Supplementary Material:
Learning to Remove Wrinkled Transparent Film with Polarized Prior

Jiaqi Tang[1,2,3] Ruizheng Wu[4] Xiaogang Xu[5,6] Sixing Hu[4] Ying-Cong Chen[1,2,3]*
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology [3]HKUST(GZ) – SmartMore Joint Lab
[4]SmartMore Corporation [5]The Chinese University of Hong Kong [6]Zhejiang University
jtang092@connect.hkust-gz.edu.cn, {ruizheng.wu, david.hu}@smartmore.com
xiaogangxu00@gmail.com, yingcongchen@ust.hk

In this supplementary material, we provide further clarification regarding whether this paper is the first work presented as film removal. Besides, we elaborate on some details of our network structure, including the angle estimation network (A-Net) and the reconstruction network (R-Net). In the experiment, we further explain the role of k-fold cross-validation and prove the effectiveness of the Prior. In addition, we show more visual comparisons between our method and baselines, including image reconstruction and downstream industrial tasks (Text OCR and QR code reading). Finally, we discuss the limitations of our model from three different perspectives and suggest possible directions for future research.

1. Is this "First" Paper for Film Removal?

Wang et al. [2] designed a hardware-based method to remove the "specular reflection" from the film. In contrast, the target of our proposed "Film Removal" task goes beyond that by simultaneously eliminating not only the specular reflection I_h but also addressing other degradations present in the original diffuse reflection ($I_{md} - I_m$), as depicted in Fig. 2 (B)(C) in the main paper.

It is significant to cope with the degradation besides I_h, since it has a negative impact on the downstream recognition (comparing results between ours and Polar-HR [4] / SHIQ [1] in Figs. 9 and 10 of the main paper, which only deal with the specular reflection).

2. Details of Lighting Environment

Drawing from our prior experience with industrial pipelines, we employed a single constant, ring-shaped industrial light source. The industrial assembly line is situated within a darkroom, which effectively minimizes the influence of ambient light. Industrial lighting is non-polarized, and it generates a polarization effect when it passes through optically active materials, such as wrinkled plastic films.

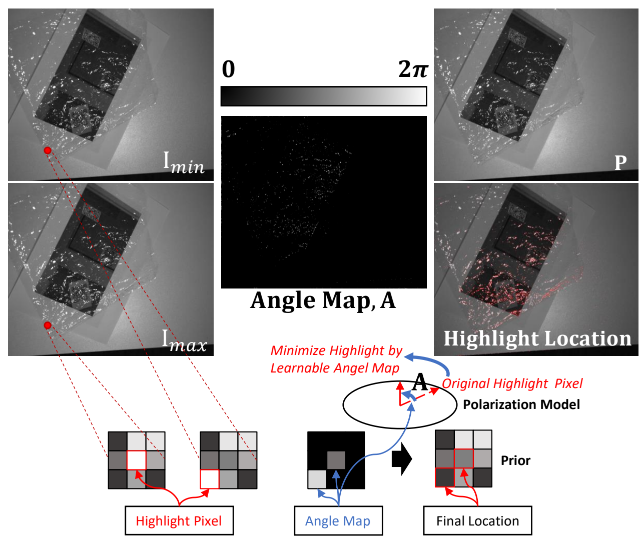


Figure 1. Explanation and Visualization of Angle Map, **A**. (Please zoom in to see the details.)

3. Explanation of Angle Map, A.

As shown in Figure 1, the Angle Map is utilized to adjust the polarized angle for each pixel. Thus, we learn the angle map by A-Net from four polarization images and then input its values into the polarization model (Eq. 9, Line 512) to obtain a prior P. Although there is no direct supervision, the optimization of the parameters in A-Net is constraint by I_{max}, I_{min} and the pre-defined polarization-based location model (Eq. 9), so the final loss will optimize the parameter of A-Net on the space of this pre-defined model, so as to only locate highlights accordingly. Furthermore, we visualize one example of I_{max}, I_{min}, Angle Map, Prior and final

*Corresponding author.

