely, when our synthetic data is used with the multi-task cascade transformer, there is a notable improvement in performance, with an increase of 2.4% for ICDAR13 FST, 1.1% for Total Text, and 1.6% for TextSeg, compared to the model that does not use synthetic data. So, this proves the effectiveness of our synthetic data. As a result, our proposed method gets the new state-of-the-art method. It is because of the realistic synthetic data and the effectively designed Multi-task Cascade Transformer that benefits from training on large data of our synthetic data and has the ability to explicitly learn distinctive text attributes for supporting text segmentation tasks.

Our result shows a significant increase of 1.9% for the F-score compared to the ARMNet [13] for the testing data of the ICDAR13 FST dataset. This indicates that the synthetic data supports the scene text segmentation network to learn the small dataset effectively (229 original training images) because the training set of the ICDAR13 FST dataset participated in generating synthetic data whose distribution is close to the testing set of the ICDAR13 FST dataset, compared to other datasets. The generated images are based on the ICDAR13 FST dataset, but the approaches in [4] and [11] use other datasets to generate weak ground truth. In [3], the result is relatively high because the model uses extra bounding boxes and recognition of characters for training. If we only utilize the combination of datasets (ICDAR13 FST, Total-Text and TextSeg) without synthetic data for training the proposed Multi-task Cascade Transformer before fine-tuning the selected dataset, it shows highly competitive results to the other methods. However, its capability can be thoroughly exploited when it is trained on synthetic data and then

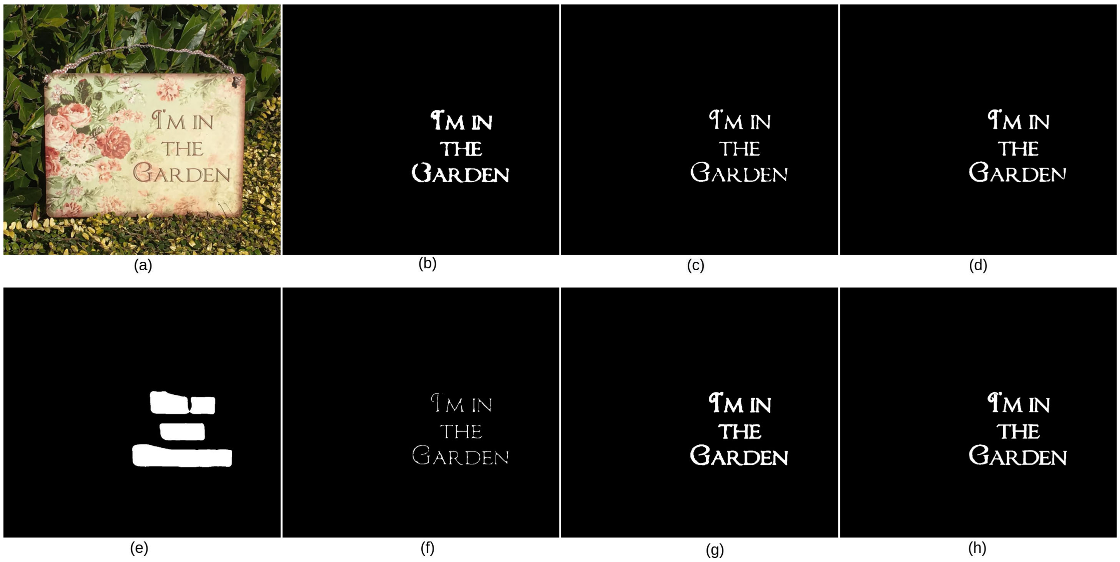


FIGURE 11. Illustration of ablation study on the effectiveness of our method on the sample in the TextSeg dataset. (a) original image, (b) Baseline, (c) Baseline + Polygon-level Segmentation task, (d) Baseline + Polygon-level Segmentation task + Text Skeleton Prediction task, (e) Polygon-level Segmentation of our proposed method, (f) Text Skeleton Prediction of our proposed method, (g) Baseline + Polygon-level Segmentation task + Text Skeleton Prediction task + Synthetic data (our proposed method), (h) Ground truth.

