Don't Forget Me: Accurate Background Recovery for Text Removal via Modeling Local-Global Context

Chongyu Liu[1], Lianwen Jin*[1,4,5], Yuliang Liu[2], Canjie Luo[1], Bangdong Chen[1], Fengjun Guo[3], and Kai Ding[3]

[1] South China University of Technology, Guangzhou, Guangdong, China
{liuchongyu1996, lianwen.jin, canjie.luo}@gmail.com
[2] Huazhong University of Science and Technology, Wuhan, Hubei, China
ylliu@hust.edu.cn
[3] IntSig Information Co. Ltd, Shanghai, China
{fengjun_guo, danny_ding}@intsig.net
[4] Pazhou Laboratory (Huangpu), Guangzhou, Guangdong, China
[5] Peng Cheng Laboratory, Shenzhen, Guangdong, China

Abstract. Text removal has attracted increasingly attention due to its various applications on privacy protection, document restoration, and text editing. It has shown significant progress with deep neural network. However, most of the existing methods often generate inconsistent results for complex background. To address this issue, we propose a Contextual-guided Text Removal Network, termed as CTRNet. CTRNet explores both low-level structure and high-level discriminative context feature as prior knowledge to guide the process of text erasure and background restoration. We further propose a Local-global Content Modeling (LGCM) block with CNNs and Transformer-Encoder to capture local features and establish the long-term relationship among pixels globally. Finally, we incorporate LGCM with context guidance for feature modeling and decoding. Experiments on benchmark datasets, SCUT-EnsText and SCUT-Syn show that CTRNet significantly outperforms the existing state-of-the-art methods. Furthermore, a qualitative experiment on examination papers also demonstrates the generalizability of our method. The code of CTRNet is available at https://github.com/lcy0604/CTRNet.

Keywords: GAN, Text Removal, Context Guidance, Transformer

1 Introduction

In recent years, text removal has attracted increasing research interests in the computer vision community. It aims to remove the text and fill the regions with plausible content. Text removal can help avoid privacy leaks by hiding some private messages such as ID numbers and license plate numbers. Besides, it can be widely used for document restoration in the field of intelligent education. It is also a crucial prerequisite step for text editing [49, 52, 53, 20, 37] and has wide applications in areas such as augmented reality translation.

(a) Input (b) w/o Context (c) w Context (d) GT

Fig. 1. Examples of scene text removal, which also show the comparison of the results with and without context guidance and feature modeling. Zoom-in for best view.

Recent text removal methods [56, 23, 48, 38, 40] have achieved significant improvements with the development of GAN [10, 27, 28]. Though the state-of-the-art methods [23, 48, 38, 39] have reported promising performance, the restoration for complex backgrounds still remains a main challenge. To solve this problem, some researchers propose to directly predict the text stroke [38, 3] and focus text region inpainting only on these stroke regions. However, text stroke prediction is an another challenging problem to be addressed, especially on image-level (with more than one text) [51, 44]. Inspired by previous image inpainting methods [25, 35, 32], we consider that directly transforming the raw image to a final text-erased image in a unified framework is one of the major causes of inconsistent results for text removal. This is due to the imbalance between text erasure and the subsequent background restoration. The corruption of text region while erasing may mislead the reconstruction of the high-frequency textures. The results with blur and artifacts are shown in Fig. 1 (b). To address this issue, we propose to mine more efficient context guidance from the existing data in a step-by-step manner to reduce the artifacts of text-erased regions and produce plausible content.

Specifically, we propose a novel text removal model, termed as CTRNet. CTRNet decouples the text removal task into four main components: Text Perception Head, Low-level/High-level Contextual Guidance blocks (LCG, HCG), and a Local-global Content Modeling (LGCM) block, as shown in Fig. 2. Text Perception Head is firstly introduced to detect the text regions and generate text masks. Subsequently, the LCG predicts the structure of text-erased images to provide low-level contextual priors, which is represented by the edge-preserved smoothing method RTV [50]. Besides, we incorporate an HCG block to learn the high-level discriminative context in latent feature space as another guidance. Structure is served as a local guide for the image encoder, while high-level context provides global knowledge. As the filling of text regions not only focuses on the information of their own and surroundings, but also uses the global affinity as reference, CTRNet introduces LGCM by the cooperation of CNNs and Transformer-Encoder [41] to extract local features and establish the long-term global relationship among the pixels, meanwhile incorporates context guidance for both feature modeling and decoding phase. Through such designs, CTRNet

can capture sufficient contextual information to remove the text more thoroughly and restore backgrounds with more visually pleasing textures, as shown in Fig. 1 (c).

Extensive experiments on the benchmark datasets, SCUT-EnsText [23] and SCUT-Syn [56] are conducted to verify the effectiveness of CTRNet. Additionally, qualitative experiment is conducted on an in-house examination paper dataset to verify the generalizability of our model.

Text removal takes complete text image as input and aims to preserve the original background of text regions, whereas image inpainting will directly mask the regions for restoration based only on the surrounding texture. Simply applying image inpainting methods to text removal will cause inaccurate background generation. We conduct experiments to compare our method with the state-of-the-art image inpainting models in Sec.4.5/4.6, which practically illustrates the difference between these two tasks.

We summarize the contributions of this work as follows:

- We propose to learn both Low-level and High-level Contextual Guidance (LCG, HCG), which we find are important and useful as prior knowledge for text erasure and subsequent background texture synthesis.
- We propose Local-global Content Modeling blocks (LGCM) to extract local features and capture long-range dependency among the pixels globally.
- The context guidance is incorporated into LGCM for the feature modeling and decoding phase, which further promotes the performance of CTRNet.
- Extensive experiments on the benchmark datasets demonstrate the effectiveness of CTRNet not only in removing the text but recovering the background textures as well, significantly outperforming existing SOTA methods.

