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

Abstract

Document images captured by smartphones and digital cameras are often subject to photometric distortions, including shadows, non-uniform shading, and color shift due to the imperfect white balance of sensors. Readers are confused by an indistinguishable background and content, which significantly reduces legibility and visual quality. Despite the fact that real photographs often contain a mixture of these distortions, the majority of existing approaches to document illumination correction concentrate on only a small subset of these distortions. This paper presents ScannerNet, a comprehensive method that can eliminate complex photometric distortions using deep learning. In order to exploit the different characteristics of shadow and shading, our model consists of a sub-network for shadow removal followed by a sub-network for shading correction. To train our model, we also devise a data synthesis method to efficiently construct a large-scale document dataset with a great deal of variation. Our extensive experiments demonstrate that our method significantly enhances visual quality by removing shadows and shading, preserving figure colors, and improving legibility.

1 Introduction

A rapid improvement in mobile phones and their built-in cameras has made the process of digitizing documents much easier than in the past. Rather than relying on bulky equipment such as scanners, people can take photos of documents with their mobile phones in order to obtain digital copies. The visual quality of document images captured by phone cameras, despite the advantage of ready accessibility and ease of manipulation, is usually inferior to images captured by scanners due to uncontrolled lighting conditions during the photographing process. The captured document images may be affected by dark shadows caused by the occlusion of key lights, uneven shading due to uneven ambient illumination, and surface

Figure 1: **Photometric distortion correction.** Our method is effective for correcting complex photometric distortions in document images. The top row shows the images captured by cameras, and the bottom row shows the enhanced images produced using the proposed method, which corrects shadows, shading, and color shift simultaneously.

normal disturbances due to paper bends or folds, or an overall color shift due to the camera's insufficient white balance.

In addition to affecting the documents' visual quality, photometric distortion also reduces their legibility and impairs the quality of subsequent processes, such as content analysis or optical character recognition (OCR). The top row of Figure 1 gives examples of document images captured by real cameras. There are often multiple photometric distortions in each image, including shadows, non-uniform shading, and color shifts. Most methods only focus on addressing one type of distortion, such as removing shadows [1, 12, 13, 17] or correcting non-smooth shading due to paper folds and bends [16, 19, 24]. Addressing only one distortion may result in a suboptimal solution. If shadows are removed without considering shading, the background color may be inaccurately estimated and residual shadows remain; if shading is corrected without removing shadows first, the shading map may be misjudged.

Our goal is to obtain document images with scanner quality by correcting photometric distortions. We observe that shadows and shading exhibit different characteristics. Shadows tend to produce clear silhouettes and transitions, while shading is often characterized by smoother transitions and color shifts. Thus, we propose *ScannerNet*, a two-stage network that removes shadows and corrects shading at different stages. In order to provide supervisory signals, we propose a flexible and effective scheme for synthesizing triplets of training images. Experiments demonstrate that, despite being trained with synthetic images, the learned model generalizes very well to real-world images. Our main contributions include:

- We design a fast, flexible, and effective data synthesis process for generating synthetic training data. This process has a low computational cost and can be controlled easily in terms of the types and variations of the photometric distortions produced. Therefore, it is much easier to enrich the dataset with a wide variety of illumination distortions.
- With the help of synthetic data, we propose a lightweight and effective deep network for removing shadows, uneven shading, and overall color shift on document images.
- Experiments show that our method produces scanner-quality document images with improved shadows and shading removal performance, color preservation of figures, and improved perceptual quality and readability in terms of OCR.

2 Related Work

Document shading rectification. Many algorithms focus on removing the shading of the spine region of a scanned book [15, 19, 23, 25, 26]. In order to reconstruct the surface geometry and its illumination, one common strategy is to use the idea of shape from shading [23, 25, 26]. Because these methods rely on the prior geometry setting between a scanner and a book, they cannot be applied to camera-based document images. Other methods, such as Lee *et al*. [15], detect foreground objects by fusing multiple edge maps and computing a shading map by interpolating background colors. Meng *et al*. [19] define a convex hull based on the background color and use it to estimate shading of the document.