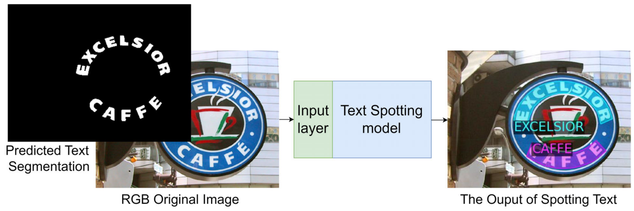


FIGURE 12. Overview of our text spotting network.

fine-tuned alternatively on the training set of ICDAR13 FST, Total-Text and TextSeg. By combining them, the result from our approach outperforms existing methods. While the results of the existing text segmentation methods are reported in the corresponding papers, we train and test the DeeplabV3+ [48] using public source code to get the result. Furthermore, We also use the official public source code of TextRNet to produce the result. TexRNet (with our synthetic data) is trained on our synthetic data and fine-tuned alternatively on the training set of ICDAR13 FST, Total-Text and TextSeg. Then, the trained TexRNet is tested on the testing set of the real dataset to get the final results.

To ensure fair comparisons, we have provided the model parameter numbers in the last column of Table 2. Our model design focuses on efficiency by utilizing a backbone [15] that delivers simple, efficient, and effective output features for semantic segmentation. Additionally, our decoders are specifically tailored to handle the distinct characteristics of text while maintaining a lightweight and efficient architecture. The parameter numbers of previous models were calculated using the published source code, ensuring consistent and accurate comparisons. By incorporating these considerations, we aim to comprehensively assess our method's efficiency and demonstrate a transparent evaluation process.

D. ABLATION STUDY

This section performs ablation studies on polygon-level segmentation, text skeleton prediction, and synthetic data. Table 3 and Figure 11 show the effectiveness of our method. Specifically, the baseline model consists of a transformer-based encoder [15], and a simple MLP-based decoder. The result is improved on F-score and IoU metrics when, as shown in Table 3, we incrementally added polygon-level segmentation and text skeleton prediction task to the model. The final model achieves the best performance, with around 2.7 and 3.4 % increase in F-score and IoU, compared to the baseline when we use our synthetic data for training the model, before finetuning on the real dataset.

E. ENHANCING TEXT SPOTTING THROUGH SCENE TEXT SEGMENTATION

Scene text segmentation has various applications in optical character recognition (OCR), including text recognition [8], text spotting [54], text removal and text style transfer [3]. In this study, we focused on utilizing scene text segmentation to enhance the performance of an existing text-spotting model. We observed that scene text segmentation can effectively serve as an attention map for guiding the text-spotting process. To investigate this phenomenon, we selected DeepSolo [52] as the state-of-the-art network for text spotting.

Our approach involved modifying the input layer of the text spotting system. Instead of the original 3-channel format,