2 Related work

Deep learning-based text removal can be categorized into one-stage methods and two-stage methods. One-stage methods are implemented in an end-to-end manner, requiring models to automatically detect the text regions and remove them in a unified framework. Nakamura et al. [29] proposed a patch-based auto-encoder [2] with skip connections, termed as SceneTextEraser. It was also the first DNN-based text removal method. Text removal can be also regarded as image-to-image translation. Following the idea of Pix2pix [14], EnsNet [56] adopted four refined losses and employed a local-aware discriminator to maintain the consistency of text-erased regions. Liu et al. [23] proposed EraseNet by introducing a coarse-to-refinement architecture and an additional segmentation head to help locate the text. MTRNet++ [39] shared the same spirit with EraseNet, but separately encoded the image content and text mask in two branches. Cho et al. [6] proposed to jointly predict the text stroke and inpaint the background, allowing the model to focus only on the restoration of text stroke regions. Wang et al. [48] presented PERT, which contained a novel progressive structure with shared parameters to remove text more thoroughly, and a region-based modification strategy to effectively guide the erasure process only on text regions.

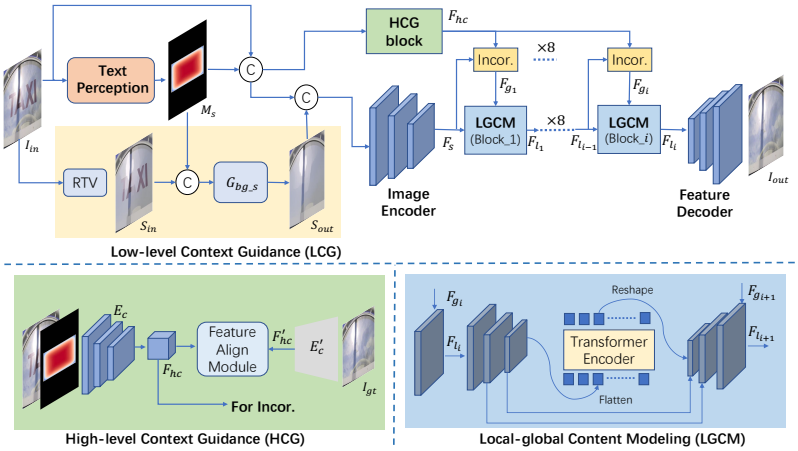


Fig. 2. The overview of the proposed CTRNet.

Two-stage methods follow the procedure of detecting the text, removing it, and then filling the background with plausible content. We further divide them into word-level and image-level. Word-level methods first crop the text regions according to the detected results, then operate the text removal process with single text [34, 38]. Qin et al. [34] utilized cGAN [10, 27] with one encoder and two decoders for both text stroke prediction and background inpaint. Tang et al. [38] proposed to predict the text strokes on word images, then both strokes and images were fed into an image inpainting network with Partial Convolution [24] to generate the text-erased results. For image-level methods, after obtaining the text mask through detection, they directly predict the results on the entire images. MTRNet [40], based on Pix2pix, implemented a text mask as an extra input. The method proposed by Keserwani et al. [18] was similar to MTRNet, but employed an additional local discriminator for better prediction. Zdenek et al. [55] considered the lack of pixel-wise training data and proposed a weak supervision method by introducing a pretrained PSENet [46] to detect the text, and then inpainted the text regions through another pretrained image inpainting method [58]. Conrad et al. [7] borrowed the concept developed by Zdenek et al. [55], but they proposed to further predict the text stroke before the application of a pretrained EdgeConnect [32] for background inpainting. Bian et al. [3] proposed a cascaded generative model, which decoupled text removal into text stroke detection and stroke removal.

3 Proposed Method

Fig. 2 shows the pipeline of the proposed CTRNet. First, we introduce text perception head to detect the text regions and generate the text masks. To better restore the backgrounds of text regions, we propose to learn more contextual priors from the existing data, including low-level background structure with LCG and high-level context features with HCG. Structure information is served as a

local guide and directly fed into the image encoder, while the high-level context feature is embedded into the high-dimensional feature space as a global guide with the Incor operation. Finally, we propose LGCM blocks to capture both local features and long-term correlation among all the pixels, so that CTRNet can make full use of different levels of information for feature decoding.

3.1 Text Perception Head

For scene text removal on image-level, purely feeding a text image into a model without any positional indication results in failed, mistaken, and incomplete erasures of text regions [56, 23, 48]. Therefore, we introduce a text perception head to help localize the text regions. With the detected results, we generate the corresponding masks and send them together with original images into the subsequent network. We propose to replace the original 0-1 mask (hard mask) with soft mask to help eliminate the defects and discontinuities between text regions and non-text regions. The procedure for soft mask generation is as follows: (1) The vanilla bounding boxes B are shrunk using the Vatti clipping algorithm [42] with the ratio of 0.9 to obtain B_s, meanwhile dilated with the same offset to B_d; (2) The soft border of text regions is defined as the minimum distance between the pixel in B_s and B_d. Fig. 3 (c) displays the example of soft-mask. Only the pixels in B_s are set to 1, while the range of pixels between B_s and B_d is $(0, 1)$. The effectiveness of the soft mask is verified in Section 4.3.

3.2 Contextual Guidance Learning

Low-level Contextual Guidance (LCG) block: Scene text removal aims to not only erase the text, but also restore the backgrounds of text regions and synthesize their corresponding textures. Previous methods [56, 23, 39, 48] follow an end-to-end training and inference procedure by directly predicting the results with scene text images as input. However, they suffer from some texture artifacts when dealing with complicated backgrounds, as shown in Fig. 4 and 5. We propose to first predict the low-frequency structure of the image, and take it as low-level guidance for the subsequent network. Inspired by Ren et al. [35] and Liu et al. [25], the structure image is constructed by the edge-preserved smooth method RTV [50], which removes high-frequency textures with only sharp edges and smooth structure remain. RTV consists of a pixel-wise windowed total variation measure and a windowed inherent variation to remove image texture. Fig. 3 (d) and (e) display an example of the structure image S_{in} and its ground-truth S_{gt} generated from I_{in} and I_{gt}, respectively. Learning a mapping between two low-frequency structures, S_{in} and S_{gt}, is much easier than removing text directly. The structural clues for text regions can effectively simplify texture generation and enhance the performance by indicating the structure semantic of text regions, as shown in Fig. 4 (e) (f) in the ablation study.