Layer Type	Group	Activation	Kernel	Stride	Padding	Output Size
Input Feature	-	-	-	-	-	$H \times W \times 6$
Convolution	1	-	3	1	1	$H \times W \times 64$
Convolution	1	-	3	1	1	$H \times W \times 64$
Lightweight RDB	64	ReLU	3	1	1	$H \times W \times 128$
Lightweight RDB	64	ReLU	3	1	1	$H \times W \times 192$
Lightweight RDB	64	ReLU	3	1	1	$H \times W \times 256$
Convolution	1	ReLU	1	1	1	$H \times W \times 64$
Lightweight RDB	64	ReLU	3	1	1	$H \times W \times 128$
Lightweight RDB	64	ReLU	3	1	1	$H \times W \times 192$
Lightweight RDB	64	ReLU	3	1	1	$H \times W \times 256$
Convolution	1	ReLU	1	1	1	$H \times W \times 64$
Lightweight RDB	64	ReLU	3	1	1	$H \times W \times 128$
Lightweight RDB	64	ReLU	3	1	1	$H \times W \times 192$
Lightweight RDB	64	ReLU	3	1	1	$H \times W \times 256$
Convolution	1	ReLU	1	1	1	$H \times W \times 64$
Concatenation	-	-	-	-	-	$H \times W \times 192$
Convolution	1	-	3	1	1	$H \times W \times 64$
Convolution	1	-	3	1	1	$H \times W \times 64$
Convolution	1	-	3	1	1	$H \times W \times 1$

Table 1. Architecture of A-Net. The Lightweight RDB is formulated as $y = x \oplus ReLU(f(x))$, where $f(\cdot)$ is a convolutional layer grouped by 64 and $ReLU(\cdot)$ is the ReLU activation function, and $\oplus$ is the concatenation operator.

Layer Type	Group	Activation	Kernel	Stride	Padding	Output Size
Input Feature	-	-	-	-	-	$H \times W \times 5$
Convolution	1	-	3	1	1	$H \times W \times 64$
Convolution	1	-	3	1	1	$H \times W \times 64$
RDB	1	ReLU	3	1	1	$H \times W \times 128$
RDB	1	ReLU	3	1	1	$H \times W \times 192$
RDB	1	ReLU	3	1	1	$H \times W \times 256$
Convolution	1	ReLU	1	1	1	$H \times W \times 64$
RDB	1	ReLU	3	1	1	$H \times W \times 128$
RDB	1	ReLU	3	1	1	$H \times W \times 192$
RDB	1	ReLU	3	1	1	$H \times W \times 256$
Convolution	1	ReLU	1	1	1	$H \times W \times 64$
RDB	1	ReLU	3	1	1	$H \times W \times 128$
RDB	1	ReLU	3	1	1	$H \times W \times 192$
RDB	1	ReLU	3	1	1	$H \times W \times 256$
Convolution	1	ReLU	1	1	1	$H \times W \times 64$
Concatenation	-	-	-	-	-	$H \times W \times 192$
Convolution	1	-	3	1	1	$H \times W \times 64$
Convolution	1	-	3	1	1	$H \times W \times 64$
Convolution	1	-	3	1	1	$H \times W \times 64$
Convolution	1	-	3	1	1	$H \times W \times 1$

Table 2. Architecture of the R-Net. The RDB is formulated as $y = x \oplus ReLU(f(x))$, where $f(\cdot)$ is a convolutional layer and $ReLU(\cdot)$ is the ReLU activation function, and $\oplus$ is the concatenation operator.

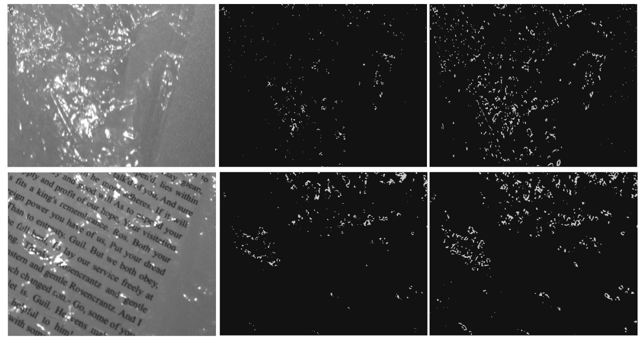
highlight location in Figure 1, showing that Angle Map can localize highlights by adjusting the polarization Angle on each pixel.

4. Details of Network Implementation

There are two learning-based modules in our framework, including the angle estimation network and the reconstruction network. The details of these two parts will be described separately in the following sections.

