Supplementary Material for ScannerNet: A Deep Network for Scanner-Quality Document Images under Complex Illumination

Chih-Jou Hsu[1]
sophia@cmlab.csie.ntu.edu.tw
Yu-Ting Wu[2]
yutingwu@mail.ntpu.edu.tw
Ming-Sui Lee[1]
mslee@csie.ntu.edu.tw
Yung-Yu Chuang[1]
cyy@csie.ntu.edu.tw

[1] National Taiwan University, Taipei, Taiwan
[2] National Taipei University, New Taipei City, Taiwan

In this document, we first provide the implementation details and hyperparameters in Section 1. Next, we present ablation studies for model design and dataset synthesis in Section 2. After that, we provide image comparisons of models trained with SDSRD [8] and our SSQD dataset in Section 3. Finally, Section 4 showcases more results and provides discussions.

1 Implementation details

For training the model, we use Adam optimizer [5] and set the parameters β for controlling the decay rates to (0.5, 0.999). The learning rate is set to 1×10^{-4}. The batch size is 4, and training runs 400 epochs. All images are resized to 256×256 using bicubic interpolation for training.

1.1 Architecture details of DeshadowNet

Our deshadow model is adapted from a sub-module of MPRNet [9], which employs an encoder-decoder architecture with channel attention mechanism. We denote the convolutional block with channel attention mechanism as Channel Attention Block (CAB). A CAB comprises a channel attention layer and two convolutional layers. The channel attention layer consists of an average pooling layer that helps to aggregate channel information of a feature map, followed by two convolutional layers and a sigmoid layer. The channel attention layer outputs an attention map that can refine features in the CAB. We use three CABs at every encoder-decoder level, with downsampling and upsampling layer in-between. There are skip connections between the encoder and decoder levels.

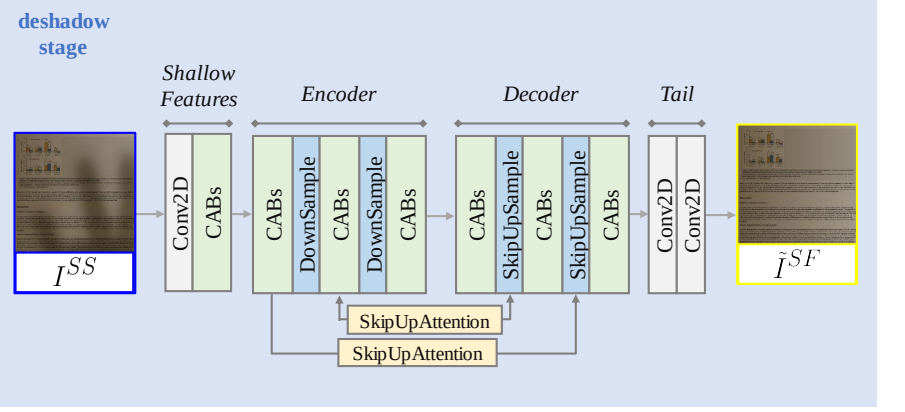


Figure 1: The network architecture of our DeshadowNet.

2 Ablation study

2.1 Network architecture

To validate the two-stage design, we train our model without two stages and remove the supervisory signals in the middle. We found that two stage design is more stable than single stage architecture. Figure 2 demonstrates the results of some real-world examples using our method with different network architecture designs. Although the DeshadowNet can remove shadows, unwanted shading effects remain in its results. Training our network as a one-stage monolithic network without the intermediate supervisory signal cannot generate satisfactory results. The two-stage design exploits the characteristics of shadows and shading separately, effectively reducing the interference between them. It validates the necessity of the two-stage design.