As shown in Fig. 2, LCG block consists of RTV method and a background structure generator G_{bg_s}. G_{bg_s} is an encoder-decoder architecture that takes both the structure S_{in} of scene text images and the soft mask M_s as input, and

(a) Input (b) HM (c) SM (d) S_{in} (e) S_{gt} (f) GT

Fig. 3. The basic elements of CTRNet. HM and SM denote hard mask and soft mask, respectively. S_{in} and S_{gt} represent the structure of the input and ground-truth. Zoom in for best view.

predicts the background structure S_{out} with text-erased. We take S_{out} as local guidance, and directly feed it into the image encoder with I_{in} to encode image features $F_s \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$.

High-level Contextual Guidance (HCG) block: In addition to the low-level structure priors, we propose to explore potential high-level contextual guidance in latent feature spaces. Previous study [25, 35, 23] with Perceptual/Style Loss [15, 9] demonstrates the effectiveness of high-level contextual supervision for image generation and translation. Therefore, we make our CTRNet to utilize such discriminative context as additional guidance information for both text removal and background restoration, instead of taking it merely as supervision for optimization. Inspired by Zhang et al. [57], we incorporate an HCG block into our CTRNet to learn high-level context features.

The architecture of HCG block is illustrated in the left-bottom of Fig. 2. The block consists of two feature encoders ($E_c(\cdot)$ and $E_c^{'}(\cdot)$), and a Feature Align Module (FAM), as done in [57]. $E_c(\cdot)$ encodes the concatenation of the original image I_{in} and its soft-mask M_s to obtain the features $F_{hc} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$, whereas $E_c^{'}(\cdot)$ extracts the context features $F_{hc}^{'} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$ from the paired labels I_{gt}. Here, $E_c^{'}(\cdot)$ is a classification model, termed as TResNet [36]. We directly use its pretrained model on the OpenImages datasets [21] to extract $F_{hc}^{'}$ with frozen weights during the training procedure. After feature dimension mapping with 1×1 convolution layers in FAM, feature align loss L_{align} is applied to approximate the distribution of F_{hc} to $F_{hc}^{'}$. The process can be formulated as

$$\begin{aligned} F_{hc}^{'} = E_c^{'}(I_{gt}); F_{hc} = E_c(I_{in}, M_s) \\ L_{align} = \left\| F_{hc} - F_{hc}^{'} \right\|_1 * (1 + \alpha M_s) \end{aligned} \tag{1}$$

α is set to 2.0. In this way, F_{hc} based on I_{in} can be transferred to contain context information of background I_{gt}, which can provide a high-level global guidance for feature modeling and decoding.

3.3 Local-global Content Modeling (LGCM)

While erasing text regions and filling them with reasonable textures as background, beyond considering text regions as a reference, it is necessary for a text

removal method to use the pixel information from the surrounding and global backgrounds. Therefore, we propose a feature content modeling block for both local (text regions) and global (surrounding and the entire background) levels. As shown in Fig. 2, the image content features $F_s \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$, incorporated with the high-level discriminative feature guidance, F_{hc} are sent to LGCM to model the local-global contextual features and enhance their representations. And the right-bottom of Fig. 2 displays the architecture of a single LGCM block.

CNNs operate locally at a fixed size (e.g. 3×3) to effectively extract features of specific regions and establish the relationship between the pixels in each local window. Therefore, four stacked vanilla 4×4 convolutions layers are utilized for local content modeling. In addition, features can be downsampled by CNNs to reduce the computation required for the subsequent global modeling operation. For global content modeling, we apply Transformer-Encoder as our basic module. Transformer-Encoder [41], which can effectively capture global interactions between pixels among the whole features and model their long-range dependency. Then two deconvolution layers are applied to upsample the modeled features and bring the inductive bias of CNN [22]. LGCM follows an iterative process with k stages ($k = 8$ empirically [57]). At the final convolution of each stage, F_{hc} are incorporated into the LGCM with ResSPADE [33, 57]. The details for LGCM and ResSPADE are presented in supplement materials. The output of the $i-th$ LGCM is denoted as F_{l_i} ($F_{l_0} = F_s$).

Finally, Feature Decoder reconstructs the final text-erased output by decoding both features F_{l_8} from the final LGCM ($8th$) block and shadow content features F_s. , which can be formulated as

$$I_{out} = H_{fd}(F_{l_8} + F_s) \tag{2}$$

3.4 Training Objective

We adopt the following losses to train our text removal network, including structure loss, multi-scale text-aware reconstruction loss, perceptual loss, style loss, and adversarial loss.

Structure loss The structure loss is used to measure the L_1 distance between the background structure output S_{out} and the ground truth S_{gt}, which is defined by:

$$L_{str} = \|S_{gt} - S_{out}\|_1 * (1 + \gamma M_s) \tag{3}$$

$(1 + \gamma M_s)$ denotes higher weight for text region. γ is set to 3.0.

Multi-scale text-aware reconstruction loss The $L_1 - norm$ difference is proposed to measure the output and the ground truth. We first predict multi-layer outputs with text removed in different sizes, then assign higher weight to text regions when computing the loss:

$$L_{msr} = \sum_n \|(I_{out_n} - I_{gt_n})\|_1 * (1 + \theta_n M_s) \tag{4}$$

n denotes n-th output in the scales of $\frac{1}{16}$, $\frac{1}{4}$ and 1 of the input. $\theta_1, \theta_2, \theta_3$ is set to $2, 3, 4$, respectively.

