

PERT: A Progressively Region-based Network for Scene Text Removal

Yuxin Wang, Hongtao Xie, Shancheng Fang, Yadong Qu and Yongdong Zhang

University of Science and Technology of China

{wangyx58, qqqyd}@mail.ustc.edu.cn, {htxie, fangsc, zhyd73}@mail.ustc.cn

Abstract

Scene text removal (STR) contains two processes: text localization and background reconstruction. Through integrating both processes into a single network, previous methods provide an implicit erasure guidance by modifying all pixels in the entire image. However, there exists two problems: 1) the implicit erasure guidance causes the excessive erasure to non-text areas; 2) the one-stage erasure lacks the exhaustive removal of text region. In this paper, we propose a ProgEssively Region-based scene Text eraser (PERT), introducing an explicit erasure guidance and performing balanced multi-stage erasure for accurate and exhaustive text removal. Firstly, we introduce a new region-based modification strategy (RegionMS) to explicitly guide the erasure process. Different from previous implicitly guided methods, RegionMS performs targeted and regional erasure on only text region, and adaptively perceives stroke-level information to improve the integrity of non-text areas with only bounding box level annotations. Secondly, PERT performs balanced multi-stage erasure with several progressive erasing stages. Each erasing stage takes an equal step toward the text-erased image to ensure the exhaustive erasure of text regions. Compared with previous methods, PERT outperforms them by a large margin without the need of adversarial loss, obtaining SOTA results with high speed (71 FPS) and at least 25% lower parameter complexity. Code is available at <https://github.com/wangyuxin87/PERT>.

1. Introduction

As one of the most important mediums in information interaction, scene text contains quite a lot of sensitive and private information [5, 15, 2, 9]. To prevent these private messages from being used in illegal ways, Scene Text Removal (STR) task aims to remove the texts in the scene images and fill in the background information correspondingly. Benefiting from the development of Generative Adversarial Networks (GANs) [6, 7], recent STR methods achieve promising results with various solutions [4, 33]. However, there

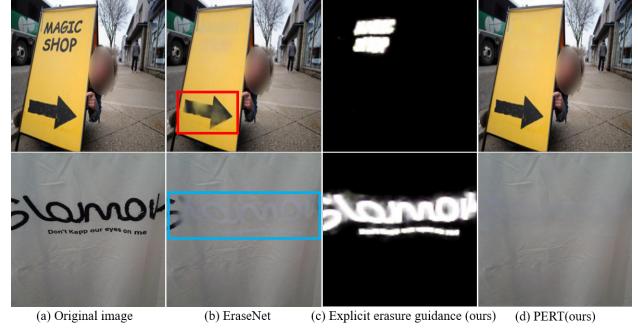


Figure 1. The comparison between PERT and EraseNet [10]. Compared with previous methods, PERT effectively handles the excessive erasure problem (red box) and inexhaustive erasure problem (blue box). Bounding box level annotations are used to supervise the explicit erasure guidance.

are still two problems to be solved.

The first problem is the excessive erasure problem, which makes the poor integrity of non-text regions. Recent methods [33, 4] simply use paired images to train the model, integrating text localization and background reconstruction into a single network. However, since the text instances are sparse and exist in partial areas of scene images, such implicit erasure guidance that modifies all pixels in the entire image is not suitable for STR task. Though EraseNet [10] additionally uses a mask branch to enhance the text location perceiving, the pixel-wise reconstruction on the entire image is still performed under an implicit erasure guidance. As shown in red box of Fig. 1 (b), such implicit erasure guidance has limited capability to maintain the integrity of non-text areas. In this paper, we are **the first** to argue that *the explicit erasure guidance is the key*, which provides targeted and regional erasure on only text regions to prevent modifying non-text areas.

The second problem is the inexhaustive erasure problem, resulting in the remnants of text traces. Early methods [13, 33] achieve text removal through a one-stage erasure, which has limited capability to obtain exhaustive erasure of text regions. Though recent methods [17, 10] design a multi-stage eraser to refine the coarse erased image, due

to the different learning difficulty caused by the same text-erased image supervision, it is difficult to balance the network architecture between the coarse and refinement stage. Thus, such imbalanced multi-stage erasure will leave some traces of text regions in the removal result (blue box of Fig. 1 (b)). Based on the above analyses, how to construct the balanced multi-stage erasure needs to be explored.