4.1. Angle Estimation Network

To accurately estimate the optimal polarized angle for each pixel, we adopt the backbone of Residual Dense Network (RDN) [6] for formulating A-Net, which can extract and merge helpful information from different layers. Moreover, we find that A-Net can be built as a lightweight module with grouped convolutions. This is because our proposed estimation strategy (predicting the optimal angle with the input of polarized images, AoP, and DoP) significantly reduces the difficulty of prediction. Table 1 shows the detail of A-Net.



Input Polar-HR (Location) Ours (Location)

Figure 2. Comparison with Polar-HR to locate highlights. Our predictions can include more overlap in highlight regions of input.

4.2. Reconstruction Network

The target of R-Net is to reconstruct the image without reflective film. We also employ the backbone of RDN to implement R-Net, whose details are shown in Table 2.

5. K-fold Cross Validation and Overfitting.

The diversity of industrial scenarios in our dataset guarantees the difference between the training and testing split when conducting K-fold cross-validations. Also, we take a manual check to ensure that each split is varied from the other. Moreover, K-fold cross-validation can avoid the limitations of the single split setting. Considering the significant difference between the training and testing set, as well as the SOTA performance on testing benchmarks, we can demonstrate the excellent generalization capacity of our model.

6. Effectiveness of *P*

The objective of P is to provide the location and amplitude priors for highlights I_h. As shown in Fig. 2, compared with one of the SOTA methods in Polarization-based Highlight Location, our prior has more overlap with the highlight regions of the input image. This is because our strategy is optimized according to the polarization modeling strategy, which has better pertinency for the "Film Removal" task with polarized inputs. Meanwhile, we should also note the efficiency of our lightweight A-Net and PLM to obtain P.

7. More Visual Results

More qualitative comparison results (Fig. 3) demonstrate the superiority of our framework over existing baselines to complete FR in terms of the generalization ability in various scenarios. Furthermore, we provide more experimental results (Fig. 4 & 5 and Fig. 6 & 7) to demonstrate that our framework is applicable to industrial downstream tasks by thoroughly removing the reflective films.

8. Limitations

Specular Highlight in Industrial Material Our model assumes that the material surface does not produce the specular highlight. However, in industrial scenarios, specular highlights may occur, which cannot be accurately described by our physics model.

Diffuse Reflection in Wrinkled Film We assume that the film is completely transparent, but in reality, it still has diffuse reflection. Although the reconstruction network can separate these degradations through end-to-end learning, our model cannot fully capture this situation.

Saturated Regions If structured features are not totally hidden in saturated areas (e.g., Figs. 4 and 5 in supp.), our network can reconstruct the original information; if the region is over-saturated, the polarized signal cannot help to obtain accurate structural information for reconstruction, and some failures could be caused. Dealing with these challenging cases will be our future work.

Generalization to Other Datasets This work focuses on industrial environments where lighting conditions are assumed to be fixed. Nevertheless, we agree that studying the robustness of lightning would be very interesting, and this will be investigated in our future work.

9. Future Work

Our current work focuses on the task of eliminating film interference in a limited number of scenarios. In order to apply our approach to more industrial scenarios, it is necessary to obtain more reliable data that covers a broader range of real-world conditions. This is one of the potential directions for future research.

In addition, it should be noted that our study only addresses the specific case of wrinkled films and thus may not be suitable for scenarios involving other types of films, such as smoother films. Therefore, future research can explore the applicability and limitations of our approach in such scenarios and investigate the feasibility of modifying our method to account for different types of films. Such investigations could improve the versatility and practicality of our approach in various real-world situations.

References

[1] Gang Fu, Qing Zhang, Lei Zhu, Ping Li, and Chunxia Xiao. A multi-task network for joint specular highlight detection and removal. In *CVPR*, 2021. 1, 4, 5, 6

[2] Chengjie Wang and Sachiko Kamata. Removal of transparent plastic film specular reflection based on multi-light sources. *2012 Symposium on Photonics and Optoelectronics*, 2012. 1

[3] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, 2022. 4, 5, 6, 7

[4] Sijia Wen, Yinqiang Zheng, and Feng Lu. Polarization guided specular reflection separation. *IEEE TIP*, 2021. 1, 4, 5, 6, 7

