Template-guided Illumination Correction for Document Images with Imperfect Geometric Reconstruction

Felix Hertlein and Alexander Naumann

FZI Research Center for Information Technology
Haid-und-Neu-Str. 10-14, Karlsruhe 76131, Germany

hertlein@fzi.de, anaumann@fzi.de

Abstract

To facilitate the transition into the digital era, it is necessary to digitize printed documents such as forms and invoices. Due to the presence of diverse lighting conditions and geometric distortions in real-world photographs of documents, document image restoration typically consists of two stages: first, geometric unwarping to remove the displacement distortions and, second illumination correction to reinstate the original colors. In this work, we tackle the problem of illumination correction for document images and, thereby, enhance downstream tasks, such as text extraction and document archival. Despite the recent state-of-the-art improvements in geometric unwarping, the reliability of those models is limited. Hence, we aim to reduce lighting impurity under the assumption of imperfectly unwarped documents. To reduce the complexity of the task, we incorporate a-priori known visual cues in the form of template images, which offer additional information about the perfect lighting conditions. In this work, we present a novel approach for integrating prior visual cues in the form of document templates. Our extensive evaluation shows a 15.0 % relative improvement in LPIPS and 6.3 % in CER over the state-of-the-art. We made all code and data publicly available at https://felixhertlein.github.io/illtrtemplate.

1. Introduction

To facilitate the transition into the digital era, it is necessary to digitize printed documents such as forms and invoices. There are different approaches to accomplish this objective: through manual data entry employing human labor, by means of automated analysis of scanned documents, or by automated analysis of photographs captured via smartphones. The method using human labor may yield the highest level of accuracy; however, it is accompanied by substantial associated expenses. The process of analyzing scanned documents entails the requirement of specialized hardware such as a scanner, which imposes extra costs on a company. Additionally, this approach exhibits limited flexibility as it mandates physical presence at a location equipped with a scanner. Given the notable limitations associated with the first two methods, this work focuses specifically on the third approach.

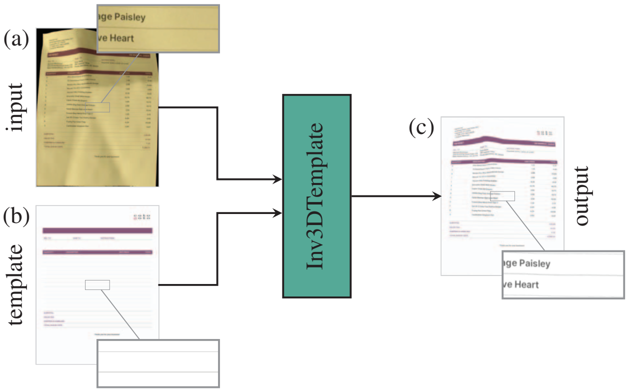


Figure 1. Our transformer-style model leverages templates (b) in addition to the input image (a) to achieve high quality results (c) for illumination correction in document images with imperfect geometric reconstruction. Our architecture IllTrTemplate extends state of the art IllTr [5].

Information retrieval and document enhancement from photographs depicting printed materials present several challenges compared to scanned documents. The capturing process introduces numerous external factors, including geometric deformations such as curls, creases, and perspective transformations, as well as illumination influences such as ambient light and shadows. The restoration process in this context is typically divided into two distinct stages. The first stage is geometric unwarping, which aims to accurately map all pixel locations to their correct positions. The second stage involves illumination correction, where lighting artifacts are removed to ensure optimal document quality.

In this work, we tackle the problem of illumination correction for information retrieval and document enhancement. An overview of our approach is provided in Figure 1. Given the imperfect results of state-of-the-art geometric unwarping models such as GeoTrTemplateLarge [8], we focus on partially unwarped documents to resemble the real-world application closely.

To reduce the complexity of the task, our model leverages a-priori information in the form of a template image [8], which provides valuable insights into the actual illumination conditions. A template is an RGB image that depicts the overall structure of a document while omitting its specific content. Although the requirement of a known document template for inference may impose certain limitations on the applicability, we argue that in many contexts, acquiring the document template is feasible. For instance, such a use case would be a manufacturer who maintains a list of suppliers with consistent invoice templates. A worker can select the correct supplier and implicitly the known template associated with the said supplier before photographing a document.

Our main contributions are as follows:

- We propose a novel approach for integrating prior visual cues in the form of document templates to reduce the complexity of the illumination correction task.
- We present a detailed evaluation of our model against state-of-the-art and achieve 15 % relative improvement in LPIPS and 6.3 % in CER.
- We present multiple ablation studies to gain deep insights into the proposed approach.

The paper is structured as follows: Section 2 provides an overview of related works in the field. The formal definition of the problem is presented in Section 3. Our proposed architecture is described in detail in Section 4, while Section 5 outlines the evaluation procedure. Section 6 presents the results, followed by the concluding remarks in Section 7.

2. Related Work

The process of document normalization typically involves two fundamental components: geometric unwarping and illumination correction. The latter can be subdivided into two categories: Document Image Binarization (DIB) and Document Image Enhancement (DIE). DIB maps all colors to a binary signal, effectively reducing the image to black and white representation. On the other hand, DIE aims to enhance the document image by attempting to restore the original colors.