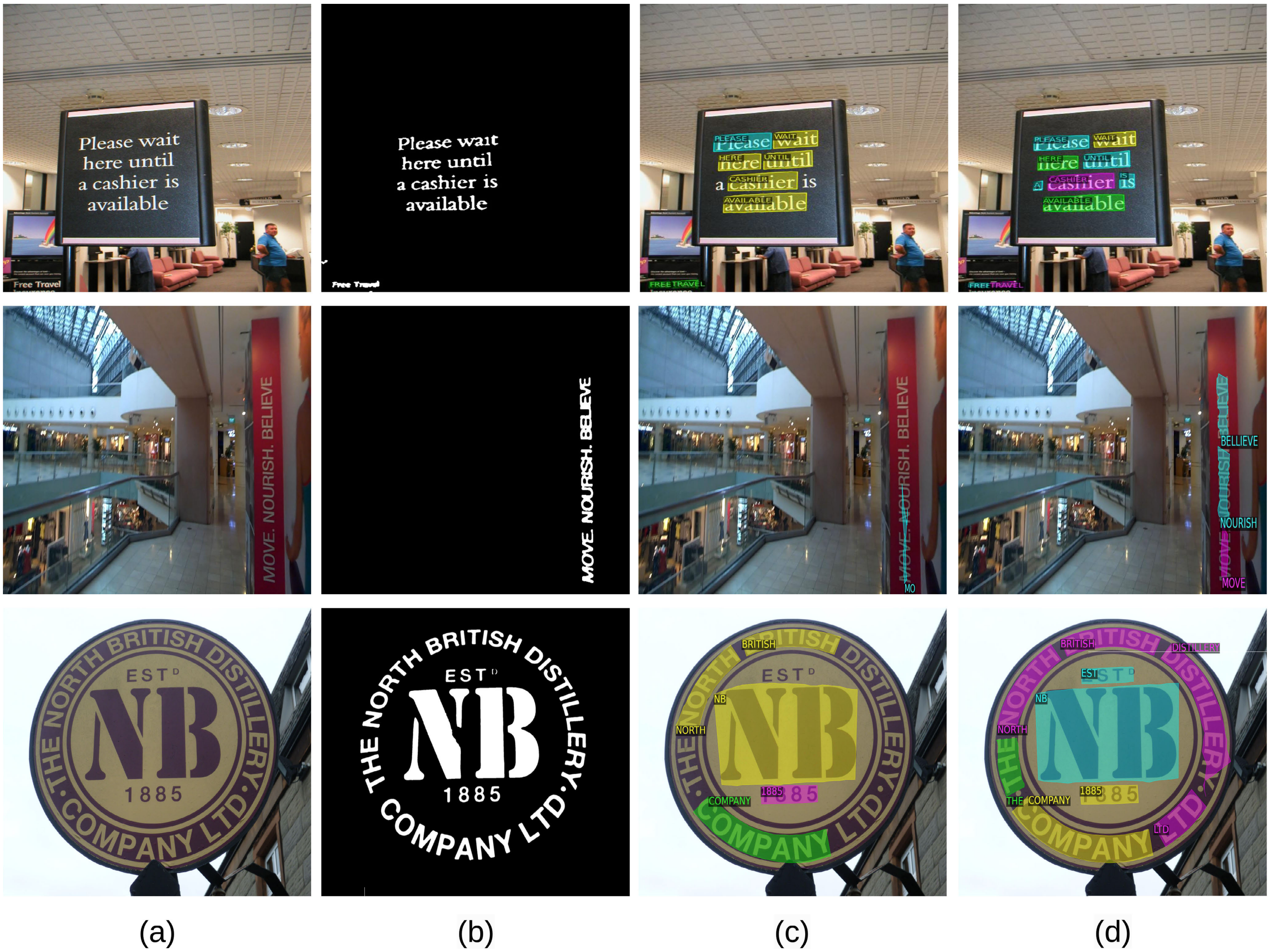


FIGURE 13. The qualitative results of our proposed Text Spotting approach are presented in (a), (b), (c), and (d), showcasing the original images, our predicted mask as an assistant, the output of the original DeepSolo, and our text spotting result, respectively. The first, second, and third rows represent samples from ICDAR13, ICDAR15, and Total-Text, respectively.

we introduced a 4-channel input comprising RGB images and an additional text segmentation mask channel. To implement this modification, we duplicated the original network and transferred the weights from a pre-trained model, excluding the replaced input layer. The modified network was then fine-tuned until convergence, as depicted in Figure 12.

Through our experiments, we discovered that incorporating the mask channel resulted in a significant improvement in text-spotting accuracy. In Table 4, we compare the performance of our modified network with the original text-spotting network that used 3-channel images as input. The results demonstrated a clear advantage in favour of our approach. We attribute this improvement to the mask channel acting as an attention map, effectively guiding the text-spotting network to focus on relevant areas of the scene. To validate our findings, we conducted thorough evaluations using state-of-the-art networks such as DeepSolo [52]. Performance metrics and illustrations are in Table 4 and Figure 13.

Following DeepSolo [52], we trained our model on the Synth150k synthetic text dataset and evaluated its performance on the ICDAR13, ICDAR15, and Total Text datasets. Initialization of DeepSolo involved using an officially pre-trained model, with the exception of the replaced input layer, which underwent fine-tuning. The results, presented in Table 4, demonstrate the superiority of our text spotting model compared to the original DeepSolo (ResNet-50), achieving higher accuracy in the ICDAR13, ICDAR15, and Total Text datasets by 1.3%, 1.2%, and 1.6% respectively.

Figure 13 provides visual examples to support our findings. It showcases three representative samples where our predicted masks effectively corrected mis-spotted words. For instance, in the ICDAR15 sample, the original DeepSolo network failed to identify all words. However, with the assistance of our predicted mask, these words were accurately spotted. Similar improvements were observed in other examples as well.

V. CONCLUSION

Pixel-level scene text segmentation has recently become an emerging topic and has proven quite challenging. We rethink its problem by introducing Text Image-conditional GANs that increase new paired data to the available limited datasets. Furthermore, we also propose the Multi-task Cascade Transformer to effectively learn the collected data, including the real-world and generated synthetic data. The framework opens a new approach to scene text segmentation and other fields in cases lacking data. Experiments show that our method is superior to state-of-the-art models. However, it is not an end-to-end framework.

REFERENCES

[1] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. I. Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. de las Heras, "ICDAR 2013 robust reading competition," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 1484–1493.

[2] C. K. Ch'ng and C. S. Chan, "Total-text: A comprehensive dataset for scene text detection and recognition," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Nov. 2017, pp. 935–942.

[3] X. Xu, Z. Zhang, Z. Wang, B. Price, Z. Wang, and H. Shi, "Rethinking text segmentation: A novel dataset and a text-specific refinement approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12040–12050.

[4] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "COCO-text: Dataset and benchmark for text detection and recognition in natural images," 2016, *arXiv:1601.07140*.