Perceptual loss Except for low-level image-to-image supervision with reconstruction loss, we also adopt perceptual loss [15] to capture high-level semantics and try to simulate human perception of image quality. Both the straight output I_{out} and the original image with text-removed I_{com} are included as loss terms. Besides, the structure output S_{out} is also taken into consideration.

$$I_{com} = I_{in} * (1 - M_s) + I_{out} * M_s \tag{5}$$

$$L_{per} = \sum_i \sum_j \left\| \phi_j(I_i) - \phi_j(I_{gt}) \right\|_1 + \sum_j \left\| \phi_j(S_{out}) - \phi_j(S_{gt}) \right\|_1 \tag{6}$$

where I_i represent I_{out} and I_{com}. $\phi_j(.)$ denotes the activation maps of the j-th ($j = 1, 2, 3$) pooling layer of VGG-16 pretrained on ImageNet [8].

Style loss We also utilize style loss to release the artifacts of the generated results. Style loss [9] construct a Gram matrix $Gr(.)$ from each selected activation map in perceptual loss. Style loss can be defined as

$$L_{style} = \sum_i \sum_j \frac{\left\| Gr_j(I_i) - Gr_j(I_{gt}) \right\|_1}{H_j W_j C_j} + \sum_j \frac{\left\| Gr_j(S_{out}) - Gr_j(S_{gt}) \right\|_1}{H_j W_j C_j} \tag{7}$$

Adversarial loss The adversarial loss encourages our model to generate more plausible details for the final results with text removed. Here we defined our adversarial loss as:

$$L_{adv} = E_{x \sim P_{\text{data}}(x)} \left[\log D(x) \right] + E_{x \sim P_{z(z)}} \left[\log \left(1 - D(G(z)) \right) \right] \tag{8}$$

z is the input I_{in} and x represents the corresponding ground-truth I_{gt}.

Total loss The overall loss function for our text removal network is defined as:

$$\begin{aligned} L_{total} = \lambda_{al} L_{align} + \lambda_{str} L_{structure} + \lambda_m L_{msr} \\ + \lambda_p L_{per} + \lambda_s L_{style} + \lambda_a L_{adv} \end{aligned} \tag{9}$$

$\lambda_{al}, \lambda_{str}, \lambda_m, \lambda_p, \lambda_s, \lambda_a$ are the trade-off parameters. In our implementation, we empirically set $\lambda_{al} = 1.0, \lambda_{str} = 2.0, \lambda_m = 10.0, \lambda_p = 0.01, \lambda_s = 120, \lambda_a = 1.0$.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets To evaluate the effectiveness of our proposed CTRNet, we conduct experiments on the two widely used benchmarks, SCUT-Syn [56] and SCUT-EnsText [23].

(1) SCUT-Syn: SCUT-Syn contains a training set of 8,000 images and a testing set of 800 images. It is a synthetic dataset with [12]. The background images are mainly collected from ICDAR-2013 [17] and MLT-2017 [30], and the text instances are manually erased.

(2) SCUT-EnsText: SCUT-EnsText is a comprehensive real-world dataset with 2,749 images for training and 813 images for testing. These images are

Table 1. Ablation Study on SCUT-EnsText. MSSIM and MSE are represented by % in the table.

	Components				Evaluation on I_{out}				Evaluation on I_{com}			
	HCG	LGCM	SM	LCG	PSNR	MSSIM	MSE	FID	PSNR	MSSIM	MSE	FID
baseline	-	-	-	-	32.39	95.45	0.13	20.75	33.21	95.52	0.11	22.15
Ours+	✓				32.90	96.62	0.11	17.40	34.88	97.09	0.10	19.42
Ours+	✓	✓			35.10	97.36	0.09	14.36	35.30	97.20	0.09	17.91
Ours+	✓	✓	✓		35.16	**97.38**	0.09	14.33	35.83	**97.42**	0.09	15.02
Ours+	✓	✓	✓	✓	**35.20**	97.36	**0.09**	**13.99**	**35.85**	97.40	**0.09**	**14.57**

collected from public scene text benchmark, including ICDAR-2013 [17], ICDAR-2015 [16], MS COCO-Text [43], SVT [45], MLT-2017 [30], MLT-2019 [31], and ArTs [4], which consists of SCUT-CTW1500 [54] and Total-Text [5]. All the images are carefully annotated with Photoshop.

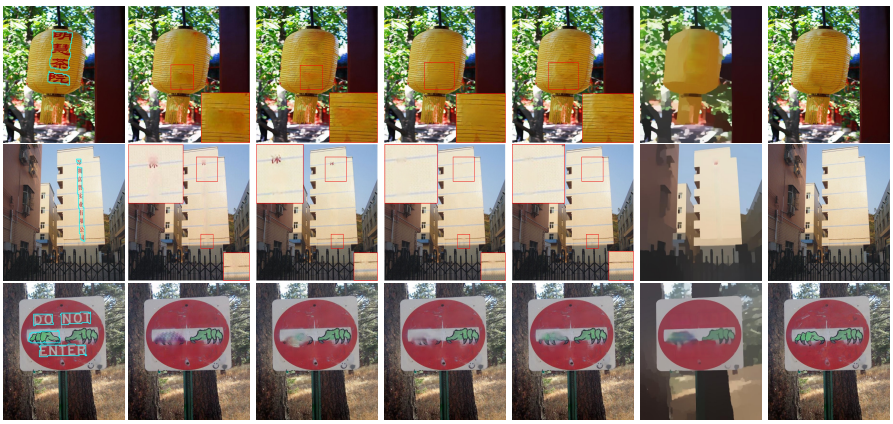
Evaluation metrics: To comprehensively evaluate the performance of our CTRNet, we utilize both Image-Eval and Detection-Eval as used in EraseNet [23]. (1) Image-Eval includes the following metrics for image quality evaluation. (1) Peak signal to noise ratio (PSNR); (2) Multi-scale Structural Similarity (MSSIM); (3) Mean Square Error (MSE); (4) Fréchet Inception Distance (FID) [13]. A higher PSNR, MSSIM and lower MSE, FID denotes better results. (2) Detection-Eval evaluates the Recall (R), Precision (P), F-measure (F), TIoU-Recall (TR), TIoU-Precision (TP), and TIoU-F-measure (TF) for the results under the protocols of ICDAR 2015 [16] and T-IoU [26]. CRAFT [1] is served as the text detector for evaluation. The lower R, P and F indicate that more text can be removed.