In this paper, we propose a novel *ProgrEssively Region-based scene Text eraser* (PERT) to handle above two problems from following two aspects: introducing an explicit erasure guidance and constructing the balanced multi-stage erasure. As shown in Fig. 2, PERT consists of several erasing blocks. Instead of integrating text localization and background reconstruction into a heavy network, we construct a lightweight decoupled structure of text localization network (TLN) and background reconstruction network (BRN) in each erasing block (shown in Fig. 3). **1) The explicit erasure guidance.** As the text region predicted by TLN is a natural erasure guidance, we propose a new region-based modification strategy (RegionMS) to explicitly guide the BRN to only modify the predicted text regions. As background textures are directly inherited from the original image, RegionMS regards scene text removal as a targeted and regional erasure process to prevent the modification of non-text areas. Since the reconstruction learning on the final erased image aims to learn the stroke-level reconstruction rules, RegionMS enables TLN to adaptively perceive stroke-level information to further ensure the integrity of non-text areas (Fig. 1 (c)) with only bounding box level annotations (Fig. 4 (d)). **2) The balanced multi-stage erasure.** To balance the network architecture and learning difficulty among different stages, we firstly construct a balanced erasure structure by sharing the parameters of each erasing block (shown in Fig. 2). Then, through only supervising the output of last erasing stage, PERT learns to adaptively balance the learning difficulty among different erasing stages, where each erasing stage aims to take an equal step toward text-erased image (shown in Fig. 5). As parameters are shared among all erasing blocks, PERT is able to achieve exhaustive erasure with a light structure. In addition, to further improve the erasure performance, we propose a new Region-Global Similarity Loss (RG loss) to consider the feature consistency and visual quality of erasure results from both local and global perspective. Compared with previous methods, PERT obtains more exhaustive erasure of text regions while maintaining the integrity of non-text areas. Without the need of adversarial loss, PERT outperforms existing methods by a large margin with high speed (71 FPS) and at least 25% lower parameter complexity.

The proposed method has several novelties and advantages: 1) To best of our knowledge, we are **the first** to propose an explicit erasure guidance in STR task. Fur-

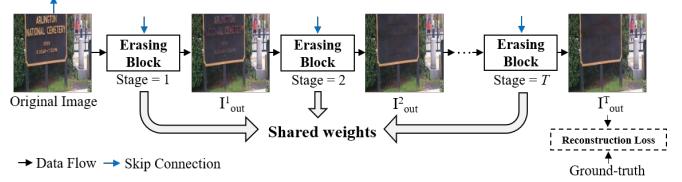


Figure 2. The pipeline of PERT. I_{out}^t is the erased image from t^{th} erasing stage.

thermore, we also provide qualitative visualization to prove how it prevents the erasure of non-text regions and why it is more suitable for STR task. 2) Through designing a balanced erasure structure and supervising only the last erasing stage, PERT effectively balances the learning difficulty and network structure among different erasing stages, obtaining exhaustive erasure of text regions with a light architecture. 3) A new RG loss is proposed to improve the feature consistency and visual quality of erasure results. The SOTA results on both synthetic and real-world datasets demonstrate the effectiveness of our method.

2. Related Work

Early STR methods [8, 20] mainly cascade the conventional text localization and background reconstruction processes for text removal. With the deep learning emerging as the most promising machine learning tool [19, 30, 21, 14], recent STR methods try to integrate the two independent processes into a single architecture by end-to-end training the network with paired images. Nakamura *et al.* [13] implement a patch-based skip-connected auto-encoder for text removal. Under the premise of reducing the consistency of removal image, the coarse erasure result is obtained by taking the patch-level images as inputs. MTRNet [18] concatenates the text location map with original image to improve the erasure performance. As texts are sparse in the scene image [16, 22, 29], such implicit erasure guidance by modifying all the pixels in the entire image will cause the excessive erasure to non-text regions. Benefiting from the development of GANs [1, 35, 32], recent methods attempt to adopt adversarial loss to improve the erasure visuality. Though the adversarial loss significantly increases the training difficulty, the impressive improvement in visual quality makes it popular in STR task. Following the structure of cGAN [12], Ensnet [33] designs a local-aware discriminator to ensure the consistency of the erased regions. EraseNet [10] and MTRNet++ [17] further construct a refinement-network to optimize the coarse output from the first erasure stage. Following these methods [10, 17], our method falls into multi-stage eraser, but we try to balance the learning difficulty and network structure among different stages. Compared with previous methods, PERT effectively handles the excessive and inexhaustive erasure problems from

