# BEDSR-Net: A Deep Shadow Removal Network from a Single Document Image

Yun-Hsuan Lin Wen-Chin Chen Yung-Yu Chuang*
National Taiwan University

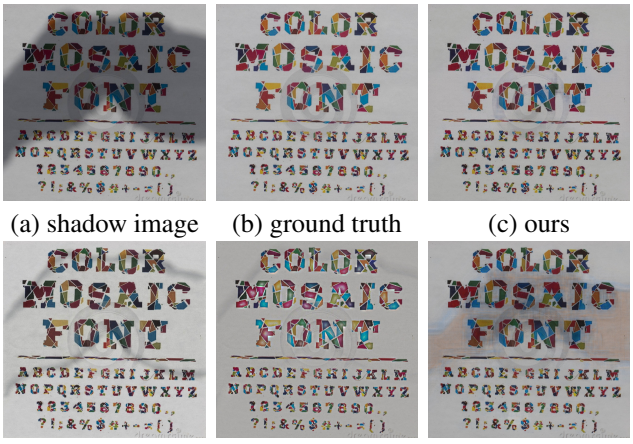
## Abstract

*Removing shadows in document images enhances both the visual quality and readability of digital copies of documents. Most existing shadow removal algorithms for document images use hand-crafted heuristics and are often not robust to documents with different characteristics. This paper proposes the Background Estimation Document Shadow Removal Network (BEDSR-Net), the first deep network specifically designed for document image shadow removal. For taking advantage of specific properties of document images, a background estimation module is designed for extracting the global background color of the document. During the process of estimating the background color, the module also learns information about the spatial distribution of background and non-background pixels. We encode such information into an attention map. With the estimated global background color and attention map, the shadow removal network can better recover the shadow-free image. We also show that the model trained on synthetic images remains effective for real photos, and provide a large set of synthetic shadow images of documents along with their corresponding shadow-free images and shadow masks. Extensive quantitative and qualitative experiments on several benchmarks show that the BEDSR-Net outperforms existing methods in enhancing both the visual quality and readability of document images.*

## 1. Introduction

Documents are indispensable and ubiquitous in our daily life. Examples include newspapers, receipts, papers, reports, and many others. There are often needs to obtain digital copies of documents. In the past, scanners were commonly used for digitizing documents with superior quality. Along with the prevalence of mobile phones and the improvement of their cameras, more and more people tend to use phone cameras in place of scanners for obtaining digital copies of documents because of their easy availability.

*This work was supported by FIH Mobile Limited and Ministry of Science and Technology (MOST) under grants 107-2221-E-002-147-MY3 and 109-2634-F-002-032.



(a) shadow image (b) ground truth (c) ours
(d) Kligler [16] (e) Bako [1] (f) ST-CGAN [28]

Figure 1. **An example of document shadow removal**. Previous methods, Kligler *et al*.'s method [16], Bako *et al*.'s method [1], and ST-CGAN [28], exhibit artifacts such as shadow edges (d), color washout (e) and residual shadows (f) in their results. Our result (c) has much fewer artifacts and is very close to the ground-truth shadow-free image (b).

Compared with scanners, there are often two problems with capturing documents using phone cameras. First, the geometry of the document could be distorted and not rectangular due to the camera perspective. Besides, the document could be folded or curved. There exist methods for rectifying and unwarping the captured documents so that they become rectangular in shape [18, 27, 19]. Second, the captured document images are vulnerable to shadows because the light sources are often blocked by the camera or user's hand. Even without occluders, the illumination is often uneven over the document when taking pictures in the real world. Thus, the document images taken by phone cameras often exhibit shadows and uneven shading, leading to bad visual quality and readability. Users usually prefer documents with uniform illumination, similar to what they can obtain using scanners, which take pictures in a well-controlled lighting environment. This paper addresses the shadow removal problem of document images for improving the quality and legibility of the captured documents.

Shadow removal is an important computer vision problem because shadows often degrade the performance of vi-

sion algorithms. Although most shadow removal methods are proposed for natural images, some are specifically designed for document images. Most existing document shadow removal algorithms use some heuristics for exploring specific characteristics of document images [3, 30, 22, 21, 1, 16, 15]. Unfortunately, due to limitations of the hand-crafted heuristics, they often work well for some document images but fail for others. Thus, their results often exhibit different types of artifacts for different types of document images. Figure 1 gives an example, where Figure 1 (a) is the input shadow image and Figure 1(b) is the corresponding shadow-free image. Figure 1(d) shows the result of Kligler *et al*.'s method [16], where some shadows remain around the boundary between the shadow and non-shadow regions. Figure 1(e) shows the result of Bako *et al*.'s method [1], in which colors are washed out, and some light shadow edges remain in the deshadowed result.