Document Image Enhancement. Significant advancements have been made in the field of Document Image Enhancement (DIE) in recent times. Das et al. [3] introduced a refinement network based on a stacked U-Net architecture, as outlined in their work titled DewarpNet. The proposed network aims to predict a shading map, which is subsequently applied to the unwarped image using element-wise division. Li et al. [13] proposed a convolutional network with residual layers to directly infer the illumination-corrected image. In 2021, Feng et al. [5] introduced a transformer encoder-decoder structure for document image enhancement called IllTr. Their approach splits the input document image into a sequence of patches, which then are processed individually and afterward stitched together. Most recently, Xue et al. [29] proposed to remove illumination artefacts by transforming the input image and a blank paper to the Fourier space and then replacing the lower frequencies in the input image with the frequencies of the blank paper. The blank paper in this context functions as a template, providing a-priori information about the expected visual output.

Document Image Binarization. In the past, various works have been conducted on document image binarization [18, 27, 14, 1, 15, 2, 10, 24, 7, 25]. Early work was presented by Lu et al. [18] and Su et al. [27]. A diverse array of binarization algorithms was assessed by Lins et al. [14]. Almeida et al. [1] proposed an approach for image binarization inspired by Otsu's method [21] for pixel thresholding. In 2019, the International Conference on Document Analysis and Recognition (ICDAR) hosted a binarization competition that included a benchmark of thirty distinct binarization algorithms [15]. Calvo-Zaragoza and Gallego [2] presented a convolutional auto-encoder architecture. DE-GAN by Souibgui and Kessentini [26] uses conditional Generative Adversarial Networks for multiple document enhancement tasks. Kang et al. [10] proposed a document binarization method using cascading UNets to cope with the limited number of training images. Recently, Souibgui et al. [24] presented DocEnTr, an encoder-decoder architecture based on vision transformers without any convolutional layers. Gonwirat and Surinta [7] proposed DeblurGAN, a CNN-GAN hybrid, for enhancing noisy handwritten characters. Finally, Souibgui et al. [25] introduced Text-DIAE, a transformer-based model that utilizes self-supervised pretraining to enhance document binarization.

Geometric Dewarping. Ma et al. [20] published the seminal work DocUNet for geometric dewarping using deep neural networks. The authors proposed a stacked U-Net architecture to directly infer the unwarped document image. The study conducted by Feng et al. [5] presented the DocTr model, a transformer encoder-decoder architecture designed to infer the backward map. Recently, Feng et al. [4] further enhanced the model's performance by incorporating a hierarchical encoder-decoder framework. An iterative refinement approach was introduced by Feng et al. [6] in their publication on DocScanner. Jiang et al. [9] formulate the unwarping process as an optimization problem based on recognized shapes. Ma et al. [19] proposed a two-step approach by segmentation-based unwarping followed by a fine-grained texture module. Recently, Liu et al. [16] present an approach for self-supervised learning using an encoder-decoder structure. In their recent work, Hertlein et al. [8] introduced a novel approach that incorporates templates into the document unwarping process. They demonstrate the advantages of utilizing prior visual information for the specific task at hand.

In our work, we aim to leverage full document templates, as proposed by Hertlein et al. [8] for document image enhancement and leverage additional information about the anticipated visual output. Our research builds upon the partially dewarped documents generated by the model GeoTrTemplate [8]. Our illumination correction architecture is derived from the transformer-style model IllTr [5].

3. Problem definition

Given an imperfectly geometrically unwarped document image $\mathbf{I_{uwp}} := \mathbf{B}^*(\mathbf{W})$, where $\mathbf{B}^*$ is an imperfect backward mapping as defined in [8] and a warped document image $\mathbf{W}$, our goal is to find a mapping ϕ from $\mathbf{I_{uwp}}$ to $\mathbf{I_{ill}}$ such that all illumination effects are removed from the image:

$$\phi\colon\ \mathbb{R}^{\mathrm{h}\times\mathrm{w}\times 3} \to \mathbb{R}^{\mathrm{h}\times\mathrm{w}\times 3}$$
$$\mathbf{I_{uwp}} \mapsto \mathbf{I_{ill}}$$

More specifically, our model is supposed to predict the unwarped albedo map $\mathbf{I_{true}} := \mathbf{B}^*(\mathbf{A})$ for an albedo image $\mathbf{A}$ which corresponds pixel-wise to the warped image $\mathbf{W}$. Note that we want the model to learn solely the illumination correction task and not to complete the partial geometric unwarping. Therefore, we define the ground truth image as the partially unwarped albedo map $\mathbf{B}^*(\mathbf{A})$ instead of the perfect flat document.

To facilitate the learning task for our model, we leverage a-priori known information about the image structure given as a template image $\mathbf{T} \in \mathbb{R}^{\mathrm{h}\times\mathrm{w}\times 3}$. Formally, this is described as follows:

$$\phi_{\mathbf{T}}\colon\ \mathbb{R}^{\mathrm{h}\times\mathrm{w}\times 3} \times \mathbb{R}^{\mathrm{h}\times\mathrm{w}\times 3} \to \mathbb{R}^{\mathrm{h}\times\mathrm{w}\times 3}$$
$$(\mathbf{I_{uwp}}, \mathbf{T}) \mapsto \mathbf{I_{ill}}$$

See Figure 2 for a visualization of the illumination correction problem.