Some methods aim to correct the shading of documents due to paper deformation [3, 4, 14, 16, 24, 31]. Using a watershed transform, Fan [9] corrects the uneven shading and color shift in digitized books by segmenting background regions. Zhang *et al*. [32, 33] correct the illumination for camera-based document images using the notion of intrinsic image. The performance of their method is dependent on the accuracy of the inpainting mask derived from edge detection and morphological operators.

Recently, data-driven approaches [6, 7, 16] have been explored to enhance document images. Both Li *et al*. [16] and Das *et al*. [6] address the problem of geometric and illumination correction of document images. Their methods first estimate the 3D distortion or shape by means of a neural network in order to rectify the geometry of documents. A second network is then used to correct the illumination. Focusing on illumination correction, Das *et al*. [7] estimate per-pixel white-balance kernels and remove shading using intrinsic decomposition. Due to the fact that their shading synthesis processes do not take into account occluders, these methods are less effective for documents containing dark shadow boundaries.

Document shadow removal. Some approaches can remove the dark shadows of document images. Jung *et al*. [12] construct a topographic surface using pixel luminance and estimate the shading artifacts on the documents by simulating the immersion process using diffusion equations. Their method tends to produce results with an overall color shift. Kligler *et al*. [13] represent image pixels as a 3D point cloud and proposed a visibility test algorithm for detecting shadows and stains in document images. The generated visibility map can improve the results of other algorithms, including document binarization and shadow removal. There are, however, often residual shadow edges observed in their results. Bako *et al*. [1] estimate the shadow map of a document image by computing the ratio of local and global background colors. The shadow map is then applied to the original document for removing shadows. To improve the robustness of background estimation, Lin *et al*. [17] recently proposed a data-driven approach called BEDSR-Net. Their approach shows a significant improvement in robustness over previous heuristic approaches; however, the proposed 3D shadow synthesis process for training data creation is not only time-consuming but also limited in terms of variation. We propose a simpler and more extendable procedure for generating synthetic data that is effective for a wide range of shadows as compared with their work.

Document images with the scanner-quality. Data-driven methods have recently been proposed to produce scanner-quality document images [2, 8]. In Bogdan *et al*.'s method [2], the training set is constructed based on the results of a previous method [10] as the ground truth. In order to prevent adding poor enhancements as ground truth, the authors had to manually select images. In addition, the learned model would be limited by the capabilities of the enhancement method used to create the ground truth. This problem can be addressed by our data synthesis method.

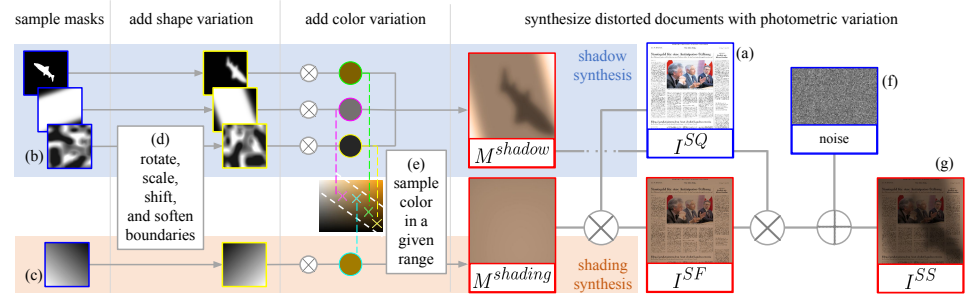

Figure 2: **The proposed data synthesis process.** Starting with some shadow maps created using silhouette maps, large-scale soft maps, and Perlin shadow map, we transform shapes and modulate colors in order to obtain a simulated shadow map M^{shadow}. We obtain the shading map $M^{shading}$ by following a similar procedure. A shadow-free image I^{SF} can be obtained by multiplying $M^{shading}$ and a scanned image I^{SQ}. By multiplying with M^{shadow} and adding noise, we obtain a simulated image I^{SS} with both shadow and shading.

