Intrinsic Decomposition of Document Images In-the-Wild

Sagnik Das[1]
sadas@cs.stonybrook.edu
Hassan Ahmed Sial[2]
hasial@cvc.uab.es
Ke Ma[1]
kemma@cs.stonybrook.edu
Ramon Baldrich[2]
ramon@cvc.uab.es
Maria Vanrell[2]
maria.vanrell@uab.cat
Dimitris Samaras[1]
samaras@cs.stonybrook.edu

[1] Computer Vision Lab
Stony Brook University
New York, USA

[2] Computer Vision Center
Universitat Autònoma de Barcelona
Barcelona, Spain

Abstract

Automatic document content processing is affected by artifacts caused by the shape of the paper, non-uniform and diverse color of lighting conditions. Fully-supervised methods on real data are impossible due to the large amount of data needed. Hence, the current state of the art deep learning models are trained on fully or partially synthetic images. However, document shadow or shading removal results still suffer because: (a) prior methods rely on uniformity of local color statistics, which limit their application on real-scenarios with complex document shapes and textures and; (b) synthetic or hybrid datasets with non-realistic, simulated lighting conditions are used to train the models. In this paper we tackle these problems with our two main contributions. First, a physically constrained learning-based method that directly estimates document reflectance based on intrinsic image formation which generalizes to challenging illumination conditions. Second, a new dataset that clearly improves previous synthetic ones, by adding a large range of realistic shading and diverse multi-illuminant conditions, uniquely customized to deal with documents in-the-wild. The proposed architecture works in two steps. First, a white balancing module neutralizes the color of the illumination on the input image. Based on the proposed multi-illuminant dataset we achieve a good white-balancing in really difficult conditions. Second, the shading separation module accurately disentangles the shading and paper material in a self-supervised manner where only the synthetic texture is used as a weak training signal (obviating the need for very costly ground truth with disentangled versions of shading and reflectance). The proposed approach leads to significant generalization of document reflectance estimation in real scenes with challenging illumination. We extensively evaluate on the real benchmark datasets available for intrinsic image decomposition and document shadow removal tasks. Our reflectance estimation scheme, when used as a pre-processing step of an OCR pipeline, shows a 21% improvement of character error rate (CER), thus, proving the practical applicability. The data and code will be available at: https://github.com/cvlab-stonybrook/DocIIW .

1 Introduction

Paper documents contain and provide valuable information that is essential in our everyday life. By digitizing the documents, we can archive, retrieve, and share contents with utmost convenience. With the increasing ubiquity of mobile camera devices, nowadays, capturing document images is the most common way of digitizing them. Such a document image is only useful if it can be used for further processing, such as text extraction and content analysis. Therefore, it is desirable to capture the image so that most of the document content is preserved. However, casual document images captured in-the-wild often contain substantial distortions due to non-uniform illumination conditions from multiple colored lights resulting in diverse shading and shadow effects. These physical artifacts hurt the accuracy and reliability of automatic information extraction methods. As an example, part of the text in the leftmost image of figure 1 remains undetected due to shadows and shading. Recent

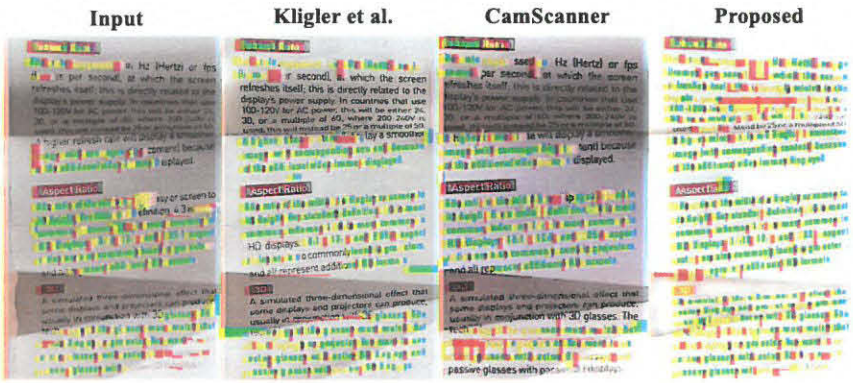

Figure 1: OCR accuracy (by Tesseract [34]) comparison on document images pre-processed with our proposed method vs. Kligler et al. (state-of-the-art document shadow removal method [25]) vs. the commercially available document capturing application CamScanner.

deep learning image-to-image translation methods [21] are widely applicable but require large sets of training images to perform well in realistic scenarios. Enforcing additional domain-specific physical ground-truth helps to improve the generalization performance of such models [9, 40]. Alas, acquiring such ground-truth is often very costly. To model document illumination correction using deep learning, one would need a large document dataset with images, reflectance and shading captured in a large number of real scenes with various illumination conditions. It is very time-consuming and costly to collect such a dataset. In this paper, we propose a weakly-supervised method to separate reflectance and shading, and train the network by using physical constraints of image formation. Moreover, to facilitate the training we carefully curate a multi-illumination dataset, Doc3DShade, for document images in-the-wild that consists of realistic illumination and can be augmented to a much larger scale using rendering engines [8].