[5] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 4, 5, 6, 8

[6] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE TPAMI*, 2020. 2

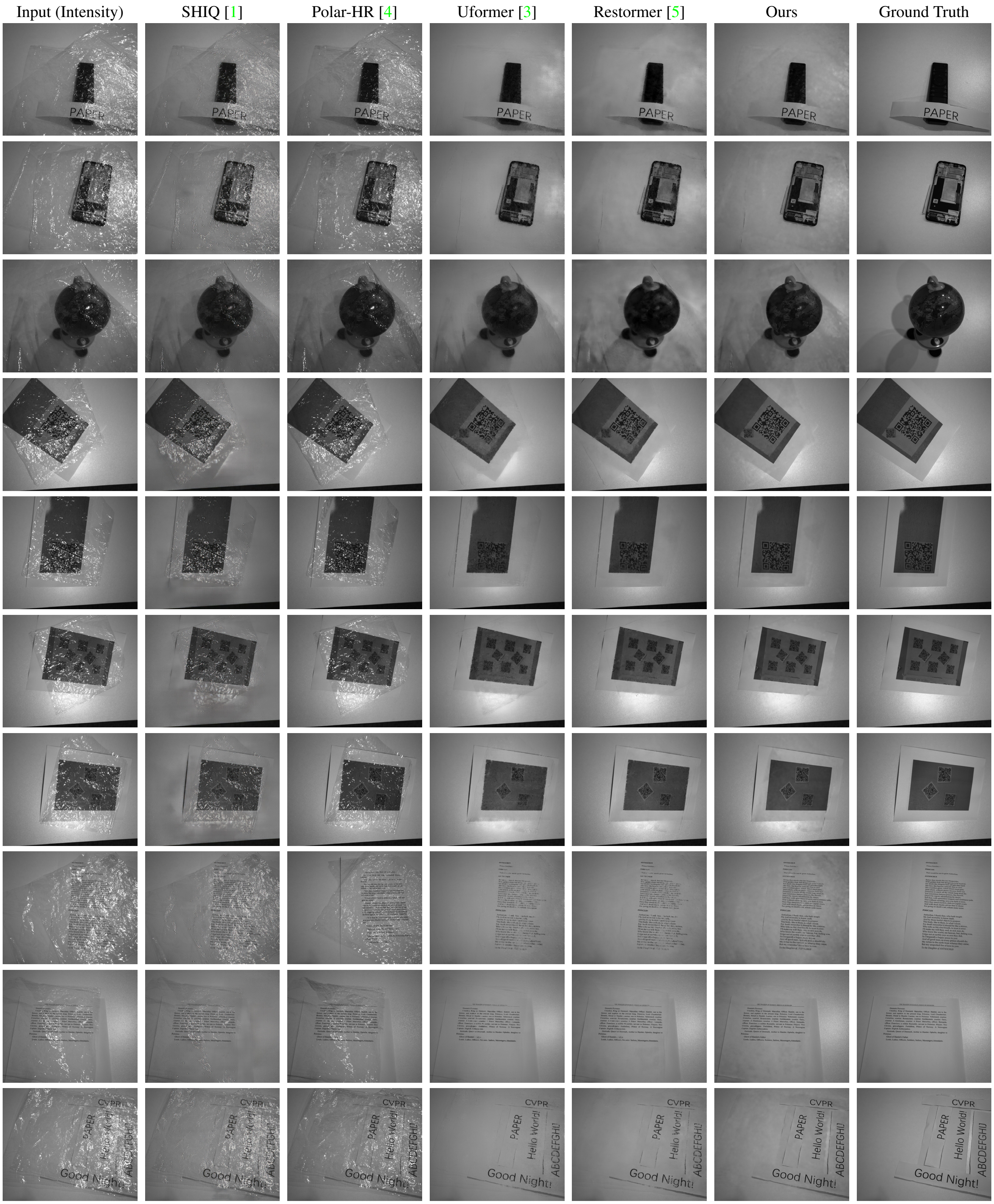


Figure 3. Qualitative Evaluation. Compared with other baselines, our model can reconstruct more realistic details in highlight regions instead of fake artifacts. Please zoom in for more details.