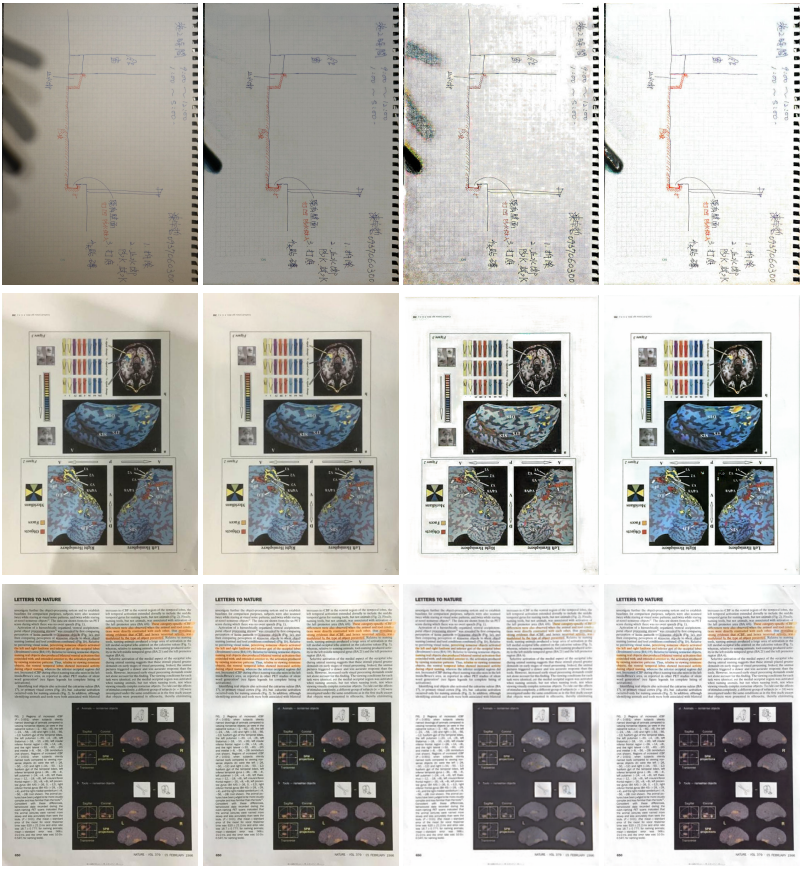
We also perform a quantitative evaluation using synthetic input images with ground truth. Although synthetic images give our method unfair advantages, it is fair to use them for ablation study. Table 1 further confirms the effectiveness of the two-stage design. We also validate the necessity of the components in the data synthesis process.

Method	SSIM ↑	PSNR ↑	RMSE ↓
single stage	0.87	25.65	19.43
ours	**0.93**	**28.11**	**11.35**

Table 1: Ablation study on model components.

2.2 Variations in training data

To look into the effects of different shadow variations in our synthesis process, we test different image synthesis settings. There are three sets. Set1 only contains small cast shadows, while Set2 adds large cast shadows. Set3 includes more shadow variations and color shifts by adding color modulation and noise. These datasets are used to train a simple ResNet-based generator for removing shadows.



input DeshadowNet one-stage two-stage

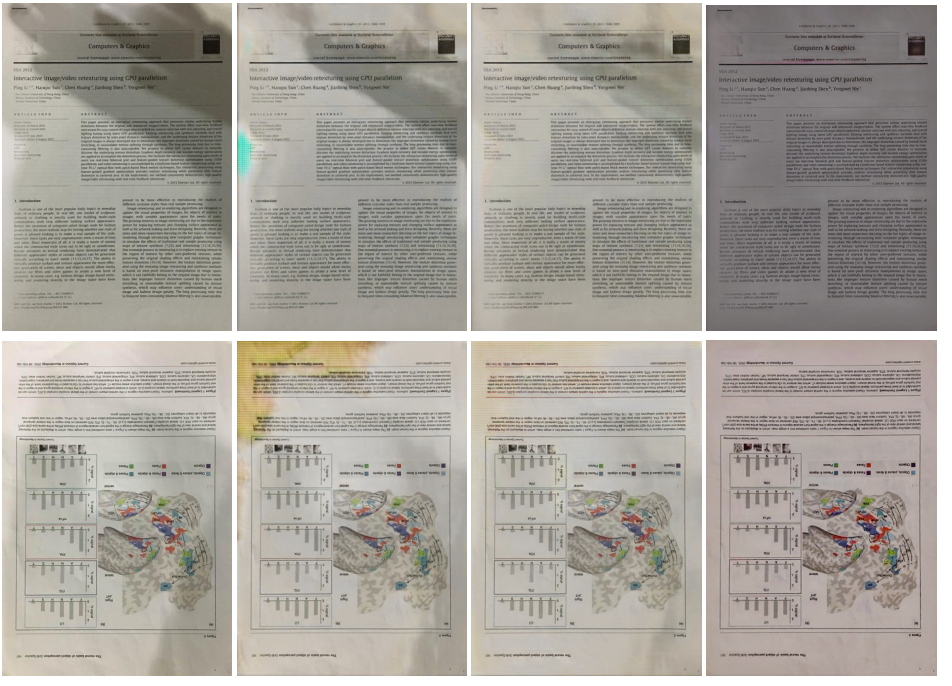
Figure 2: Ablation study using one-stage and two-stage models. The two-stage model can handles photometric correction better than the single-stage model, especially for cases with color figures and complex backgrounds.

Figure 3 shows estimated shadow-free images using the models trained on different datasets. We found that more shadow variations in the synthesized shadows lead to better shadow removal results, while they could degrade background color prediction performance. Simpler shadow masks could alleviate color shift a bit but at the expense of worse shadow removal. In our application, shadow removal is more important, and the background color has to be adjusted for better white balance eventually. Thus, more shadow variations lead to more visually pleasing results.