Prior document shadow and shading removal methods [1, 43] mostly rely on local color statistics, thus imposing strong assumptions such as uniform background color of the document. Therefore, the application of these methods is generally limited to documents of a particular type with minimal colors, or graphics. Moreover, these methods assume the paper shape is almost flat [23, 25], which also makes these approaches inapplicable to in-the-wild images of warped documents under challenging illumination conditions. To demonstrate the inability of existing methods in handling non-uniform illumination on complex paper shapes, we compare Optical Character Recognition (OCR) results in figure 1. We compute OCR after pre-processing the input image (figure 1) with the method of Kligler et al. [25] (current state-of-the-art for document shadow removal), a commercially available document

processing application, and our proposed reflectance estimation scheme. Our method shows a significantly higher number of detected characters proving the effectiveness of the intrinsic image-based modeling of document images over traditional approaches.

We address the aforementioned problems with a unified architecture based on two learning-based modules, WBNet and SMTNet, trained in a self-supervised manner on a new multi-illuminant dataset, Doc3DShade, built on top of the public Doc3D [9] dataset. We formulate the problem of document reflectance estimation based on the physics of image formation under the Lambertian assumption, i.e., an image I, is a composition of the shading S and reflectance R, $I = R \otimes S$, where $\otimes$ denotes the Hadamard product. Images in the Doc3Dshade are created by following this assumption. In Doc3DShade we capture realistic shading MS of a deformed non-textured document under a large range of realistic colors, and diverse multi-illuminant conditions. Note that MS contains both the shading and the color of the paper material. We render real document textures T in Blender [8] and create I by combining $I = T \otimes MS$. Our formulation is similar by considering $R = T \otimes M$. Since the shading component is physically captured in controlled environments, these images are clearly superior in terms of physical illumination effects than the existing Doc3D [9] images. We have generated 90,000 images in Doc3DShade which effectively double the size of Doc3D.

The WBNet and SMTNet address white-balancing and shading removal respectively. WBNet inputs an RGB image and estimates a white-balanced image where the illuminant color is removed. We can train WBNet with Doc3DShade which contains a white-balanced image for every input. In the second module, SMTNet takes the white-balanced image and regresses a per-pixel shading image and paper material image. The use of the white-balanced image allows us to train SMTNet in a self-supervised manner using a physical constraint, the chromatic consistency of the white-balanced and the reflectance image. Chromatic consistency states that the white-balanced output's chromaticity is same as true reflectance of the paper and the texture. We thus, eliminate the costly need to physically capture explicit shading and paper reflectance ground truth. Our method shows strong results on challenging illumination conditions and generalizes across different document types. The practical significance of the proposed method is demonstrated by a 21% decrease in OCR error rate when our shading removal is used as pre-processing for OCR.

In summary our contributions are: First, a unified learning-based architecture that directly estimates document reflectance based on an intrinsic image formation model, that generalizes across challenging illumination conditions for different types of documents as well as for warped documents.
Second, we created Doc3Dshade, a new dataset that clearly improves previous synthetic or hybrid ones, by physically acquiring a range of realistic color diverse multi-illuminant conditions and using them to synthesize a multitude of complexly illuminated document images. This dataset is uniquely customized to deal with documents in-the-wild.

2 Related Work

Color constancy(CC) is the visual phenomenon of perceiving the true object color, independently of ambient lighting. Computational color constancy aims to estimate the color of the image scene, used for white-balancing, i.e. a corrected version of the image under an achromatic illuminant. It is a standard pre-processing step in document analysis tasks [25]. Both statistics based methods[14, 42] and deep learning based methods [6, 17, 29] have used a uniform illumination model to solve the CC task. More recent methods focus on the multi-illuminant CC problem, working on pairs of images taken under different illuminants [18, 20] or using two-branch networks to describe local illuminants [38]. Our

Figure 2: Data Creation Pipeline: (a) Shows the hardware setup. The captured shapes are textured in Blender and combined with the captured shading image by Hadamard product ($\otimes$) to create I. The 'orange' and 'blue' arrows denote the rendering and the combining.

white-balancing network (WBNet) is a single image color constancy model targeted to document images.

Intrinsic image decomposition is an inverse optics formulation that can be helpful in removing shading and shadows effects in document images.Early work in this domain [3, 12, 16, 41, 46] is dominated by Retinex theory [26]. Later work added more physical priors on reflectance and shading [2, 5, 13, 37]. Recent supervised deep learning approaches [24, 33] have used end-to-end regression models to predict all the intrinsic modalities. More recently constrains such as physical intrinsic losses [4, 11, 39] and human judgments [35, 51] were imposed on the network to solve this problem. Unsupervised and semi-supervised models generally use multiple illumination varying sequences of the same scene to learn this decomposition [22, 27, 28, 31]. We only use single images and manage to train our decomposition framework using synthetic texture as a weak-supervision signal.

Shadow and shading removal of lighting variations has been widely studied. Most shadow removal methods, formulated for natural outdoor images, do not perform well on document images because of the specific characteristics of printed text. Earlier methods used an intrinsic approach to remove shading effects from the images [7, 49], document colour homogeneity [1], neighbouring pixel values [43], or combined visibility detection with existing state-of-the-art methods [25] or finally, applied diffusion equations on pixel intensity [23].

Self-supervised learning alleviates the need of large-scale labeled data. It exploits unlabeled image/video attributes, automatically generating pseudo labels for training. Self-supervised learning tasks include image colorization (an RGB image as the target and its grayscale version as the input) [50], image jigsaw puzzle (patches from one image as the input and their spatial relation as the target) [10], temporal order verification (frames from one video as the input and their temporal relation as the target) [45]. Drawing ideas from these prior methods, we train the shading decoder of SMTNet to produce a shading image consistent with the shading image estimated from the material decoder output.

