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. In this supplementary material, we first introduce the details of our model, including Local-global Content Modeling block and ResSPADE. Meanwhile, we present the supplementary experiments to further demonstrate that our model performs favorably against state-of-the-art approaches. Moreover, we process more document restoration examples on examination papers to verify the generalizability of CTRNet.

1 The details of CTRNet

1.1 The Details of ResSPADE

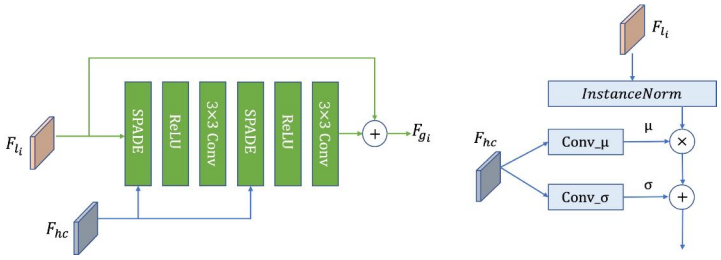
Spatially-Adaptive Normalization (SPADE) and ResSAPDE are proposed to synthesize images with semantic guidance [6]. It is proved to be effective in image inpainting and background restoration [9], thus we introduce ResSPADE to spatially incorporate the learned high-level context guidance F_{hc} into LGCM blocks for feature modeling and decoding. The architecture of ResSPADE is shown in Fig. 1.

1.2 The Details of Local-global Content Modeling (LGCM)

The architecture of LGCM block is shown in Fig. 2. A single stage (i-th) for LGCM can be formulated as follows:

Given the modeled features $F_{l_i} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$ ($F_{l_0} = F_s$, F_s is features from Image Encoder), our local modeling module (CNNs) first obtains the local features and downsamples the feature maps twice as

$$F_{local} = H_{conv}(F_{l_i}) \tag{1}$$



(a) ResSPADE (b) Detail of SPADE

Fig. 1. The architecture of ResSPADE [6].

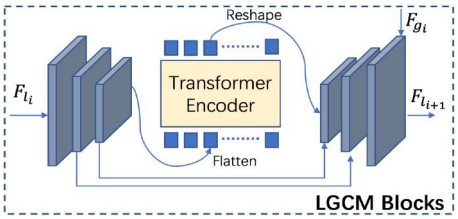


Fig. 2. The architecture of Local-global Content Modeling block.

where $H_{conv}(\cdot)$ consists of two 4×4 convolution layers and 2 residual blocks. Then $F_{local} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C}$ is fed into the global modeling module as

$$\begin{aligned} F_p &= \text{pos}(F_{local}) \\ Q_j, K_j, V_j &= (W_{Q_j}, W_{K_j}, W_{V_j}) * F_p \end{aligned} \tag{2}$$

Here, $\text{pos}(\cdot)$ denotes Position Encoding function. $W_{Q_j}, W_{K_j}, W_{V_j}$ are projection matrices for *query*, *key* and *value* in a single head self-attention. And $j = 0, 1, ...N$, N denotes N-head self-attention layer ($N = 6$ as default). Given these Q, K, V, the multi-head attention map (MHA) and the global features $F_{global} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C}$ can be calculated as

$$\begin{aligned} AM_i &= \text{Softmax}(\frac{Q_j * K_j^T}{\sqrt{d}}) * V_j \\ MHA &= \text{Concat}(AM_1, ...AM_j, ...) \\ F_{global} &= \text{Project}(MHA) \end{aligned} \tag{3}$$

$\text{Project}(\cdot)$ contains LayerNorm and Multi-Layer Perceptron in series. Then F_{global} and F_{local} are aggregated through lateral connection and upsampled to the same dimension as $F_s \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$ with CNNs, denoted as $H_{deconv}(.)$. The operation can also bring the inductive bias of CNN [3].

Meanwhile, our CTRNet incorporates the high-level contextual guidance F_{hc} with ResSPADE [6] as in Zhang et al. [9] to attain F_{g_i}. The architecture of

Table 1. Architecture details. "SN" denotes Spectrum Normalization [5].

Model	Layer	Kernel, Stride
G_{bg_s}	Conv + SN + LeakyReLU	(4×4), (2×2)
	Conv + SN + LeakyReLU	(3×3), (1×1)
	Conv + SN + LeakyReLU	(4×4), (2×2)
	Residual Block $\times 2$	(3×3), (1×1)
	Residual Block (downsample)	(3×3), (2×2)
	Residual Block	(3×3), (1×1)
	Residual Block (downsample)	(3×3), (2×2)
	Residual Block	(3×3), (1×1)
	Residual Block (downsample)	(3×3), (2×2)
	DeConv + SN + LeakyReLU	(3×3), (2×2)
	DeConv + SN + LeakyReLU	(3×3), (2×2)
	DeConv + SN + LeakyReLU	(3×3), (2×2)
	DeConv + SN + LeakyReLU	(3×3), (2×2)
	DeConv + SN + LeakyReLU	(3×3), (2×2)
	Conv	(3×3), (1×1)
Image Encoder	Conv + SN + LeakyReLU	(4×4), (2×2)
	Residual Block	(3×3), (1×1)
	Conv + SN + LeakyReLU	(4×4), (2×2)
	Residual Block	(3×3), (1×1)
	Conv + SN + LeakyReLU	(3×3), (1×1)
Feature Decoder	Conv	(3×3), (1×1)
	DeConv + SN + LeakyReLU	(4×4), (2×2)
	Conv	(3×3), (1×1)
	DeConv + SN + LeakyReLU	(4×4), (2×2)
	Conv	(3×3), (1×1)

ResSPADE is shown in Fig. 1. The final output $F_{l_{i+1}}$ of one LGCM block is

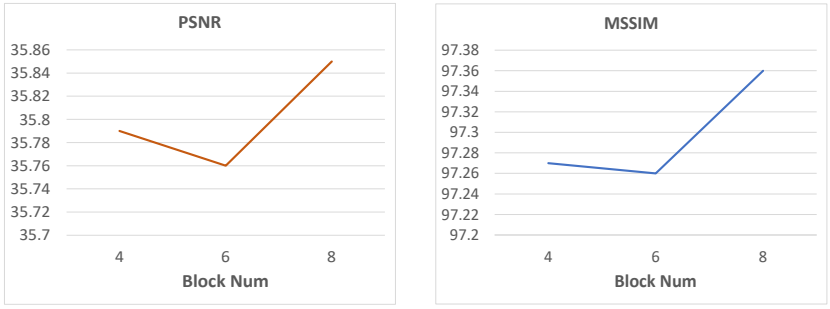
$$F_{l_{i+1}} = F_{g_i} + H_{deconv}(F_{local} + F_{global}) \tag{4}$$

1.3 The Architectures of CTRNet

In this section, we present the detail architectures of the background structure generator G_{bg_s} , image encoder and feature decoder in Table 1.