4.2 Implement details

We utilize Pixel Aggregation Network (PAN) [47] as text perception head for CTRNet. The input size is set to 512×512. Adam solver [19] is used to optimized our method, and the β is set to (0.0, 0.9) as default.The batch size is set to 2. All experiments are conducted on a workstation with two NVIDIA 2080TI GPUs. More training details and the architectures of each component are provided in the supplementary materials.

4.3 Ablation Study

In this section, we conduct experiments on SCUT-EnsText to verify the contributions of different components in CTRNet. Our baseline model is implemented by a Pix2pix-based model, which takes both the images and the corresponding masks as input. As the text perception head is frozen when training the other components, the detected text regions remain the same in each experiment during inference; thus, we only employ Image-Eval to evaluate the performance. Apart from the direct output I_{out}, we also paste the erased text regions back to the input images based on the detected results to obtain I_{com}. The quantitative



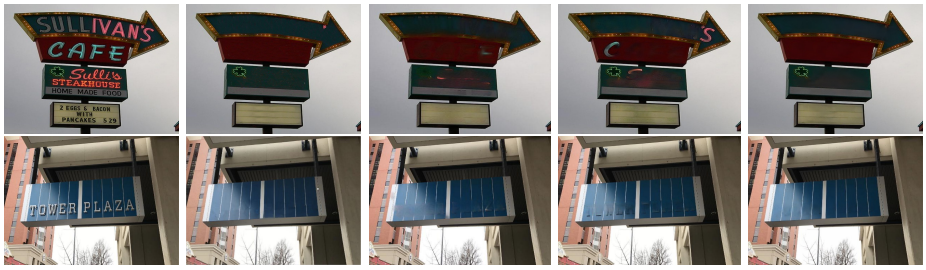
(a) (b) (c) (d) (e) (f) (g)

Fig. 4. Qualitative results for ablation studies on HCG, LGCM, LCG. (a) The input images; (b) Baseline results; (c) Baseline + HCG; (d) Baseline + HCG + LGCM; (e) Baseline + HCG + LGCM + LCG; (f) The structure output from LCG for the model (e);(g) The ground-truth. Zoom in for best view.

results for both outputs are presented in Table 1. The qualitative results are displayed in Fig. 4. Besides, we evaluate each loss item and their corresponding hyper-parameters, and the results are presented in our supplement materials.

HCG: HCG block aims to learn high-level discriminative context feature representation, which can effectively guide the process of feature modeling and decoding. As shown in Table 1, the incorporation of HCG into the modeling and decoding phase with ResSPADE blocks yields significant improvement on all metrics, with the increases of 0.51, 1.17, 0.02, 3.35 for I_{out} and 1.67, 1.57, 0.01, 2.73 for I_{com} in PSNR, MSSIM, MSE, and FID, respectively. The qualitative results shown in Fig. 4 also illustrate the effect of this component. Comparing with the results of the baseline model in Fig. 4 (b), the results in Fig. 4 (c) indicate that the HCG block can help generate a more plausible background and release more artifacts in the output.

LGCM: As shown in Table 1, the incorporation of our LGCM significantly facilitates performance improvement of 2.20, 0.74, 0.02, 3.04 for I_{out} in PSNR, MSSIM, MSE, and FID, respectively, while 0.42, 0.11, 0.01, 1.51 for I_{com}. Such a remarkable promotion benefits a lot from both the local and global modeling for the features and the learned context prior, which can capture not only the long-range dependency among pixels around the feature maps but their relationship at a fixed window as well. Therefore, LGCM enables our CTRNet to take advantage of both local and global information. The qualitative results are presented in Fig. 4 (d). In comparison, the outputs of our model without LGCM exhibit some obvious defects on the text regions (the up/bottom row of Fig. 4 (c)), while those with LGCM are more favorable, though there still exist mistaken erasure (e.g. the bottom of Figure 4 (d), but the restored background is smoother than (c)). Besides, with the long-rang dependency, the text can be removed more



(a) Input (b) GT (c) EN[23] (d) Stroke[38] (e) Ours

Fig. 5. Qualitative results on SCUT-EnsText for comparing our model with previous scene text removal methods. EN denotes EraseNet [23] and Stroke denotes the method proposed by Tang et al. [38]. Zoom in for best view.

Table 2. Comparison with state-of-the-art methods on SCUT-EnsText. The methods with "*" denote that the text mask are generated with the GT instead of the detected results. MSSIM and MSE are represented by % in the table. Bold indicates SOTA, while Underline indicates second best.

Methods	Image-Eval				Detection-Eval					
	PSNR	MSSIM	MSE	FID	R	P	H	T-R	T-P	T-H
Original	-	-	-	-	69.5	79.4	74.1	50.9	61.4	55.7
Pix2pix[14]	26.70	88.56	0.37	46.88	35.4	69.7	47.0	24.3	52.0	33.1
STE[29]	25.46	90.14	0.47	43.39	5.9	40.9	10.2	3.6	28.9	6.4
EnsNet[56]	29.54	92.74	0.24	32.71	32.8	68.7	44.4	50.7	22.1	30.8
EraseNet[23]	32.30	95.42	0.15	19.27	4.6	53.2	8.5	2.9	37.6	5.4
PERT[48]	33.25	96.95	0.14	-	2.9	52.7	5.4	1.8	38.7	3.5
Ours (I_{out})	**35.20**	**97.36**	**0.09**	**13.99**	**1.4**	**38.4**	**2.7**	**0.9**	**28.3**	**1.7**
Tang et al.[38]	35.34	96.24	0.09	-	3.6	-	-	-	-	-
Ours (I_{com})	**35.85**	**97.40**	**0.09**	**14.57**	**1.7**	**40.1**	**3.3**	**1.1**	**29.4**	**2.1**
Tang et al.*[38]	37.08	96.54	**0.05**	-	-	-	-	-	-	-
Ours* (I_{com})	**37.20**	**97.66**	0.07	**11.72**	-	-	-	-	-	-

thoroughly with incorrect detection results, which is presented in the middle row of Fig. 4 (c) and (d). Furthermore, we discuss the number of LGCM blocks in the supplementary materials.