3 Training Dataset: Doc3DShade

We collected a large-scale documents dataset, Doc3DShade, which combines diverse, realistic illumination scenarios with natural paper textures. Doc3DShade increases the physical correctness of Doc3D, a recent document dataset [9]. The contributions of Doc3DShade are two-fold; first, it captures physically accurate and realistic shading under complex illumination conditions, which is impossible to obtain using rendering engines (which use approximate illumination models to simulate light transport). Second, it contains various paper materials with different reflectance properties whereas a synthetically rendered dataset like Doc3D uses purely diffuse material to render images. Doc3DShade is thus physically more accurate and can be used to model scene illumination in a physically grounded way.

Capturing 3D Shape and Shading: Our capturing setup (figure 2 (a)) consists of a rotatory

platform, 8 directional lights and a depth camera. More details are in the supplementary material. To capture $3D$ shapes of documents, we randomly deformed textureless papers and placed them on the rotatory platform. The platform is randomly rotated and each document is captured with a random combination of single and multiple color directional lights. Light chromaticity was constrained to be around the Planckian locus [47] to simulate natural lighting conditions. We also captured each document under white lights to use it later as ground-truth for white balancing. To include various material properties in the shading images, we have used 9 diffuse paper materials that are used in various types of documents such as magazines, newspapers and printed papers, etc. In summary, we obtain the following data for each warped paper: a 3D point cloud and multiple images with shading under single and multiple lights of different color temperatures. An illustration of the captured shapes and corresponding shading images are shown in Fig. 2. Note that the captured shading (MS) map in this setup is a combined form of the paper material (M) and the shading (S) component.

Image Rendering: We create the 3D mesh for each point cloud following [15]. Each mesh is textured with a random document image and rendered with diffuse white material in Blender [8] (Fig. 2). In total, we have used $\sim$ 5000 textures collected from various documents such as magazines, books, flyers, etc. Each texture image is combined with a randomly selected single light shading image of the same mesh. To create more variability, we uniformly sample and linearly combine two single light shading images to simulate a fake multi-light shading image: $I = T \bigotimes (a.MS_1 + (1-a).MS_2)$, with $a \in [0,1]$. Additionally, for each image, we also render the white balanced image by combining the synthetic texture image and the shading image captured under white light. We have created 90K images with diffuse texture and white-balance images as ground-truths, for training and testing.

4 Proposed Method

Our reflectance estimation framework for document images captured under non-uniform lighting in a real-world scenario follows a two-step approach for illuminant and shading correction and leverages the physical properties of the image formation model. In the first step, we estimate a white-balanced image to neutralize the color of the scene illuminants. In the second step, we disentangle shading, texture, and material of the document, which allows us to obtain the shading-free image.

4.1 Image Formation Model.

We assume the scene is Lambertian, and illuminated by n light sources l. The color of each light source l_i is represented as a three channel vector(l_i^r, l_i^g, l_i^b). λ_i is the shading induced by the i-th light. Under the Lambertian assumption with no inter-reflections, image intensity I at pixel p can be modeled as:

$$I^c(p) = R^c(p) \sum_i \lambda_i(p) l_i^c \tag{1}$$

where $R^c(p)$ is the reflectance at pixel p and $c \in \{r,g,b\}$ denotes the color channel. In the case of documents we can assume the document reflectance is a combination of a *diffuse texture* component, i.e. the content of the documents and a *material* component, i.e. the color of the paper. Therefore we can rewrite the image formation model (Eq. 1) as:

$$I^c(p) = M^c(p) T^c(p) \sum_i \lambda_i(p) l_i^c \tag{2}$$

with M and T the reflectance of *material* and *texture* respectively. The shading-free image we want to recover is the product of M and T. T is available as ground-truth in Doc3DShade, but M is combined with the shading term thus not explicitly available as ground-truth.

4.2 Problem Formulation.

Considering the image model given in Eq. 2 we will first eliminate illuminant color effects through white-balancing and then disentangle material and texture.

We estimate a white-balanced image version I_{wb} of the input image. It has the same reflectance properties as the original images, given by Eq. 2, but under multiple achromatic lights, i.e. $l_i = (\eta_i, \eta_i, \eta_i)$. We follow similar formulation used by Hui et al. in [19]:

$$I^c_{wb}(p) = M^c(p)T^c(p)\sum_i \lambda_i(p)\eta_i \tag{3}$$

To obtain the white-balanced image, I_{wb}, we follow the formulation of [19], where I_{wb} is estimated in terms of a per-pixel white-balance kernel $WB(p) \in \mathcal{R}^3$ which corrects the color of the incident illumination at every single pixel, p, thus:

$$I^c_{wb}(p) = WB^c(p)I^c(p) \tag{4}$$

Since the shading λ_i is invariant to the light color, per pixel chromaticity[1] for I_{wb} is:

$$C^c_{wb}(p) = \frac{R^c(p)\sum_i \lambda_i(p)\eta_i}{\sum_c R^c(p)\sum_i \lambda_i(p)\eta_i} = \frac{R^c(p)}{\sum_c R^c(p)} \tag{5}$$