To combat the problems of hand-crafted heuristics, recently, deep learning has been employed in many vision problems. However, it has not been explored for document shadow removal, although there are quite a few deep-learning-based shadow removal methods for natural images [23, 28, 13]. ST-CGAN is a state-of-the-art natural image shadow removal method [28]. Given a set of training triplets of shadow images, shadow-free images, and shadow masks, it uses an end-to-end architecture for performing the shadow detection and removal tasks simultaneously. In principle, shadow removal methods for natural images can also be used for document images. However, there are two issues with applying deep learning to document shadow removal. First, it requires a large set of paired document images for training. Second, the performance would be sub-optimal since these methods do not take advantage of specific properties of document images. As an example, even after training with shadow/shadow-free pairs of document images along with shadow masks, ST-CGAN still fails to recover a proper shadow-free image, as shown in Figure 1(f). The shadow region remains, although it becomes lighter. Although recent shadow removal methods [6, 17] perform better than ST-CGAN, they often use pre-trained models on ImageNet and do not consider the characteristics of document images. Thus, they share the same problems with ST-CGAN on document images.

This paper proposes the first deep-learning-based approach for document image shadow removal. For addressing the first issue about training data, we propose to use synthetic images. This way, it is much easier to obtain a large-scale training set with great variations. Through extensive experiments, we show that the deep models trained on synthetic images remain effective for real-world images. For taking advantage of specific characteristics of document images, inspired by Bako *et al*. [1], we propose a network module for estimating the global background color of the document since most documents have a single background color, often the color of the paper. By exploring the global property, the background estimation module also discovers information about the locations of shadows in the form of an attention map. With both the estimated background color and the attention map, our shadow removal module can perform the shadow removal task much better. Extensive experiments show that our method not only outperforms existing methods in visual quality but also improves the readability of documents. As shown in Figure 1(c), our method is more robust with fewer artifacts. Our contributions include:

- We propose the first deep learning approach for shadow removal of document images, which outperforms state-of-the-art methods. By exploring the specific properties of document images, our model estimates the background color and an attention map as the first step. The information proves useful in improving image quality and reducing model parameters. Also, by exploring the attention map, the proposed model does not require shadow masks for training, alleviating the effort for collecting training data and reducing the risk with inaccurate masks.
- We provide a large-scale dataset of image triplets consisting of the shadow image, the corresponding shadow-free image, and shadow mask. The images are synthesized with a graphics renderer. The source code, datasets, and pre-trained models will be released.
- We show that the deep model trained on synthetic images remain effective for real images via thorough experiments with real images of different characteristics, collected by different research groups.

## 2. Related work

### 2.1. Shadow removal for natural images

Finlayson *et al*. proposed illumination invariant methods which remove shadows well for high-quality images [8, 7]. Guo *et al*. proposed a method for removing shadows by finding the relation between shadow and non-shadow regions with similar materials and removing shadows by relighting [11]. Gong *et al*. presented an interactive approach for high-quality shadow removal with two rough user inputs [9]. Gryka *et al*. focused on removing soft shadows using a learning-based approach with user-provided strokes [10]. Recently, several deep-learning-based methods have been proposed for natural image shadow removal, and they achieve state-of-the-art performance in the area [23, 28, 13, 17]. Qu *et al*. proposed the DeshadowNet that harnesses multi-context information for removing shadows from images [23]. Hu *et al*. proposed the direction-aware spatial context (DSC) module [13] that applies the spatial Recurrent Neural Network model [2] for

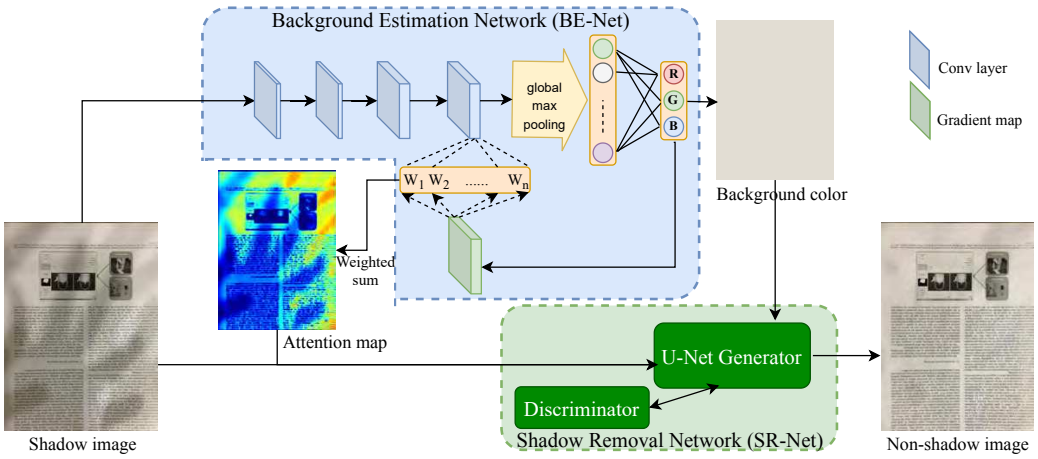



Figure 2. **The architecture of the BEDSR-Net**. It consists of two sub-networks: BE-Net for estimating the global background color of the document and SR-Net for removing shadows. Given the input shadow image, BE-Net predicts the background color. As a side product, it generates an attention map, which depicts how likely each pixel belongs to the shadow-free background. With the help of the attention map, our model removes the typical requirement of ground-truth shadow masks for training. Along with the input shadow image, the estimated background color and the attention map are fed into the SR-Net for determining the shadow-free version of the input shadow image.

obtaining 2D spatial context from four directions. However, since these models [23, 13] use VGG, pre-trained on natural images, as their backbone models, they are less suitable for document images. Wang *et al*. introduced ST-CGAN, which is trained to perform both the shadow detection and removal tasks simultaneously [28]. ST-CGAN uses the architecture of stacked conditional generative adversarial networks, where two cGANs would facilitate each other to improve the performance of both tasks. Self-supervised learning has also been introduced to shadow removal for natural images recently, such as ARGAN [6] and Mask-ShadowGAN [12]. Although effective for removing shadows in natural images, they are not specifically designed for document images. Thus, their performance on document image shadow removal is sub-optimal even after being retrained on document images, as shown in the experiments.