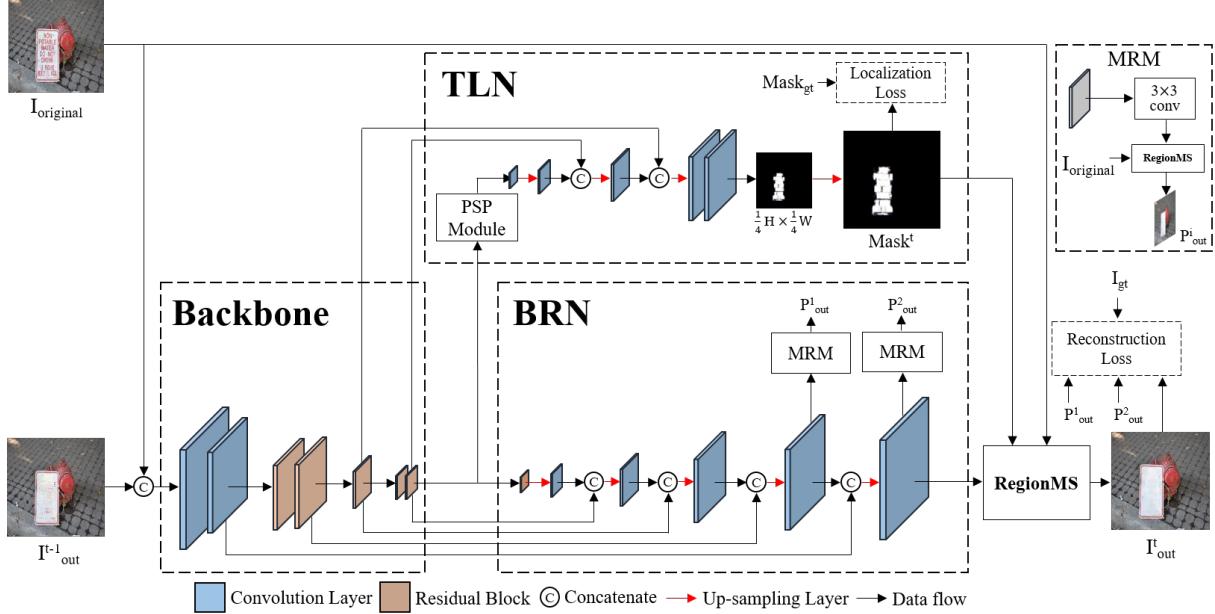


Figure 3. The architecture of erasing block. I_{out}^t means the output from stage t , which has the same resolution of the original image $I_{original}$. $Mask^t$ is the mask map generated from TLN in stage t . MRM means Multi-scale Reconstruction Module.

the aspects of introducing explicit erasure strategy and constructing balanced multi-stage erasure.

3. Our Approach

3.1. Pipeline

The pipeline of PERT is shown in Fig. 2, which cascades T lightweight erasing blocks. Through iteratively implementing erasing block on the erased image from previous stage, the output of the last erasing block is used as the final erasure result. To eliminate the effect of text characteristic losing, we concatenate the original image $I_{original} \in R^{H \times W \times 3}$ with $I_{out}^{t-1} \in R^{H \times W \times 3}$ (H and W are height and width respectively) to generate the input of t^{th} erasing block (shown in Fig. 3). For the first erasing stage, we concatenate $I_{original}$ with itself as the input.

3.2. Lightweight Erasing Block

Inspired by the manual erasure process that firstly locating text regions and then performing regional modification, we design a decoupled structure in each erasing block. As shown in Fig. 3, the erasing block contains a text localization network (TLN) for text location prediction and a background reconstruction network (BRN) for background textures reconstruction. Backbone network are shared by both TLN and BRN for feature extraction, which contains two convolution layers and five residual blocks [3].

As text instances have large-variance scales, generating the robust representation of multi-scale texts is necessary for accurate scene text detection [23, 24]. Inspired by PSP

Module [34] which obtains promising performance in long-range dependency capture, we implement a PSP Module in TLN for robustly representing multi-scale texts. To reduce the computation cost, we only gradually up-sample the feature map to 1/4 size of the original image (bilinear interpolation is employed), and directly resize it to the same size of input image. Mask map $Mask^t \in [0, 1]$ is generated through a sigmoid layer, where t is the stage number. In the training stage, we supervise $Mask^t$ with the bounding box level annotations.

In order to learn the robust reconstruction rules of background textures, BRN is expected to contain two characteristics: 1) Modeling low-level texture information for background and foreground texture perception. 2) Capturing high-level semantics for feature representation enhancement. Thus, we construct BRN with skip connections to perceive low-level structures from shallow layers, and employ a relative deep network for high-level semantics capture. As the background reconstruction is a more challenging task than text localization, we use the deconvolution operation in the up-sampling layer to enhance the feature representation ability. Furthermore, the Multi-scale Reconstruction Module (MRM) is constructed to predict multi-scale erasure results (P_{out}^1 and P_{out}^2), which is only implemented in the training stage for performance boosting [33].

Instead of integrating the text localization and background reconstruction into an heavy network, our decoupled structure effectively reduces the parameter size, which only needs two lightweight network for localization and reconstruction respectively (detailed in Sec. 4.4).

3.3. Region-based Modification Strategy

Different from previous implicit erasure guidance methods [33, 10], which modify all the pixels in the entire image, we provide an explicit erasure guidance for targeted and regional erasure. Such explicit erasure guidance is achieved by a region-based modification strategy (RegionMS).

RegionMS is formulated in Eq. 1. Firstly, to sample the candidate erasure areas, we use the text mask map $Mask^t \in [0, 1]$ to filter reconstructed image I^t from BRN. Thus, BRN will only modify the pixels in the predicted text regions. Relatively, to prevent the erasure of non-text areas, we preserve the background textures by performing element wise product between $(1 - Mask^t)$ and original image $I_{original}$. It is worth mentioning that RegionMS is also employed in MRM to generate multi-scale predictions (P_{out}^1 and P_{out}^2), and the corresponding mask map is directly resized from $Mask^t$.

$$I_{out}^t = Mask^t \times I^t + (1 - Mask^t) \times I_{original} \quad (1)$$

To better illustrate the significance of RegionMS in ensuring the integrity of non-text areas, we provide some visualization of $Mask^t$. The details about this part is available in Sec. 4.5.1

3.4. Balanced Multi-stage Erasure

As the deep network can erase a large-range step for difficult cases while the shallow network provides a small-range one, due to the different learning difficulty caused by the same text-erased image supervision, it is difficult to balance the learning difficulty and network architecture in the multi-stage erasure [17, 10]. Thus, we adopt the most straightforward approach by dividing the overall learning difficulty and total network structure into T equal parts, where T is an ablation study in Sec. 4.3.