4. Architecture

We base our architecture on the state-of-the-art document image enhancement model IllTr [5]. It was published by Feng et al. [5] as part of the image dewarping and illumination correction model named DocTr. We briefly summarize the model architecture of IllTr in the following:

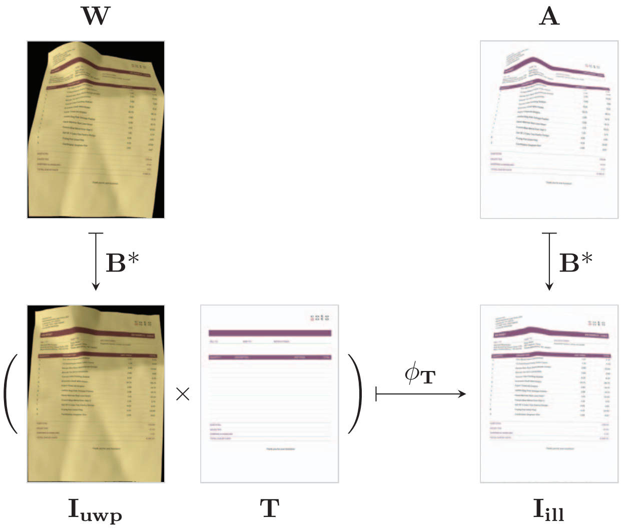


Figure 2. Visualization of the illumination correction problem.

Since the partially unwarped documents $\mathbf{I_{uwp}}$ contain high-frequency signals IllTr avoids scaling the input to a fixed size. Instead, the document is split into slightly overlapping patches with a fixed size of $p \times p$ pixels and processed individually before being stitched together afterward. Each patch $\mathbf{P_{img}} \in \mathbb{R}^{p\times p\times 3}$ is then preprocessed by a convolutional module called *Illumination Head*. This module extracts visual features from a single patch $\mathbf{P_{img}}$ by convoluting and downsampling. The resulting feature vector is then flattened into a sequence of tokens $f_i \in \mathbb{R}^{N\times c}$ with $c = 512$ and $N = \frac{p}{8} * \frac{p}{8}$. Using a transformer encoder-decoder structure, IllTr encodes the global relationship between the features f_i and generates global-aware representations before decoding them to a low-resolution prediction $f_j \in \mathbb{R}^{\frac{p}{8}\times\frac{p}{8}\times c}$. Finally, a learnable module called *Illumination Tail* upsamples the low-resolution features f_j to generate the final high-resolution patch prediction $f_k \in \mathbb{R}^{p\times p\times 3}$. For more details see the original work [5].

Our architecture processes the input image similarly to IllTr in patches of size $p \times p$ before stitching them together in the end. In contrast to the prior work, we have a-priori visual information in from templates $\mathbf{T}$ available. To exploit this information we propose two different variants:

1. The template $\mathbf{T}$ is scaled to a fixed size of $p \times p$ pixels. We refer to this template representation as $\mathbf{TP_{full}}$. Since $\mathbf{TP_{full}}$ is created using the full template, it contains all low-frequency signals but misses out on the fine-grained details.

2. We crop a window of $(p+2m)\times(p+2m)$ pixels from the template $\mathbf{T}$ for a margin of m pixels such that the $p \times p$ center region corresponds to the image patch region $\mathbf{P_{img}}$. We then scale the cropped window to our fixed size of $p\times p$ pixels. We refer to the scaled patch as

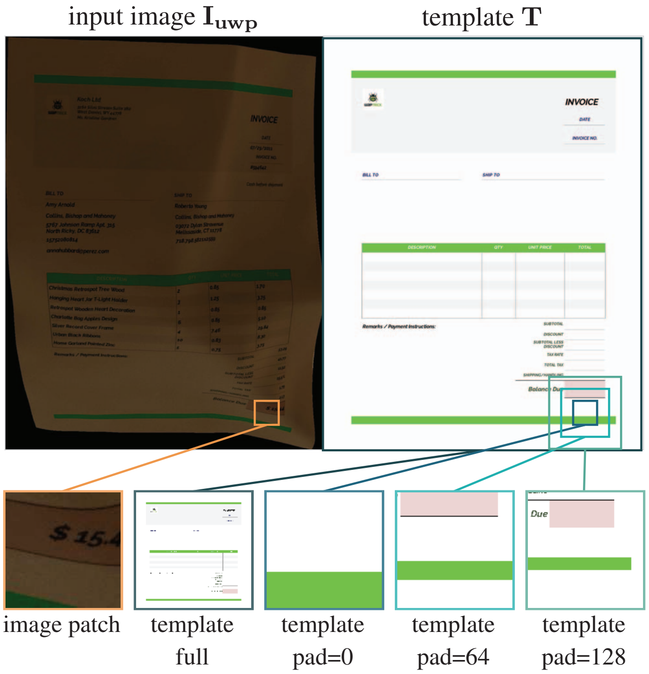


Figure 3. Visualization of a single image patch and the corresponding template patches for the different variants. For the padded crop variant, the graphic shows three template patches with margin m = {0, 64, 128}, respectively.

$\mathbf{TP_{pad=m}}$. This method of cropping captures the local template context for a given image patch. Since the alignment of $\mathbf{I_{uwp}}$ and $\mathbf{T}$ is not pixel-wise correct due to the imperfect unwarping $\mathbf{B}^*$, cropping at the exact same coordinates might not contain the relevant visual features for the given image patch. We introduced the margin m to tackle this problem.

Each variant encodes information about the visual template structure in a $p \times p$ patch referred to as $\mathbf{TP_x}$. See Figure 3 for a visualization of the patch extraction variants.

Given an image patch $\mathbf{P_{img}}$ and a template patch $\mathbf{TP_x}$, we apply one independent *Illumination Head* per patch which yields two sequences of features $f_i \in \mathbb{R}^{N \times c}$ and $t_i \in \mathbb{R}^{N \times c}$. We concatenate the sequences f_i and t_i to $c_i \in \mathbb{R}^{2N \times c}$, before applying the same encoder structure as IllTr. This allows the model to attend inbetween the image features, as well as cross relations between image and template features. That way the model is conceptually capable of integrating prior visual cues provided by the template image in the intermediate feature representation.