### 2.2. Shadow removal for document images

Some methods have been designed specifically for shadow removal of document images. One thread of methods is based on the concept of intrinsic images and remove shadows through reducing the luminance contrast in regions with similar reflection components [29, 3, 30]. Jung *et al*. proposed the water-filling method inspired by the immersion process of a topographic surface with water [15]. The method however tends to cause color shift and have results much brighter than they should be. Kligler *et al*. proposed a method for enhancing the quality of document images by representing an image as a 3D point cloud and using visibility detection methods to choose pixels to recover [16]. However, there are often shadow edges remaining in their results. By assuming a constant background color, direct interpolation has been used for document image shadow removal by Oliveira *et al*. [22] and Bako *et al*. [1]. Bako *et al*.'s method obtains a shadow map by calculating the ratio between the global and local background colors for each patch and then adjusts the input shadow image with the shadow map [1]. Since these methods detect the background regions and perform interpolation in the rest, they fail when the documents contain large regions of figures or are covered by a large area of shadows. This paper proposes the first deep learning method for removing shadows from a single document image. By harnessing the power of data, our method is more robust than existing methods.

## 3. Method

This paper proposes BEDSR-Net (Background Estimation Document Shadow Removal Network) for removing shadows from a single document image. The training set, $D=\{S_i, N_i\}_{i=1}^{N}$, consists of $N$ image pairs $(S_i, N_i)$ where $S_i$ is a shadow image while $N_i$ is its corresponding non-shadow image. After training, the BEDSR-Net forms a function $\Psi_{BESR}(S)$ which accepts a shadow image $S$ and returns a predicted non-shadow image $\tilde{N}$ that approximates the real non-shadow image $N$. Figure 2 illustrates the architecture of the BEDSR-Net, consisting of two sub-networks, BE-Net (Background Estimation Network) and SR-Net (Shadow Removal Network). Given the input shadow document image $S$, the BE-Net, $(\tilde{b}, \tilde{H}) = \Psi_{BE}(S)$, estimates the global background color of the document $\tilde{b}$ and an estimated attention map $\tilde{H}$ which depicts the probability of each pixel belonging to the shadow-free background of the document. Given the shadow image $S$ and the outputs of BE-Net, $(\tilde{b}, \tilde{H})$, the SR-Net, $\tilde{N} = \Psi_{SR}(S, (\tilde{b}, \tilde{H}))$, predicts the shadow-free image $\tilde{N}$ as the final output.

### 3.1. Background Estimation Network (BE-Net)

Inspired by Bako *et al*. [1], we also attempt to recover the background color of the document and uses it to assist the shadow removal process. Bako *et al*.'s method uses heuristics for estimating the background color by analyzing color distributions. It tends to cause a color shift or color fading when the document is covered by a large area of shadows or color figures. To address these problems, we use a deep network for a more robust estimation of the background color. The proposed BE-Net takes a shadow image $S$ as the input and estimates the predicted background color $\tilde{b}$.

For training the BE-Net, we need to identify the ground-truth background color $b_i$ for each document image pair $(S_i, N_i)$ in the training set $D$. It can be achieved by asking users to manually pick up a region belonging to the background in the non-shadow image $N_i$ and calculating the average color of the picked region. For relieving the manual effort, we used the following procedure for obtaining the ground-truth background $b_i$ automatically. First, we cluster pixels of the non-shadow image $N_i$ into two groups according to their intensity values. For clustering, we employed Gaussian mixture models (GMM) with Expectation Maximization (EM). The two groups often correspond to the content and the background, respectively. The background color of the document is often brighter. Thus, the cluster with a higher intensity is taken as the background cluster. Since the image $N_i$ contains no shadow, we can use the mean color of the background cluster as the background color $b_i$ for the $i$-th image pair in the training set. We found that the procedure works well empirically. Thus, for each image pair in the training set, we obtain its background color $b_i$ using this procedure. The procedure is for speeding up ground truth collection. The result can be corrected by users if the heuristic fails, for example, when the document has a dark background color. The brighter background color is not an assumption of our model itself.

Given the shadow image $S_i$ and its background color $b_i$, we can train our BE-Net in a supervised manner by minimizing the following cost,

$$\mathcal{L}_{color} = \sum_{i=1}^{N} \|b_i - \Psi_{BE}(S_i)\|_1, \tag{1}$$

so that the predicted background color $\tilde{b}_i = \Psi_{BE}(S_i)$ estimated from the shadow image $S_i$ approximates the true background color $b_i$. As shown in Figure 2, our BE-Net consists of four convolution layers, followed by a global max pooling layer and a fully connected layer. The convolution layers extract proper features from the input shadow image. We adopt the global pooling mechanism to summarize each feature map into a value. By using global pooling to bridge the convolution layers and the fully connected layer, our network can deal with images with different sizes.

