LEARNING TO REMOVE REFLECTIONS FOR TEXT IMAGES

Ce Wang[†], *Renjie Wan*[◇], *Feng Gao*[‡], *Boxin Shi*[†,♮], *Ling-Yu Duan*[†,♮,*]

[†]National Engineering Lab for Video Technology, Peking University, Beijing, China
[◇]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[‡]The Future Lab, Tsinghua University, Beijing, China
[♮]Peng Cheng Laboratory, Shenzhen, China
{wce, shiboxin, lingyu}@pku.edu.cn, rjwan@ntu.edu.sg, gaofeng2018@mail.tsinghua.edu.cn

ABSTRACT

Text images taken behind a piece of glass in the wild are largely contaminated by reflections. Directly applying existing reflection removal methods on text images with reflections cannot recover clear and correct text contents due to the ignorance of special characteristics of texts. This paper proposes a stacked framework to solve the text image reflection removal problem by specifically considering the regional properties of reflection and embedding the specific text priors into the estimation process in a unified manner. Experiment results on a newly collected dataset demonstrate that the proposed method outperforms state-of-the-art methods in recovering visually pleasant reflection-free images and recognizable text features.

Index Terms— Text image, reflection removal, GAN, text recognition

1. INTRODUCTION

With the widespread usage of imaging sensors, various text images are taken by different devices (*e.g.*, the cameras in surveillance devices or autonomous vehicles). Under the unconstrained scenarios, the text images are prone to be degraded by different distortions. Especially when the text images are taken behind a piece of glass (*e.g.*, traffic signs taken inside a car), they are mostly contaminated by reflections, which unpleasantly affects the human perception and downgrades the performance of many text recognition algorithms. It is therefore of great interest to remove these reflections and enhance the visibility of the text contents.

Utilizing the non-learning image statistics [1, 2] and the deep learning framework [3, 4], much progress has been achieved in recent years for reflection removal problems. Despite their success in images with general scenes, the state-of-the-art approaches are barely suitable for images with rich text information, since the constraints used for general images are not tailored for text images, where the text characters usually show high contrasts against nearby regions and each character has a near-uniform color. On the other hand, due to the regional properties [5], the reflections usually occupy limited regions of the whole images. If simply applying existing reflection removal methods (e.g., CEILNnet [3]) by forcing the global consistency without carefully considering localization of regions with reflections and integrating specific text characteristics, obvious artifacts such as blur and distortion of character shape can be observed on the reflection removal results, as shown in Figure 1.

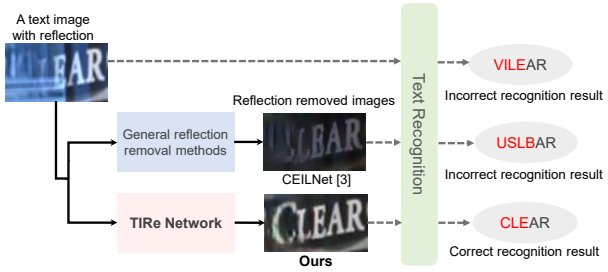

Fig. 1. Given a text image with reflection, our proposed TIRe Network can more clearly remove the reflections, which leads to a more correct text recognition result, than existing reflection removal method designed for a general scene (*e.g.*, CIELNet [3]).

To recover clear and correct text contents from text images with reflections, we first explore the advantages by developing a stacked framework to integrate the reflection localization and reflection removal in an end-to-end pipeline, where the localization information benefits the removal process by correctly pointing to regions covered by reflections. Then we embed the specific text priors into the whole estimation process by comparing the text similarity in a compact feature space instead of the pixel level similarity adopted by previous methods [3]. Our **T**ext **I**mage **Re**flection Removal Network (TIRe Network) is shown in Figure 2, which includes two components: the text reflection localization network to roughly indicate the reflection location, and the text reflection removal network to recover the reflection-free texts. Our major contributions are summarized as follows:

*Ling-Yu Duan is the corresponding author.