Before applying the decoder module as in IllTr, we discard half of the intermediate features from the encoder which correspond to the template features since we are only interested in a prediction for the image patch. Ultimately, we apply the *Illumination Tail* similar to IllTr to upsample the output features.

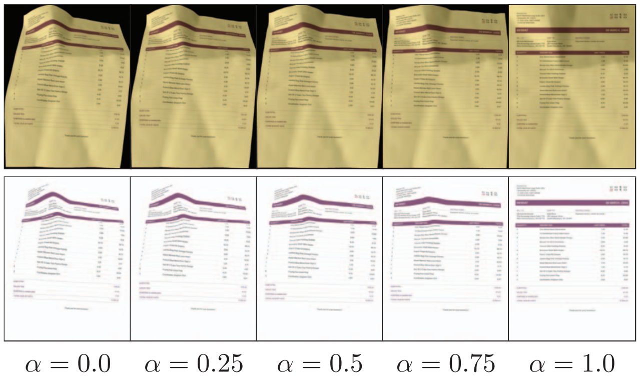


Figure 4. Depiction of the Inv3D dataset with partial unwarpings $\hat{\mathbf{B}}_\alpha$. The upper row shows the input images, while the lower row illustrates the corresponding ground truth illumination-corrected counterparts.

As for the loss function, we adhere to the original and combine the L1 loss with the perceptual loss, also known as VGG loss, [22] using a weighting factor λ.

$$\mathcal{L}_{total} = \mathcal{L}_1 + \lambda \mathcal{L}_{VGG} \quad (1)$$

5. Evaluation

In this section, we explain how we train and evaluate our model. First, we discuss the datasets we use, as described in Section 5.1. Then, we review the evaluation metrics we employ, which are explained in Section 5.2. Finally, we provide details on the implementation of our approach, as outlined in Section 5.3.

5.1. Datasets

5.1.1 Training Dataset

We train the models on the Inv3D dataset introduced by Hertlein et al. [8]. In Inv3D, each sample has its distinct template, leading to a one-to-one match of warped images and templates during training. Since the dataset contains fully warped document images and our target domain is partially unwarped document images, we apply the ground truth backward transformation $\hat{\mathbf{B}}$ partially to the warped image $\mathbf{W}$ and the warped albedo map $\mathbf{A}$. The parameter $\alpha \in [0, 1]$ scales the amplitude of the backward map. $\hat{\mathbf{B}}_0(\mathbf{W})$ corresponds to the input image $\mathbf{W}$ and $\hat{\mathbf{B}}_1(\mathbf{W})$ equates to the perfectly unwarped document image containing solely illumination effects. See Figure 4 for an example of various unwarping progressions. Note that we explicitly use random values α drawn from a uniform distribution between 0 and 1 during training to simulate the imperfect geometric unwarping from the preceding unwarping stage instead of using perfectly unwarped documents.

5.1.2 Evaluation Dataset

For evaluation, we use the real-world dataset Inv3DReal [8] and geometrically unwarp it using the state-of-the-art model GeoTrTemplateLarge [8]. This way, we can evaluate our models in a realistic setting as our models are intended to be applied after the geometric unwarping step. We refer to the unwarped dataset as Inv3DRealUnwarp in the following.

5.2. Evaluation Metrics

All metrics employed compare the model output $\mathbf{I_{ill}}$ with a reference image of identical resolution. For the synthetic evaluations, the reference image is the identically warped ground truth albedo image $\mathbf{I_{true}}$. Due to the absence of a ground truth backward map for the real-world dataset Inv3DRealUnwarp, we are unable to compare the model output against a pixel-aligned and perfectly illuminated image. Instead, we evaluate the model's performance by comparing it to the perfectly unwarped and illuminated image, which serves as the closest available approximation. All images have a resolution of 2200×1700 pixels.

We assess the models using four metrics, namely MS-SSIM [28], LPIPS [30], ED, and CER, which we describe below in detail.

MS-SSIM Multiscale structural similarity metric [28] is an established image similarity metric that calculates statistical properties at multiple scales and thus, ensures scale-invariance. The MS-SSIM score ranges from 0 to 1, with 1 indicating the highest attainable score.

LPIPS Learned perceptual image patch similarity (LPIPS) was introduced by Zhang et al. [30]. The authors train AlexNet [11] to learn image similarity based on human similarity perception. LPIPS scores range from 0 to infinity, with 0 representing the optimal score.

ED The edit distance (ED) is a text-based metric to measure the similarity between two texts. To retrieve the texts from our output and reference image, we apply Tesseract 4.0.0 [23] to each image independently. The ED is specified as the minimum changes (insertions, deletions, and substitutions) needed, to convert the output text into the reference text. This metric is also commonly denoted as the Levenshtein distance [12]. The optimal value is 0.

CER Character error rate (CER) is defined as the Levenshtein distance [12] over the total number of characters within the reference. The optimal value is 0.

5.3. Implementation Details

We attempt to keep the hyperparameters as close as possible to the original work IllTr [5]. The patch size p is set to 128 pixels and the overlap between two patches to 16 pixels similar to IllTr. The loss weight λ from Equation 1 is 10^{-5} and we set the batch size to 24. For training we employed the AdamW optimizer [17] with an initial learning rate of 10^{-4} and a StepLR scheduler [1] with a step size of 20 and a gamma of 0.3. Contrary to IllTr, we do not stop training after 35 epochs and continue it until there is no further improvement for 25 continuous epochs measured by the loss $\mathcal{L}_{total}$ in Equation 1 on the validation data split. To improve the resilience to different lighting variations, we employ a random color jitter during training with a random brightness, contrast, saturation, and hue.