Along the way of estimating the background, BE-Net also learns knowledge about spatial distributions of shadow-free background and others, which provides additional cues for indicating potential locations of shadows. To utilize the information, we extract an attention map by applying the Grad-CAM method [25] to the feature map of the last convolution layer in BE-Net. As shown in Figure 2, the attention map does capture well where the shadow-free background (red color) and other (blue color) pixels reside in the shadow image. The inferred attention map also reveals cues about shadow locations and can play the role of the shadow mask. With its help, unlike other shadow removal networks such as ST-CGAN, our model does not require the ground-truth shadow masks for training. It provides the advantages of saving the effort for preparing shadow masks and avoiding the potential errors in the shadow masks. Note that a shadow mask is often derived from the shadow and non-shadow images using some heuristics since it cannot be captured directly. Thus, it could contain errors.

### 3.2. Shadow Removal Network (SR-Net)

For recovering a shadow-free image from a shadow image, we employ the conditional generative adversarial networks (cGANs) [20], which have been shown effective in many tasks such as image-to-image translation. A cGAN model consists of two players: a generator $G$ and a discriminator $D$. Given a conditioning variable, the generator $G$ aims to produce realistic images to fool the discriminator $D$, while the discriminator $D$ attempts to distinguish the images generated by $G$ from the real ones in the dataset. The competition enhances the generator in producing the result indistinguishable from real images.

For the generator $G$, we adopt the U-Net model [24], which is a fully convolutional neural network, consisting of an encoder and a decoder. The features from the decoder will be combined with those from the encoder through skip connections at each spatial resolution. We used a five-level hierarchy for both the encoder and decoder. The generator $G$ takes the concatenation of the shadow image $S_i$, the predicted background color $\tilde{b}_i$ and the attention map $\tilde{H}_i$ as input, and then predicts the non-shadow image $\tilde{N}_i = G(S_i, \tilde{b}_i, \tilde{H}_i)$. For the discriminator $D$, we employ Markovian discriminator (PatchGAN) [14]. The input of $D$ is the 6-channel concatenation of a shadow image $S_i$ and the paired non-shadow image $N_i$. For training the SR-Net, the following loss is used,

$$\mathcal{L} = \lambda_1 \mathcal{L}_{data} + \lambda_2 \mathcal{L}_{GAN}, \tag{2}$$

where $\mathcal{L}_{data}$ measures the deviation of the predicted non-shadow image from the real one,

$$\mathcal{L}_{data} = \mathbb{E}_{S_i, N_i \sim P_{data}(S_i, N_i)} ||N_i - \tilde{N}_i||. \tag{3}$$

| Dataset | #pairs | Characteristics |
|---|---|---|
| Bako [1] | 81 | light shadows/text only |
| Kligler [16] | 300 | dark shadows/colorful symbols |
| Jung [15] | 87 | multi-cast shadows |
| RDSRD | 540 | complex content/shadows |
| SDSRD | 8,309 | synthetic, diverse lighting and contents |

Table 1. **Overview of the document shadow removal datasets**: Bako [1], Kligler [16], Jung [15], our RDSRD and SDSRD, in terms of numbers of image pairs and characteristics.

and $\mathcal{L}_{GAN}$ is the GAN loss for the competition of $G$ and $D$,

$$\mathcal{L}_{GAN} = \mathbb{E}_{S_i,N_i\sim P_{data}(S_i,N_i)}[\log D(S_i,N_i)]+ \\ \mathbb{E}_{S_i\sim P_{data}(S_i)}[\log(1-D(S_i,\tilde{N}_i))]. \quad (4)$$

Adam is used for optimization. The parameters are set empirically as $\lambda_1$=1 and $\lambda_2$=0.01. After training, the generator $G$ is used to generate the output of SR-Net, *i.e.*, $\Psi_{SR} \equiv G$.

## 4. Datasets

Although there exist a few datasets for document image shadow removal, they are only used for evaluation and of small scale. Table 1 summarizes datasets for document shadow removal. Previous datasets do not have a large number of images. Training deep models, however, requires sufficient data. Our model requires paired shadow and shadow-free images. Capturing such pairs in the real world is possible but time-consuming as it requires careful control. Also, limited by human effort, it is less likely to provide images with great variations in document contents, lighting conditions, and shadow complexity. Since current graphics algorithms already can render shadows realistically, we explore the possibility of using synthetic images for training.

### 4.1. Synthetic document shadow removal dataset

For having a large set of document images with great variations, we synthesize document images using Blender [5] and Python scripts. For providing variations in document types and contents, we collected 970 document images, most from the PRImA Layout Analysis dataset [4]. For each document, we synthesized several shadow images using different lighting conditions and occluders. Since the images are synthesized, shadow-free images and shadow masks can be obtained with ease. We synthesized a total of 8,309 triplets of shadow images, shadow-free images, and shadow masks. They are divided into two groups, 7,533 for training and 776 for testing. We call it Synthetic Document Shadow Removal Dataset (SDSRD). The training set of SDSRD is used for training BEDSR-Net. Note that training BEDSR-Net does not require shadow masks. We generate shadow masks because training ST-CGAN requires them. Figure 3 gives examples of SDSRD.