2 Implement Details

For fair comparison, we train our CTRNet only on the training set of SCUT-EnsText and SCUT-Syn, then evaluate the performance on their corresponding testing set, respectively. For text perception head, we separately train PAN [8] using the official losses and obtain the text detection results. It is frozen in the training of other components of CTRNet. Besides, we first pre-train background



(a) PSNR (b) MSSIM

Fig. 3. Qualitative results for ablation studies on the number of LGCM blocks.

structure generator G_{bg_s} at 100 epochs with both the adversarial loss and structure loss $L_{structure}$ in Low-level Contextual Guidance block, then G_{bg_s} is jointly optimized with the whole system. The input size is 512×512. Adam solver [2] is used and the β is set to (0.0, 0.9) as default. The initial learning rate is 0.0001 for all experiments. For SCUT-EnsText, we decay the learning rate at 100 epochs and further finetune for 50 epochs. For SCUT-Syn, the learning rate decays in 50 epochs and the model is finetuned for another 50 epochs. The batch size is 2. All the experiments are conducted on a workstation with two NVIDIA 2080TI GPUs.

3 Ablation Study

3.1 The ablation study of the number of LGCM blocks

We also conduct experiments on the number of LGCM blocks used in CTRNet. The results are shown in Fig. 3.

3.2 The ablation study of Soft Mask

Qualitative results for the ablation study on soft-mask are shown in Fig. 4.

3.3 The ablation study of loss items

CTRNet incorporates 6 loss items for training, including L_{align}, $L_{structure}$, L_{msr}, L_{per}, L_{style}, and L_{adv}. Among them, L_{adv} is the basic loss in our model, while L_{align} and $L_{structure}$ are corresponding to our HCG and LCG. In this section, therefore, we conduct experiments to evaluate the effectiveness of L_{msr} and L_{per}/L_{style}. The results are shown in Table 2. We apply L_{msr} to improve $L1$ loss with higher weights for text regions in different scales to capture more information, contributing an increase of 0.56 on PSNR. Besides, without L_{per} and L_{style}, the PSNR for CTRNet drops 0.14. These two loss can effectively

Table 2. Ablation study on the effectiveness of L_{msr} and L_{per}/L_{style}.

Methods	PSNR	
	I_{out}	I_{com}
Ours (- L_{msr})	34.64	35.56
Ours (- L_{per}/L_{style})	35.06	35.80
Ours	**35.20**	**35.85**

Table 3. Ablation study on different parameter setting for L_{align} and $L_{structure}$.

λ_{al}	λ_{str}	PSNR
5	2	35.77
3	2	35.80
1	2	**35.85**
1	4	35.78

supervise the output in a high-dimension feature space to capture high-level semantics and improve the quality of our results.

We also conduct ablation study on the hyper-parameters of each loss item. $\lambda_{style}, \lambda_{per}, \lambda_m, \lambda_a$ follow the setup of commonly used. We conduct experiments on $\{\lambda_{al}, \lambda_{str}\}$ for L_{align} and $L_{structure}$, and the results are presented in Table 3. When $\lambda_{al} = 1$ and $\lambda_{str} = 2$, CTRNet can obtain the best performance for text removal.

4 Failure Cases

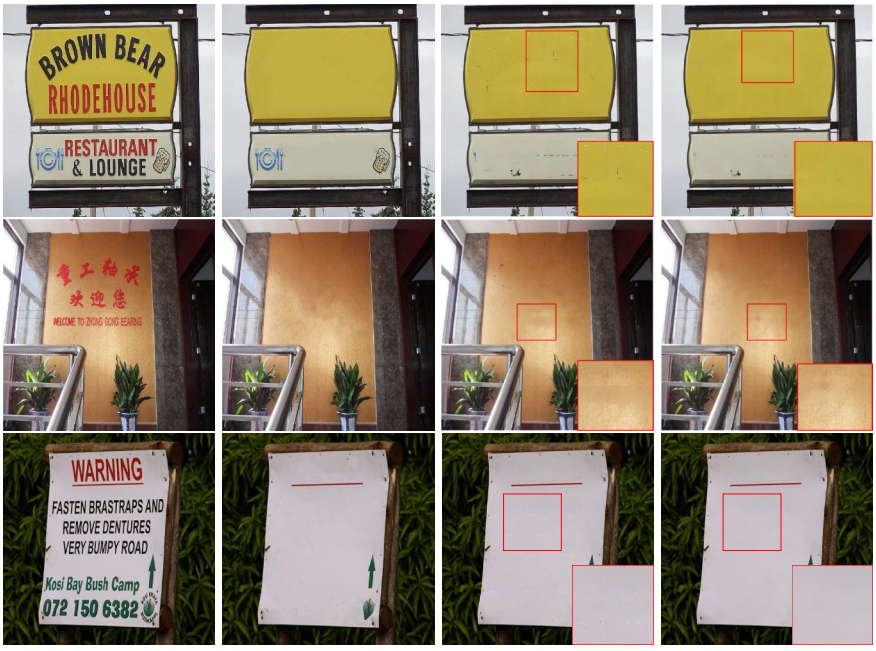
Our model has some limitation, as shown in Fig. 5. CTRNet fails in handling text in large scale and can not effectively recover the background with multiple pattern styles.

5 More comparisons on SCUT-EnsText and SCUT-Syn

This section shows more qualitative comparisons with Pix2pix, EnsNet, EraseNet and our CTRNet on SCUT-EnsText and SCUT-Syn. For SCUT-EnsText, the results are referred to Fig. 6, Fig. 7, Fig. 8, Fig. 9. For SCUT-Syn, the results are referred to Fig. 10, Fig. 11, Fig. 12, Fig. 13.

6 More results on SCUT-EnsText and Examination Papers

This section shows more results on SCUT-EnsText and Examination Papers generated by our model. For SCUT-EnsText, the results are referred to Fig. 14, Fig. 15, Fig. 16. For Exam papers, the results are referred to Fig. 17, Fig. 18, Fig. 19.



(a) Input (b) GT (c) With hm (d) With sm

Fig. 4. Qualitative results for ablation studies on the soft-mask. HM and SM denotes hard-mask (0-1) and soft-mask, respectively. Best viewed with zoom-in.

(a) (b)

Fig. 5. Some failure cases from our CTRNet. Left: input; Right: result.

(a) Input (b) GT (c) Pix2pix[1]

(d) EreaseNet[4] (e) Tang et al.[7] (f) Ours

(a) Input (b) GT (c) Pix2pix[1]

(d) EraseNet[4] (e) Tang et al.[7] (f) Ours

Fig. 6. Qualitative results on SCUT-EnsText for comparing our model with previous scene text removal methods.