3 Data synthesis

The training of deep networks requires sufficient data. Unfortunately, it is challenging for our applications to collect an adequate number of image pairs, each of which contains two images of the same document: the first is taken by a camera under complex lighting, and the second is scanned with a scanner. It is very time-consuming to collect such image pairs and introduce a wide variety of photometric distortions during the photographing process. It is also difficult to align the captured and scanned images accurately. Thus, existing datasets of real image pairs often contain only a small set of images with limited photometric distortions, such as the datasets collected by Bako *et al*. (81 pairs) [1], Kligler *et al*. (300 pairs) [13] and Jung *et al*. (87 pairs) [12], that only contain shadows.

A number of previous studies have demonstrated that photometric distortions of document images can be effectively simulated using image synthesis [6, 16, 17]. They also demonstrate that deep networks trained on synthetic data are capable of generalizing well to real-world images. However, they use 3D rendering to synthesize images. Although the synthesis can be more physically accurate, it is more time-consuming to synthesize and more difficult to generate images with specific photometric distortions. In addition, they don't often have complicated shadows and shading since it's hard to set up 3D scenes. Inspired by portrait shadow manipulation [34], we design a 2D image synthesis method for simulating complex shadows, shading, and color shift. In comparison with 3D rendering, our process is more flexible and efficient. The process is easier and more intuitive when it comes to generating specific photometric distortions, such as complex multi-cast shadows, soft shadows, dark shadows, shading patterns, and color shifts.

3.1 Synthesis process

Figure 2 illustrates our synthesis process. We follow the idea of intrinsic images for data synthesis. In the notion of intrinsic images, $I = R \times L$, where I is the image, R is the reflectance map, and L is the lighting map. In this case, the scanned image I^{SQ} (Figure 2(a)) can be treated as the reflectance map, since it is the image when the lighting is uniform, that is, $L = 1$. As a result, we can model the photometrically distorted image as the product of the

Figure 3: **Two examples of our SSQD.** In each example, the images from left to right are the scanned image I^{SQ}, the shading distribution map $M^{shading}$, the shadow-free image I^{SF}, the shadow distribution map M^{shadow}, and the image with both shadows and shading I^{SS}.

scanned image and the lighting distribution map. There are two types of lighting distribution maps, the shadow distribution map and the shading distribution map, which respectively model the occlusion of key lights and other effects caused by non-uniform lighting.

As shown in Figure 2(b), there are three sources for simulating shadow maps: silhouette maps, large-scale soft maps, and Perlin shadow maps. Similar to Zhang *et al.* [34], we use silhouette masks to model the shapes of shadows cast by occluders. We collected 1,400 silhouette masks from the binary shape database [27], which consists of 70 shape categories. Among them, 47 categories are used in synthesizing the training set, 8 categories are used in the validation set, and 15 categories are used in the test set. In order to model large-scale shadows, we manually synthesize several maps. In order to add more shadow variations, we also use Perlin noise to generate Perlin shadow maps. Multi-cast shadows can be modeled by using a shadow distribution map derived from a mixture of shadow maps. Each selected shadow map is scaled and shifted randomly for adding more variations (Figure 2(d)). Also, to simulate the softness of shadows, we use Gaussian filters with different scales to blur the shadow maps. For modeling light's color and color shift, the shadow distribution map is modulated by a color sampled within a plausible range in the HSV space (Figure 2(e)). This way, we obtain a shadow distribution map denoted by M^{shadow} in Figure 2.

For the shading distribution map, we begin with a smooth gradient map (Figure 2(c)). Similarly, we randomly transform the map and modulate its color. With the generated shading distribution map $M^{shading}$, we multiply it with the scanned image I^{SQ} in order to synthesize a shadow-free image I^{SF}. Afterward, we multiply I^{SF} with the generated shadow distribution map M^{shadow}, and optionally add noise (Figure 2(f)) to the generated image for obtaining the image I^{SS} with both shadows and shading (Figure 2(g)).

3.2 The SSQD dataset

We collected 1,014 scanned document images from the DSSE Layout Analysis Dataset [29], PRIMA Layout Analysis dataset [5], DocUnet [18] and Li *et al.*'s dataset [16]. There is a wide variety of documents contained in them, including papers, receipts, textbooks, notes, and magazines. All of them are well-illuminated with clean and sharp content. Some of them contain only texts, while others are grayscale.