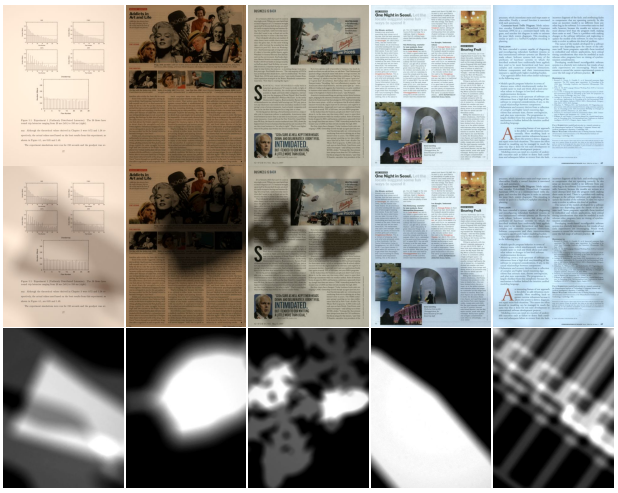


Figure 3. **Example triplets from SDSRD.** It provides images with complex shadows in both shape and intensity. From top to bottom, the images are shadow-free images, shadow images, and shadow masks, respectively.

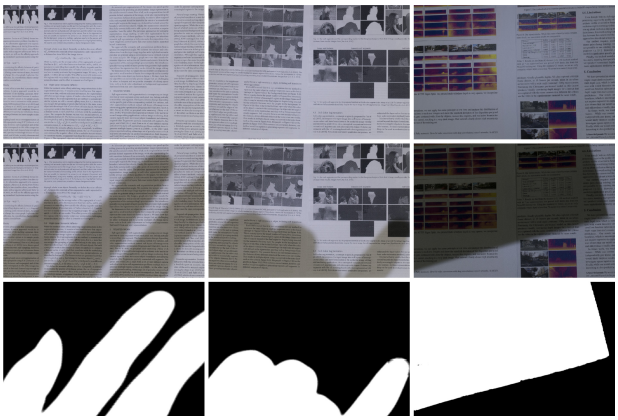


Figure 4. **Example triplets from RDSRD**. The images contain intricate shadows in terms of shape. From top to bottom, the images are shadow-free images, shadow images, and shadow masks.

### 4.2. Real document shadow removal dataset

For evaluating on real images with more variations, we have also collected the Real Document Shadow Removal Dataset (RDSRD). The images were captured using Sony RX100 m3 and flashlights, all on tripods for ensuring fixed locations. The camera is triggered using a remote through WiFi to avoid touching the camera during capture. The dataset consists of 540 images of 25 documents, including paper, newspaper, and slides, under different lighting conditions and occluders. Figure 4 gives examples of RDSRD. This dataset is used only for evaluation.

## 5. Experiments

We introduce compared methods and metrics, and then compare them on visual quality and content preservation.

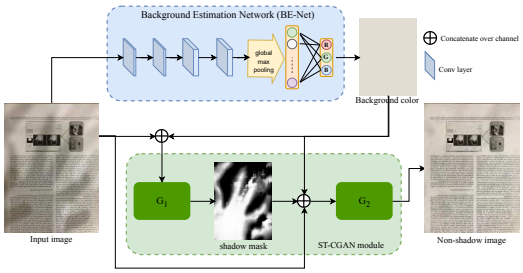



Figure 5. **The architecture of ST-CGAN-BE.**

### 5.1. Compared methods and evaluation metrics

We compare our BEDSR-Net with four state-of-the-art methods, including three traditional document shadow removal methods by Bako *et al*. [1], Kligler *et al*. [16] and Jung *et al*. [15], and one state-of-the-art deep-learning-based natural image shadow removal method, ST-CGAN [28]. For a fair comparison, we used the publicly available source codes provided by the authors whenever available. Among the four methods, ST-CGAN is the only one without source code released. Thus, we reproduce it on our own. The implementation has been validated using their dataset and reaches similar performance as reported in their paper. For showing the importance of the background estimation module, we incorporate our BE-Net into ST-CGAN and name it ST-CGAN-BE. Figure 5 shows its architecture. ST-CGAN has two generators, $G_1$ for shadow detection and $G_2$ for shadow removal. Training $G_1$ and $G_2$ requires shadow masks and shadow-free images, respectively. The estimated background color is fed to both generators in ST-CGAN-BE. Note that the attention map is not included in ST-CGAN-BE. All learning-based approaches are trained with the training set of SDSRD.

We evaluate the compared methods from two perspectives, visual quality, and content preservation. For visual quality, we use the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) index as the metrics. For evaluating content preservation, we test the performance of Optical Character Recognition (OCR) techniques on the recovered shadow-free images. In general, if the content is better recovered, OCR should be able to recognize more content.