$C^c_{wb}(p)$ is identical with the chromaticity of the Reflectance component $R^c(p)$ which is the physical constraint used in our self-supervised loss. Eq. 5 can also be expressed in terms of the texture and material components as following:

$$C^c_R(p) = \frac{R^c(p)}{\sum_c R^c(p)} = \frac{M^c(p)T^c(p)}{\sum_c M^c(p)T^c(p)} \tag{6}$$

In what follows, we are going to use the derived physical constraints, $C^c_{wb} = C^c_R$, to estimate the intrinsic components of a document image given a single image I^c. Our approach is based on using two sub-networks- WBNet and SMTNet (Fig. 3). The first network WBNet estimates the white-balance kernel, $WB^c(p)$, whose chromaticity, C_{wb}, is used in the subsequent network, SMTNet, that in turn will estimate the shading, $\lambda_i(p)$ and the material, $M^c(p)$ in a self-supervised fashion. We choose to employ self-supervised training since separate ground-truths are not available for $\lambda_i(p)$ and $M^c(p)$.

4.3 WBNet: White-balance Kernel Estimation.

The first sub-network is designed to estimate the per-pixel white-balance kernel WB. Since the number of illuminants, their colors, and the individual shading terms are unknown, estimating the white-balanced image becomes an ill-posed problem given a single image. We treat this task as an image-to-image translation problem. Given an input color image, $I \in \mathcal{R}^{h\times w\times 3}$ the white-balance kernel, $WB \in \mathcal{R}^{h\times w\times 3}$ is estimated using a UNet [36] style encoder-decoder architecture with skip connections. The predicted white-balanced image, $\hat{I}_{wb}$ is then estimated using Eq. 4, by Hadamard product of $\hat{WB}$ and I. Note that although it is possible to directly estimate the white-balanced image $\hat{I}_{wb}$ from I, convolutional encoder-decoder architectures cannot preserve sharp details such as text. On the other hand, WB is a smooth per-pixel map that is much easier to learn in reality.

Loss Function: To train the WBNet we essentially use two loss terms, one on the predicted white-balance kernel ($\mathcal{L}_{wb}$) and another on the chromaticity of the estimated white-balanced

[1] Chromaticity of a pixel is given as: $C_r = r/(r+g+b), C_g = g/(r+g+b), C_b = b/(r+g+b)$

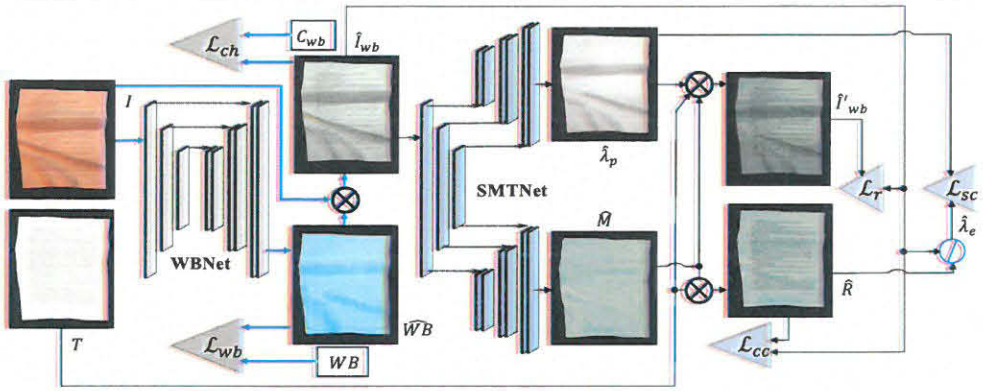

Figure 3: Proposed framework: The WBNet takes RGB image, I as input and produces the white-balance kernel ($\hat{WB}$). The white-balanced image, $\hat{I}_{wb}$ is then forwarded to the SMTNet which regresses the material ($\hat{M}$) and the shading, $\hat{\lambda}_p$. The $\otimes$ denote the Hadamard product, $(/)$ denote division and the triangles denote the loss functions.

image ($\mathcal{L}_{ch}$). As seen from Eq. 5, chromaticity is a shading invariant term, as is also the white-balance kernel, WB. Thus, it is intuitive to learn WB using a loss that is shading invariant. Therefore, we prefer to use the $\mathcal{L}_{ch}$ instead of a standard reconstruction loss on the predicted white-balanced image, $\hat{I}_{wb}$. We validate our choice in an ablation study in the supplementary material.

Additionally, we force intensity at each pixel to be preserved after white-balancing, which acts as a constraint on the predicted image. The combined loss is:

$$\mathcal{L}_{wbn} = \mathcal{L}_{wb}(\hat{WB}, WB) + \alpha_1 \mathcal{L}_{ch}(\hat{C}_{wb}, C_{wb}) + \alpha_2 \mathcal{L}_{int}(\hat{In}_{wb}, In) \tag{7}$$

where $\hat{WB}$, $\hat{C}_{wb}$, $\hat{In}_{wb}$ are the predicted white-balance kernel, chromaticity and intensity image of the predicted white-balanced image respectively. Corresponding ground-truths are available in our training dataset. Per-pixel intensity image is $\hat{In}_{wb}(p) = \sum_c \hat{I}^c_{wb}(p)$. The α's are the weights associated with each loss term. We use L_1 distance for each loss term. For $\mathcal{L}_{wb}$ and $\mathcal{L}_{ch}$ losses we use a mask to avoid pixels where ground-truth pixel values are zero.

