SyntheticDoc: A Large Synthetic Dataset for Document Unwarping and Illumination Correction

Daniel Woortmann*[1], Tanguy Magne*[1], and Olga Sorkine-Hornung[1]

ETH Zurich, Department of Computer Science,
Universitätstrasse 6, 8092 Zurich, Switzerland
d.woortmann@gmail.com, {tanguy.magne,olga.sorkine}@inf.ethz.ch

Abstract. Deep learning models have become the standard tool for document rectification and illumination correction, yet their performance is fundamentally bound by their training data. For nearly a decade, the community has heavily relied on Doc3D, a pioneering but increasingly limited document unwarping dataset in terms of scale and quality. To address this bottleneck, we introduce SyntheticDoc, a massive, high-quality dataset designed to push the boundaries of document unwarping. SyntheticDoc is composed of 1,000,000 high-resolution procedurally generated training samples, alongside extensive validation and test sets. Each sample is paired with rich, pixel-perfect annotations, including UV maps, normal maps, albedo and shading. To ensure physical accuracy and photorealism, the paper geometries are generated via a physics-based simulator and rendered using a path tracer. To demonstrate the benefit of our dataset, we train a simple baseline model on SyntheticDoc and report on its performance in comparison to state-of-the-art methods on both document unwarping and illumination correction tasks. Our dataset is available at https://igl.ethz.ch/projects/SyntheticDoc/ and the code used to generate it at https://github.com/tanguymagne/SyntheticDoc.

Keywords: Document unwarping · Illumination correction · Dataset

1 Introduction

The digitization of physical documents has become an essential task in bridging the gap between legacy paper formats and modern digital workflows. Because of their ease of use and flexibility, mobile devices are increasingly utilized for this purpose. However, in contrast to flatbed scanners, which capture images of perfectly flat documents under controlled lighting, photos of documents taken with mobile devices such as smartphones suffer from multiple visual degradations. Since the paper is rarely perfectly planar and the camera viewpoint can vary significantly, geometric distortions are almost always present. In addition, under unconstrained and non-uniform lighting, these 3D deformations cause unwanted shadows and variable shading. These artifacts reduce the visual quality of the captured document and make its automatic processing significantly more difficult.

* Equal contribution.

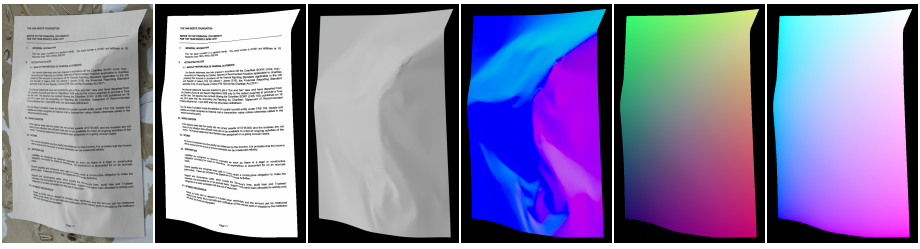

Fig. 1: Example of a sample from our SyntheticDoc dataset, with all its annotations.

To obtain a visually appealing and practically useful scan, the captured document needs to undergo geometric unwarping and illumination correction. Historically, document unwarping was tackled through model-based approaches, which attempt to estimate the document's 3D geometry using constrained representations and flatten it by solving an optimization problem. However, these methods are limited in the types of deformations they can handle. Deep learning-based approaches have become the standard way to solve both document unwarping and illumination correction. Yet, the performance of these techniques is fundamentally bounded by the quality and scale of their training datasets. Generating training data for these tasks is challenging, as networks require pixel-perfect annotations, such as mappings from the document photograph to the unwarped image or shading maps, which are nearly impossible to acquire for real-world photographs of documents. The most commonly used dataset, Doc3D [7], containing 100,000 samples, is synthetic, and relies on manually scanned 3D meshes, rendered at a relatively low image resolution (448×448 pixels). An alternative dataset, UVDoc [47] (20,000 samples), employs a hybrid approach, compositing document textures with real photographs of blank deformed paper, but it provides only coarse annotations for the unwarping mapping. Both datasets require extensive manual labor to create, making them inherently difficult to scale.

To overcome this scalability and quality bottleneck, we present SyntheticDoc, the first large-scale dataset for document unwarping and illumination correction. SyntheticDoc contains 1,000,000 training samples, 100,000 validation samples and more than 38,000 test samples. To achieve this unprecedented scale, we rely on rendering and use scalable processes at each step of the data generation pipeline. The deformed 3D document meshes are simulated using ArcSim [35, 36], a physical simulation engine for which we design multiple parameterized scenarios to ensure diverse outputs. The document textures and background materials are sourced from various repositories with permissive licenses. Finally, the samples are rendered using Blender [3], enabling the procedural generation of lighting environments, camera poses and paper material properties. Each sample is rendered at a high resolution (1024×1440 pixels) and includes pixel-perfect annotations, such as albedo, shading, normal maps, UV maps and 3D coordinates. A sample from our dataset, along with its annotations, is presented in Fig. 1.

To demonstrate the practical benefits of our dataset, we train a lightweight baseline model to simultaneously perform document unwarping and illumination correction. We evalute the inference speed and performance of this model on the standard DocUNet benchmark [31], comparing it to the state-of-the-art.