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00222

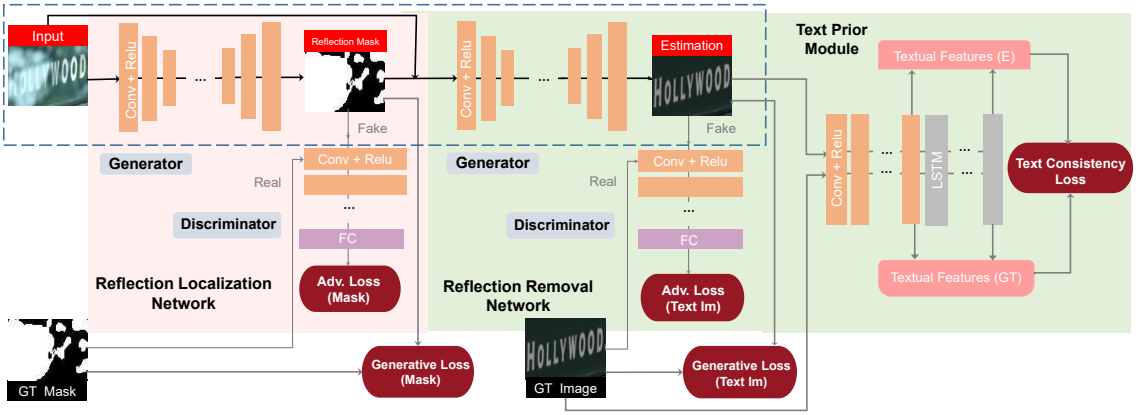

Fig. 2. The framework of TIRe Network. Our network includes two parts: the reflection localization network to locate the reflection regions and the reflection removal network to remove the reflections based on the localization information. Text priors are embedded into the whole estimation process based on a text recognition model to guarantee correct text recovery. When using the pre-trained framework to make inference, only the part inside the dashed bounding box needs to be loaded.

- We propose the first reflection removal framework which is *specially* designed for text images and build the first text image dataset with reflections to facilitate the corresponding research.
- We design a stacked framework to sophisticatedly localize and *clearly* remove text reflections in a unified manner.
- We *correctly* recover important and recognizable text features by employing the feature level similarity of text contents.

2. RELATED WORK

Reflection removal. Reflection removal has been discussed for decades, the investigation of which also witnesses a breakthrough from non-learning stage to deep learning era. Previous works of the non-learning category focus on exploiting different priors from the properties of reflection, *e.g.*, the sparsity prior [6, 7], the different blur levels [1, 2, 8]. These handcrafted priors based approached above are all derived from visible differences of certain properties between the background and reflection, however the observation of such difference is usually restricted to some specific assumptions, thus the methods often fail in more general scenes.

Another category addresses the problem in a deep learning manner, benefiting from its comprehensive capability of modeling implicit data features, and has achieved promising results. Chandramouli *et al.* [9] and Fan [3] utilized the classical two-staged learning framework to remove reflection in a data-driven manner. Wan *et al.* [4] proposed a concurrent model to address the limitations existed in the two-stage framework. Zhang *et al.* [10] introduced a novel perceptual loss to realize the more reliable recovery. Though these learning based methods can obtain visually better results, they often ignore the insights of regional difference in reflection, hence we propose a reflection localization scheme to incorporate regional weights of attention. Besides, no existing reflection removal methods are specifically aimed at the recovery of text images.

Text image enhancement. Text sequence is a common occurrence in everyday scenes, whose recognition from images is thus of great interest. However, the text images in the wild are often contaminated by various degradations then unrecognizable, particularly, the text contents are often covered by glass reflection like the showcase. Previous works on enhancing the visibility of text images mainly focus on the deblurring and shadow removal of low quality text images to implement better binarization. For example, Pan *et al.* [11] proposed a method to restore a blurry text image with L_0 regularization on intensity and statistical prior on gradient. Based on the unique properties of text sequences in images, Choo *et al.* [12] proposed a text-specific deblurring approach. As for shadows in text images, Bako *et al.* [13] matched the local background color with a global reference to produce a shadow-free output. Kligler *et al.* [14] proposed a more general method to realize unshadowing and text binarization via visibility detection of point-set from the image. The text image enhancing methods above have achieved visually pleasing results for their specific tasks, however it is unclear whether they are still effective with reflection-contaminated text images.