Firstly, we construct a balanced erasure structure by implementing the erasing block with same structure in each erasing stage. To reduce the parameter complexity, we share the parameters among all erasing blocks. As the generation of text-erased image in T different erasure degrees requires a lot of human costs, we simply supervise the output of only last erasing block, guiding the network to adaptively balance the learning difficulty among different stages. By doing these, a new balanced multi-stage erasure is obtained, where each erasing stage aims to take an equal step toward the text-erased image for exhaustive erasure (detailed in Sec. 4.5.1).

3.5. Training Objective

We use the original image $I_{original}$, text-removed ground-truth I_{gt} and localization map $Mask_{gt}$ in bounding-box level (0 for non-text regions and 1 for text regions) to

train PERT. The total loss contains two parts: localization loss and reconstruction loss. In order to adaptively balance the learning difficulty among different erasing stages, we only implement the reconstruction loss to the erased images (I_{out} , P_{out}^1 and P_{out}^2) in the final erasing stage. In contrast, we supervise the mask map ($Mask^t$) in each erasing stage to guarantee an accurate erasure guidance.

3.5.1 Localization loss

We use dice loss defined in [25] to guide the learning process of TLN. As shown in Eq. 2, p_i is the prediction and y_i is the ground-truth.

$$L_{loc} = 1 - \frac{2 \sum_i p_i y_i}{\sum_i p_i + \sum_i y_i} \quad (2)$$

3.5.2 Reconstruction loss

Benefiting from the RegionMS and balanced multi-stage erasure, the simple similarity losses are sufficient to train PERT.

1) Region-Global Similarity Loss (RG loss). The RG loss is newly proposed in this paper to consider the feature consistency and visual quality of erasure results from both local and global perspective. RG loss contains two parts (shown in Eq. 3): region-aware similarity (RS) loss and global-aware similarity (GS) loss.

$$L_{RG} = L_{RS} + L_{GS} \quad (3)$$

$$\begin{aligned} L_{RS} = & \sum_{i=1}^2 \alpha_i \| (P_{out}^i - I_{i,gt}) * Mask_{i,gt} \|_1 + \sum_{n=1}^2 \beta_i \| (P_{out}^i \\ & - I_{i,gt}) * (1 - Mask_{i,gt}) \|_1 + \alpha \| (I_{out} - I_{gt}) \\ & * Mask_{gt} \|_1 + \beta \| (I_{out} - I_{gt}) * (1 - Mask_{gt}) \|_1 \end{aligned} \quad (4)$$

As shown in Eq. 4, RS loss takes multi-scale predictions P_{out}^i into consideration. $Mask_{i,gt}$ and $I_{i,gt}$ are generated by directly resizing the mask map $Mask_{gt}$ and text-erased image I_{gt} . The RS loss assigns the pixels in text region with a higher weight. To be specific, we set $\alpha, \alpha_1, \alpha_2 = 13, 10, 12$ and $\beta, \beta_1, \beta_2 = 2, 0.8, 1$ respectively.

Different from RS loss, GS loss aims to penalize the consistency and enhance the visual quality from a global view. We firstly down-sample $n = 3$ activation maps $\phi_n(I_{out}) \in R^{H_n \times W_n}$ from the 4th, 9th, 16th layer of pre-trained VGG16 network to $F_{out}^n \in R^{S_m \times S_m}$ through max-pooling (the same process to I_{gt}). Inspired by the pair-wise Markov random field, which is widely used to improve the spatial labeling contiguity, we compute the pair-wise similarities between ground-truth and predicted features. Let γ_{ij}^n denotes the similarity between the j^{th} pixel and i^{th} pixel in

feature F^n (shown in Eq. 5). $\gamma_{ij}^{n,out}$ means the similarity from F_{out}^n .

$$\gamma_{ij}^n = (F_i^n)^T F_j^n / (\|F_i^n\|_2 \|F_j^n\|_2) \quad (5)$$

In our experiments, we set $S_m = 8, 4$ and 1 when $m = 1, 2$ and 3 , which means we calculate the pair-wise similarities in three different scales ($\gamma_{ij}^{n,m}, m = 1, 2, 3$) for each feature $\phi_n(I)$. Finally, the GS loss is formulated in Eq. 6. We choose the squared difference to compute the pair-wise similarity.

$$L_{GS} = \sum_{n=1}^3 \sum_{m=1}^3 \left(\frac{1}{S_m \times S_m} (\gamma_{ij}^{n,m,out} - \gamma_{ij}^{n,m,gt})^2 \right) \quad (6)$$

2) Negative SSIM Loss [28]. This loss is used to analyse the degradation of structural information:

$$L_{ssim} = -SSIM(I_{out}, I_{gt}) \quad (7)$$

3) VGG Loss. Inspired by previous STR methods [33, 10], we also adopt VGG loss (L_{vgg}) to improve the erasure results. The details can be obtained in the previous methods [33, 10].

Finally, the total loss function is formulated in Eq. 8.

$$L = L_{loc} + L_{RG} + L_{ssim} + L_{vgg} \quad (8)$$

4. Experiments

4.1. Datasets and Evaluation

4.1.1 Datasets

We conduct the experiments following the setup of [10]. We train PERT on the only official training images of SCUT-Syn [33] or SCUT-EnsText [10], and then evaluate the model on the corresponding testing sets, respectively. Details of these two datasets can be found in the previous works [33, 10].