Based on these scanned images, we generate 7,000 triplets for training our network, 1,000 for validation, and 2,000 for testing. We call our dataset *Synthetic Scanner-Quality Dataset (SSQD)*. Figure 3 gives two examples from our SSQD dataset. On the left is an example with simple hard shadows obtained from a single silhouette. In the example on the right, Perlin shadow masks are used to produce more complex shadows.

In comparison to the 3D rendering process [16, 17], our procedure is more efficient and flexible. It takes Lin *et al.* [17] 11 days to synthesize around 8,000 images whereas our process only takes 2 hours to synthesize 10,000 images. Controlling the lighting and

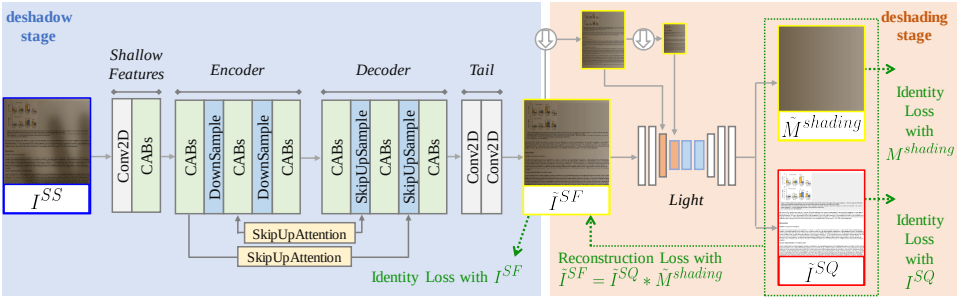

Figure 4: **The architecture of the proposed ScannerNet.** It has two sub-networks, DeshadowNet and DeshadingNet, which respectively exploit characteristics of shadows and shading effects for their removal.

occluders in 3D makes manipulating shadows and shading more difficult. Our 2D process makes adding complex shadows and shading patterns easier and more intuitive. In Section 5, it is demonstrated that the more diverse variations in the synthesized training dataset improve the performance of deep neural networks.

4 ScannerNet

ScannerNet is designed to correct various types of photometric distortions in document images. Figure 4 illustrates its architecture. For exploiting shadows and shading separately, it consists of two stages: The deshadow stage (DeshadowNet) first eliminates shadows with sharper intensity changes, while the deshading stage (DeshadingNet) corrects shading and color shifts that occur slowly. The model is more stable to train when the challenging problem is decomposed into two sub-problems and supervisory signals are provided to each of them. Training ScannerNet requires N triplets of document images, $\{I_i^{SS}, I_i^{SF}, I_i^{SQ}\}_{i=1}^{N}$ where I_i^{SS} denotes the i-th image with both *Shadows* and *Shading*, I_i^{SF} the corresponding *Shadow-Free* image, and I_i^{SQ} the corresponding *Scanner-Quality* image. Section 3 has described the procedure for synthesizing these triplets.

4.1 DeshadowNet

DeshadowNet learns to eliminate shadows at the first stage. With an input image I_i^{SS}, the network predicts a shadow-free image $\widetilde{I}_i^{SF}$ by minimizing the difference with the label I_i^{SF}:

$$L_{Deshadow} = \sum_{i=1}^{N} \left\| I_i^{SF} - \widetilde{I}_i^{SF} \right\|_1 . \tag{1}$$

Similar to BEDSR-Net [17], we use the attention mechanism to extract the long-range dependencies within a document image in order to solve the shadow removal problem. Unlike BEDSR-Net, which employs Grad-CAM [21] to extract the attention map, we apply the self-attention mechanism proposed by CBAM [28]. Inspired by the MPRNet [30], which uses self-attention to restore degraded images, we have designed our network with an encoder-decoder architecture built with channel attention blocks (CABs) to refine features as they propagate, and with convolutional layers at the head and tail of the network. Compared with

BEDSR-Net, self-attention and CABs allow for a significantly smaller model size, making it more suitable for mobile devices. As a result of the reduced model size, the training process is also faster.