6. Results

In this section, we present the results of our experiments. The sections 6.1 and 6.2 compare our models with the state-of-the-art quantitatively and qualitatively, respectively. Section 6.3 presents three ablation studies for detailed insights into the best performing model.

6.1. Quantitative Results

Table 1 lists the quantitative results of our approach in comparison to the state-of-the-art model IllTr [5] and the identity baseline. First of all, we observe that all models trained on Inv3D outperform the baseline in all metrics. The IllTr model trained on DocProj yields a lower MS-SSIM value than the baseline method and, thus, indicates a degradation in visual similarity. When comparing IllTr trained on DocProj [13] with the same model trained on Inv3D, it is apparent that the training on Inv3D is superior in all metrics. This could be attributed to the smaller domain gap for Inv3D to our evaluation dataset and the training process with partially unwarped documents. When comparing IllTr with our model IllTrTemplate, we find that IllTrTemplate surpasses IllTr in all variants and metrics. Within our four variants, there is no clear best model based on the set of all metrics. The visual metrics MS-SSIM and LPIPS indicate that a padding of 128 pixels works best, while the text metrics ED and CER favor a padding of 0 pixels as the most favorable choice. With an LPIPS of 0.221, the variant with 128 pixel padding achieves a 15 % relative improvement in contrast to the original model Inv3D trained on the same data. Thus, this model is recommended for document archival and retrieval. For the text metrics, the variant with 0 pixels padding achieves a relative improvement of 6.3 % for ED and CER and is therefore beneficial for the task of information extraction.

6.2. Qualitative Results

Figure 5 shows randomly selected images from the evaluation dataset Inv3dRealUnwarp. Looking at the illumination correction results, we observe that all models, IllTr and IllTrTemplate, generate patchy artefacts to some degree. More precisely, the individual patches do not always

[1] https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.StepLR.html

Model	Template	Train Dataset	↑MS-SSIM	↓LPIPS	↓ED	↓CER
Identity	—	—	0.711 (0.094)	0.324 (0.111)	329.1 (184.6)	0.512 (0.264)
IllTr [5]	—	DocProj [13]	0.651 (0.133)	0.306 (0.118)	306.4 (151.2)	0.477 (0.213)
IllTr [5]	—	Inv3D [8]	0.718 (0.154)	0.260 (0.095)	264.6 (161.0)	0.412 (0.229)
IllTrTemplate (ours)	full	Inv3D [8]	0.736 (0.109)	0.234 (0.086)	257.9 (159.7)	0.402 (0.227)
IllTrTemplate (ours)	pad=0	Inv3D [8]	0.731 (0.107)	0.231 (0.085)	**247.9 (156.3)**	**0.386 (0.221)**
IllTrTemplate (ours)	pad=64	Inv3D [8]	0.760 (0.144)	0.226 (0.082)	251.4 (161.7)	0.391 (0.226)
IllTrTemplate (ours)	pad=128	Inv3D [8]	**0.762 (0.137)**	**0.221 (0.082)**	251.3 (159.8)	0.392 (0.227)

Table 1. Evaluation of our model IllTrTemplate on the Inv3dRealUnwarp dataset. Values in brackets denote the standard deviation across all test samples.

agree on a common background color which leads to visible patches within the stitched image. This finding is likely due to the independent illumination correction for each patch before stitching them together in the end. The illumination corrections of IllTr trained with DocProj [13] (column (b)) compared to those trained on Inv3D show strong artefacts around the shadow borders, which indicates a lack of hard shadows in the DocProj dataset.

The comparison of IllTr and IllTrTemplate demonstrates the effectiveness of adding template information for color reconstruction. All IllTrTemplate variants seem to incorporate the original colors as provided by the templates in their illumination correction output. Thus, the reconstructed images appear to be more similar to the original. Note, since there are no ground truth backward mappings $\hat{\mathbf{B}}$ available for the real-world dataset Inv3DReal, the reference image does not contain any warping.

When considering the last two rows, we observe that all models struggle with removing the fine-grained creases. This is likely caused by the domain gap between the synthetically generated dataset Inv3D and the real-world evaluation dataset Inv3DReal.

6.3. Ablation Studies

In the following, we conduct a series of ablation studies. We consider only the model IllTrTemplate with a padding of 128 pixels as it is the best-performing model according to the LPIPS metric.

6.3.1 Ablation 1: Categorization

We split the Inv3dRealUnwarp dataset samples into their different categories depending on the type of document sheet modification and environment setting during recording. Table 2 shows the results for IllTrTemplate with 128 pixels padding trained on Inv3D and evaluated on Inv3dRealUnwarp. For the document modification type, we observe that *crumpleseasy* improves most according to all metrics. *Crumpleshard* seems to be the hardest modification type, which coincides with the qualitative findings that hard creases are not corrected properly. When considering the dataset split by environment setting, it becomes apparent that the majority of metrics, except for MS-SSIM, collectively affirm that the *color* environment is comparatively less challenging, whereas the *shadow* setting poses the greatest difficulty. The latter also aligns with the observations of the qualitative analysis, wherein the presence of harsh shadows resulted in the generation of more pronounced artifacts.