### 5.2. Quantitative evaluation

Table 2 reports quantitative comparisons of the compared methods on five datasets and the average in terms of PSNR and SSIM. Our BEDSR-Net outperforms others on most datasets. For RDSRD, SDSRD, and Kligler's datasets, our method reaches the best performance. In particular, our models beat other methods significantly on Kligler's dataset, which contains very dark shadows and color texts. For Bako's and Jung's datasets, their methods achieve the best performance. However, their performance on other datasets can be bad. For example, Bako *et al*.'s method performs poorly on Jung's dataset since it can only handle light shadows. On the other hand, Jung *et al*.'s method is the worst among all compared methods on Bako's dataset as it tends to wash out colors. These methods are derived from heuristics, and their datasets often better match the characteristics of their heuristics. Our model is very competitive, with only a small margin behind the best methods on Bako's and Jung's datasets. Overall, our method is more robust than previous methods as it provides stable and good results for images with different characteristics.

Our model is based on U-Net. As an ablation study, Table 2 reports the performance of U-Net trained on SDSRD with a supervised setting. BEDSR-Net outperforms U-Net by a large margin, showing our performance comes from more than the architecture of U-Net and training data. As another ablation study, the superior performance of our ST-CGAN-BE compared to ST-CGAN shows the importance of the background estimation module. Also, the notable performance gain from ST-CGAN-BE to BEDSR-Net indicates that the predicted attention map provides more useful information than the shadow masks generated by the first generator of ST-CGAN-BE. Finally, BEDSR-Net achieves better performance than ST-CGAN with fewer parameters, 19.8M for BEDSR-Net, and 38.9M for ST-CGAN.

### 5.3. Qualitative evaluation on visual quality

For visual comparisons, Figure 6 shows several shadow removal results of the compared methods. Although Bako *et al*.'s method performs well in quantitative evaluation, it fails to recover the image with color texts (example #7) or large area of shadows (example #3). Both Bako *et al*.'s method and Kligler *et al*.'s method exhibit remaining shadow edges when there are strong shadows (example #3 and #4). Jung's method often incurs a severe color shift in the results. Its results are often dramatically brighter than the ground-truth shadow-free images. The color is washed out, and the contrast is reduced. ST-CGAN runs into problems when there are large dark shadows (example #3 and #4).

Although our model is derived from the assumption of single dominant background color, its utility is not as restricted as it appears. Since the whole document is captured as an image as a whole, there is no clear distinction between the content and background. Taking example #7 of Figure 6 as an example, it can be interpreted in two ways: (1) ten colorful numbers and a color gradient area on a white paper or (2) ten colorful numbers on a white paper with color gradients. For the second interpretation, there are multiple background colors, and our method still obtains a good result. As long as there is a dominant uniform color in the document image, our method can still work well. We argue that documents with such a characteristic represent a significant portion of the real world. As evidence, we extensively tested our method on existing document datasets, indepen-