4.4 SMTNet: Separating Material, Texture and Shading.

Given the estimated white-balanced image $\hat{I}_{wb}$ the second module, SMTNet estimates the material $\hat{M} \in \mathcal{R}^{h \times w \times 3}$ and shading $\hat{\lambda} = \sum_i \lambda_i \in \mathcal{R}^{h \times w \times 1}$. The structure of the network is illustrated in Fig. 3. The network consists of one encoder and two identical decoder branches, MNet and SNet. MNet and SNet regress the $\hat{M}$ and $\hat{\lambda}$ from the input image respectively. Ideally, it is possible to separately learn $\hat{M}$ and $\hat{\lambda}$ in a supervised manner if the corresponding ground-truths are available. But captured shading (MS) in our dataset is a combined representation of $M.\lambda$ which is further combined with the diffuse textures T to generate our training images $I = M.T.\lambda$ (Eq. 3). Therefore, we resort to a self-supervised approach to train SMTNet for disentangling M and λ given I, assuming T is available as ground-truth.

MNet: Given the predicted material image, $\hat{M}$ from the MNet branch, we can estimate the reflectance image as $\hat{R} = \hat{M}.T$ and estimate the shading image as $\hat{\lambda}_e = I/(\hat{M}.T)$. To train this branch, we use the consistency property of Eq. 5 and Eq. 6. If the predicted material $\hat{M}$, is correct the chromaticity of the $\hat{R}$ and chromaticity of the input white-balanced image should be the same. We use this chromatic consistency loss to train the MNet branch.

SNet: The other branch, SNet predicts the shading image $\hat{\lambda}_p$. To ensure the consistency of predicted material and predicted shading, $\hat{\lambda}_p$ should be equal to the estimated shading from the MNet branch, $\hat{\lambda}_e$. We use this shading consistency loss to train the SNet branch.

The SMTNet is a modified UNet architecture with a single encoder and two separate decoders. Decoders share encoded features but use different skip connections to the encoder.

Loss Function: The two primary loss functions used to train the SMTNet are the chromatic consistency ($\mathcal{L}_{cc}$) and the shading consistency ($\mathcal{L}_{sc}$) loss. Additionally, we use the reconstruction loss to ensure the predicted material image, $\hat{M}$ and shading image, $\hat{\lambda}_p$ reconstruct the input image when combined with the input diffuse texture, T. We also add smoothness constraints on the predicted material and shading. Specifically, the L_1 norm of the gradients for the material, $||\nabla\hat{M}||$ and L_1 norm of the second order gradients for the shading, $||\nabla^2\hat{\lambda}_p||$ are added as regularizers. Whereas the first term ensures piecewise smoothness of the $\hat{M}$, the second term ensures a smoothly changing shading map. The combined loss function is:

$$\mathcal{L}_{smt} = \mathcal{L}_{cc}(\hat{C}_{wb}, \hat{C}_R) + \beta_1\mathcal{L}_{sc}(\hat{\lambda}_p, \hat{\lambda}_e) + \beta_2\mathcal{L}_r(\hat{I'}_{wb}, \hat{I}_{wb}) + \beta_3||\nabla^2\hat{\lambda}_p|| + \beta_4||\nabla\hat{M}|| \quad (8)$$

Where $\hat{C}_{wb}$ and $\hat{C}_R$ are the chromaticities of input white-balanced image and estimated reflectance image. The $\hat{\lambda}_p$ and $\hat{\lambda}_e$ are the predicted and estimated shading. $\hat{I'}_{wb}$, $\hat{I}_{wb}$ are the reconstructed and input white-balanced image. The $\hat{I'}_{wb}$ is estimated as: $\hat{M}.T.\hat{\lambda}_p$, the hadamard product of predicted material image, shading image and input diffuse texture. The β's are the weights associated with each loss term. For each loss term we use the L1 loss.

5 Evaluation

For the quantitative and qualitative evaluation of our approach, we have experimented with multiple datasets that are captured under varying illumination conditions. Our evaluation contains three comparative studies. First, we show how the proposed method behave in intrinsic decomposition of document images. Second, we use the proposed approach as a post-processing step in a document unwarping pipeline [9] to show the practical applicability of our method in terms of OCR errors. Third, we show an extensive qualitative comparison with current state-of-the-art document shadow removal methods [1, 23, 25, 43]. From these experiments, we show a 21% improvement in OCR when used as a pre-processing step and also show strong qualitative performance in intrinsic document image decomposition and shadow removal tasks.

5.1 Intrinsic Document Image Decomposition.

In this section we evaluate the performance of our method in decomposing intrinsic light components on two datasets e.g., DocUNet benchmark [30] and MIT Intrinsics [5]. The DocUNet benchmark contains shading and shadow-free flatbed scanned version. Practically, scanned images are the best possible way to digitize a document and can be considered as the reflectance ground-truth. For this experiment, we apply our method on the benchmark images, then use the pre-computed unwarping maps from [9] to unwarp each image which aligns each image with its scanned reflectance ground-truth. The quantitative results for image quality before and after applying our approach is reported in table 1, in terms of the the unwarping metrics: multi-scale structural similarity (MS-SSIM) [44] and Local Distortion (LD) [48]. The qualitative results are shown in figure 5. Improvement of the MS-SSIM and LD are limited since these metrics are more influenced by the quality of the unwarping. Qualitative images show significant improvement over shading

DewarpNet results	**Image Quality**		**OCR Errors**	
	MS-SSIM	LD	CER	WER
with shading	0.4692	8.98	0.3136	0.4010
w/o shd [9]	0.4735	8.95	0.2692	0.3582
w/o shd (Ours)	**0.4792**	**8.74**	**0.2453**	**0.3325**

Table 1: Quantitative comparison of [9]'s unwarping quality on DocUNet benchmark dataset [30] when our proposed approach is applied as a pre-processing step before OCR.

removal applied in [9]. It is due to explicit modeling of the paper background and illumination, which enables our model to retain a consistent background color. Few additional qualitative results on real images captured under complex illumination are shown in figure4. Intermediate output of WBNet in figure4 demonstrates good generalization to real scenes.