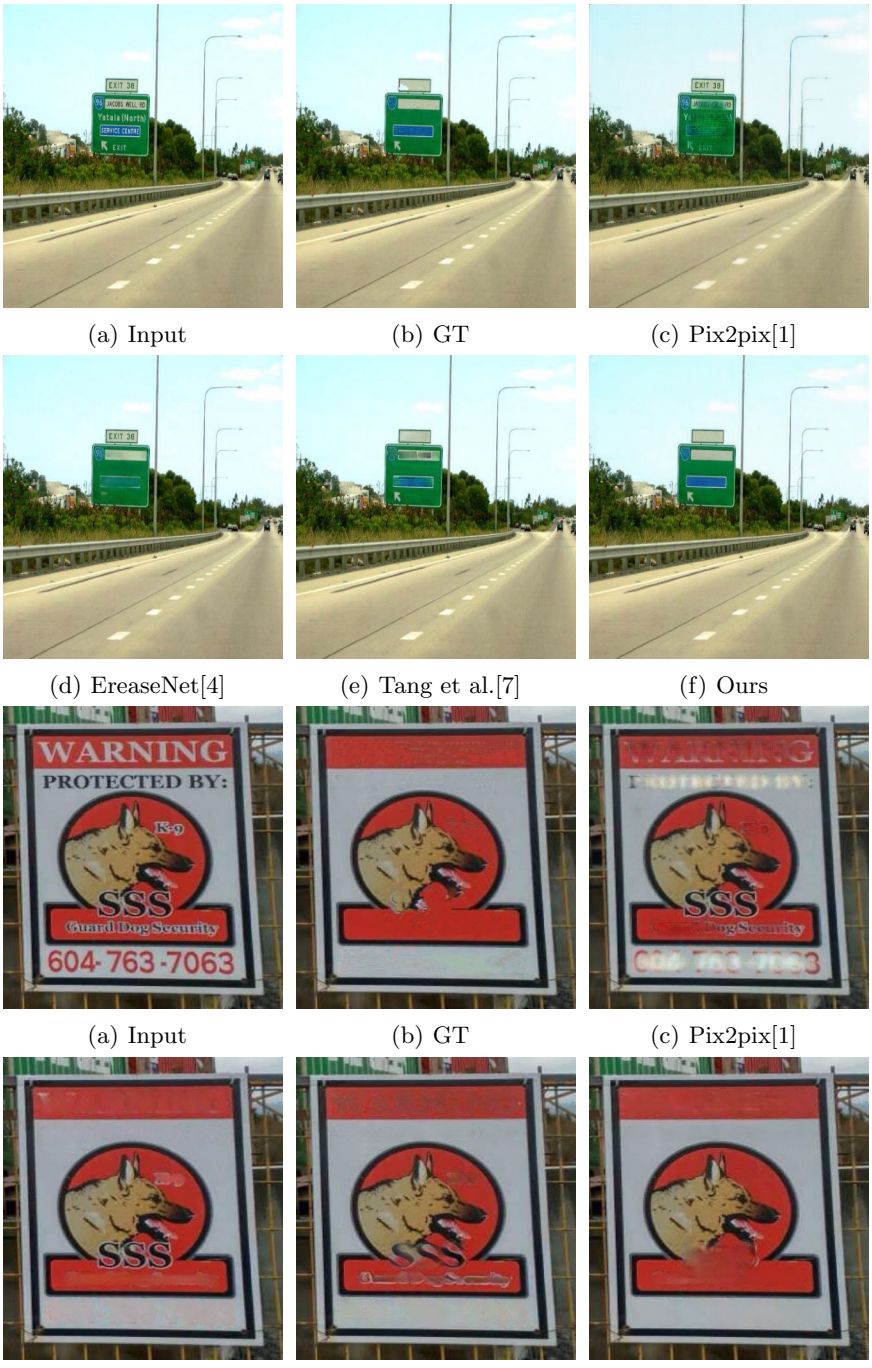
(a) Input (b) GT (c) Pix2pix[1]

(d) EreaseNet[4] (e) Tang et al.[7] (f) Ours

(a) Input (b) GT (c) Pix2pix[1]

(d) EraseNet[4] (e) Tang et al.[7] (f) Ours

Fig. 7. Qualitative results on SCUT-EnsText for comparing our model with previous scene text removal methods.



(a) Input (b) GT (c) Pix2pix[1]

(d) EreaseNet[4] (e) Tang et al.[7] (f) Ours

(a) Input (b) GT (c) Pix2pix[1]

(d) EraseNet[4] (e) Tang et al.[7] (f) Ours

Fig. 8. Qualitative results on SCUT-EnsText for comparing our model with previous scene text removal methods.

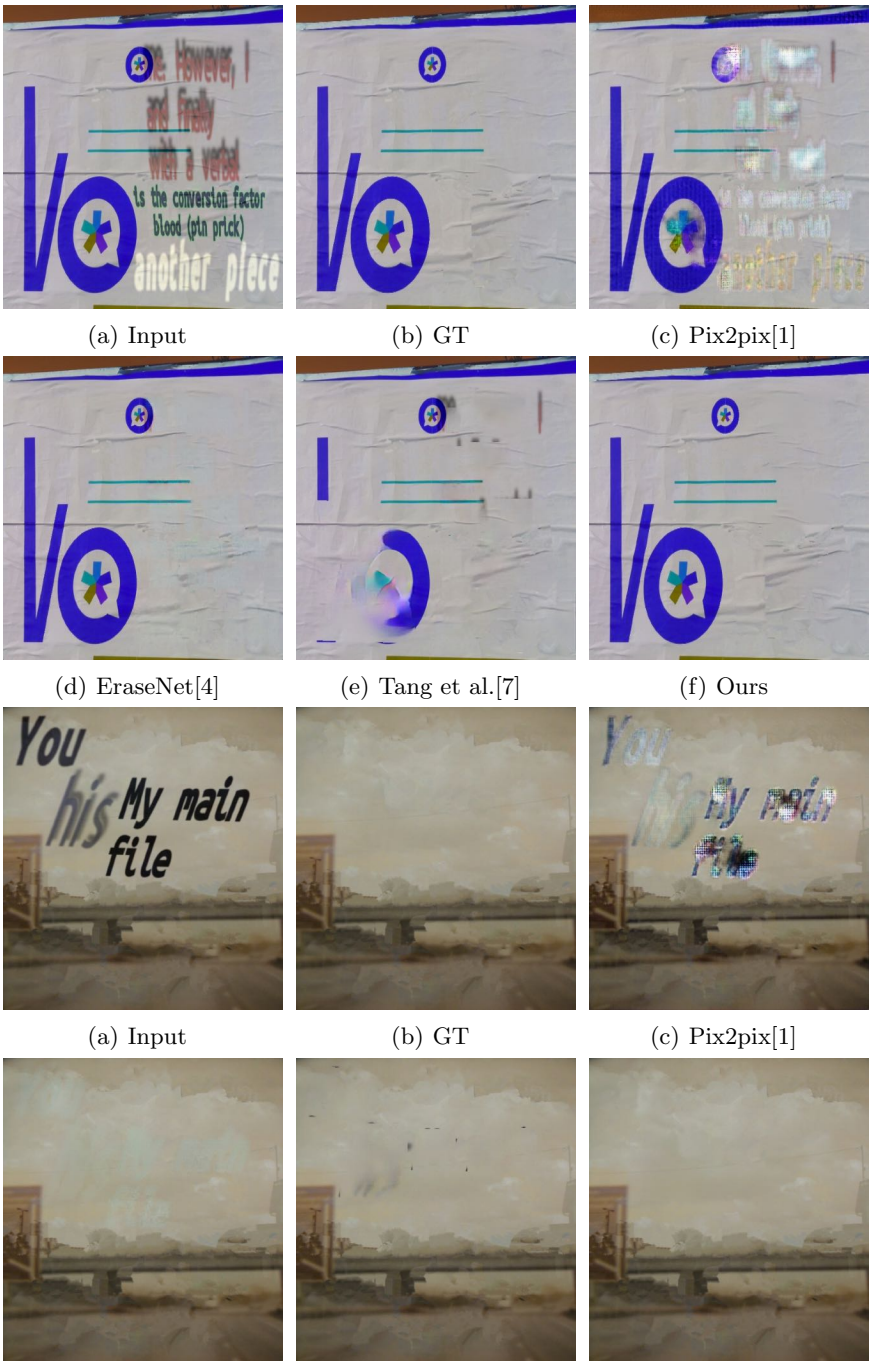
(a) Input (b) GT (c) Pix2pix[1]