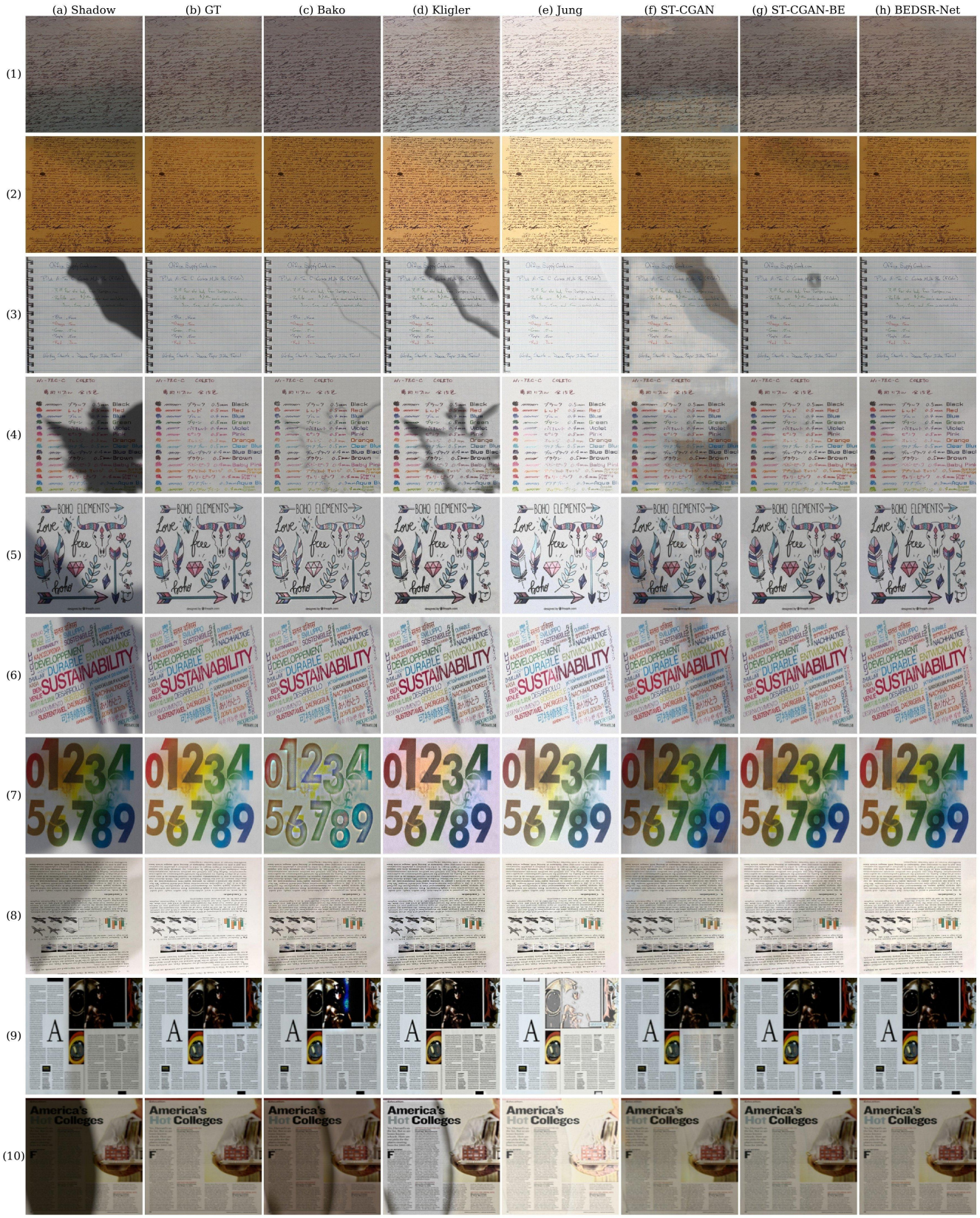



Figure 6. **Visual comparison of competing methods**: Bako [1], Kligler [16], Jung [15], ST-CGAN [28], our ST-CGAN-BE and BEDSR-Net, on ten images in which (1)-(2) are from Bako, (3)-(7) from Kligler, (8) from Jung and (9)-(10) from the testing set of SDSRD.

| | Average | | SDSRD | | RDSRD | | Bako's dataset | | Kligler's dataset | | Jung's dataset | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| input shadow images | 22.03 | 0.8652 | 22.80 | 0.8992 | 21.73 | 0.8093 | 28.45 | 0.9742 | 19.31 | 0.8429 | 20.35 | 0.8850 |
| Bako [1] | 30.01 | 0.9231 | 31.55 | 0.9658 | 28.24 | 0.8664 | **35.22** | **0.9823** | 29.66 | 0.9051 | 23.70 | 0.9015 |
| Kligler [16] | 23.25 | 0.8081 | 22.03 | 0.8435 | 22.53 | 0.7056 | 26.50 | 0.8381 | 26.45 | 0.8481 | 24.45 | 0.8332 |
| Jung [15] | 17.04 | 0.7990 | 17.06 | 0.8226 | 14.45 | 0.7054 | 13.88 | 0.8059 | 19.21 | 0.8724 | **28.49** | 0.9108 |
| U-Net [24] | 29.47 | 0.8985 | 33.63 | 0.9728 | 28.35 | 0.8676 | 26.68 | 0.8833 | 23.33 | 0.7829 | 23.09 | 0.8399 |
| ST-CGAN [28] | 33.14 | 0.9408 | 39.38 | 0.9834 | 30.31 | 0.9016 | 29.12 | 0.9600 | 25.92 | 0.9062 | 23.71 | 0.9046 |
| ST-CGAN-BE | 36.77 | 0.9521 | 42.98 | **0.9938** | 32.32 | 0.9054 | 33.90 | 0.9801 | 32.50 | 0.9338 | 26.45 | 0.9080 |
| BEDSR-Net | **37.55** | **0.9534** | **43.59** | 0.9935 | **33.48** | **0.9084** | 35.07 | 0.9809 | **32.90** | **0.9354** | 27.23 | **0.9115** |

Table 2. **Quantitative comparisons of visual quality using PSNR and SSIM**. We compare our models, BEDSR-Net and ST-CGAN-BE, with four competitive methods. The best scores of each dataset are marked in red bold, while the second best ones are marked in blue.

(a) shadow image (b) our result

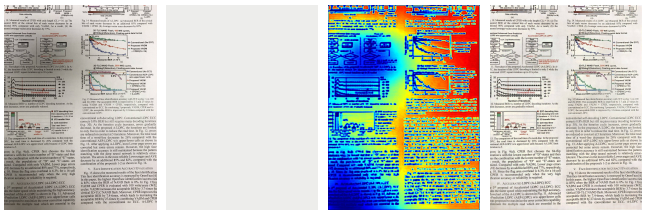
Figure 7. **An example with a tan background and a large figure.**

dently collected by several groups. Our method obtains excellent performance on all datasets, even if most of them are not collected by us. Figure 7 gives an example with a tan background and a substantial figure. Our method could run into problems if there is no single dominant color, such as a paper entirely with a color gradient. However, it is a rare case, and most existing methods could also fail. Also, our method could fail when the document is entirely in shadow, or there are complicated shadows cast by multiple lights.

Figure 8(a) gives a document image captured by a mobile phone in an uncontrolled environment. Figure 8(b) shows the estimated background color. Figure 8(c) displays the predicted attention map in which red color indicates shadow-free background while blue color denotes shadowed background and non-background regions. Both faithfully capture the real characteristics of the input image. With their help, BEDSR-Net successfully recovers the shadow-free image in Figure 8(d).