3 Dataset

3.1 Dataset comparison

Figure 4 shows the results of BEDSR-Net [8] and our DeshadowNet trained with SDSRD [8] and our SSQD. The top row is for BEDSR-Net, while the bottom for DeshadowNet. Owing



input Set1 Set2 Set3

Figure 3: Ablation study on data synthesis. It is essential to include a variety of variations in the dataset. Set1 only contains small shadows. Set2 adds large shadows. Set3 contains all variations in our synthesis process.

to the richer shadow variations in SSQD, the models trained with SSQD significantly outperform those trained with SDSRD for shadow removal. It also shows that our dataset is agnostic to deep models.

3.2 Links to data materials

We provide links to the images used in the synthesis process in this section. For scanned images, we use DSSE Layout Analysis Dataset , PRIMA Layout Analysis Dataset, DocUNet and Li et al.'s Dataset. For silhouette masks used to generate shadow maps, we collected 1,400 silhouette masks from the Binary Shape Database.

4 Results.

4.1 Comparison with shading rectification methods.

Some recent data-driven approaches aim to correct illumination for document images [2, 3, 7]. However, they only deal with shading and do not consider shadows cast by occluders. Figure 5 compares with some of them. With the help of the encoded global information, our method removes shadows and shading more successfully with better content color preservation.

Figure 4: Qualitative comparison of the SDSRD dataset [8] using 3D synthesis and our SSQD dataset with 2D synthesis for shadow removal. The models trained with our SSQD dataset obtains superior shadow removal results compared to those trained with SDSRD, owing to the richer shadow variations offered by 2D synthesis.

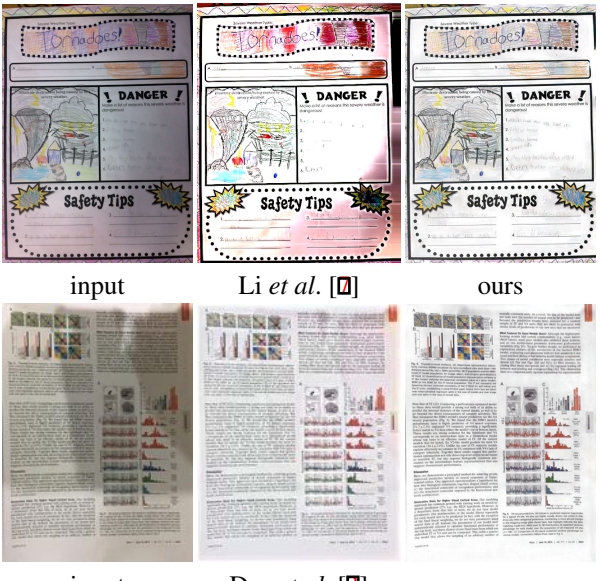


Figure 5: Qualitative comparisons with previous data-driven methods for illumination correction [3, 7]. Our method can remove dark shadows better than previous methods while preserving the content and colors.

4.2 More results

Figure 6 shows more results and comparisons with previous methods: Bako [1], Jung [4], Kligler [6], BEDSR-Net [8] and BEDSR-Scan, BEDSR-Net trained with our Synthetic Scanner-Quality Dataset (SSQD). The test cases contain various document and shadow types, including documents with (a) complex background, (b) large shadows, (c) complex figures, (d) dark shadows, and (e) (f) (g) multi-cast shadows. Conventional methods often have difficulties with removing shadows completely. Jung performs better than Bako and Klinger, but still suffers from residual shadows due to dark shadows in Figure 6(d). Deep-learning-based methods such as BEDSR-Net generally perform a better job on shadow removal than conventional methods. With the help of SSQD, BEDSR-Scan outperforms BEDSR-Net, showing the effectiveness of the SSQD dataset. In all cases, our method demonstrates the best performance among all competitors. Compared to BEDSR-Scan, our model removes shadows more completely (Figure 6(a)) and preserves the document contents better, particularly figures (Figure 6(c)) and color (Figure 6(e)).

4.3 Execution time

The average inference rates over 5 runs for our model and BEDSR-Net on the same machine are 50.5 and 60.4 images per second, respectively.

References

[1] Steve Bako, Soheil Darabi, Eli Shechtman, Jue Wang, Kalyan Sunkavalli, and Pradeep Sen. Removing shadows from images of documents. In *Proc. Asian Conference on Computer Vision (ACCV)*, pages 173–183, 2016.

[2] Sagnik Das, Ke Ma, Zhixin Shu, Dimitris Samaras, and Roy Shilkrot. DewarpNet: Single-image document unwarping with stacked 3D and 2D regression networks. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 131–140, 2019.

[3] Sagnik Das, Hassan Ahmed Sial, Ke Ma, Ramon Baldrich, Maria Vanrell, and Dimitris Samaras. Intrinsic decomposition of document images in-the-wild. In *Proc. the British Machine Vision Conference (BMVC)*, 2020.