(d) EreaseNet[4] (e) Tang et al.[7] (f) Ours

(a) Input (b) GT (c) Pix2pix[1]

(d) EraseNet[4] (e) Tang et al.[7] (f) Ours

Fig. 9. Qualitative results on SCUT-EnsText for comparing our model with previous scene text removal methods.



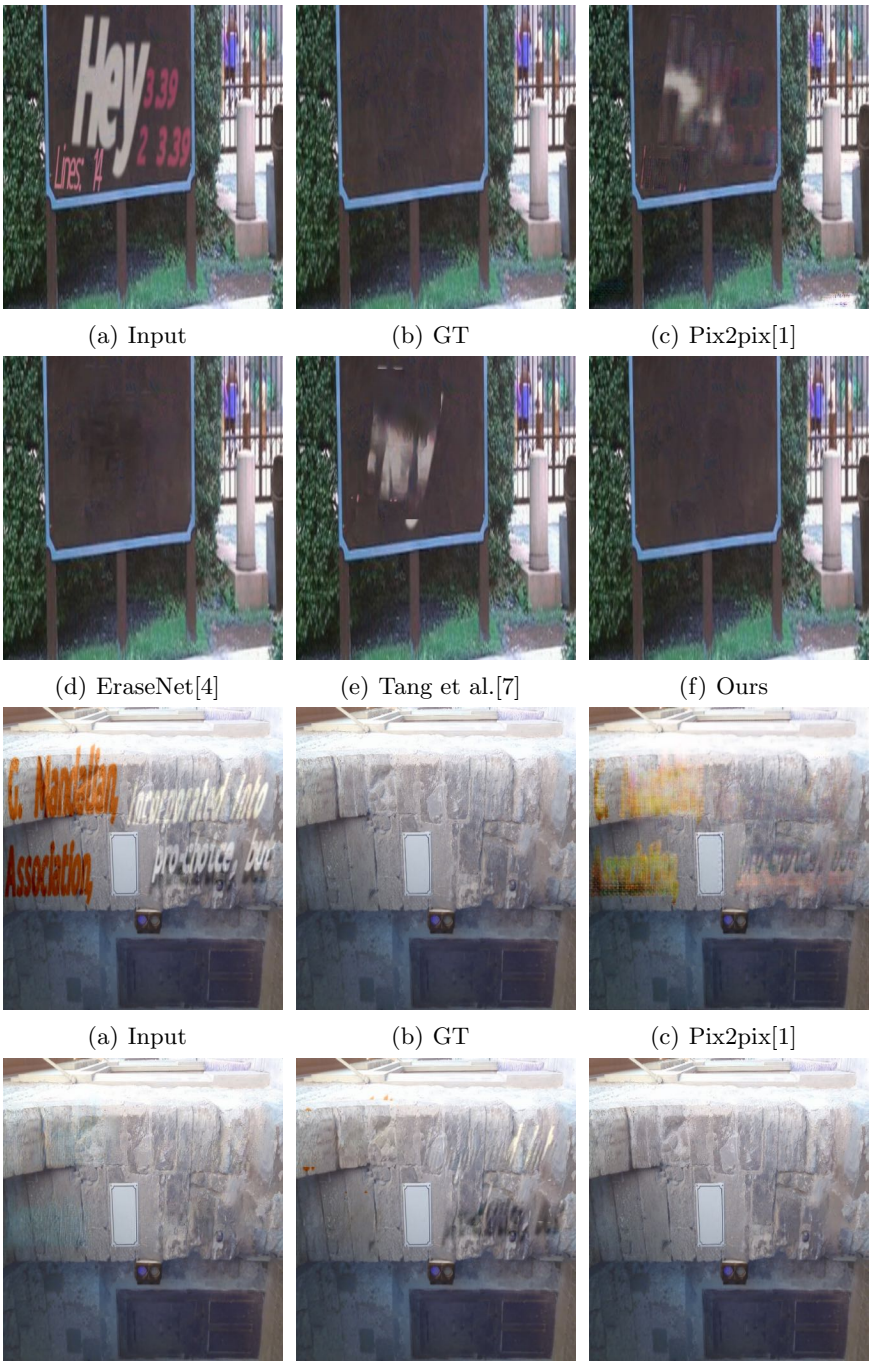
(a) Input (b) GT (c) Pix2pix[1]

(d) EraseNet[4] (e) Tang et al.[7] (f) Ours

(a) Input (b) GT (c) Pix2pix[1]

(d) EraseNet[4] (e) Tang et al.[7] (f) Ours

Fig. 10. Qualitative results on SCUT-Syn for comparing our model with previous scene text removal methods.