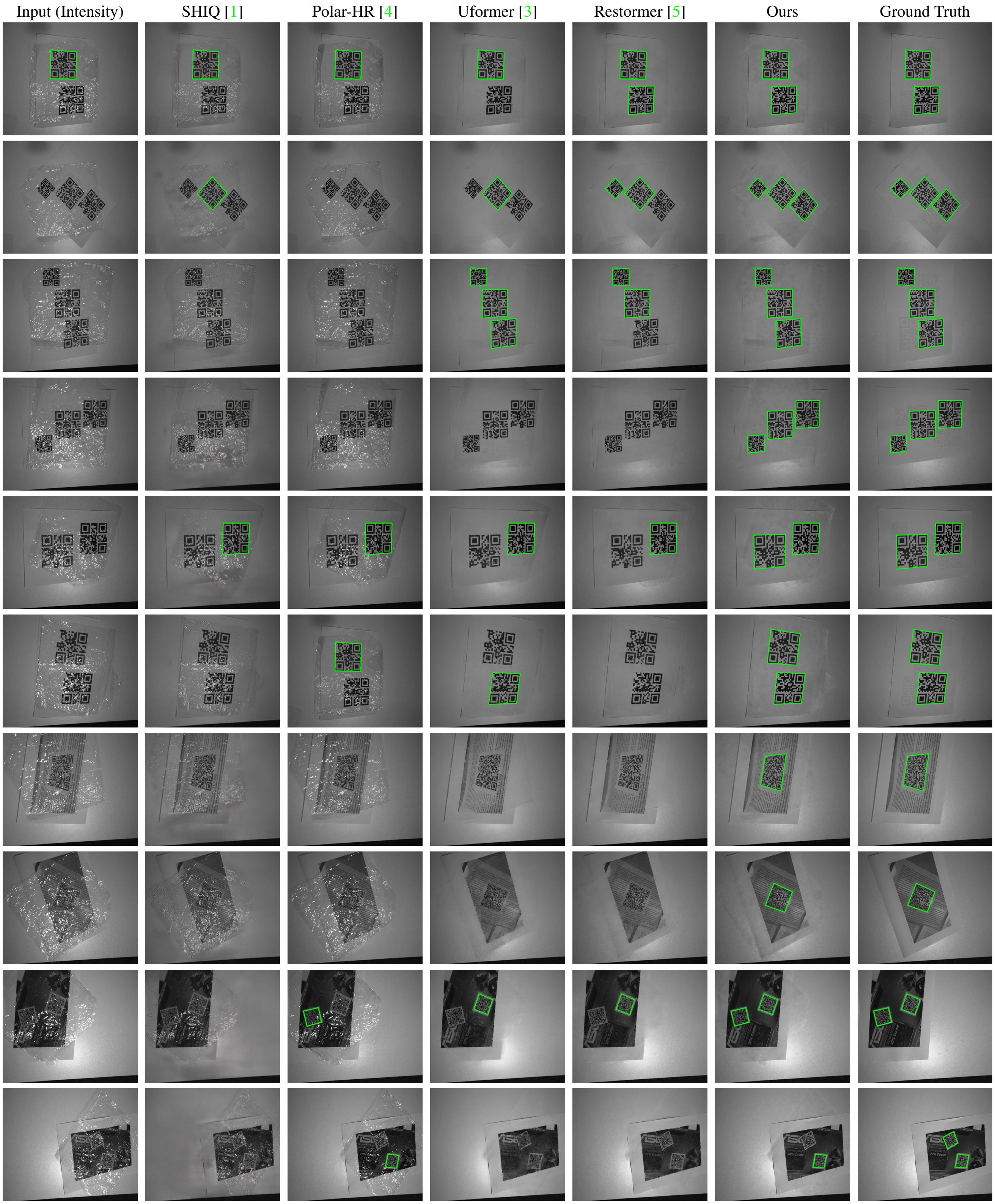


Figure 4. QR code reading (1). After the processing of our framework, the QR code reading accuracy can be improved markedly.