3. PROPOSED METHOD

From text images with reflections, TIRe Network recovers clear text by focusing on the removal process mainly within reflection-contaminated regions through localizing and correct text by embedding comprehensive text-specific priors.

3.1. Network architecture

As shown in Figure 2, TIRe Network is composed of two parts, reflection localization and reflection removal networks

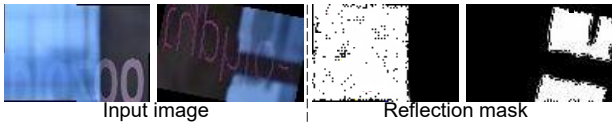

Fig. 3. The reflection masks localized by the reflection localization network.

with the guidance of textual feature prior, respectively. Except that the part for text prior is largely based on the existing text recognition network, the two networks share a similar mirror-like generative framework with the encoder contracting the feature channels step by step to capture the context information and the decoder to obtain the final results.

3.1.1. Reflection localization network

Due to the various regional transmission properties of the reflections, the contamination caused by reflection usually occupies a limited part of the whole image. The missing information in these regions cannot be well reconstructed by solely minimizing the global loss on the whole image, which may lead to unfaithful solutions such as blurry region, especially for the text images that have higher requirements for recognition reliability of the final estimated results. Previous work [5] has already shown the success of weighted regions scheme in reflection removal problems. It not only reveals where the removal process should focus, but also improves the representation of attention interests. We integrate such a mechanism into our framework by embedding the reflection localization network into the whole estimation process as follows:

$$\mathbf{M} = \mathcal{G}_l(\mathbf{I}), \tag{1}$$

where $\mathcal{G}_l$ and $\mathbf{M}$ denote the reflection localization network and the estimated reflection map, respectively. As shown in Figure 3, our reflection localization network can well locate the regions covered by reflections in most regions, and the sparse mis-classified pixels will not negatively affect the next step processing.

Adversarial loss. To roughly estimate the missing information covered by the reflections, we employ the Conditional Wassertein GAN [15] as follows:

$$\begin{aligned}\mathcal{L}^m_{\mathbf{adv}} = \min_{\mathcal{G}_l} \max_{D_l \in \mathcal{D}} \; & E_{\mathbf{M},\mathbf{M}^* \sim \mathbb{P}_r}[D_l(\mathbf{M}, \mathbf{M}^*)] \\ & - E_{\mathbf{M} \sim \mathbb{P}_r}[D_l(\mathbf{M}, \mathcal{G}_l(\mathbf{M}))],\end{aligned} \tag{2}$$

where D_l denotes the discriminator network, D is the set of 1-Lipschitz functions and $\mathbb{P}_r$ is the real data distributions. Our discriminator takes an input image with a size of 64 times 96 , composed of 6 strided convolutional layers followed by the ReLU activation function. In the last layer, we use the sigmoid function to generate the final classification result on data authenticity.

Besides the adversarial loss, we also adopt the $\mathcal{L}_2$ distance to encourage a straightforward but accurate reconstruction for the target mask domain. Then by combining the aforementioned terms, the loss functions for the reflection localization network are concluded as follows:

$$\mathcal{L}_{RLN} = \lambda_l \mathcal{L}_2(\mathbf{M}^*, \mathbf{M}) + \beta_l \mathcal{L}^m_{\mathbf{adv}}. \tag{3}$$

3.1.2. Reflection removal network

To make the reflection removal network focus on the regions covered by reflections, the reflection map $\mathbf{M}$ estimated in the reflection localization network and the original mixture image with reflection $\mathbf{I}$ are concatenated into a 6 channel tensor as the input to the reflection removal network as:

$$\mathbf{B} = \mathcal{G}_R([\mathbf{M}, \mathbf{I}]), \tag{4}$$

where $\mathcal{G}_R$ and $[\cdot, \cdot]$ denote the reflection removal network and the concatenation operation, respectively.