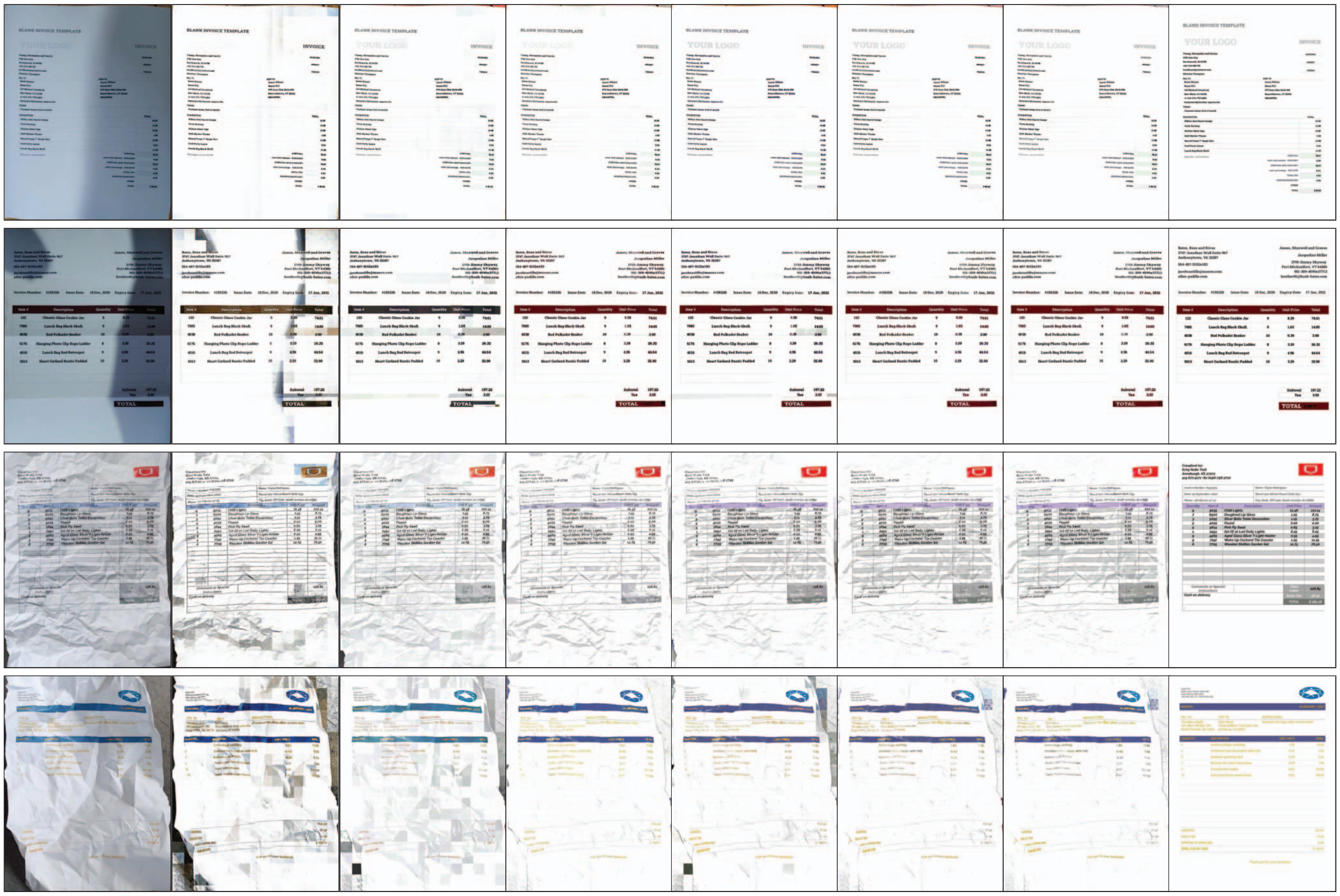
6.3.2 Ablation 2: Unwarping importance

To gain insights into the importance of the quality of the preceding geometric unwarping step on the illumination correction, we evaluate the test split of Inv3D with varying degrees of unwarping. See Figure 4 for an example of various unwarping progressions.

Table 3 shows the results of this ablation study. The absolute values of the visual metrics MS-SSIM and LPIPS exhibit a remarkable closeness to their respective optimum. This indicates the near-perfect illumination correction of the test split of Inv3D. Meanwhile, the text metrics ED and CER continue to exhibit considerably high values. This implies an imprecise reconstruction of high-frequent signals within the image since the fine-grained details are crucial for text recognition. In all metrics except for MS-SSIM, the best results were achieved using the perfect geometric unwarping with $\alpha = 1$. Since the visual metrics are already near their optimum, there is no steep decrease in performance when considering $\alpha < 1$. Note that the extremely high CER values for low α values are due to the poor OCR performance of Tesseract in the reference image $\mathbf{I}^{\alpha}_{\mathbf{true}}$.

6.3.3 Ablation 3: Error distribution

In a third ablation study, we examine the error distribution over all samples in the evaluation dataset Inv3dRealUnwarp. Since the imperfections in the geometric unwarping step affect the illumination step, we classified all input samples in three categories *extremely flawed*, *flawed*, and *moderate* depending on the severity of the un-



(a) Input (b) IllTr @ DocProj (c) IllTr @ Inv3d (d) IllTrTemplate Inv3d full (e) IllTrTemplate Inv3d pad=0 (f) IllTrTemplate Inv3d pad=64 (g) IllTrTemplate Inv3d pad=128 (h) Original

Figure 5. Qualitative results of state-of-the-art IllTr [5] and our model IllTrTemplate. The samples were drawn randomly from Inv3dRealUnwarp. The left column shows the input images $\mathbf{I_{uwp}}$. The rightmost column shows the optimal image $\hat{\mathbf{B}}(\mathbf{A})$. The center columns depict the illumination corrected images per model $\mathbf{B}^*(\mathbf{W})$.