[5] N. Nayef, F. Yin, I. Bizid, H. Choi, Y. Feng, D. Karatzas, Z. Luo, U. Pal, C. Rigaud, J. Chazalon, W. Khlif, M. M. Luqman, J. Burie, C. Liu, and J. Ogier, "ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification—RRC-MLT," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Nov. 2017, pp. 1454–1459.

[6] N. Nayef, Y. Patel, M. Busta, P. N. Chowdhury, D. Karatzas, W. Khlif, J. Matas, U. Pal, J. Burie, C. Liu, and J. Ogier, "ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition—RRC-MLT-2019," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 1582–1587.

[7] C. K. Chng, Y. Liu, Y. Sun, C. C. Ng, C. Luo, Z. Ni, C. Fang, S. Zhang, J. Han, E. Ding, J. Liu, D. Karatzas, C. S. Chan, and L. Jin, "ICDAR2019 robust reading challenge on arbitrary-shaped text—RRC-ArT," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 1571–1576.

[8] C. Wang, S. Zhao, L. Zhu, K. Luo, Y. Guo, J. Wang, and S. Liu, "Semi-supervised pixel-level scene text segmentation by mutually guided network," *IEEE Trans. Image Process.*, vol. 30, pp. 8212–8221, 2021.

[9] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *Int. J. Comput. Vis.*, vol. 116, no. 1, pp. 1–20, Jan. 2016.

[10] Y. Tang and X. Wu, "Scene text detection and segmentation based on cascaded convolution neural networks," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1509–1520, Mar. 2017.

[11] S. Bonechi, M. Bianchini, F. Scarselli, and P. Andreini, "Weak supervision for generating pixel–level annotations in scene text segmentation," *Pattern Recognit. Lett.*, vol. 138, pp. 1–7, Oct. 2020.

[12] S. Bonechi, P. Andreini, M. Bianchini, and F. Scarselli, "COCO_TS dataset: Pixel–level annotations based on weak supervision for scene text segmentation," in *Proc. Int. Conf. Artif. Neural Netw.* Cham, Switzerland: Springer, 2019, pp. 238–250.

[13] Y. Ren, J. Zhang, B. Chen, X. Zhang, and L. Jin, "Looking from a higher-level perspective: Attention and recognition enhanced multi-scale scene text segmentation," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 3138–3154.

[14] Q.-V. Dang and G.-S. Lee, "Scene text segmentation by paired data synthesis," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2023, pp. 1–16.

[15] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–15.

[16] L. Wu, C. Zhang, J. Liu, J. Han, J. Liu, E. Ding, and X. Bai, "Editing text in the wild," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1500–1508.

[17] Y. Yang, Y. Zhuang, and Y. Pan, "Multiple knowledge representation for big data artificial intelligence: Framework, applications, and case studies," *Frontiers Inf. Technol. Electron. Eng.*, vol. 22, no. 12, pp. 1551–1558, Dec. 2021.

[18] S. Lee, M. Cho, K. Jung, and J. H. Kim, "Scene text extraction with edge constraint and text collinearity," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3983–3986.

[19] J. Zhang and R. Kasturi, "Character energy and link energy-based text extraction in scene images," in *Proc. 10th Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2011, pp. 308–320.

[20] A. Mishra, K. Alahari, and C. V. Jawahar, "An MRF model for binarization of natural scene text," in *Proc. Int. Conf. Document Anal. Recognit.*, Sep. 2011, pp. 11–16.

[21] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn. (ICML)*. San Francisco, CA, USA: Morgan Kaufmann, 2001, pp. 282–289.

[22] Y. Gu, Y. Wang, H. Zhang, J. Wu, and X. Gu, "Enhancing text classification by graph neural networks with multi-granular topic-aware graph," *IEEE Access*, vol. 11, pp. 20169–20183, 2023.

[23] Q. Dang and G. Lee, "Document image binarization with stroke boundary feature guided network," *IEEE Access*, vol. 9, pp. 36924–36936, 2021.

[24] S. Li, "Fall detection with wrist-worn watch by observations in statistics of acceleration," *IEEE Access*, vol. 11, pp. 19567–19578, 2023.

[25] Q.-V. Dang and G.-S. Lee, "Document image binarization by GAN with unpaired data training," *Int. J. Contents*, vol. 16, no. 2, pp. 8–18, 2020.

[26] C. Tiago, S. R. Snare, J. Šprem, and K. McLeod, "A domain translation framework with an adversarial denoising diffusion model to generate synthetic datasets of echocardiography images," *IEEE Access*, vol. 11, pp. 17594–17602, 2023.