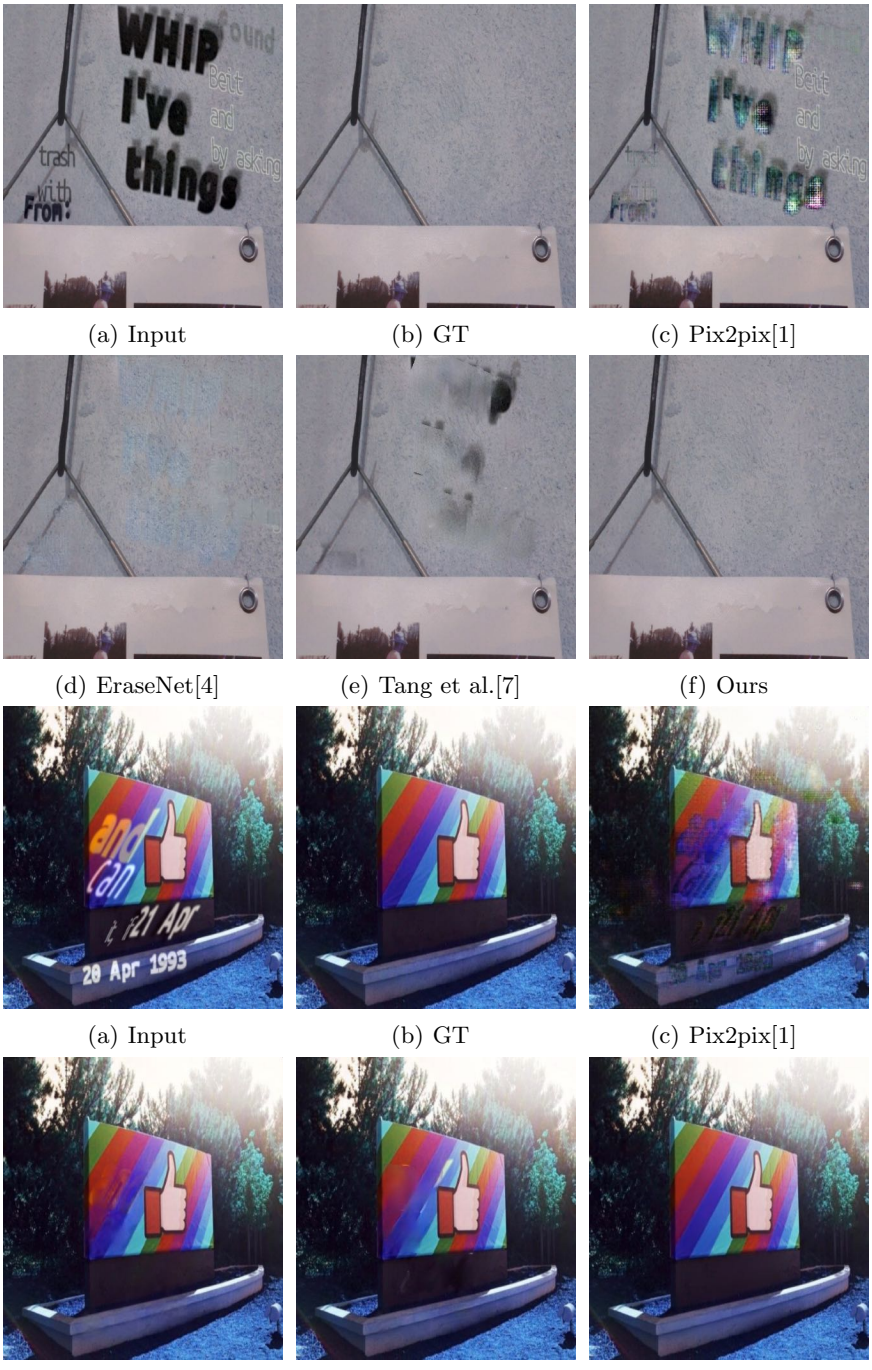
(a) Input (b) GT (c) Pix2pix[1]

(d) EraseNet[4] (e) Tang et al.[7] (f) Ours

(a) Input (b) GT (c) Pix2pix[1]

(d) EraseNet[4] (e) Tang et al.[7] (f) Ours

Fig. 11. Qualitative results on SCUT-Syn for comparing our model with previous scene text removal methods.



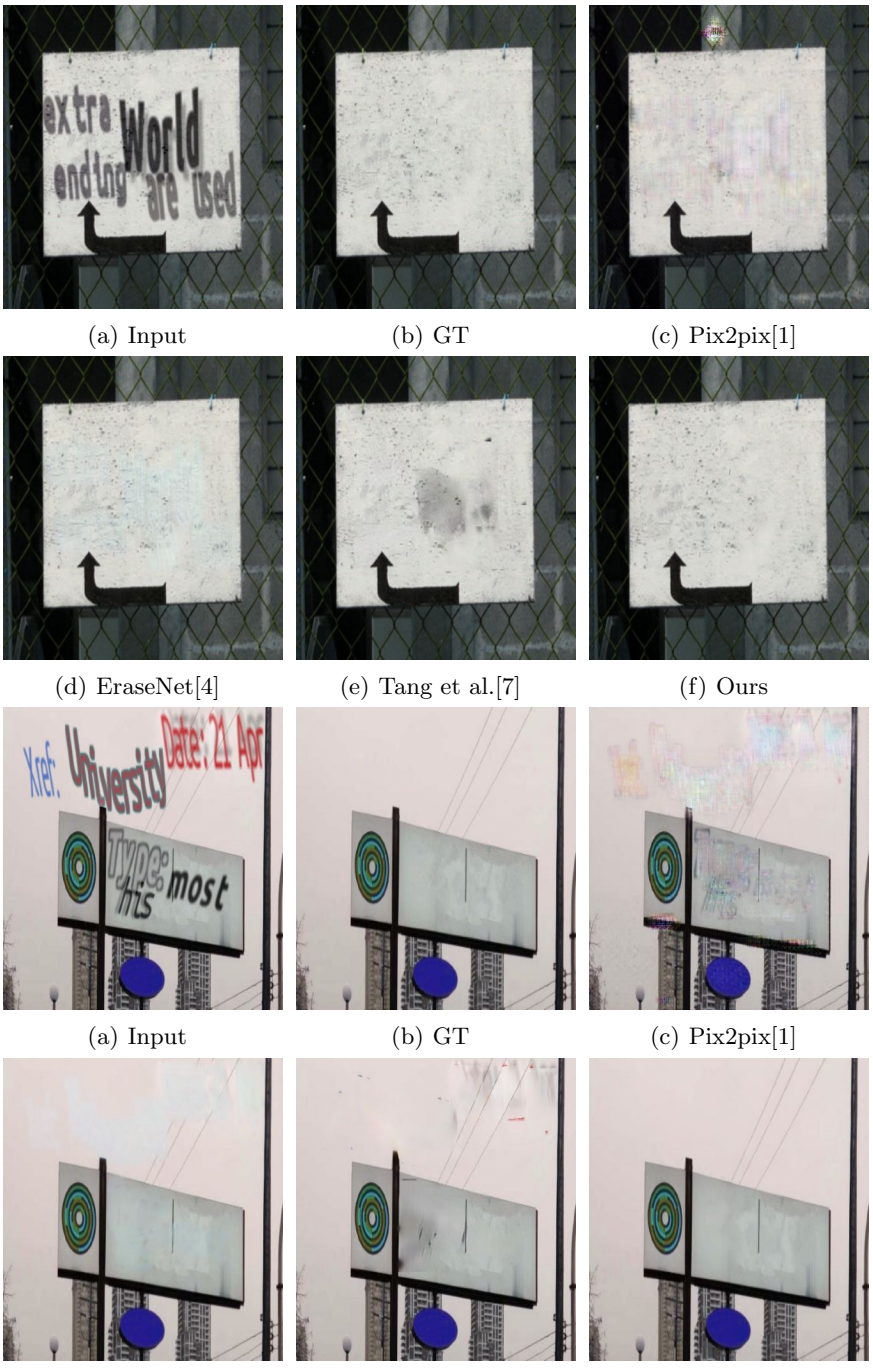
(a) Input (b) GT (c) Pix2pix[1]

(d) EraseNet[4] (e) Tang et al.[7] (f) Ours

(a) Input (b) GT (c) Pix2pix[1]

(d) EraseNet[4] (e) Tang et al.[7] (f) Ours

Fig. 12. Qualitative results on SCUT-Syn for comparing our model with previous scene text removal methods.