LCG: According to the results shown in Table 1, under the same experimental setting, CTRNet with LCG yields slightly improvement in PSNR and FID by approximately 0.03 and 0.40 on average, respectively for both I_{out} and I_{com}, while the other metrics remain comparable. One of the basic challenges of scene text removal is the background restoration, the generated structure can indicate the region clues of text-erased so that provide the guidance for texture synthesis of the background. Fig. 4 (e) and (f) show the outputs of CTRNet with LCG and the generated background structure. With the background structure, our text removal network can restore a more natural background texture, as indicated by the red boxes in the figures. Besides, as shown in the bottom row of Fig. 4 (e), there exist wrong detected results, but LCG can still help retain the corresponding region and predict more reasonable contents than others.

(a) Input (b) GT (c) EN[23] (d) Stroke[38] (e) Ours

Fig. 6. Qualitative results on SCUT-Syn for comparing our model with previous scene text removal methods. EN denotes EraseNet [23] and Stroke denotes the method proposed by Tang et al. [38]. Zoom in for best view.

Table 3. Comparison with state-of-the-art methods on SCUT-Syn. MSSIM and MSE are represented by (%) in the table. Bold indicates SOTA, while Underline indicates second best.

Methods	Image-Eval			
	PSNR	MSSIM	MSE	FID
Pix2pix[14]	26.76	91.08	0.27	47.84
STE[29]	25.40	90.12	0.65	46.39
EnsNet[56]	37.36	96.44	0.21	-
EnsNet (reimplemented)[56]	36.23	96.76	0.04	19.96
EraseNet[23]	38.32	97.67	0.02	9.53
MTRNet++[39]	34.55	98.45	<u>0.04</u>	-
Zdenek et al.[55]	37.46	93.64	-	-
Conrad et al.[7]	32.97	94.90	-	-
PERT[48]	<u>39.40</u>	<u>97.87</u>	0.02	-
Tang et al.[38]	38.60	97.55	0.02	-
Ours	**41.28**	**98.50**	**0.02**	**3.84**

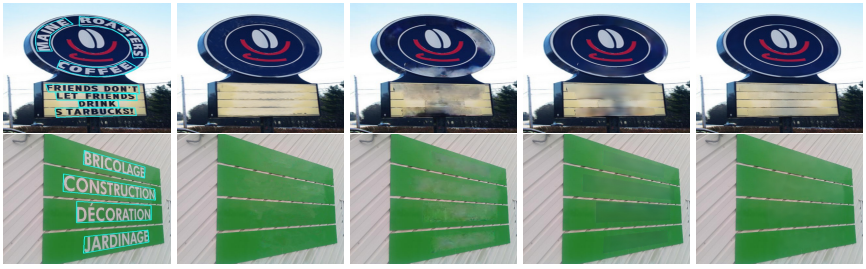
Soft Mask: The application of soft-mask mainly aims to eliminate the discontinuity and inconsistency at the junction of text/non-text regions on the output. Soft mask only achieves only a slight improvement on I_{out}, but a significant promotion on I_{com} by 0.53 in PSNR and 0.22 in MSSIM, respectively. Qualitative results of I_{com} for the comparison on hard-mask (0-1) and soft-mask are shown in the supplement file. With soft-mask, the output can preserve smoother edges between text and non-text regions. Besides, as the soft-mask is expanded, the texts are removed more completely and thoroughly.

4.4 Comparison with the State-of-the-arts

In this section, we conduct experiments to evaluate the performance of CTRNet and the relevant SOTA methods on scene text removal on both SCUT-EnsText and SCUT-Syn. The quantitative results on SCUT-EnsText are given in Table 2, and the quantitative results on SCUT-Syn are given in Table 3. The results for

Table 4. Comparison with state-of-the-art image inpainting methods on SCUT-EnsText. MSSIM and MSE are represented by (%) in the table.

Methods	Image-Eval			
	PSNR	MSSIM	MSE	FID
CTSDG[11]	33.10	95.55	0.14	20.01
SPL[57]	35.41	97.39	0.07	17.85
Ours	**37.20**	**97.66**	**0.07**	**11.72**



(a) Input (b) GT (c) CTDSG[11] (d) SPL[57] (e) Ours

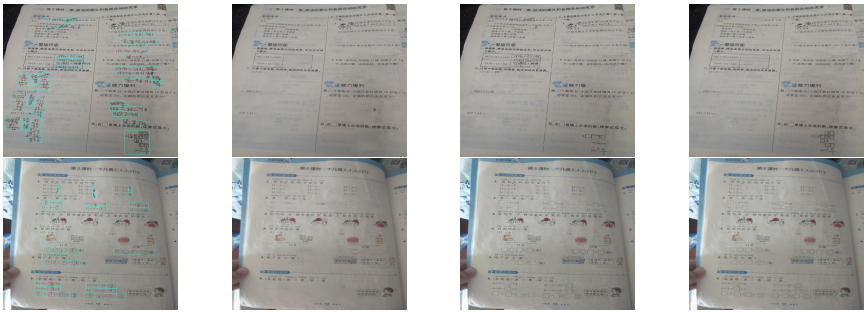
Fig. 7. Qualitative results on SCUT-EnsText for comparing our model with state-of-the-art image inpainting methods. Zoom in for best view.

these two datasets demonstrate that our proposed model outperforms existing state-of-the-art methods on both Image-Eval and Detection-Eval, indicating that our model can effectively remove the text on th images and meanwhile restore more reasonable background textures. Only the results of Tang et al. [38] preserve non-text regions of the input (i.e. I_{com}) while the others are direct model output I_{out}, we evaluate all their performance for fair comparisons.