4.2 DeshadingNet

We use intrinsic decomposition to remove shading since the scanner-quality image is similar to the document's reflectance map, and the shading distribution serves as the lighting. Given the estimated shadow-free image $\widetilde{I}_i^{SF}$ as the input and the corresponding scanner-quality image I_i^{SQ} as the label, DeshadingNet simultaneously predicts the estimated scanner-quality image $\widetilde{I}_i^{SQ}$ and the estimated shading map $\widetilde{M}_i^{shading}$ by minimizing the following loss,

$$L_{Deshading} = L_{SQ} + L_{Reconst}. \tag{2}$$

The first term L_{SQ} measures the difference between the estimated scanner-quality image $\widetilde{I}_i^{SQ}$ and the label I_i^{SQ},

$$L_{SQ} = \sum_{i=1}^{N} \left\| I_i^{SQ} - \widetilde{I}_i^{SQ} \right\|_1. \tag{3}$$

The second term requires that the product of the estimated reflectance $\widetilde{I}_i^{SQ}$ and lighting $\widetilde{M}_i^{shading}$ is equal to the input image $\widetilde{I}_i^{SF}$, that is

$$L_{Reconst} = \sum_{i=1}^{N} \left\| \widetilde{I}_i^{SF} - \widetilde{I}_i^{SQ} \times \widetilde{M}_i^{Shading} \right\|_1. \tag{4}$$

We have built our deshading network on top of UNet [20]. However, as in previous research [16], we encode the global information of shadow-free images by adding two down-sample layers and concatenating the outputs with the encoder layers. The encoded global information helps preserve content and improve visual quality. Details on the implementation and hyperparameters are given in the supplementary materials.

5 Experiments

The proposed method is evaluated on real-world images in terms of shadow removal and scanner quality. Additionally, we conduct experiments to investigate the performance gain that can be achieved by using our SSQD dataset. We also report the OCR performance to demonstrate that our method improves the legibility of the document in addition to visual quality. In the supplementary materials, we present additional results, ablation studies, and limitations.

Shadow removal. We first compare our DeshadowNet with document shadow removal methods, Kligler [13], Jung [12], and BEDSR-Net [17] on real-world images. The BEDSR-Net is a deep-learning-based method, while others are conventional methods. Since these methods only remove shadows, we compare them with our DeshadowNet. The second to fifth columns of Figure 5 demonstrate that our method has the best performance among all competitors. Complicated multi-cast shadows are present in Figure 5(b)(c), and only Jung [12] and DeshadowNet can handle them successfully. Figure 5(a) illustrates an example with extremely dark shadows that can only be removed completely by our method.