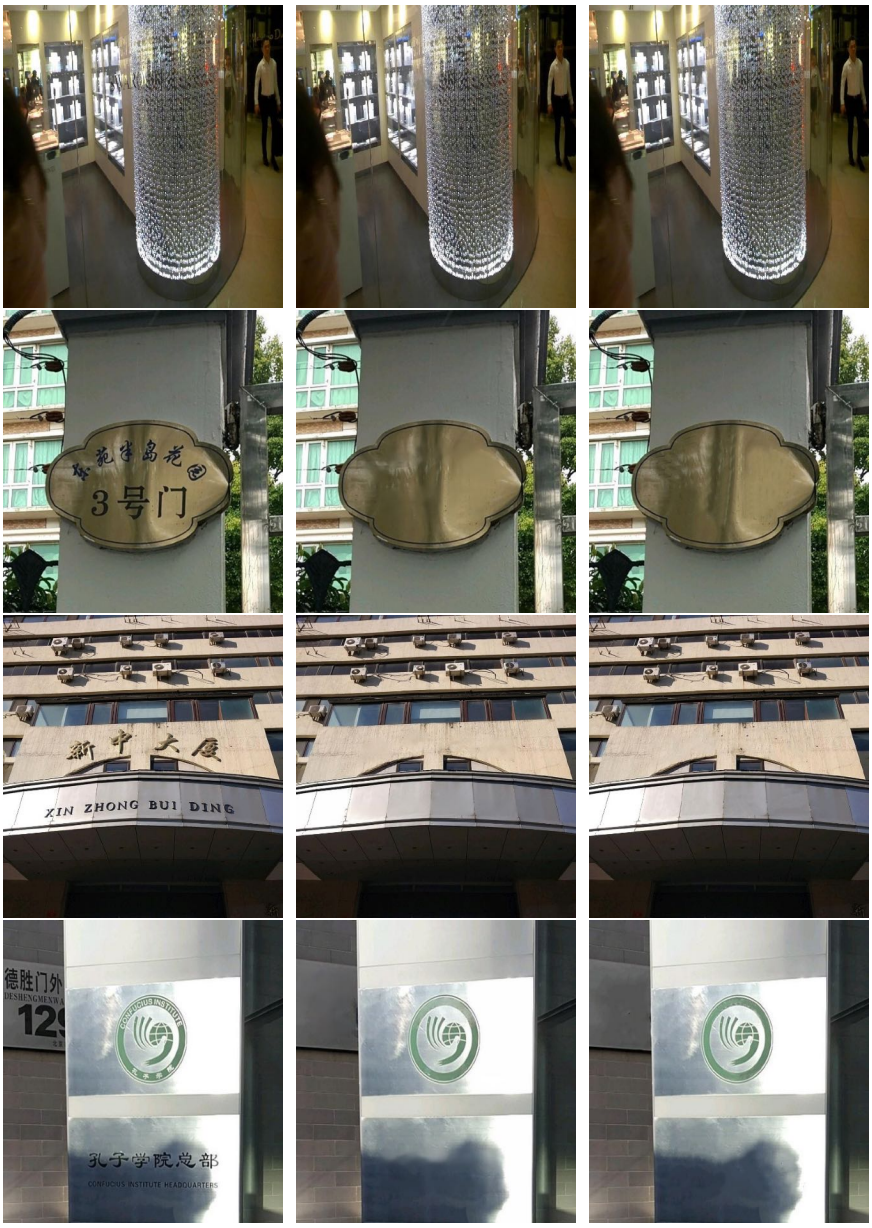
(a) Input (b) GT (c) Pix2pix[1]

(d) EraseNet[4] (e) Tang et al.[7] (f) Ours

(a) Input (b) GT (c) Pix2pix[1]

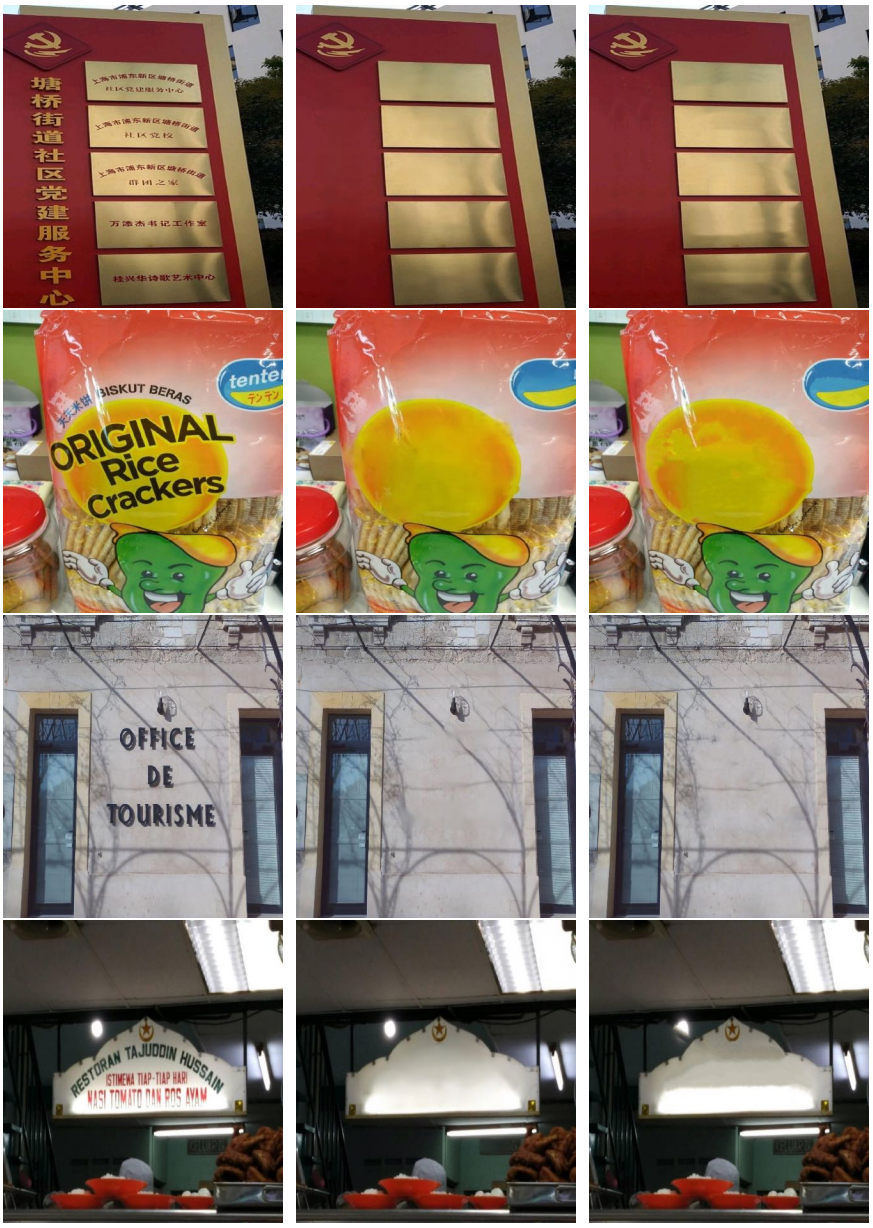
(d) EraseNet[4] (e) Tang et al.[7] (f) Ours

Fig. 13. Qualitative results on SCUT-Syn for comparing our model with previous scene text removal methods.