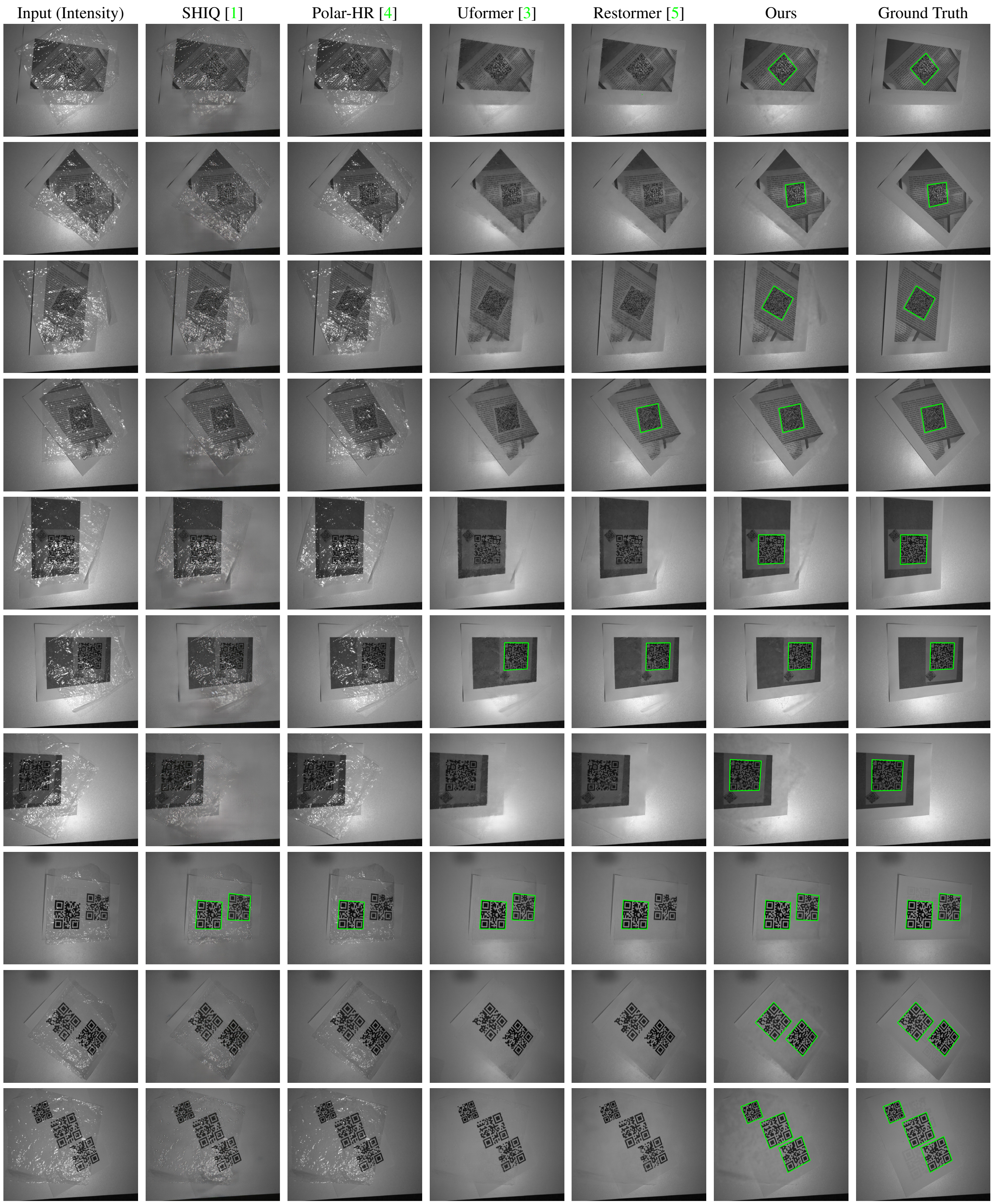
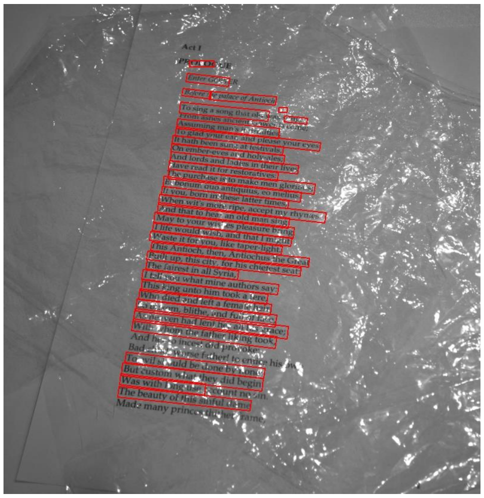


Figure 5. QR code reading (2). After the processing of our framework, the QR code reading accuracy can be improved markedly.