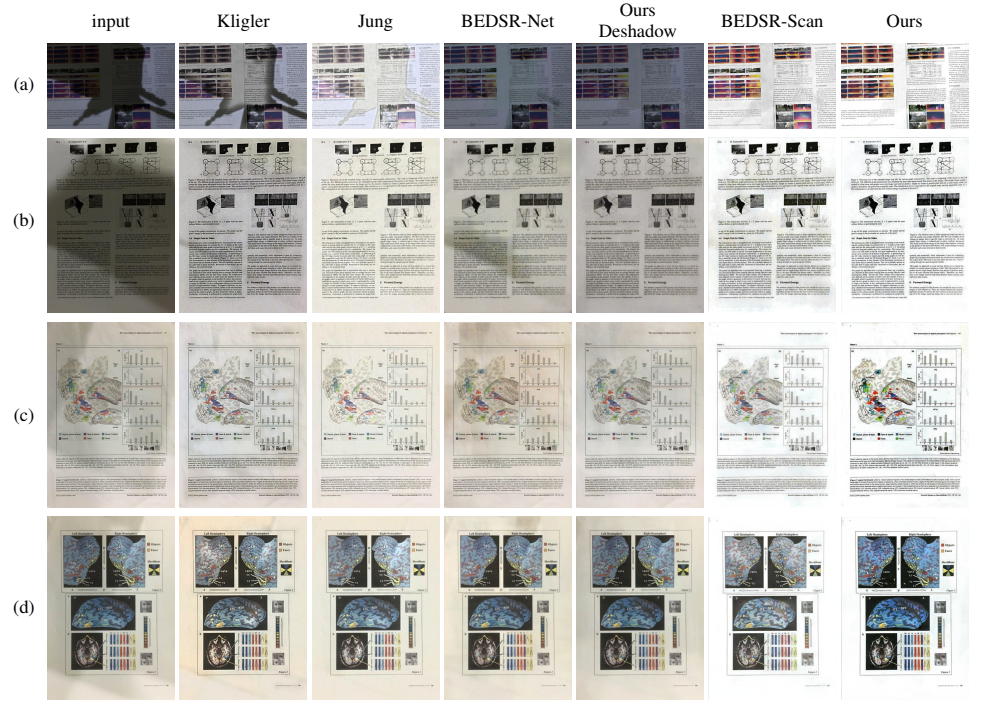

Figure 5: **Visual comparisons on real images.** From the 2^{nd} to 5^{th} columns, we compare the shadow removal results of Kligler [13], Jung [12], BEDSR-Net [17], and our DeshadowNet. For the last two columns, we compare the scanner-quality results of our method with those of BEDSR-Net trained on our dataset (BEDSR-Scan).

The top section of Table 1 reports the quantitative comparisons of document deshadow methods. The peak signal-to-noise ratio (PSNR) metric measures the error relative to the ground truth signal, whereas the structural similarity index (SSIM) metric emphasizes the structure of the image. Our method consistently achieves the 1^{st} or 2^{nd} highest PSNR and SSIM scores across most datasets, demonstrating that our method not only has smaller errors but also visually resembles the ground truth better. While BEDSR-Net achieves comparable performance with our method, our DeshadowNet has a significantly smaller model size (3.6M parameters) than theirs (19.8M) and is therefore more suitable for mobile devices.

The effectiveness of the SSQD dataset. One of our contributions is the data synthesis procedure that enriches the training dataset's variations. With the objective of studying the effectiveness of the SSQD dataset with more variations, we trained the U-Net [20] and our DeshadowNet using the SSQD dataset and the SDSRD dataset generated by 3D synthesis [17]. In Table 1, the bottom section compares the performance of models trained on different datasets. Across all test datasets, models trained using SSQD generally perform significantly better than those trained using SDSRD. This shows that our SSQD dataset is not only more effective but also agnostic to deep learning models.

Scanner quality. Only our method is capable of handling both deshadowing and deshading tasks for obtaining scanner-quality images. Despite its purpose of removing shadows, Jung's method [12] tends to produce bright, scanned-like images. According to Figure 5, our method generates images with the highest visual quality similar to scanned images since

	Model Size	Bako's Dataset		Jung's Dataset		Lin's RDSRD		Kligler's Dataset		Our SSQD	
Methods		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Input Shadow Image		28.45	0.974	20.35	0.885	21.73	0.809	19.31	0.843	18.93	0.692
Bako [1]		35.22	**0.982**	23.70	0.902	28.24	0.866	29.66	0.905	24.39	0.850
Jung [12]		13.88	0.806	**28.49**	0.911	14.45	0.705	19.21	0.872	14.66	0.768
BEDSR-Net [17]	19.8M	35.07	0.981	27.23	0.912	**33.48**	**0.908**	**32.90**	**0.935**	25.37	0.821
DeshadowNet (ours)	**3.6M**	**35.86**	0.981	26.39	**0.914**	33.00	0.905	31.92	0.932	**32.09**	**0.908**
U-Net [10] (SDSRD)	17.3M	32.44	0.979	25.66	0.910	30.10	0.893	30.15	0.929	26.82	0.867
U-Net [10] (SSQD)	17.3M	33.68	0.977	25.82	0.915	30.92	0.895	31.71	0.935	30.27	0.900
DeshadowNet (SDSRD)	3.6M	35.38	0.982	26.28	0.903	31.53	0.902	28.76	0.918	26.55	0.846
DeshadowNet (SSQD)	3.6M	35.86	0.981	26.34	0.914	33.00	0.905	31.92	0.932	32.09	0.908

Table 1: **Quantitative comparisons of different datasets.** The top portion of the table compares our DeshadowNet with previous shadow removal methods [1, 12, 17]. In each dataset, the best score is indicated in red bold, while the second best score is indicated in blue. The bottom portion shows that training U-Net [10] and DeshadowNet on our SSQD dataset can boost the models' performance, owing to the richer shadow variations in SSQD.