4.1.2 Evaluation

To comprehensively evaluate the erasure results of our method, we use both Image-Eval (PSNR, MSSIM, MSE, AGE, pEPs and pCEPs) and Detection-Eval (precision (P), recall (R), F-measure(F), TIoU-precision (TP), TIoU-recall (TR), TIoU-F-measure(TF)). Details about Image-Eval and Detection-Eval can be found in the previous works [11, 10]. A higher PSNR and SSIM or lower MSE, AGE, pEPs, pCEPs, P, R, F, TP, TR and TF represent better results.

4.2. Implementation Details

Data augmentation includes random rotation with maximum degree of 10° and random horizontal flip with a probability of 0.3 during training stage. PERT is end-to-end

trained using Adam optimizer. The learning rate is set to $1e-3$. The model is implemented in Pytorch and trained on 2 NVIDIA 2080Ti GPUs.

To share the parameters among different stages, we iteratively use the same erasing block in all stages. As the gradient will accumulate on the same erasing block, the simple `loss.backward()` & `optimizer.step()` are used for parameter updating. Details are available in our submitted code.

4.3. Ablation Study

4.3.1 The region-based modification strategy

As shown in Tab. 1, through introducing an explicit erasure guidance, the proposed PERT significantly improves the erasure performance by 0.66, 0.33, 0.0002, 0.3931, 0.0016 and 0.0008 in PSNR, MSSIM, MSE, AGE, pEPs and pCEPs respectively. We attribute this remarkable improvement to two reasons: 1) the RegionMS provides targeted and regional modification on only text-region textures without changing pixels in non-text areas, ensuring the integrity of text-free regions. 2) The RegionMS reduces the learning difficulty of reconstruction process, helping BRN to focus on targeted reconstruction rules of text regions without considering non-text areas. The qualitative visualization are detailed in Sec. 4.5.1.

4.3.2 The balanced multi-stage erasure

As shown in Tab. 2, the one-stage erasure ($T = 1$) has a limited capability to reconstruct background textures. When we increase the number of erasing stages step-by-step, the balanced multi-stage erasure increases the performance on all metrics, and the relative increases are 1.85, 0.59, 0.0003, 0.5361, 0.0092 and 0.0034 in PSNR, MSSIM, MSE, AGE, pEPs and pCEPs respectively. Benefiting from sharing parameters among all erasing blocks, PERT step-by-step refines the erasure result with ZERO parameter size increase.

4.3.3 The Region-Global similarity loss

The RG loss penalizes the feature consistency and enhances the visual quality from both local and global views. As shown in Tab. 3, the RG loss achieves the improvement by 0.52, 0.1, 0.0001, 0.1556, 0.0026 and 0.0020 in PSNR, MSSIM, MSE, AGE, pEPs and pCEPs respectively. Benefiting from the balanced multi-stage erasure and RegionMS, the simple similarity losses are sufficient to train PERT without the need of adversarial loss.

4.4. Comparison with State-of-the-Art Methods

The quantitative results on SCUT-EnsText dataset are shown in Tab. 4. The state-of-the-art performance demonstrates that the proposed PERT outperforms existing meth-

Model	PSNR MSSIM	MSE AGE	pEPs	pCEPs
w/o RegionMS	32.59 96.62	0.0016 2.5764	0.0152	0.0096
w/ RegionMS	33.25 96.95	0.0014 2.1833	0.0136	0.0088

Table 1. Ablation studies about the RegionMS.

T	PSNR MSSIM	MSE AGE	pEPs	pCEPs	Param size
T = 1	31.40 96.36	0.0017 2.7194	0.0228	0.0122	14.0M
T = 2	32.72 96.83	0.0016 2.3750	0.0160	0.0104	14.0M
T = 3	33.25 96.95	0.0014 2.1833	0.0136	0.0088	14.0M

Table 2. Ablation studies about the number of erasing stages. Param means parameter.

Model	PSNR MSSIM	MSE	AGE	pEPs	pCEPs
w/o RG loss	32.73 96.85	0.0015	2.3389	0.0162	0.0108
w/ RG loss	33.25 96.95	0.0014	2.1833	0.0136	0.0088

Table 3. Ablation studies about the RG loss.

ods on all metrics. Though EraseNet constructs a mask branch to enhance the perception of text appearance, the explicit erasure guidance in PERT effectively improves the erasure performance and achieves a new state-of-the-art result in both Detection-Eval and Image-Eval. For further fair comparison, we carefully reimplement EraseNet [10] with the same training setting as ours, the improvement is also consistent. Compared with GAN-based methods [4, 33, 10], our method obtains impressive performance with much simpler training objective, using only simple similarity losses to guide the learning process. In addition, we quantitatively compare the erasure results with recent approaches on SCUT-Syn dataset. As shown in Tab. 5, the proposed PERT obtains the dominant performance compared with both GAN-based [4, 33, 10] and GAN-free methods [13].