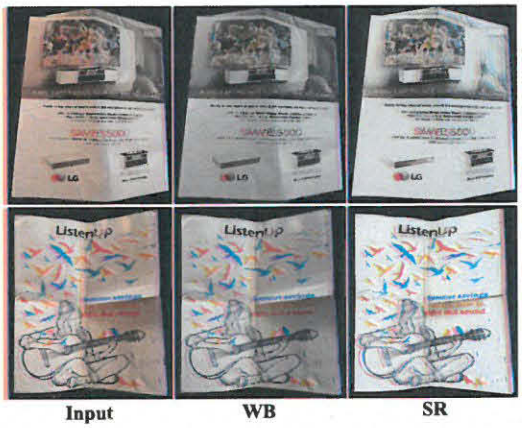

Figure 4: Qualitative results on real images after applying white balancing (After WB) and shading removal (After Shd. Rem.). Input images are non-uniformly illuminated with two lights.

Figure 5: Results on real-world images from [30]. [9]'s shading removal method fails to retain the background since shading, illumination and document background is modeled as a single modality.

Additionally, in figure 7 (a), we show that our method generalizes for 'paper-like' objects of MIT-Intrinsics [5], e.g., paper and teabag objects without any fine-tuning. Interestingly, IIW results show that it fails to preserve the text on the 'teabag', which proves general intrinsic methods are probably not suitable for document images and calls for further research attention to specially design intrinsic image methods for documents.

5.2 Pre-processing for OCR.

To demonstrate practical application of our approach, we compare OCR performance on the DocUNet benchmark images using word error rate (WER) and character error rate (CER) detailed in [9]. At first, the same unwarping step is applied as described in section 5.2. Then we perform OCR (Tesseract) on the DewarpNet OCR dataset before and after applying the shading removal and compare the proposed approach with the shading removal scheme presented in [9]. The results are reported in Table 1, where we can see a clear reduction of the error in character recognition that represents a 21% performance increase.

5.3 Shadow Removal & Non-Uniform Illumination.

Although our method is not explicitly trained for shadow removal tasks it shows competitive performance when compared to previous shadow removal approaches [1, 23, 25, 43]. We show the qualitative comparison in figure, 6 (a), 6 (b) and 6 (c) on the real benchmark datasets provided in the works [1], [23], and [43] respectively. Our method consistently performs well on different examples with soft shadows and shading but fails on hard shadow cases such as the image at the bottom row of 6 (c). Additionally, in figure 7 (b) we show the performance of our method in challenging multi-illuminant conditions with shadows and shadings available in a recent OCR dataset [32]. We can see the results of after our white-balancing (WBNet) in the WB column of 7 (b), and in column SMR we can see how SMTNet gracefully handles strong illumination conditions in real scenarios.

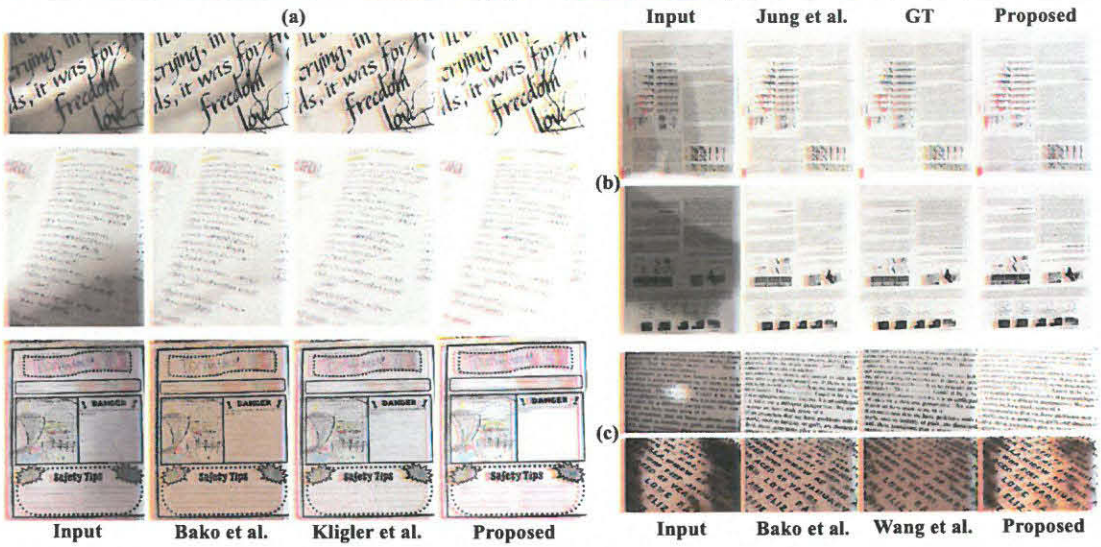

Figure 6: Comparison with existing shadow removal methods [1 ,23 ,25 ,43] on real image sets: (a), (b), (c) are obtained from [1 ,23 ,43] respectively. These comparisons show our method well generalizes on soft shadows. We report a fail case on hard shadows at the bottom row of (c).