it handles both shadows and shading. In addition to having more uniform lighting, our generated images also display cleaner and more vibrant content. As illustrated in Figure 5(c), our method is capable of handling crumpled paper and color shifts well. In addition, our method better preserves the colors of figures in Figure 5(a)(c)(d). For comparison, we also train BEDSR-Net using our SSQD dataset and denote it as BEDSR-Scan. With the same training data, our model outperforms BEDSR-Scan for fewer background residuals and better color preservation, demonstrating that the proposed architecture is more effective. As an alternative method of obtaining scanner-quality images, one can use our DeshadingNet to remove shading from previous shadow removal methods [1, 12, 17]. Table 2 shows such comparisons quantitatively using the testing set of our SSQD dataset. Our method achieves the best SSIM and PSNR scores.

Method	Bako *et al.* [1]	Jung *et al.* [12]	BEDSR-Net [17]	BEDSR-Scan	Ours
SSIM ↑	0.86	0.91	0.83	**0.93**	**0.93**
PSNR ↑	20.19	27.30	22.81	27.97	**28.11**

Table 2: **Quantitative evaluation on scanner quality.** In comparison with all other methods, our method yields the best SSIM and PSNR values.

OCR performance. In addition to improving the visual quality of the document, our method also enhances its legibility. For validation, we use tesseract-OCR [22], an open-source tool maintained by Google, to perform OCR on the enhanced images using the RDSRD dataset [17]. Documents with little or no text are removed. For the remaining 384 images, we calculate the edit distances and report the average distance for each method. Table 3 summarizes the results. The proposed method achieves the best OCR performance. Note that our method even outperforms the OCR results using the ground-truth non-shadow images in RDSRD which has a distance of 808.9. This is due to the fact that our method also removes shading, which further enhances OCR.

Method	Input	Bako *et al.* [1]	Jung *et al.* [12]	Kligler *et al.* [13]	BEDSR-Net [17]	Ours
Dist ↓	2170.8	1026.3	821.3	1105.4	881.6	**765.2**

Table 3: **OCR comparisons using the edit distance.** The smaller the distance, the better the legibility. We achieve the best OCR results with the images enhanced by our method.

Figure 6: One failure case of our method. The results of our method might still contain residual color and shadow boundaries when the shadows are extremely dark. However, our method still produces the most visually pleasing results compared to previous approaches.

Limitations. Figure 6 illustrates one failure case of our method. It is possible that the results of our method will still contain residual color or shadow boundaries if the document contains extremely dark shadows. Additionally, the model may have problems when there are multiple shadow colors caused by light sources that are of different colors. It should be noted, however, that our method still produces the most visually appealing results when compared to previous methods.

6 Conclusion

This paper proposes ScannerNet, a comprehensive method for rectifying complex illumination distribution of document images using deep models. Our method consists of two sub-networks to progressively remove shadows and shading. First, DeshadowNet creates a uniformly lit intermediate image by removing shadows; then, DeshadingNet corrects the residual shading and color shifts to produce a scanning-quality document image. We propose a method of generating training data using mask composition in order to train our model with a great deal of diversity. This synthesizing flow has a relatively low computational cost. As a result, the dataset can be easily extended. Our experiments demonstrate that our method is more effective at removing document shadows and correcting shading, thereby improving visual quality in a significant way.

Acknowledgments. This work was supported in part by FIH Mobile Limited and National Science and Technology Council under grants MOST 110-2634-F-002-051, MOST 110-2221-E-002-124-MY3 and 107-S-C76. We thank to National Center for High-performance Computing (NCHC) for providing computational and storage resources.

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] Karina O. M. Bogdan, Guilherme A. S. Megeto, Rovilson Leal, Gustavo Souza, Augusto C. Valente, and Lucas N. Kirsten. DDocE: Deep document enhancement with

multi-scale feature aggregation and pixel-wise adjustments. In *International Conference on Document Analysis and Recognition (ICDAR) Workshop*, pages 229–244, 2021.