To qualitatively compare the erasure results, we visualize some examples in Fig. 4. Specially, we choose the latest and the most related approach [10] for a detailed comparison. 1) **The integrity of non-text regions.** Though EraseNet [10] introduces a mask subnetwork to enhance the perception of text appearance, the implicit erasure guidance may modify the pixels belonging to non-text regions (the first row in Fig. 4). Benefiting from the targeted and regional erasure provided by RegionMS, PERT modifies only

text regions while ensuring the integrity of non-text areas.

2) **The exhaustive erasure of text regions.** Through balancing the learning difficulty and network structure among different erasing stages, PERT effectively achieves more exhaustive erasure by progressively taking an equal step toward text-erased images. As shown in the second row and (g) in Fig. 4, PERT achieves exhaustive erasure and provides text-erased image with high visual quality.

4.4.1 Model size and speed

We compare the parameter size between PERT and existing methods in Tab. 6. We attribute our low parameter complexity to following two reasons: 1) the decoupled erasure structure reduces the network learning difficulty, which only needs two lightweight networks for detection and reconstruction without constructing a heavy network to achieve both functions simultaneously. 2) PERT shares the parameters among all erasing blocks. As shown in Tab. 6, PERT effectively reduces the model size by at least 25% from existing multi-stage erasers [17, 10] without performance decrease. The speed comparison is shown in Tab. 7. Without the need of adversarial loss, PERT effectively reduces the training time. In the testing stage, PERT also obtains the comparable speed and achieves real-time inference. Based on the above analyses, PERT obtains a better balance among the erasure performance, parameter complexity and inference speed.

4.5 Quantitative Analysis

4.5.1 The significance of explicit erasure guidance

We summarize the reasons why RegionMS is more suitable for STR task into two aspects: 1) As the reconstruction learning (Eq. 3) on final erasure result I_{out}^t aims to learn the stroke-level reconstruction rules, RegionsMS promotes TLN to perceive stroke-level information (Fig. 4 (e)) with only bounding box level annotations (Fig. 4 (d)). On the one hand, for text areas with large character spacing, TLN generates the stroke-level mask map to further reduce the modification of background textures. On the other hand, TLN predicts region-level mask map for the text areas with small character spacing, as the region-level mask map only covers little background textures. Thus, such explicit erasure guidance is able to guide an accurate erasure by implementing targeted and regional modification on stroke-level text regions. 2) RegionMS reduces the learning difficulty in BRN, where texture reconstruction of text-free areas is not considered, resulting in a coarse reconstruction (shown in Fig. 5).

Method	Image-Eval						Detection-Eval					
	PSNR	MSSIM	MSE	AGE	pEPs	pCEPs	P	R	F	TP	TR	TF
Original images	-	-	-	-	-	-	79.4	69.5	74.1	61.4	50.9	55.7
Pix2Pix [4]	26.6993	88.56	0.0037	6.0860	0.0480	0.0227	69.7	35.4	47.0	52.0	24.3	33.1
STE [13]	25.4651	90.14	0.0047	6.0069	0.0533	0.0296	40.9	5.9	10.2	28.9	3.6	6.4
EnsNet [33]	29.5382	92.74	0.0024	4.1600	0.0307	0.0136	68.7	32.8	44.4	50.7	22.1	30.8
EraseNet [10]	32.2976	95.42	0.0015	3.0174	0.0160	0.0090	53.2	4.6	8.5	37.6	2.9	5.4
EraseNet* [10]	32.0486	95.47	0.0015	3.2751	0.0169	0.0098	58.3	3.8	7.1	42.1	2.3	4.4
PERT	33.2493	96.95	0.0014	2.1833	0.0136	0.0088	52.7	2.9	5.4	38.7	1.8	3.5

Table 4. Comparisons between previous methods and proposed PERT on SCUT-EnsText. * means our reimplementations.

Method	PSNR	MSSIM	MSE	AGE	pEPs	pCEPs
Pix2Pix [4]	26.76	91.08	0.0027	5.4678	0.0473	0.0244
STE [13]	25.40	90.12	0.0065	9.4853	0.0553	0.0347
EnsNet [33]	37.36	96.44	0.0021	1.73	0.0069	0.0020
EraseNet [10]	38.32	97.67	0.0002	1.5982	0.0048	0.0004
EraseNet* [10]	37.70	97.34	0.0003	1.8044	0.0059	0.0009
PERT	39.40	97.87	0.0002	1.4149	0.0045	0.0006

Table 5. Comparisons between previous methods and proposed PERT on SCUT-Syn. * means our reimplementations.