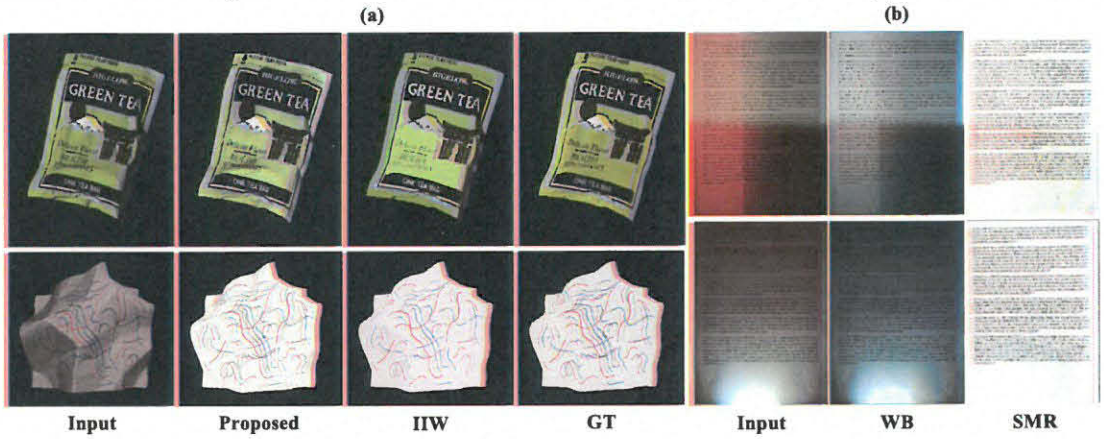

Figure 7: (a) Comparison with IIW method [5], IIW fails to accurately preserve text. (b) Results on multi-illuminant OCR dataset [32], WB is white-balanced image and SMR is output after removing material and shading.

6 Conclusions

In this work, we present a unified learning-based architecture that directly estimates document reflectance based on intrinsic image formation. We achieve compelling results in intrinsic image decomposition of documents and document shading removal under challenging non-uniform illumination conditions. Additionally, we contribute a large multi-illuminant document dataset that extends the public dataset Doc3D [9]. The primary limitation of our model is the absence of explicit modeling for shadows. Thus, Doc3DShade does not have many hard shadow examples and the few shadow instances are self-shadows of document shapes. Therefore our proposed model does not generalize well for arbitrary hard shadows. In future work, we will explicitly model shadows in our self-supervised architecture.

Acknowledgements: This work has been partially supported by project TIN2014-61068-R, FPI Predoctoral Grant (BES-2015-073722), project RTI2018-095645-B-C21 of Spanish Ministry of Economy, Competitiveness and the CERCA Programme / Generalitat de Catalunya, SAMSUNG Global Research Outreach (GRO) Program, NSF grants CNS-1718014, IIS-1763981, a gift from the Nvidia Corporation, a gift from Adobe Research, the Partner University Fund and the SUNY2020 ITSC.

References

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

[2] Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(8):1670–1687, 2015.

[3] Harry Barrow and J Tenenbaum. Recovering intrinsic scene characteristics. Comput. Vis. Syst, 2, 1978.

[4] Anil S. Baslamisli, Hoang-An Le, and Theo Gevers. CNN based learning using reflection and retinex models for intrinsic image decomposition. In Computer Vision and Pattern Recognition, 2018.

[5] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. ACM Trans. on Graphics (SIGGRAPH), 33(4), 2014.

[6] Simone Bianco, Claudio Cusano, and Raimondo Schettini. Color constancy using cnns. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 81–89, 2015.

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

[8] Blender Online Community. Blender - a 3D modelling and rendering package. www.blender.org.

[9] Sagnik Das, Ke Ma, Zhixin Shu, Dimitris Samaras, and Roy Shilkrot. Dewarpnet: Single-image document unwarping with stacked 3d and 2d regression networks. In The IEEE International Conference on Computer Vision (ICCV), October 2019.

[10] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE international conference on computer vision, pages 1422–1430, 2015.

[11] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David P. Wipf. Revisiting deep intrinsic image decompositions. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8944–8952, 2018.

[12] Brian Funt, Mark Drew, and Michael Brockington. Recovering shading from color images. In European Conference on Computer Vision, pages 124–132, 1992.

[13] Peter V. Gehler, Carsten Rother, Martin Kiefel, Lumin Zhang, and Bernhard Schölkopf. Recovering intrinsic images with a global sparsity prior on reflectance. In Neural Information Processing Systems, pages 765–773, 2011.

[14] Arjan Gijsenij and Theo Gevers. Color constancy using natural image statistics and scene semantics. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33 (4):687–698, 2010.

[15] B. Haefner, Y. Quéau, T. Möllenhoff, and D. Cremers. Fight ill-posedness with ill-posedness: Single-shot variational depth super-resolution from shading. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[16] Berthold KP Horn. Determining lightness from an image. Computer graphics and image processing, 3(4):277–299, 1974.