Local context loss. As we discussed before, the regional property of the reflections makes the global loss on the whole image less reliable during the text image recovery process. To address this problem, apart from the concatenation operation at the first layer of the reflection removal network, we also adopt the local context loss as follows:

$$\mathcal{L}_c = \|\mathbf{M} \odot (\mathbf{B} - \mathbf{B}^*)\|_1, \tag{5}$$

where $\odot$ is the element-wise product operation and $\mathbf{M}$ is the reflection map obtained from the reflection localization network. When the reflections occupy the whole image place, Equation (5) will degenerate into the common global loss.

Text consistency loss. Existing reflection removal methods usually aim at estimating the recovered images with higher PSNR and/or SSIM values by using different pixel-wise loss functions. However, the pixel-wise loss functions cannot capture the high frequency components of an image [16], which are closely related with the text edges. On the other hand, since we aim at estimating information for high level computer vision task, it is more reasonable to define the text similarity in a compact feature space rather than the image pixel space. To solve this issue, we embed the specific text priors into the estimation of the background layer by optimizing the whole network based on the feature level differences. This module further regularizes the generating process of reflection removal network only in the training stage, without additional computation cost in testing stage. We adopt the classical text recognition model as the the text prior network to extract the feature level differences as follows:

$$\mathcal{L}_{tct} = \sum_{c_1} ||\mathcal{F}_{c_1}(\mathbf{B}^\star) - \mathcal{F}_{c_1}(\mathbf{B})||_1 + \sum_{c_2} ||\mathcal{F}_{c_2}(\mathbf{B}^\star) - \mathcal{F}_{c_2}(\mathbf{B})||_1, \tag{6}$$

where $\mathcal{F}_{c_1}$ and $\mathcal{F}_{c_2}$ denote the c_1-th and c_2-th layer feature from the CRNN model [17].

Fig. 4. Samples in the training dataset.

We also adopt the adversarial loss similar to Equation (2) as follows:

$$\begin{aligned}\mathcal{L}^{r}_{\mathbf{adv}} = \min_{\mathcal{G}_r}\max_{D\in\mathcal{D}} E_{\mathbf{B},\mathbf{B}^*\sim\mathbb{P}_r}[D(\mathbf{B},\mathbf{B}^*)] \\ - E_{\mathbf{B}\sim\mathbb{P}_r}[D(\mathbf{B},\mathcal{G}_r(\mathbf{B}))].\end{aligned} \tag{7}$$

By combining the above terms in Equation (5), Equation (6), and Equation (7), the loss functions for the reflection removal network can be concluded as follows:

$$\begin{aligned}\mathcal{L}_{RRN} = \lambda_r\mathcal{L}_1(\mathbf{B},\mathbf{B}^*) + \beta_r\mathcal{L}_c(\mathbf{B},\mathbf{B}^*) \\ + \gamma_r\mathcal{L}^{r}_{\mathbf{adv}} + \theta_r\mathcal{L}_{tct}(\mathbf{B},\mathbf{B}^*),\end{aligned} \tag{8}$$

where $\mathcal{L}_1$ denote the classical pixe-wise loss function and λ_r, β_r, γ_r, and θ_r represent the weighting coefficients to balance different terms.

3.1.3. Overall loss functions

By combining $\mathcal{L}_{RLN}$ in Equation (3) and $\mathcal{L}_{RRN}$ in Equation (8), the overall loss functions for our TIRe Network becomes:

$$\mathcal{L} = \mathcal{L}_{RLN} + \mathcal{L}_{RRN}. \tag{9}$$

3.2. Implementation and training details

We have implemented our model using PyTorch. The complete training process of our network can be divided into two stages: 1) We train the reflection localization network and text prior network to convergence; 2) We then fix the text prior network and combine them with the reflection removal network, and the entire network is fine-tuned again, which grants more opportunities for these functional parts to cooperate accordingly. The learning rate for the whole network training is set to 5×10^{-5} for the first 40 epochs and then decreases to 5×10^{-6}.