[4] Seungjun Jung, Muhammad Abul Hasan, and Changick Kim. Water-filling: An efficient algorithm for digitized document shadow removal. In *Proc. Asian Conference on Computer Vision (ACCV)*, pages 398–414, 2018.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[6] N. Kligler, S. Katz, and A. Tal. Document enhancement using visibility detection. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2374–2382, 2018.

[7] Xiaoyu Li, Bo Zhang, Jing Liao, and Pedro V. Sander. Document rectification and illumination correction using a patch-based cnn. *ACM Transactions on Graphics (TOG)*, 11 2019.

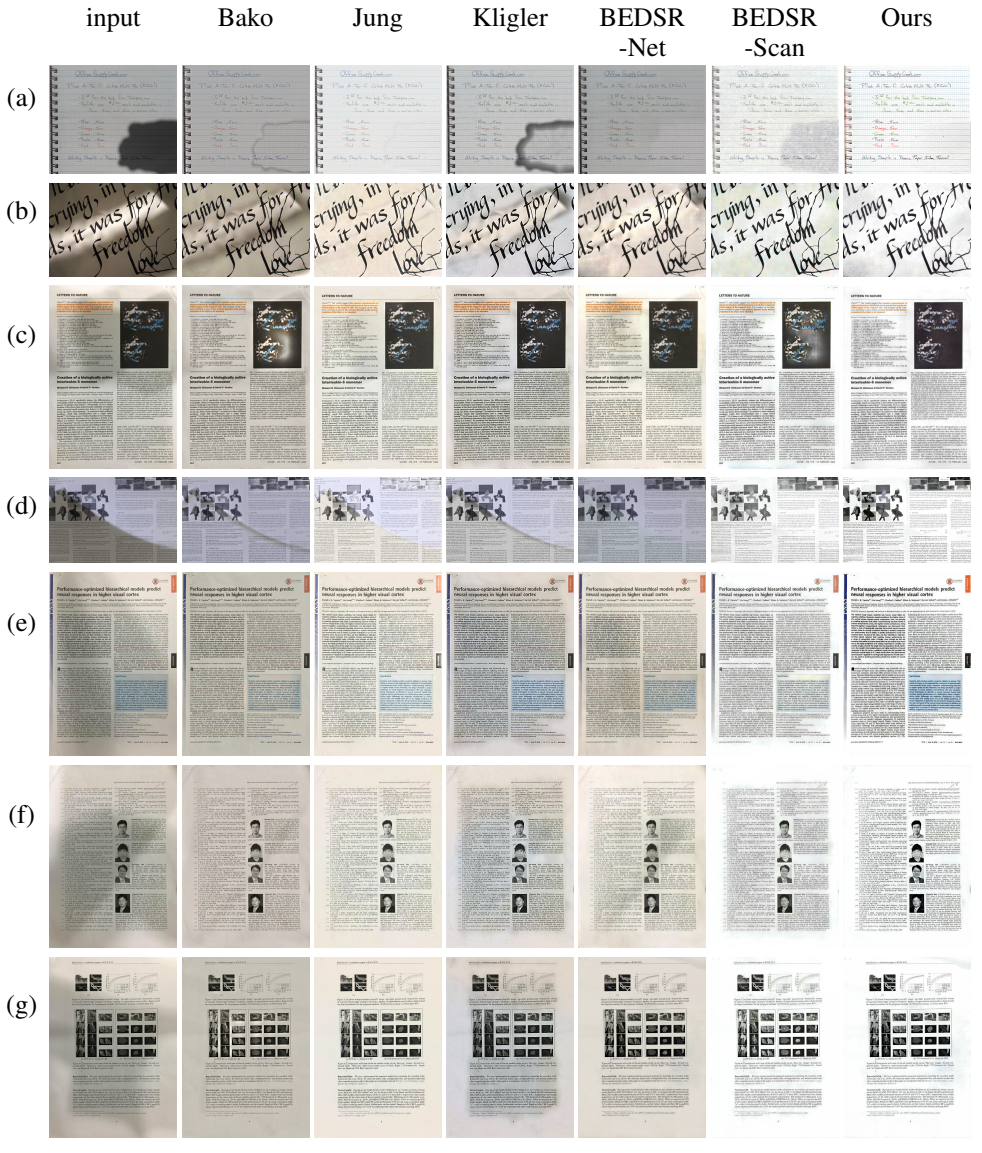


Figure 6: Qualitative comparisons with previous methods on various shadow and document types: (a) complex background, (b) large shadows, (c) complex figures, (d) dark shadows, (e) (f) (g) multi-cast shadows with various document types.

[8] Yun-Hsuan Lin, Wen-Chin Chen, and Yung-Yu Chuang. BEDSR-Net: A deep shadow removal network from a single document image. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[9] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14821–14831, 2021.