Input (Intensity)

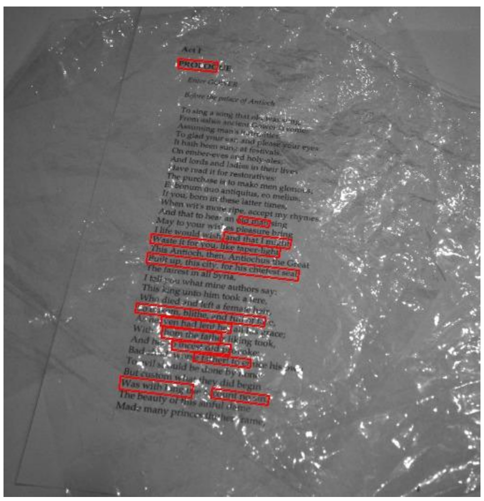


Q600
?
Enter Go
Befere tpe palace of Antioch
To sing a song that olg ?
?
Assuming man's ieities
To glad you rear, and please your eyes
It hath been sung at festivals
On ember-eves ard holy-ales
And lords and ladies in their lives
Have read it for restoratives
The purchase is to make men glorigss
E bonum quo antiquius, eo Melius
If you, born in these latter times
When wit's more ripe, accept my rhymies
And that to hear an old man sing
May to your wises pleasure bring

I life would wish, and that I might
Waste it for you, like taper-light
This Antioch, then, Antiochus he Great
Ruilt up, this city, for his chiefest seat
The fairest in all Syria
I tali ou what mine authors say
This king unto him took a jer
Who died and 'left a female heir
? soom, blithe, and fuif of fe
As neiven had lent he ai race
Wit. hom the father liking took
?
?
To vil siuld be done by ton
But custom what they did begin
Was with ing use ?
The beauty of this sinful dme
?

Reading Rate: 69.7% Error Words: 66

Polar-HR [4]

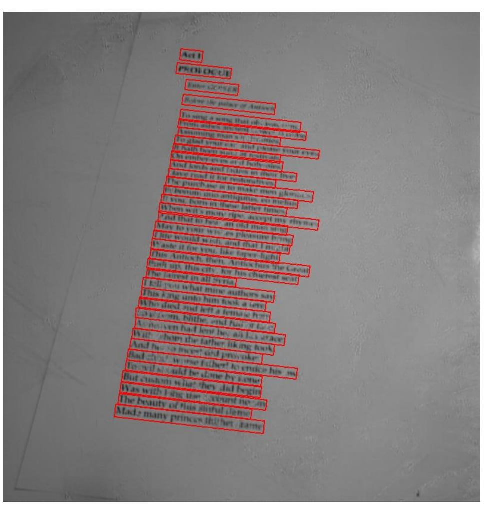


?
PROLOG
?
? Dd man ?
?

? and that I might
Waste it for you, like taper-light
?
Putft up, this city, for his chiefest seat
?
ot Lam, blithe, and fui of fs
Ven had lenthe
hom the fafhe
O incest did .pl
E frher ft oer
Was with Dngu
?
count no sin

Reading Rate: 13.3% Error Words: 189

Uformer [3]



Act I
PROLOGEL
?
?
To sing a sceng that Ins ?
Pveni ashen anct emeCha weJobe
Aasming mnshis
To glad yraur ear, ad pleaw you? eyes
N hath ben sng ar hustivals
On embep-eves atd holy-tles
And kords and fTadlos in theiv lives
Laverevhet it for rerstonwtives
The purchase in to makc men ghorocis
Pt benum guo antiquius, eo Meliaas
H vou, born in thee lather tmes
When wit's mne ripe, aaept my rhynes
Ad that to hear an old man sing
May to your wisies pleasure bring

I life wout wish, ad that I melbt
laste it for you, like faper-hight
Yhis Antioch, then, Antiochus the Great
Puih up, this city, for his chietest seat
The fairest in all Syria
I tli yrou what mine authors say
This kng unto him fook a ere
Who died and left a fema heiy
Eowlon bhthe, ond fur or lae
Au nesiven had len hey ahs upaee
With yhom the fathet liking took
Aad hes o inceet ard povoke
Bad ho wurse fixthert to encehism
To evil sixuld be dhone by one
But custom what chey did begin
Was with Iing toeCtuntner sin
The beauty of this snful deme
Mada many princes thither frame

Reading Rate: 49.5% Error Words: 110

Figure 6. Text OCR (1). The text OCR results with the processing of baselines. The text OCR results with the processing of our framework can be seen in Fig. 7.

Figure 7. Text OCR (2). The text OCR results with the processing of our framework. With our framework, the performance of OCR is enhanced observably (compare the reading rate and error words in Fig. 6 and Fig. 7).