3.3. Training dataset

Due to the data-driven nature of learning procedure, as is pointed out by previous works [4, 18], the building of a proper training dataset is of great significance to the reflection removal problem. To guarantee the diversity and training quantity of text images, we build a background text image collection integrating the existing in-the-wild IIIT 5K-word dataset [19] with 10K synthesized text images, as is shown in Figure 4. Specially, the synthesized images are generated by our own with sequences randomly selected from dictionary of commonly used words, blended into arbitrary background scenes cropped from the COCO dataset [20] with random skews. Following the synthesis procedure in [4], we capture 300 images with real-world reflections and acquire their corresponding regional masks with a threshold scheme. Then we obtain the final training dataset of 10k images by adding the reflection images to the background text images with various parameters, (example images from the training dataset are shown in Figure 4). As for the real data in testing, we capture 200 images containing text in complex real-world scenarios, where the corresponding reference images are not perfectly aligned. Then the testing dataset for both visual quality comparison and recognition experiment is built by integrating such real samples and synthetic samples with equal quantity.

4. EXPERIMENTS

We first compare the visual quality of our method and state-of-the-art reflection removal approaches, *e.g.*, WS16 [2], NR17 [8], CEILNet [3], and ZR18 [10]. Then, another experiment on the text recognition accuracy is conducted to investigate whether our proposed method can contribute to the high-level machine vision tasks. At last, we conduct an ablation study to verify the effectiveness of the reflection localization and text consistency loss in our network.

Visual quality comparison. We first show the examples of reflection-free text images recovered by our method and other four methods on real-world testing data in Figure 5, to compare their performance in terms of visual quality. As shown in Figure 5, the non-learning based methods WS16 [2] and NR17 [8] cannot effectively remove the reflection in images, with WS16 even further degrading the image sometimes (*e.g.*, the reflection region in the second row). The state-of-art data-driven method ZR18 [10] handles the reflections better but still cannot realize complete removal, and the CEILNet [3] also introduces artifacts into the text images, particularly, the deformed shape of letters caused by over smoothing and distortions (*e.g.*, the letter "m" in the fourth row). Only our method can better handle the regional reflections (*e.g.*, the first row) and recover text content with higher sharpness and contrast (*e.g.*, the last row).

Quantitative comparison. Since we mainly consider the image captured in real-world scenario and aim at removing reflections specifically to make texts in images more recognizable, the classical pixel-wise quantitative evaluation metrics for image quality (*e.g.*, PSNR and SSIM) are not suitable to comprehensively evaluate the performance of our text-preserving reflection removal. Thus, we adopt the text recognition accuracy of the estimated results to better investigate whether our method can improve the performance of high level text recognition tasks. As shown in Figure 6, for the