[3] M. S. Brown and Y. Tsoi. Geometric and shading correction for images of printed materials using boundary. *IEEE Transactions on Image Processing*, 15(6):1544–1554, 2006.

[4] Michael S. Brown, Mingxuan Sun, Ruigang Yang, Lin Yun, and W. Brent Seales. Restoring 2D content from distorted documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):1904–1916, 2007.

[5] Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. ICDAR2017 competition on recognition of documents with complex layouts-rdcl2017. In *Proc. International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1404–1410. IEEE, 2017.

[6] 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.

[7] 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.

[8] Soumyadeep Dey and Pratik Jawanpuria. Light-weight document image cleanup using perceptual loss. In *Proc. International Conference on Document Analysis and Recognition (ICDAR)*, pages 238–253, 2021.

[9] J. Fan. Robust color image enhancement of digitized books. In *Proc. International Conference on Document Analysis and Recognition (ICDAR)*, pages 561–565, 2009.

[10] Jian Fan. Enhancement of camera-captured document images with watershed segmentation. In *Proc. Int. Workshop on Camera-Based Document Analysis and Recognition*, pages 87–93, 2007.

[11] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. DSLR-quality photos on mobile devices with deep convolutional networks. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3277–3285, 2017.

[12] 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.

[13] 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.

[14] G. V. Landon, Y. Lin, and W. B. Seales. Towards automatic photometric correction of casually illuminated documents. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007.

[15] J. Lee, C. Chen, and C. Chang. A novel illumination-balance technique for improving the quality of degraded text-photo images. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(6):900–905, 2009.

[16] 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.

[17] 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.

[18] Ke Ma, Zhixin Shu, Xue Bai, Jue Wang, and Dimitris Samaras. Docunet: document image unwarping via a stacked u-net. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4709, 2018.

[19] G. Meng, S. Xiang, N. Zheng, and C. Pan. Nonparametric illumination correction for scanned document images via convex hulls. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1730–1743, 2013.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241. Springer, 2015.

[21] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[22] Raymond W Smith. Hybrid page layout analysis via tab-stop detection. In *Proc. International Conference on Document Analysis and Recognition (ICDAR)*, pages 241–245, 2009.

[23] Chew Lim Tan, Li Zhang, Zheng Zhang, and Tao Xia. Restoring warped document images through 3D shape modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):195–208, February 2006.

[24] Y. Tsoi and M. S. Brown. Multi-view document rectification using boundary. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007.

[25] T. Wada, H. Ukida, and T. Matsuyama. Shape from shading with interreflections under proximal light source-3D shape reconstruction of unfolded book surface from a scanner image. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 66–71, 1995.

[26] Toshikazu Wada, Hiroyuki Ukida, and Takashi Matsuyama. Shape from shading with interreflections under a proximal light source: Distortion-free copying of an unfolded-book. *International Journal of Computer Vision*, 24(2):125–135, September 1997.

[27] Bin Wang and Yongsheng Gao. Hierarchical string cuts: A translation, rotation, scale, and mirror invariant descriptor for fast shape retrieval. *IEEE Transactions on Image Processing*, 23, 08 2014.

[28] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proc. European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[29] Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C Lee Giles. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5315–5324, 2017.

[30] 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.

[31] A. Zandifar. Unwarping scanned image of Japanese/English documents. In *Proc. International Conference on Image Analysis and Processing (ICIAP)*, pages 129–136, 2007.

[32] L. Zhang, A. M. Yip, and C. L. Tan. Removing shading distortions in camera-based document images using inpainting and surface fitting with radial basis functions. In *Proc. International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 984–988, 2007.

[33] Li Zhang, A. M. Yip, M. S. Brown, and Chew Lim Tan. A unified framework for document restoration using inpainting and shape-from-shading. *Pattern Recognition*, 42(11):2961–2978, November 2009.

[34] Xuaner Zhang, Jonathan T. Barron, Yun-Ta Tsai, Rohit Pandey, Xiuming Zhang, Ren Ng, and David E. Jacobs. Portrait shadow manipulation. *ACM Transactions on Graphics (TOG)*, 39(4), 2020.