Model	↑MS-SSIM	↓LPIPS	↓ED	↓CER
perspective	0.770 (0.153)	0.210 (0.089)	244.9 (163.9)	0.381 (0.227)
curled	0.768 (0.124)	0.199 (0.073)	239.9 (175.4)	0.380 (0.264)
fewfold	0.770 (0.139)	0.209 (0.070)	259.4 (170.4)	0.402 (0.233)
multifold	0.757 (0.151)	0.230 (0.088)	256.9 (157.5)	0.398 (0.221)
crumpleseasy	**0.797 (0.115)**	**0.190 (0.055)**	**225.5 (166.0)**	**0.352 (0.235)**
crumpleshard	0.711 (0.124)	0.289 (0.075)	281.2 (120.7)	0.440 (0.172)
bright	0.751 (0.142)	0.221 (0.088)	251.6 (163.9)	0.392 (0.230)
color	0.766 (0.137)	**0.213 (0.083)**	**229.1 (163.0)**	**0.355 (0.225)**
shadow	**0.770 (0.131)**	0.230 (0.075)	273.3 (150.5)	0.430 (0.222)

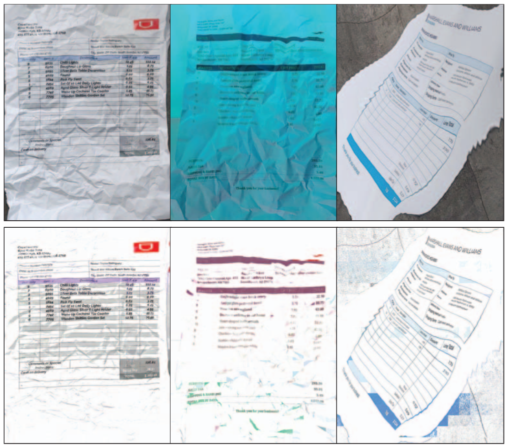
Table 2. Ablation 1: The Inv3dRealUnwarp dataset is partitioned into categories based on their respective modifications (upper part) and environment settings (lower part). The depicted results have been generated by our model IllTrTemplate with padding of 128 pixels. Values in brackets denote the standard deviation within each category.

warping errors. In the first category, substantial sections of the image remain uncovered by the document. In the *flawed* category, smaller areas are left without overlay. The last category includes all samples where, at minimum, the outline has been accurately mapped. Figure 6 (top row) shows one example per category.

Figure 7 plots the distribution of LPIPS values over the samples of Inv3dRealUnwarp after the illumination correc-

Unwarp factor α	↑MS-SSIM	↓LPIPS	↓ED	↓CER
0.0 (fully warped)	**0.992 (0.038)**	0.058 (0.019)	202.8 (149.2)	2.725 (17.091)
0.2	0.981 (0.016)	0.046 (0.019)	207.9 (145.4)	3.952 (21.332)
0.4	0.979 (0.015)	0.045 (0.020)	216.4 (137.4)	1.637 (7.349)
0.6	0.978 (0.014)	0.043 (0.021)	220.2 (150.2)	1.245 (7.544)
0.8	0.978 (0.014)	0.040 (0.022)	212.0 (158.8)	0.673 (3.270)
1.0 (fully unwarped)	0.982 (0.014)	**0.030 (0.026)**	**194.9 (171.9)**	**0.472 (1.104)**

Table 3. Ablation 2: We investigate the importance of the unwarp factor α. All results were obtained by our model IllTrTemplate with 128 pixel padding on a subset of 360 samples of Inv3DTest. Values in brackets denote the standard deviation across all test samples.



Moderate Flawed Severely flawed

Figure 6. Samples of Inv3dRealUnwarp with the highest LPIPS value per imperfection category for IllTrTemplate with 128 pixel padding. The upper row depicts input images, lower row shows the results after illumination correction.

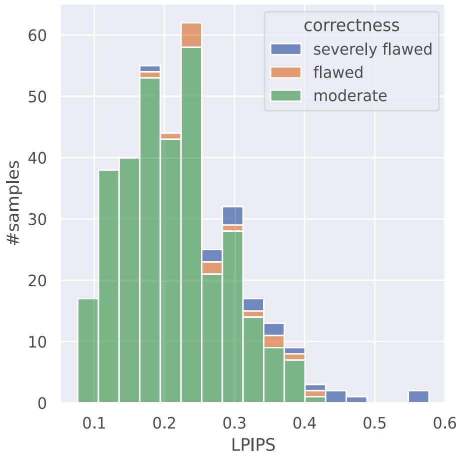


Figure 7. Distribution of the LPIPS error over all Inv3DRealUnwarp samples given by the IllTrTemplate model with a padding of 128 pixels. All samples are classified depending on the severity of the unwarping errors.

tion. The histogram shows that the highest metric values are given by severely flawed geometric unwarpings. Since our model solely corrects the illumination and not completes the partial geometric unwarping, this finding is to be expected.

7. Conclusion

In this work, we addressed the problem of illumination correction for documents with imperfect geometric reconstruction. We leveraged additional a-priori known visual cues in the form of templates to facilitate the task at hand. We presented two methods for incorporating the template information using a transformer encoder-decoder architecture. To evaluate the effectiveness of additional template images, we conducted a comparative analysis against the state-of-the-art model IllTr [5]. We assessed a total of four new template-based models, specifically the full template model and the cropped template models with paddings of 0, 64, and 128 pixels. We measured the performance using multiple metrics and observed a relative improvement of 15 % LPIPS and 6.3 % CER error compared to IllTr. Among the evaluated models, the best performing one was our IllTrTemplate with a padding of 128 pixels around the corresponding patch according to the LPIPS metric. This additional padding likely compensates for imperfections that may arise during the geometric unwarping stage, thus, capturing the relevant prior information at a high resolution. A series of ablation studies were conducted that revealed a domain gap between the synthetically generated dataset, Inv3D, and the evaluation dataset, Inv3DRealUnwarp.