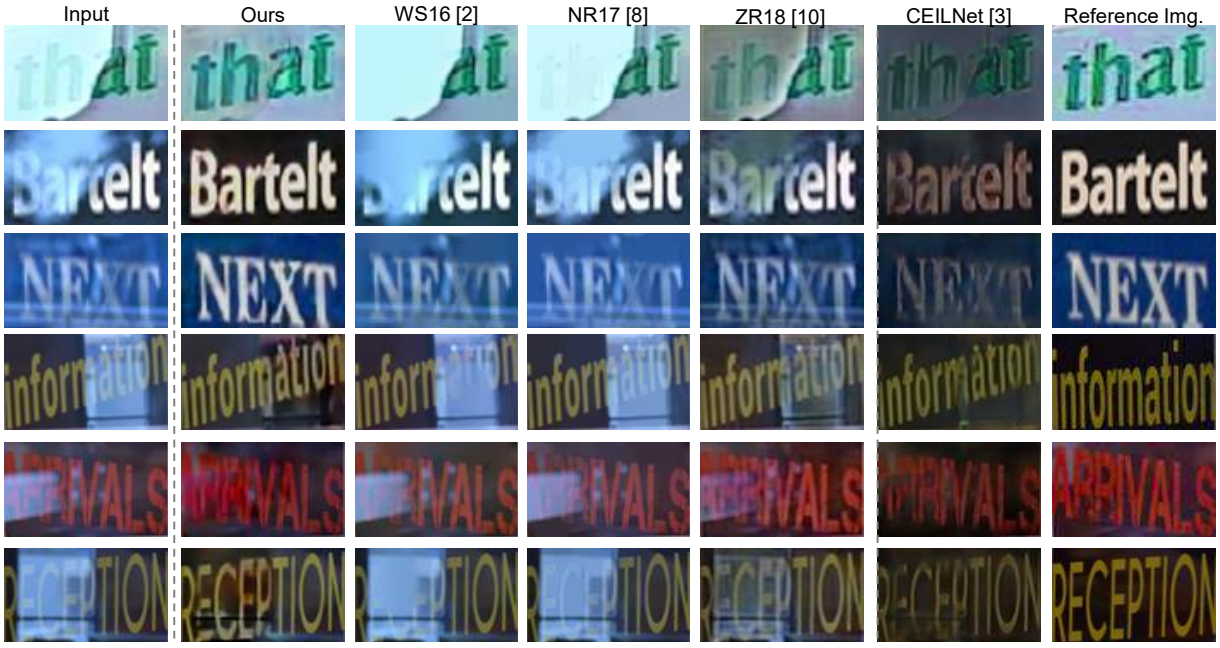

Fig. 5. Examples of reflection removal results on the evaluation dataset, compared with CEILNet [3], ZR18 [10], NR17 [8], and WS16 [2]. More results can be found in the supplementary materials.

Fig. 6. Images processed by our method with corresponding text recognition results underneath, compared with original text image with reflection, CEILNet [3] , and ZR18 [10].

word "REGENCY" in the second row, other methods cannot clearly recover the reflection covered letters "REG", which are mistaken for "FIN" or "ROM" in recognition, while our method can make more letters correctly recognized after reflection removal.

Table 1 shows the numerical result of text recognition by pre-trained CRNN [17] with no lexicon on our testing dataset, and the accuracy is calculated on word level rather than letter level. We can observe that the regional strong reflections cause great performance drop compared to the 78.2% accuracy on reflection-free IIIT-5k text image dataset. Then, our method also obtains a remarkable accuracy gain of 11.8% compared to the original inputs, higher than other general reflection removal methods, which can also be inferred from Figure 6. This gain in text recognition verifies that our method can better recover the reflection contaminated text images on the whole.

Table 1. The accuracy of text recognition on text images with reflection (TIR), reflection-removed images by the method of CEILNet [3], ZR18 [10] and ours.

	TIR	CEILNet [3]	ZR18 [10]	Ours
Accuracy	0.463	0.535	0.485	0.581

Table 2. The accuracy of text recognition on full model of TIRe Network, the model without reflection localization and model without text consistency loss.

	W/o local.	W/o text consist.	Full model
Accuracy	0.556	0.494	0.581

Ablation study. To validate the effectiveness of our proposed regional reflection localization scheme and text-specific feature-level consistency loss, we also implement ablation experiment on models without those functional modules. The quantitative results on recovered text recognition are shown in Table 2, the text consistency constraint shows more contribution than reflection localization in the improvement of recognition. Cases of estimated reflection-free images are shown in Figure 7, where we can tell that our specific designs enhance the performance of reflection removal.

5. CONCLUSIONS

We present a stacked framework to specifically solve the text image reflection removal problem. Different from existing methods which regards the mixture image as a whole, our framework can better handle the regional reflections by in-

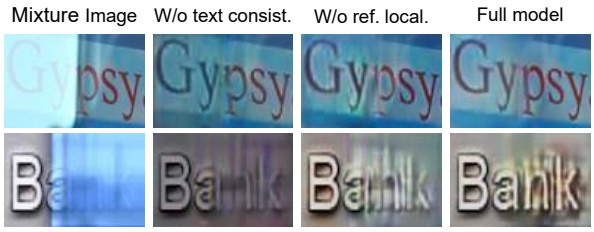

Fig. 7. Examples of reflection removal results by our full model compared with models without reflection localization or textual constraint.