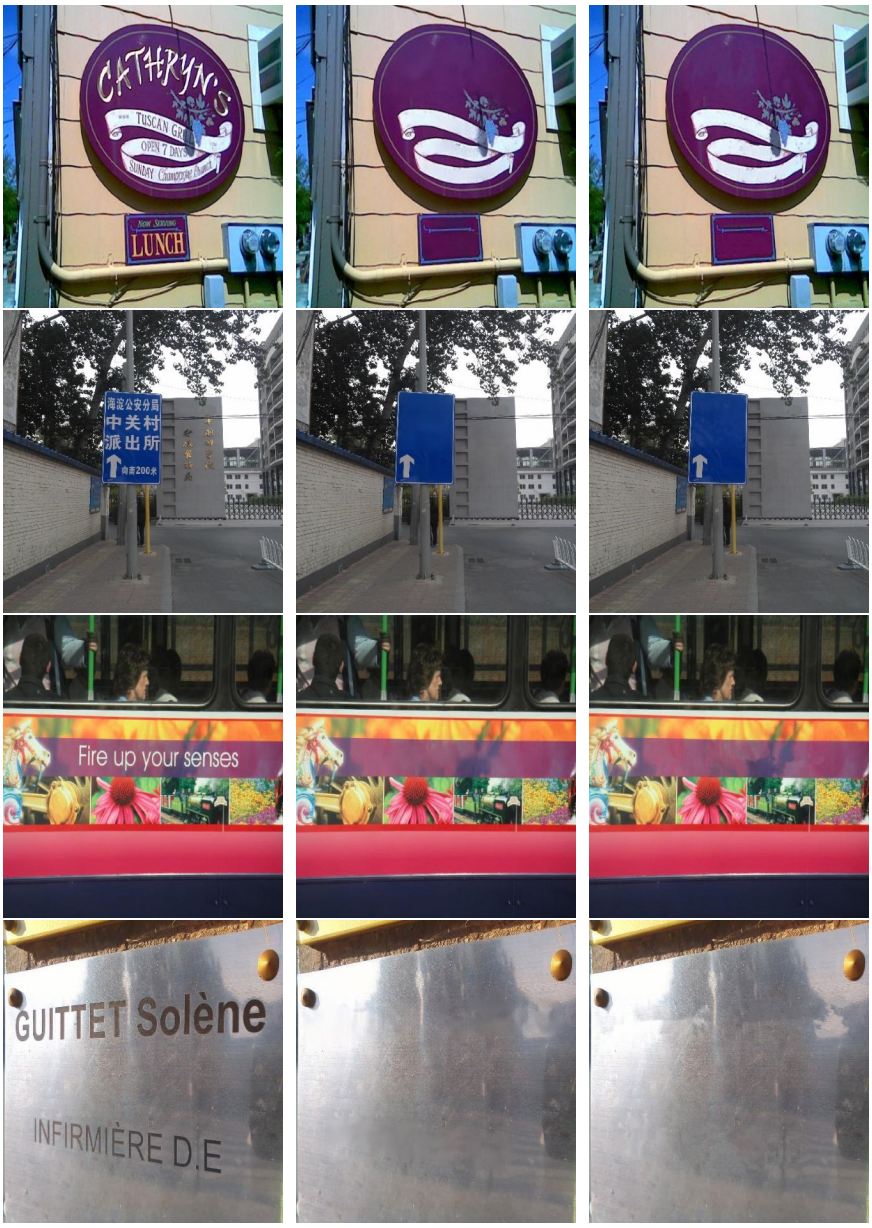
(a) Input (b) Ours (c) GT

Fig. 14. More qualitatively results on SCUT-EnsText.



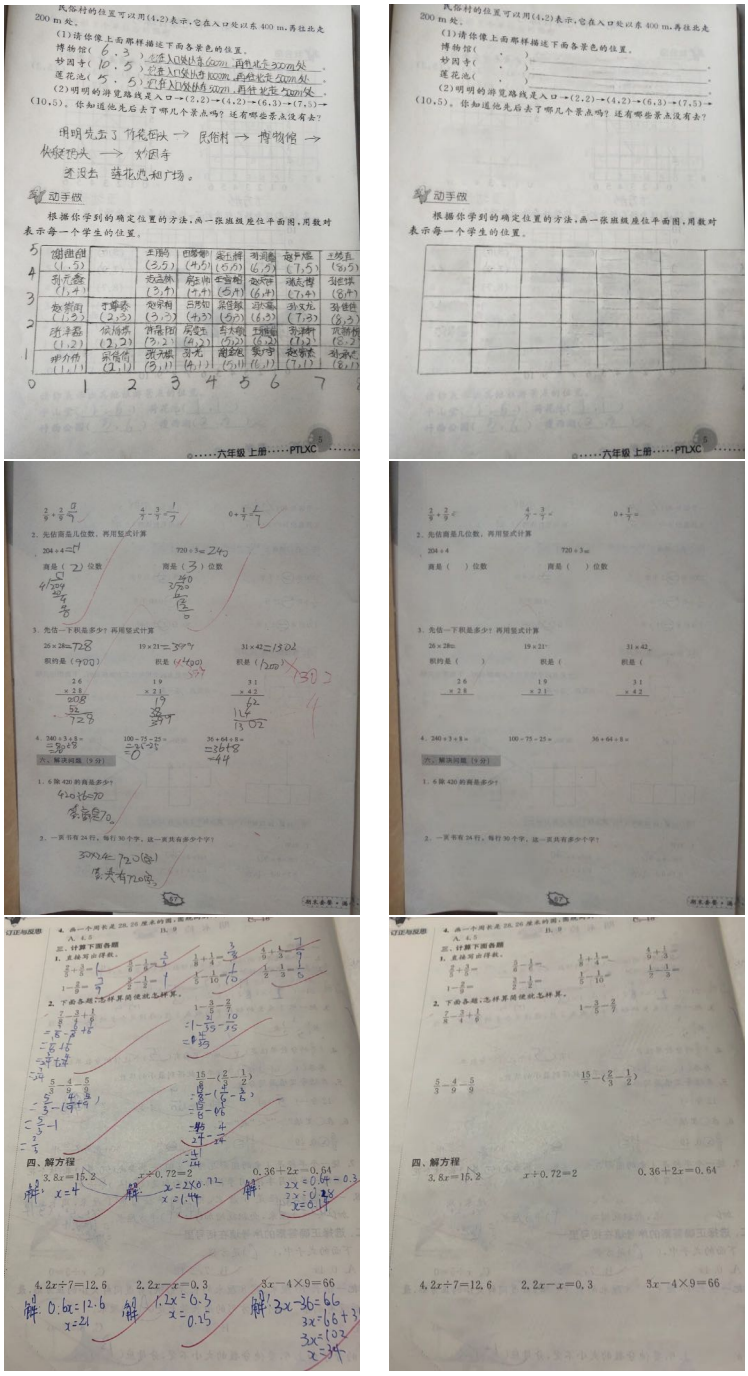
(a) Input (b) Ours (c) GT

Fig. 15. More qualitatively results on SCUT-EnsText.