[17] Yuanming Hu, Baoyuan Wang, and Stephen Lin. Fc4: Fully convolutional color constancy with confidence-weighted pooling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4085–4094, 2017.

[18] Zhuo Hui, Aswin C Sankaranarayanan, Kalyan Sunkavalli, and Sunil Hadap. White balance under mixed illumination using flash photography. In 2016 IEEE International Conference on Computational Photography (ICCP), pages 1–10. IEEE, 2016.

[19] Zhuo Hui, Aswin C. Sankaranarayanan, Kalyan Sunkavalli, and Sunil Hadap. White balance under mixed illumination using flash photography. In IEEE Intl. Conf. Computational Photography (ICCP), 2016.

[20] Zhuo Hui, Ayan Chakrabarti, Kalyan Sunkavalli, and Aswin C Sankaranarayanan. Learning to separate multiple illuminants in a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3780–3789, 2019.

[21] 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, pages 1125–1134, 2017.

[22] Michael Janner, Jiajun Wu, Tejas D Kulkarni, Ilker Yildirim, and Josh Tenenbaum. Self-supervised intrinsic image decomposition. In Advances in Neural Information Processing Systems, pages 5936–5946, 2017.

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

[24] Seungryong Kim, Kihong Park, Kwanghoon Sohn, and Stephen Lin. Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In ECCV, 2016.

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

[26] Edwin H. Land and John McCann. Lightness and retinex theory. Journal of the Optical Society of America, 61(1):1–11, 1971.

[27] Louis Lettry, Kenneth Vanhoey, and Luc Van Gool. Unsupervised Deep Single-Image Intrinsic Decomposition using Illumination-Varying Image Sequences. Computer Graphics Forum (Proceedings of Pacific Graphics), 37(10), October 2018.

[28] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In 2018 CVPR, pages 9039–9048, 2018.

[29] Zhongyu Lou, Theo Gevers, Ninghang Hu, Marcel P Lucassen, et al. Color constancy by deep learning. In BMVC, pages 76–1, 2015.

[30] Ke Ma, Zhixin Shu, Xue Bai, Jue Wang, and Dimitris Samaras. Docunet: document image unwarping via a stacked u-net. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4700–4709, 2018.

[31] Wei-Chiu Ma, Hang Chu, Bolei Zhou, Raquel Urtasun, and Antonio Torralba. Single image intrinsic decomposition without a single intrinsic image. In ECCV, 2018.

[32] Hubert Michalak and Krzysztof Okarma. Robust combined binarization method of non-uniformly illuminated document images for alphanumerical character recognition. Sensors, 20(10):2914, 2020.

[33] Takuya Narihira, Michael Maire, and Stella X. Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In International Conference on Computer Vision (ICCV), 2015.

[34] Mamata Nayak and Ajit Kumar Nayak. Odia characters recognition by training tesseract ocr engine. International Journal of Computer Applications, 975:8887, 2014.

[35] Thomas Nestmeyer and Peter V. Gehler. Reflectance adaptive filtering improves intrinsic image estimation. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1771–1780, 2017.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015.

[37] Li Shen, Chuohao Yeo, and Binh-Son Hua. Intrinsic image decomposition using a sparse representation of reflectance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12):2904–2915, 2013.

[38] Wu Shi, Chen Change Loy, and Xiaoou Tang. Deep specialized network for illuminant estimation. In European Conference on Computer Vision, pages 371–387. Springer, 2016.

[39] Hassan A. Sial, Ramon Baldrich, and Maria Vanrell. Deep intrinsic decomposition trained on surreal scenes yet with realistic light effects. J. Opt. Soc. Am. A, 37(1): 1–15, Jan 2020.

[40] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. ACM Transactions on Graphics (Proceedings SIGGRAPH), 2019.

[41] Marshall F. Tappen, Edward H. Adelson, and William T. Freeman. Estimating intrinsic component images using non-linear regression. In IEEE Conference on Computer Vision and Pattern Recognition, pages 1992–1999, 2006.

[42] Joost Van De Weijer, Theo Gevers, and Arjan Gijsenij. Edge-based color constancy. IEEE Transactions on image processing, 16(9):2207–2214, 2007.

[43] Bingshu Wang and CL Philip Chen. An effective background estimation method for shadows removal of document images. In 2019 IEEE International Conference on Image Processing (ICIP), pages 3611–3615. IEEE, 2019.

[44] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, volume 2, pages 1398–1402. Ieee, 2003.

[45] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8052–8060, 2018.

[46] Yair Weiss. Deriving intrinsic images from image sequences. In International Conference on Computer Vision, pages 68–75, 2001.

[47] Gunther Wyszecki and W. Stiles. Color science: Concepts and methods, quantitative data and formulae, 2nd edition. Color Research & Application, 07 2000.

[48] Shaodi You, Yasuyuki Matsushita, Sudipta Sinha, Yusuke Bou, and Katsushi Ikeuchi. Multiview rectification of folded documents. IEEE transactions on pattern analysis and machine intelligence, 40(2):505–511, 2017.

[49] 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 Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), volume 2, pages 984–988. IEEE, 2007.

[50] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In European conference on computer vision, pages 649–666. Springer, 2016.

[51] Tinghui Zhou, Philipp Krähenbühl, and Alexei A. Efros. Learning data-driven reflectance priors for intrinsic image decomposition. 2015 IEEE International Conference on Computer Vision (ICCV), pages 3469–3477, 2015.