Fig. 8. An example of challenging cases, where the color of text resembles the reflection color.

tegrating the reflection location into our framework. On the other hand, due to the embedding of the specific text priors, our framework can better keep the consistency of the text contents. Verified using the newly collected dataset of text images with reflections, our framework achieves better performances than existing methods for both reflection-free image recovery and text recognition.

Limitations. There still exist several extreme cases where our method achieved limited success in text recovery. Figure 8 illustrates a challenging example where characters with color hard to distinguish from the reflection. Our future work will attempt to deal with such cases and we also plan to extend to other types of scenes besides the text images.

Acknowledgement: This work was supported by the National Natural Science Foundation of China under Grant 61661146005, Grant U1611461, Grant 61872012, and in part by the National Research Foundation, Prime Ministers Office, Singapore, under the NRF-NSFC Grant NRF2016NRF-NSFC001-098.

6. REFERENCES

[1] Y. Li and M. S. Brown, "Single image layer separation using relative smoothness," in *Proc. CVPR*, 2014.

[2] R. Wan, B. Shi, A. H. Tan, and A. C. Kot, "Depth of field guided reflection removal," in *Proc. ICIP*, 2016.

[3] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf, "A generic deep architecture for single image reflection removal and image smoothing," *Proc. ICCV*, 2017.

[4] R. Wan, B. Shi, L. Duan, A. Tan, and Alex C Kot, "CRRN: Multi-scale guided concurrent reflection removal network," in *Proc. CVPR*, 2018.

[5] R. Wan, B. Shi, L. Duan, A. Tan, W. Gao, and Alex C Kot, "Region-aware reflection removal with unified content and gradient priors," *IEEE TIP*, 2018.

[6] A. Levin and Y. Weiss, "User assisted separation of reflections from a single image using a sparsity prior," *IEEE TPAMI*, 2007.

[7] R. Wan, B. Shi, A. Tan, and A. C. Kot, "Sparsity based reflection removal using external patch search," in *Proc. ICME*, 2017.

[8] N. Arvanitopoulos, R. Achanta., and S.Ssstrunk, "Single image reflection suppression," in *Proc. CVPR*, 2017.

[9] P. Chandramouli, M. Noroozi, and P. Favaro, "Convnet-based depth estimation, reflection separation and deblurring of plenoptic images," in *Proc. ACCV*, 2016.

[10] X. Zhang, Ng Ren, and Q. Chen, "Single image reflection separation with perceptual losses," in *Proc. CVPR*, 2018.

[11] J. Pan, Z. Hu, Z. Su, and M. Yang, "Deblurring face images with exemplars," in *Proc. ECCV*, 2014.

[12] H. Cho, J. Wang, and S. Lee, "Text image deblurring using text-specific properties," in *Proc. ECCV*, 2012.

[13] S. Bako, S. Darabi, E. Shechtman, J. Wang, K. Sunkavalli, and P. Sen, "Removing shadows from images of documents," in *Proc. ACCV*, 2016.

[14] N. Kligler, S. Katz, and A. Tal, "Document enhancement using visibility detection," in *Proc. CVPR*, 2018.

[15] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C Courville, "Improved training of wasserstein gans," in *Proc. NIPS*, 2017.

[16] P. Isola, J. Zhu, T. Zhou, and A. A Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, 2017.

[17] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE TPAMI*, 2017.

[18] P. Wieschollek, O. Gallo, J. Gu, and J. Kautz, "Separating reflection and transmission images in the wild," *arXiv preprint arXiv:1712.02099*, 2017.

[19] A. Mishra, K. Alahari, and C. V. Jawahar, "Scene text recognition using higher order language priors," in *BMVC*, 2012.

[20] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, P. Ramanan, D.and Dollár, and C L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. ECCV*, 2014.