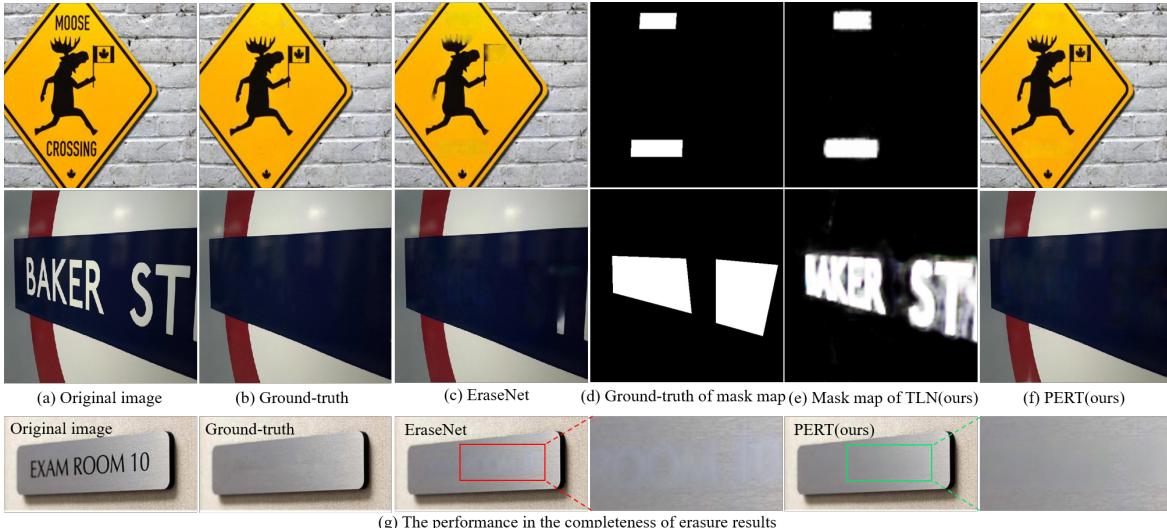


Figure 4. The visualization of erasure results on SCUT-EnsText.

Method	PSNR	MSE	Params
Pix2Pix [4]	26.8	0.0027	54.4M
MTRNet++ [17]	34.55	0.0004	18.7M
EraseNet [10]	38.32	0.0002	19.7M
PERT	39.40	0.0002	14.0M

Table 6. Comparisons between previous methods and proposed PERT on erasure results and parameter size.

	EraseNet	PERT-1	PERT-2	PERT-3
MSSIM	95.42	96.36	96.83	96.95
Training (h)	34.7	11.4	14.3	16.6
Testing (FPS)	86	204	101	71

Table 7. Comparisons between previous methods and proposed PERT on speed. PERT-i means using i erasing stages.

4.5.2 The significance of balanced multi-stage erasure

By designing a balanced erasure structure and only supervising the last erasing stage, PERT performs a balanced

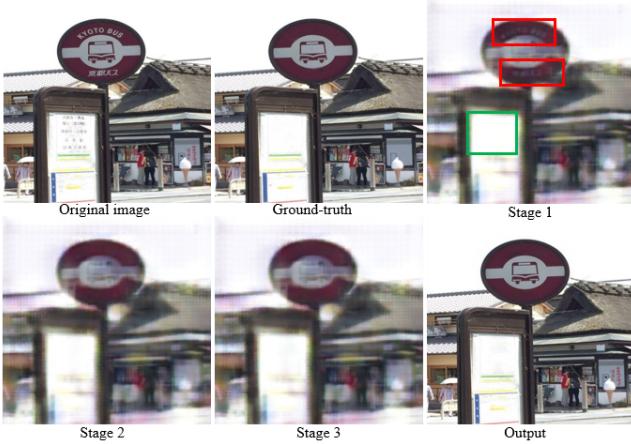


Figure 5. The visualization of reconstructed images from BRN in different stages. Output is the erased image from RegionMS in the last stage.

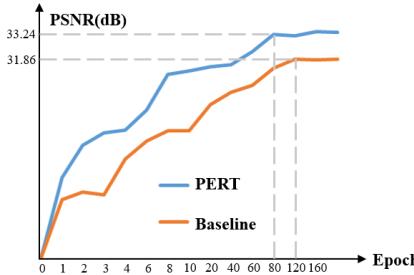


Figure 6. The comparison of convergence in the training stage between PERT and baseline model.

multi-stage erasure. To be specific, for difficult cases (red boxes in Fig. 5), BRN optimizes the erasure result from the previous stage in an equally small-range step. Thus, exhaustive erasure of difficult cases is achieved by progressively erasing. For the relatively easy cases (green box in Fig. 5), PERT tends to achieve a more exhaustive erasure in the early stages.

4.5.3 The significance in training

We visualize the PSNR value on the SCUT-EnsText during training. The baseline model is implemented with adversarial loss and constructed without balanced multi-stage erasure and RegionMS. As shown in Fig. 6, PERT obtains a better (33.24 vs 31.86) and faster (80 epoch vs 120 epoch) convergence.

4.5.4 Limitation

As shown in Fig.7, our method fails when TLN provides an inaccurate detection result (detecting non-text region or missing detecting text region). However, this problem also exists in previous methods (EraseNet [10] for example). It

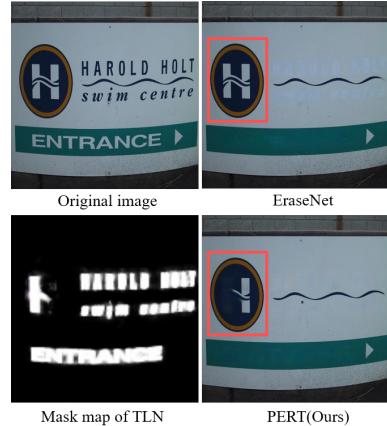


Figure 7. The example of failed cases.

is worth mentioning that our method reduces the impact of this problem to a certain extent (*e.g.* the erasure result of character "H" in Fig.7). Furthermore, the proposed PERT provides a solution for this issue by embedding the latest detection branch [31, 27] to improve the quality of mask map.