[27] G. Chai, Z. Wang, G. Guo, Y. Chen, Y. Jin, W. Wang, and X. Zhao, "Recurrent attention dense network for single image de-raining," *IEEE Access*, vol. 8, pp. 111278–111288, 2020.

[28] Y. Mi, S. Yuan, X. Li, and J. Zhou, "Dense residual generative adversarial network for rapid rain removal," *IEEE Access*, vol. 9, pp. 24848–24858, 2021.

[29] Y. Fan, Y. Shi, N. Zhang, and Y. Chu, "Image inpainting based on structural constraint and multi-scale feature fusion," *IEEE Access*, vol. 11, pp. 16567–16587, 2023.

[30] K. Shi, M. Alrabeiah, and J. Chen, "Progressive with purpose: Guiding progressive inpainting DNNs through context and structure," *IEEE Access*, vol. 11, pp. 2023–2034, 2023.

[31] W. Xu, K. Shawn, and G. Wang, "Toward learning a unified many-to-many mapping for diverse image translation," *Pattern Recognit.*, vol. 93, pp. 570–580, Sep. 2019.

[32] F. Zhan, C. Xue, and S. Lu, "GA-DAN: Geometry-aware domain adaptation network for scene text detection and recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9104–9114.

[33] F. Zhan, S. Lu, and C. Xue, "Verisimilar image synthesis for accurate detection and recognition of texts in scenes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 249–266.

[34] F. Zhan, H. Zhu, and S. Lu, "Spatial fusion GAN for image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3648–3657.

[35] L. Kang, P. Riba, M. Rusiñol, A. Fornés, and M. Villegas, "Pay attention to what you read: Non-recurrent handwritten text-line recognition," *Pattern Recognit.*, vol. 129, Sep. 2022, Art. no. 108766.

[36] A. U. Haque, S. Ghani, and M. Saeed, "Image captioning with positional and geometrical semantics," *IEEE Access*, vol. 9, pp. 160917–160925, 2021.

[37] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, and M. Chiaberge, "Action transformer: A self-attention model for short-time pose-based human action recognition," *Pattern Recognit.*, vol. 124, Apr. 2022, Art. no. 108487.

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[39] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007.

[40] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[41] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-PVT: Polyp segmentation with pyramid vision transformers," 2021, *arXiv:2108.06932*.
[42] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with Fourier convolutions," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 3172–3182.
[43] E. O. Brigham and R. E. Morrow, "The fast Fourier transform," *IEEE Spectr.*, vol. S-4, no. 12, pp. 63–70, Dec. 1967.
[44] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.
[45] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.
[46] Z. Wan, J. Zhang, D. Chen, and J. Liao, "High-fidelity pluralistic image completion with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4672–4681.
[47] Y. Jin, D. Han, and H. Ko, "TrSeg: Transformer for semantic segmentation," *Pattern Recognit. Lett.*, vol. 148, pp. 29–35, Aug. 2021.
[48] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with Atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.
[49] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 173–190.
[50] G. Bradski, "The OpenCV library," *Dr. Dobb's J. Softw. Tools Prof. Program.*, vol. 25, no. 11, pp. 120–123, 2000.
[51] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Ann. Oper. Res.*, vol. 134, no. 1, pp. 19–67, Feb. 2005.
[52] M. Ye, J. Zhang, S. Zhao, J. Liu, T. Liu, B. Du, and D. Tao, "DeepSolo: Let transformer decoder with explicit points solo for text spotting," 2022, *arXiv:2211.10772*.
[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[54] M. Liao, G. Pang, J. Huang, T. Hassner, and X. Bai, "Mask textspotter v3: Segmentation proposal network for robust scene text spotting," in *Proc. 16th Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 706–722.

QUANG-VINH DANG received the B.E. degree in mechatronics technology from the Ho Chi Minh City University of Technology and Education, in 2011, the master's degree in mechatronics engineering from the Ho Chi Minh City University of Technology, in 2013, and the Ph.D. degree in artificial intelligence convergence from Chonnam National University, South Korea, in 2022, where his research focused on developing deep learning models. He is currently a Computer Scientist specializing in computer vision and deep learning. He plans to pursue an academic career and continue his research in the field of computer vision and deep learning.

GUEE-SANG LEE received the B.S. degree in electrical engineering and the M.S. degree in computer engineering from Seoul National University, South Korea, in 1980 and 1982, respectively, and the Ph.D. degree in computer science from The Pennsylvania State University, in 1991. He is currently a Professor with the Department of Electronics and Computer Engineering, Chonnam National University, South Korea. His main research interests include image processing, computer vision, and video technology.