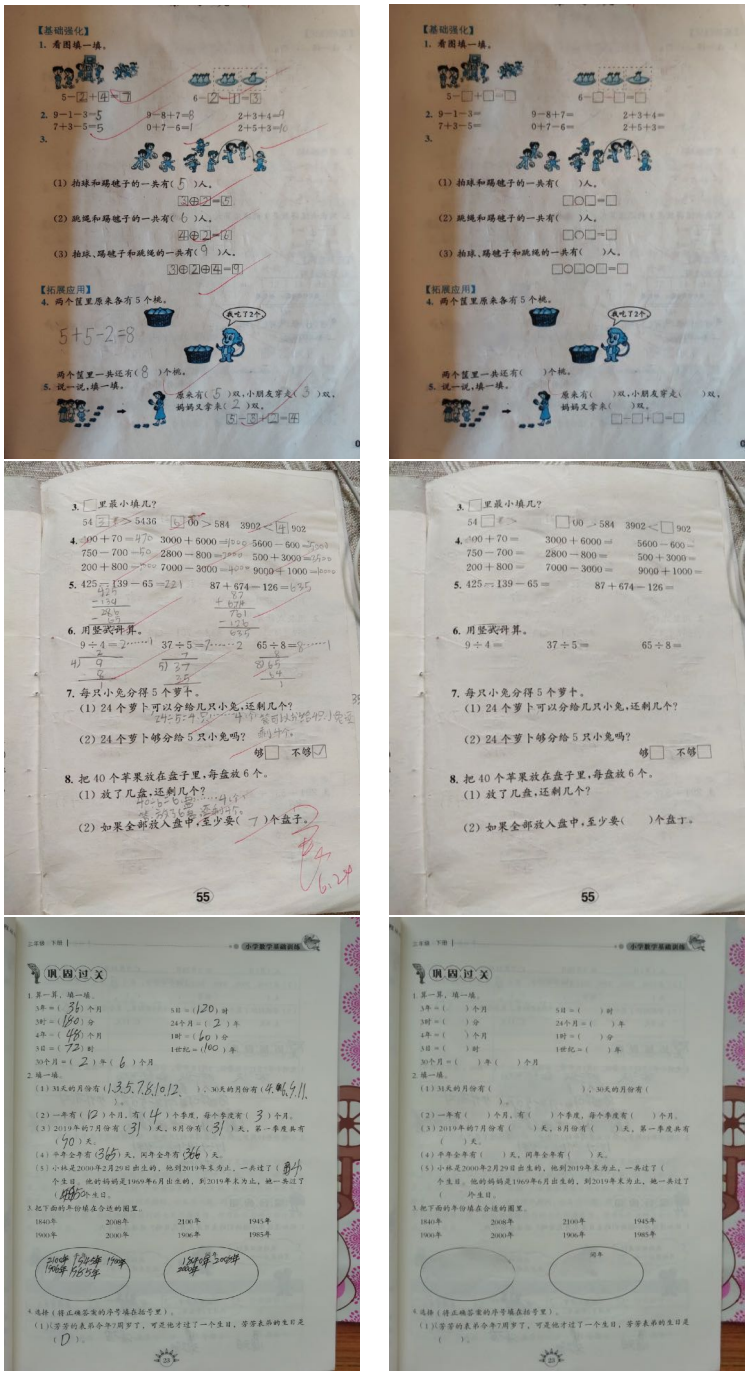
(a) Input (b) Ours (c) GT

Fig. 16. More qualitatively results on SCUT-EnsText.



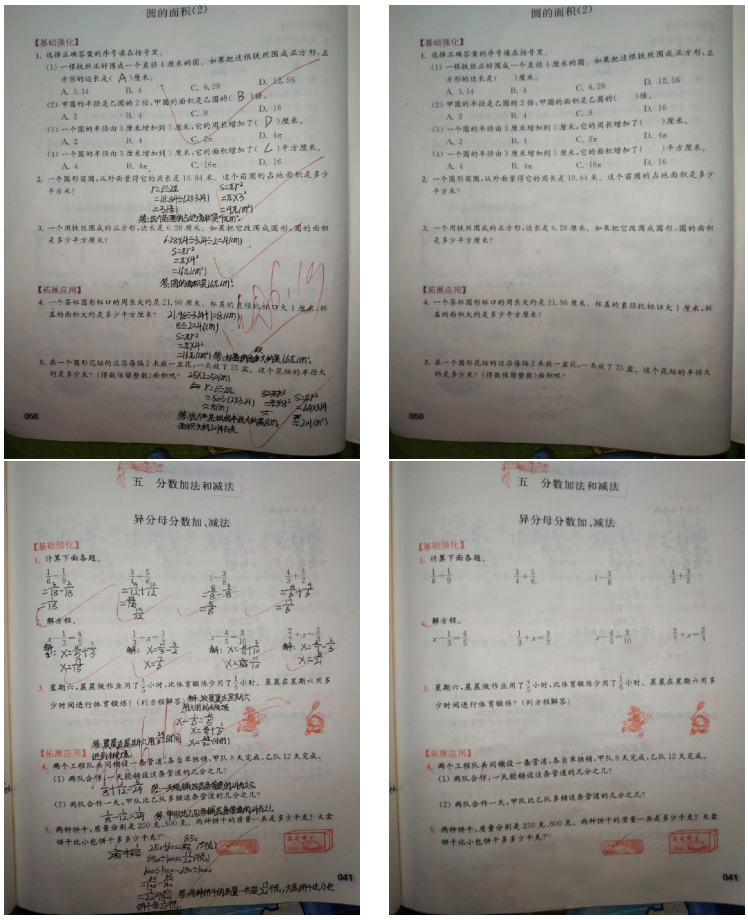
(a) Input (b) Ours

Fig. 17. More qualitatively results on Examination papers.



(a) Input (b) Ours

Fig. 18. More qualitatively results on Examination papers.



(a) Input (b) Ours

Fig. 19. More qualitatively results on Examination papers.

References

1. 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)
2. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. Proc. ICLR (2014)
3. 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)
4. 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)
5. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: Proc. ICLR (2018)
6. 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)
7. 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)
8. 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)
9. 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)