5. Conclusion

This paper proposes a simple but strong scene text eraser named PERT. Based on the explicit erasure guidance and balanced multi-stage erasure, we qualitatively and quantitatively verify that PERT effectively handles the excessive and inexhaustive erasure problems in STR task. The simplicity of PERT makes it easy to develop new scene text removal models by modifying the existing ones or introducing other network modules. Extensive experiments demonstrate that the proposed method achieves state-of-the-art performance on both synthetic and real-world datasets while maintaining a low complexity. In the future, we will develop this work to the end-to-end text edit task.

References

- [1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.
- [2] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE international conference on computer vision*, pages 5076–5084, 2017.
- [3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on*

- computer vision and pattern recognition*, pages 1125–1134, 2017.
- [5] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160. IEEE, 2015.
- [6] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.
- [7] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.
- [8] Mohammad Khodadadi and Alireza Behrad. Text localization, extraction and inpainting in color images. In *20th Iranian Conference on Electrical Engineering (ICEE2012)*, pages 1035–1040. IEEE, 2012.
- [9] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. Show, attend and read: A simple and strong baseline for irregular text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8610–8617, 2019.
- [10] Chongyu Liu, Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and Yongpan Wang. Erasenet: End-to-end text removal in the wild. *IEEE Transactions on Image Processing*, 29:8760–8775, 2020.
- [11] Yuliang Liu, Lianwen Jin, Zecheng Xie, Canjie Luo, Shuaitao Zhang, and Lele Xie. Tightness-aware evaluation protocol for scene text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9612–9620, 2019.
- [12] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [13] Toshiki Nakamura, Anna Zhu, Keiji Yanai, and Seiichi Uchida. Scene text eraser. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 832–837. IEEE, 2017.
- [14] Zhi Qiao, Yu Zhou, Dongbao Yang, Yucan Zhou, and Weiping Wang. Seed: Semantics enhanced encoder-decoder framework for scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13528–13537, 2020.
- [15] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2016.
- [16] Zhuotao Tian, Michelle Shu, Pengyuan Lyu, Ruiyu Li, Chao Zhou, Xiaoyong Shen, and Jiaya Jia. Learning shape-aware embedding for scene text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4234–4243, 2019.
- [17] Osman Tursun, Simon Denman, Rui Zeng, Sabesan Sivalan, Sridha Sridharan, and Clinton Fookes. Mtrnet++: One-stage mask-based scene text eraser. *Computer Vision and Image Understanding*, 201:103066, 2020.
- [18] Osman Tursun, Rui Zeng, Simon Denman, Sabesan Sivalan, Sridha Sridharan, and Clinton Fookes. Mtrnet: A generic scene text eraser. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 39–44. IEEE, 2019.
- [19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [20] Priyanka Deelip Wagh and DR Patil. Text detection and removal from image using inpainting with smoothing. In *2015 International Conference on Pervasive Computing (ICPC)*, pages 1–4. IEEE, 2015.
- [21] Zhaoyi Wan, Mingling He, Haoran Chen, Xiang Bai, and Cong Yao. Textscanner: Reading characters in order for robust scene text recognition. *arXiv preprint arXiv:1912.12422*, 2019.
- [22] Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. Shape robust text detection with progressive scale expansion network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [23] Wenhai Wang, Enze Xie, Xiaoge Song, Yuhang Zang, Wenjia Wang, Tong Lu, Gang Yu, and Chunhua Shen. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8440–8449, 2019.
- [24] Yuxin Wang, Hongtao Xie, Zilong Fu, and Yongdong Zhang. Dsrn: A deep scale relationship network for scene text detection. In *IJCAI*, pages 947–953, 2019.
- [25] Yuxin Wang, Hongtao Xie, Zilong Fu, and Yongdong Zhang. Dsrn: A deep scale relationship network for scene text detection. In *IJCAI*, pages 947–953, 2019.
- [26] Yuxin Wang, Hongtao Xie, Zhengjun Zha, Youliang Tian, Zilong Fu, and Yongdong Zhang. R-net: A relationship network for efficient and accurate scene text detection. *IEEE Transactions on Multimedia*, 23:1316–1329, 2020.
- [27] Yuxin Wang, Hongtao Xie, Zheng-Jun Zha, Mengting Xing, Zilong Fu, and Yongdong Zhang. Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11753–11762, 2020.
- [28] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- [29] Chuhui Xue, Shijian Lu, and Wei Zhang. Msr: multi-scale shape regression for scene text detection. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 989–995, 2019.
- [30] Fangneng Zhan and Shijian Lu. Esir: End-to-end scene text recognition via iterative image rectification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2059–2068, 2019.
- [31] Chengquan Zhang, Borong Liang, Zuming Huang, Mengyi En, Junyu Han, Errui Ding, and Xinghao Ding. Look more

- than once: An accurate detector for text of arbitrary shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10552–10561, 2019.
- [32] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.
- [33] Shuaítao Zhang, Yuliang Liu, Lianwen Jin, Yaxiong Huang, and Songxuan Lai. Ensnet: Ensconce text in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 801–808, 2019.
- [34] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.
- [35] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.