For future research, our focus will be on developing an illumination correction model that maintains a global document representation throughout the processing of all patches. This approach aims to eliminate patch-like artifacts that may arise during the image processing. By incorporating a global document representation, we anticipate achieving smoother and more coherent results, enhancing the overall quality and visual appearance of the processed documents. Another objective of our future work is to improve the illumination correction specifically with a focus on text extraction. We plan on designing a text-based loss function to set the focus of the model to the fine-grained details such as characters. Lastly, addressing the domain gap could increase the applicability in real-world scenarios.

References

[1] Marcos Almeida, Rafael Dueire Lins, Rodrigo Bernardino, Darlisson Jesus, and Bruno Lima. A new binarization algorithm for historical documents. *Journal of Imaging*, 4(2):27, 2018. 2

[2] Jorge Calvo-Zaragoza and Antonio-Javier Gallego. A selectional auto-encoder approach for document image binarization. *Pattern Recognition*, 86:37–47, 2019. 2

[3] Sagnik Das, Ke Ma, Zhixin Shu, Dimitris Samaras, and Roy Shilkrot. Dewarpnet: Single-image document unwarping with stacked 3d and 2d regression networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 131–140, 2019. 2

[4] Hao Feng, Shaokai Liu, Jiajun Deng, Wengang Zhou, and Houqiang Li. Deep unrestricted document image rectification. *arXiv preprint arXiv:2304.08796*, 2023. 2

[5] Hao Feng, Yuechen Wang, Wengang Zhou, Jiajun Deng, and Houqiang Li. Doctr: Document image transformer for geometric unwarping and illumination correction. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 273–281, 2021. 1, 2, 3, 5, 6, 7, 8

[6] Hao Feng, Wengang Zhou, Jiajun Deng, Qi Tian, and Houqiang Li. Docscanner: robust document image rectification with progressive learning. *arXiv preprint arXiv:2110.14968*, 2021. 2

[7] Sarayut Gonwirat and Olarik Surinta. Deblurgan-cnn: effective image denoising and recognition for noisy handwritten characters. *IEEE Access*, 10:90133–90148, 2022. 2

[8] Felix Hertlein, Alexander Naumann, and Patrick Philipp. Inv3d: a high-resolution 3d invoice dataset for template-guided single-image document unwarping. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–12, 2023. 2, 3, 4, 5, 6

[9] Xiangwei Jiang, Rujiao Long, Nan Xue, Zhibo Yang, Cong Yao, and Gui-Song Xia. Revisiting document image dewarping by grid regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4543–4552, 2022. 2

[10] Seokjun Kang, Brian Kenji Iwana, and Seiichi Uchida. Complex image processing with less data—document image binarization by integrating multiple pre-trained u-net modules. *Pattern Recognition*, 109:107577, 2021. 2

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 5

[12] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966. 5

[13] 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)*, 38(6):1–11, 2019. 2, 5, 6

[14] Rafael Dueire Lins, Rodrigo Barros Bernardino, Darlisson Marinho de Jesus, and José Mário Oliveira. Binarizing document images acquired with portable cameras. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 6, pages 45–50. IEEE, 2017. 2

[15] Rafael Dueire Lins, Ergina Kavallieratou, Elisa Barney Smith, Rodrigo Barros Bernardino, and Darlisson Marinho de Jesus. Icdar 2019 time-quality binarization competition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1539–1546. IEEE, 2019. 2

[16] Shaokai Liu, Hao Feng, Wengang Zhou, Houqiang Li, Cong Liu, and Feng Wu. Docmae: Document image rectification via self-supervised representation learning. *arXiv preprint arXiv:2304.10341*, 2023. 3

[17] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 5

[18] Shijian Lu, Bolan Su, and Chew Lim Tan. Document image binarization using background estimation and stroke edges. *International Journal on Document Analysis and Recognition (IJDAR)*, 13:303–314, 2010. 2

[19] Ke Ma, Sagnik Das, Zhixin Shu, and Dimitris Samaras. Learning from documents in the wild to improve document unwarping. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 3

[20] 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. 2

[21] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979. 2

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[23] Ray Smith. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE, 2007. 5

[24] Mohamed Ali Souibgui, Sanket Biswas, Sana Khamekhem Jemni, Yousri Kessentini, Alicia Fornés, Josep Lladós, and Umapada Pal. Docentr: An end-to-end document image enhancement transformer. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1699–1705. IEEE, 2022. 2

[25] Mohamed Ali Souibgui, Sanket Biswas, Andres Mafla, Ali Furkan Biten, Alicia Fornés, Yousri Kessentini, Josep Lladós, Lluis Gomez, and Dimosthenis Karatzas. Text-diae: A self-supervised degradation invariant autoencoder for text recognition and document enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2330–2338, 2023. 2

[26] Mohamed Ali Souibgui and Yousri Kessentini. De-gan: A conditional generative adversarial network for document enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1180–1191, 2020. 2

[27] Bolan Su, Shijian Lu, and Chew Lim Tan. Robust document image binarization technique for degraded document images. *IEEE transactions on image processing*, 22(4):1408–1417, 2012. 2

[28] 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. 5

[29] Chuhui Xue, Zichen Tian, Fangneng Zhan, Shijian Lu, and Song Bai. Fourier document restoration for robust document dewarping and recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4573–4582, 2022. 2

[30] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5