### 5.4. Evaluation on content preservation

We also evaluate how the readability of documents is enhanced by reporting OCR performance on the recovered shadow-free images. In the experiment, 188 images with texts are used. First, we apply an open-source OCR tool [26] to recognize texts for ground-truth shadow-free images and results of compared methods. Then, we measure the OCR performance by comparing the text strings using the Levenshtein distance, also known as edit-distance.



(a) input (b) background (c) attention (d) our result

Figure 8. **A real example taken using a phone camera under an uncontrolled environment**. Our BEDSR-Net recovers the shadow-free image well and the attention map indicates background and non-background pixels very well.

| method | input | Bako | Kligler | Jung | ST-CGAN | BEDSR-Net |
|---|---|---|---|---|---|---|
| distance | 551.9 | 50.2 | 93.2 | 92.5 | 133.1 | 38.5 |

Table 3. Average edit-distances of the input images, Bako *et al.*'s method [1], Kligler *et al.*'s method [16], Jung *et al.*'s method [15], ST-CGAN [28], and the proposed BEDSR-Net.

As reported in Table 3, BEDSR-Net outperforms others, showing that it enhances not only visual quality but also the readability of documents by better preserving structure and content. Note that the test is for validating how our method preserves the content and improves document readability, rather than reaching the state-of-the-art OCR performance.

## 6. Conclusion

This paper proposes BEDSR-Net, the first deep learning model for shadow removal of document images. For exploring specific properties of documents, we propose BE-Net for background color estimation. It also generates an attention map, which is shown effective in indicating shadow locations. With the help of the estimated background color and attention map, our model achieves state-of-the-art performance in visual quality. It also improves the readability of document images. For training with document images with great diversity, we train our model using a synthetic dataset and show that the trained model works well for real images. In the future, we would like to explore unpaired training, handling documents with more complex backgrounds, and applying the background estimation module to document layout recognition.

## References


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

[2] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-Outside Net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2874–2883, 2016.

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

[4] Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. ICDAR2017 competition on recognition of documents with complex layouts-RDCL2017. In *Proceedings of 14th International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1404–1410, 2017.

[5] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, 2018.

[6] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao. ARGAN: Attentive recurrent generative adversarial network for shadow detection and removal. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 10213–10222, 2019.

[7] Graham D. Finlayson, Mark S. Drew, and Cheng Lu. Entropy minimization for shadow removal. *International Journal of Computer Vision*, 85(1):35–57, 2009.

[8] Graham D. Finlayson, Steven D. Hordley, Cheng Lu, and Mark S. Drew. On the removal of shadows from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):59–68, 2005.

[9] Han Gong and Darren Cosker. Interactive shadow removal and ground truth for variable scene categories. In *Proceedings of British Machine Vision Conference (BMVC)*, 2014.

[10] Maciej Gryka, Michael Terry, and Gabriel J. Brostow. Learning to remove soft shadows. *ACM Transactions on Graphics (TOG)*, 34(5):153, 2015.

[11] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2956–2967, 2012.

[12] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-ShadowGAN: Learning to remove shadows from unpaired data. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 2472–2481, 2019.

[13] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7454–7462, 2018.

[14] 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 (CVPR)*, pages 1125–1134, 2017.

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

[16] Netanel Kligler, Sagi Katz, and Ayellet Tal. Document enhancement using visibility detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2374–2382, 2018.

[17] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 8578–8587, 2019.

[18] Shijian Lu, Ben M. Chen, and Chi Chung Ko. Perspective rectification of document images using fuzzy set and morphological operations. *Image and Vision Computing*, 23(5):541–553, May 2005.

[19] Ke Ma, Zhixin Shu, Xue Bai, Jue Wang, and Dimitris Samaras. DocUNet: Document image unwarping via a stacked U-Net. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[20] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[21] Daniel Marques Oliveira and Rafael Dueire Lins. A new method for shading removal and binarization of documents acquired with portable digital cameras. In *Proceedings of Third International Workshop on Camera-Based Document Analysis and Recognition*, pages 3–10, 2009.

[22] Daniel Marques Oliveira, Rafael Dueire Lins, and Gabriel de França Pereira e Silva. Shading removal of illustrated documents. In *Proceedings of International Conference on Image Analysis and Recognition (ICIAR)*, pages 308–317, 2013.

[23] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson W.H. Lau. DeshadowNet: A multi-context embedding deep network for shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4067–4075, 2017.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of International Conference on Medical image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

[25] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[26] Raymond W Smith. Hybrid page layout analysis via tab-stop detection. In *Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR)*, pages 241–245, 2009.

[27] Yuandong Tian and Srinivasa G. Narasimhan. Rectification and 3D reconstruction of curved document images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[28] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1788–1797, 2018.

[29] Qingxiong Yang, Kar-Han Tan, and Narendra Ahuja. Shadow removal using bilateral filtering. *IEEE Transactions on Image processing*, 21(10):4361–4368, 2012.

[30] Li Zhang, Andy M. Yip, and Chew Lim Tan. Removing shading distortions in camera-based document images using inpainting and surface fitting with radial basis functions. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 984–988, 2007.