The qualitative comparison with other methods on SCUT-EnsText is shown in Fig. 5, and for SCUT-Syn in Fig. 6. In Fig. 5, the results generated by EraseNet [23] and Tang et al. [38] still contain artifacts and discontinuities on the output when dealing with complex text images. By utilizing local-global content modeling and different level contextual guidance, our model can predict more realistic textures for text regions and obtain significantly fewer noticeable inconsistencies. Besides, in Fig. 6, our model can also generate results of higher quality with more visually pleasing and plausible contents for synthetic data. More cases for comparison are given in the supplement materials.

4.5 Comparison with State-of-the-art Image Inpainting Methods

We conduct experiments to compare CTRNet with existing SOTA image inpainting methods, CTSDG [11] and SPL [57] on SCUT-EnsText. The quantitative and qualitative results are given in Table 4 and Fig. 7, respectively. Our model outperforms these two methods in all metrics with a remarkable margin, meanwhile can restore the text region background with more reasonable and realistic textures. The reason is that while we simply apply image inpainting methods on scene text removal, the text regions will be directly abandoned by



(a) Input (b) SPL[57] (c) EraseNet[23] (d) Ours

Fig. 8. Qualitative results on examination papers. Zoom in for best view.

masking according to the bounding boxes (blue boxes in Fig. 7 (a)), causing that the model can not effectively deduce the background information.

4.6 Application on Handwritten Text Removal

In this section, we apply CTRNet to help restore document images to verify its generalizability. We collect 1,000 in-house examination paper images and manually annotate them by erasing the handwriting with PhotoShop. Then we train CTRNet and EraseNet [23] on the data and evaluate them with other paper images. Besides, we also train SPL [57] for comparison to further illustrate the difference between text removal and image inpainting. The visualization results are shown in Fig. 8. Our method can retain more printed words than SPL and EraseNet, which is more suitable for document restoration task. More results are given in the supplement materials.

5 Conclusion

In this paper, we propose a new text removal model called CTRNet. CTRNet introduces both low-level and high-level contextual guidance, which can effectively promote the performance on texture restoration for complex backgrounds. We further use smooth structure images and discriminative context features to represent the low-level and high-level context, respectively. Besides, the learned contextual guidance is incorporated into the image features and modeled in a local-global manner to effectively capture both sufficient context information and long-term correlation among all of the pixels. The experiments conducted on three benchmark datasets have demonstrated the effectiveness of the proposed CTRNet, outperforming previous state-of-the-art methods significantly.

Acknowledgement

This research is supported in part by NSFC (Grant No.: 61936003), GD-NSF (no.2017A030312006, No.2021A1515011870), and the Science and Technology Foundation of Guangzhou Huangpu Development District (Grant 2020GH17).

References

1. Baek, Y., Lee, B., Han, D., Yun, S., Lee, H.: Character region awareness for text detection. In: Proc. CVPR. pp. 9365–9374 (2019)
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
3. Bian, X., Wang, C., Quan, W., Ye, J., Zhang, X., Yan, D.M.: Scene text removal via cascaded text stroke detection and erasing. Computational Visual Media **8**(2), 273–287 (2022)
4. Chng, C.K., Liu, Y., Sun, Y., Ng, C.C., Luo, C., Ni, Z., Fang, C., Zhang, S., Han, J., Ding, E., Liu, J., Karatzas, D., Chan, C.S., Jin, L.: ICDAR2019 robust reading challenge on arbitrary-shaped text - RRC-ArT. In: Proc. ICDAR. pp. 1571–1576 (2019)
5. Ch'ng, C.K., Chan, C.S.: Total-Text: A comprehensive dataset for scene text detection and recognition. In: Proc. ICDAR. vol. 1, pp. 935–942 (2017)
6. Cho, J., Yun, S., Han, D., Heo, B., Choi, J.Y.: Detecting and removing text in the wild. IEEE Access **9**, 123313–123323 (2021)
7. Conrad, B., Chen, P.I.: Two-stage seamless text erasing on real-world scene images. In: Proc. ICIP. pp. 1309–1313. IEEE (2021)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proc. CVPR. pp. 248–255 (2009)
9. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proc. CVPR. pp. 2414–2423 (2016)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proc. NIPS. pp. 2672–2680 (2014)
11. Guo, X., Yang, H., Huang, D.: Image inpainting via conditional texture and structure dual generation. In: Proc. ICCV. pp. 14134–14143 (2021)
12. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proc. CVPR. pp. 2315–2324 (2016)
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Proc. NIPS pp. 6626–6637 (2017)
14. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proc. CVPR. pp. 1125–1134 (2017)
15. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Proc. ECCV. pp. 694–711 (2016)
16. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., Shafait, F., Uchida, S., Valveny, E.: ICDAR 2015 competition on robust reading. In: Proc. ICDAR. pp. 1156–1160 (2015)
17. Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i. Bigorda, L.G., Mestre, S.R., Mas, J., Mota, D.F., Almazàn, J.A., de las Heras, L.P.: ICDAR 2013 robust reading competition. In: Proc. ICDAR. pp. 1484–1493 (2013)
18. Keserwani, P., Roy, P.P.: Text region conditional generative adversarial network for text concealment in the wild. IEEE Trans. Circuits Syst. Video Technol. (2021)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. Proc. ICLR (2014)
20. Krishnan, P., Kovvuri, R., Pang, G., Vassilev, B., Hassner, T.: TextStyleBrush: Transfer of text aesthetics from a single example. arXiv preprint arXiv:2106.08385 (2021)

21. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4. Int. J. Comput. Vision **128**(7), 1956–1981 (2020)
22. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: SwinIR: Image restoration using swin transformer. In: Proc. ICCV. pp. 1833–1844 (2021)
23. Liu, C., Liu, Y., Jin, L., Zhang, S., Luo, C., Wang, Y.: EraseNet: End-to-end text removal in the wild. IEEE Trans. Image Process. **29**, 8760–8775 (2020)
24. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proc. ECCV. pp. 85–100 (2018)
25. Liu, H., Jiang, B., Song, Y., Huang, W., Yang, C.: Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In: Proc. ECCV. pp. 725–741. Springer (2020)
26. Liu, Y., Jin, L., Xie, Z., Luo, C., Zhang, S., Xie, L.: Tightness-aware evaluation protocol for scene text detection. In: Proc. CVPR. pp. 9612–9620 (2019)
27. Mirza, M., Osindero, S.: Conditional generative adversarial nets. In: arXiv preprint arXiv:1411.1784 (2014)
28. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: Proc. ICLR (2018)
29. Nakamura, T., Zhu, A., Yanai, K., Uchida, S.: Scene text eraser. In: Proc. ICDAR. vol. 01, pp. 832–837 (2017)
30. Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., Luo, Z., Pal, U., Rigaud, C., Chazalon, J., Khlif, W., Luqman, M.M., Burie, J., Liu, C., Ogier, J.: ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification - RRC-MLT. In: Proc. IDAR. vol. 01, pp. 1454–1459 (2017)
31. Nayef, N., Patel, Y., Busta, M., Chowdhury, P.N., Karatzas, D., Khlif, W., Matas, J., Pal, U., Burie, J.C., Liu, C.l., et al.: ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition–RRC-MLT-2019. In: Proceedings of ICDAR. pp. 1582–1587 (2019)
32. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M.: EdgeConnect: Structure guided image inpainting using edge prediction. In: Proc. ICCV Workshops
33. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proc. CVPR. pp. 2337–2346 (2019)
34. Qin, S., Wei, J., Manduchi, R.: Automatic semantic content removal by learning to neglect. Proc. BMVC pp. 1–12 (2018)
35. Ren, Y., Yu, X., Zhang, R., Li, T.H., Liu, S., Li, G.: StructureFlow: Image inpainting via structure-aware appearance flow. In: Proc. ICCV. pp. 181–190 (2019)
36. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. In: Proc. ICCV. pp. 82–91 (2021)
37. Shimoda, W., Haraguchi, D., Uchida, S., Yamaguchi, K.: De-rendering stylized texts. In: Proc. ICCV. pp. 1076–1085 (2021)
38. Tang, Z., Miyazaki, T., Sugaya, Y., Omachi, S.: Stroke-based scene text erasing using synthetic data for training. IEEE Trans. Image Process. **30**, 9306–9320 (2021)
39. Tursun, O., Denman, S., Zeng, R., Sivapalan, S., Sridharan, S., Fookes, C.: MTRNet++: One-stage mask-based scene text eraser. Comp. Vis. Image Understanding **201**, 103066 (2020)
40. Tursun, O., Zeng, R., Denman, S., Sivapalan, S., Sridharan, S., Fookes, C.: MTRNet: A generic scene text eraser. In: Proc. ICDAR. pp. 39–44. IEEE (2019)
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Proc. NIPS pp. 6000–6010 (2017)

42. Vatti, B.R.: A generic solution to polygon clipping. Communications of the ACM **35**(7), 56–63 (1992)
43. Veit, A., Matera, Tomas, e.a.: COCO-Text: Dataset and benchmark for text detection and recognition in natural images. In: arXiv preprint arXiv:1601.07140 (2016)
44. Wang, C., Zhao, S., Zhu, L., Luo, K., Guo, Y., Wang, J., Liu, S.: Semi-supervised pixel-level scene text segmentation by mutually guided network. IEEE Trans. Image Process. **30**, 8212–8221 (2021)
45. Wang, K., Belongie, S.J.: Word spotting in the wild. In: Proc. ECCV. pp. 591–604 (2010)
46. Wang, W., Xie, E., Li, X., Hou, W., Lu, T., Yu, G., Shao, S.: Shape robust text detection with progressive scale expansion network. In: Proc. CVPR. pp. 9336–9345 (2019)
47. Wang, W., Xie, E., Song, X., Zang, Y., Wang, W., Lu, T., Yu, G., Shen, C.: Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In: Proc. ICCV. pp. 8440–8449 (2019)
48. Wang, Y., Xie, H., Fang, S., Qu, Y., Zhang, Y.: A simple and strong baseline: Progressively region-based scene text removal networks. arXiv e-prints pp. arXiv–2106 (2021)
49. Wu, L., Zhang, C., Liu, J., Han, J., Liu, J., Ding, E., Bai, X.: Editing text in the wild. In: Proc. ACM MM. pp. 1500–1508 (2019)
50. Xu, L., Yan, Q., Xia, Y., Jia, J.: Structure extraction from texture via relative total variation. ACM Trans. Graphics **31**(6), 1–10 (2012)
51. Xu, X., Zhang, Z., Wang, Z., Price, B., Wang, Z., Shi, H.: Rethinking text segmentation: A novel dataset and a text-specific refinement approach. In: Proc. CVPR. pp. 12045–12055 (2021)
52. Yang, Q., Huang, J., Lin, W.: Swaptext: Image based texts transfer in scenes. In: Proc. CVPR. pp. 14700–14709 (2020)
53. Yu, B., Xu, Y., Huang, Y., Yang, S., Liu, J.: Mask-guided GAN for robust text editing in the scene. Neurocomputing **441**, 192–201 (2021)
54. Yuliang, L., Lianwen, J., Shuaitao, Z., Sheng, Z.: Curved scene text detection via transverse and longitudinal sequence connection. Pattern Recogn. **90**, 337 – 345 (2019)
55. Zdenek, J., Nakayama, H.: Erasing scene text with weak supervision. In: Proc. WACV (2020)
56. Zhang, S., Liu, Y., Jin, L., Huang, Y., Lai, S.: EnsNet: Ensconce text in the wild. In: Proc. AAAI. vol. 33, pp. 801–808 (2019)
57. Zhang, W., Zhu, J., Tai, Y., Wang, Y., Chu, W., Ni, B., Wang, C., Yang, X.: Context-aware image inpainting with learned semantic priors. In: Proc. IJCAI. pp. 1323–1329 (2021)
58. Zheng, C., Cham, T.J., Cai, J.: Pluralistic image completion. In: Proc. CVPR. pp. 1438–1447 (2019)