2 Related works

2.1 Document unwarping

Early model-based approaches to document unwarping rely on a two-step process: estimating the 3D document surface via specialized hardware [4, 5, 34, 61], multi-view imagery [21, 45, 56] or visual cues [33, 42], followed by simulated or geometric flattening [4, 5, 21, 26, 34, 44, 56, 61]. However, data-driven approaches have largely replaced these methods due to their superior generalization capabilities. Ma et al. [31] introduce the first deep learning-based unwarping method, while DewarpNet [7] proposes an architecture based on 3D coordinates, inspired by classical techniques. Following these foundational works, the field innovates rapidly. Several works explore patch-based [9, 24] and iterative [58] architectures, integrate text-line supervision [15, 19, 23] and employ transformers [13, 64] to capture global context. Other works focus on predicting coarse mappings [47, 53, 54], adopting multi-task frameworks to handle various document processing steps [41, 59, 63] and, more recently, leveraging diffusion models [22, 62]. While these deep learning methods show strong performance, their effectiveness strongly depends on the scale and quality of the training datasets, which motives us to develop our SyntheticDoc dataset.

Datasets. Since modern models typically predict a dense backward mapping from the warped input to the flat document, their training datasets require complex, pixel-perfect annotations that are difficult to obtain for real-world photographs. Earlier methods rely on a synthetic dataset generated via non-physically plausible 2D deformations [31], later augmented with supplementary

Table 1: Comparison of the various document unwarping datasets. *3D geometry source* characterizes the type of geometry used for rendering in a synthetic dataset, *Mapping* represents the kind of mapping annotations and *3D* signifies the inclusion of 3D annotations.

Dataset	# Samples	Resolution	Type	3D geometry source	Mapping	3D
DocUNet [31]	100,000	600 × 800	Synthetic 2D	N/A	Dense	✗
Doc3D [7]	100,000	448 × 448	Synthetic	3D scanned	Dense	✓
DocProj [24]	2,450	1800 × 2400	Synthetic	Geometric deformations	Dense	✗
DIW [30]	5,000	512 × 512	Real	N/A	—	✗
Inv3D [17]	25,000	1600 × 1600	Synthetic	3D scanned	Dense	✓
Book3D [28]	56,000	1200 × 800	Synthetic	Geometric deformations	Dense	✓
UVDoc [47]	20,000	488 × 712	Pseudo-real	Depth camera	Coarse	✓
SyntheticDoc	1,000,000	1024 × 1440	Synthetic	Physics simulation	Dense	✓

Fig. 2: Examples of data points from various document unwarping datasets.

annotations [52, 53]. For years, Doc3D [7] has served as the standard dataset for document unwarping. It has been extended with additional ground-truth labels, such as text lines [15, 32] or cropped images [12]. While rendered in Blender, similar to our SyntheticDoc, Doc3D contains far fewer samples at a significantly lower resolution. Moreover, its reliance on manual 3D scanning is not scalable. The captured 3D scans required geometric post-processing that resulted in over-smoothing, removing fine-grained details and in some cases making the surfaces non-developable. DocProj [24] uses an approach similar to ours but features simplistic geometries obtained by geometrically deforming a mesh and a very limited size. UVDoc [47] prioritizes visual and geometric realism, being the only non-rendered dataset providing mapping annotations. Inv3D [17] and Book3D [28] focus on certain document types and remain severely limited in scale and/or deformation variety. Finally, datasets of real-world documents [30] only provide the flatbed scan as ground truth, making them impractical for training mapping-based networks. As summarized in Tab. 1, SyntheticDoc stands out thanks to its massive scale and high image quality. It also contains the largest variety of document types and deformations and, as illustrated in Fig. 2, the rendering of SyntheticDoc is more realistic than previous datasets, with shadows correctly cast on the background.

2.2 Illumination correction

Beyond geometric unwarping, document restoration encompasses photometric corrections to convert camera-captured document images into clean, flatbed-style scans. In this context, deshadowing and illumination correction are two

Table 2: Comparison of the various illumination correction datasets. For RealDAE, the resolution ranges from 398×164 to 5344×5312 pixels.

Dataset	# Samples	Resolution	Type	3D geometry source	Albedo	Shading
Doc3D [7]	100,000	448×448	Synthetic	3D scanned	✓	✗
DocProj [24]	2,450	1800×2400	Synthetic	Geometric deformations	✓	✗
Doc3DShade [8]	90,000	640×480	Synthetic	3D scanned	✓	✗
RealDAE [57]	450	Various	Real	N/A	✓	✗
SyntheticDoc	1,000,000	1024×1440	Synthetic	Physical simulation	✓	✓

distinct but frequently confounded tasks. Deshadowing removes shadows cast on the document by external objects [25, 27], while illumination correction rectifies shading. Since SyntheticDoc focuses on the latter, we review the relevant illumination correction literature here.

Classical methods rely on estimating a shading map by interpolating background border colors [6] or inpainting document content [60]. Modern data-driven approaches largely outperform these techniques. Initial deep learning methods address the problem sequentially, correcting colorimetry only after rectifying the document geometry [7, 13, 24]. Subsequent works focus specifically on dedicated illumination networks [8], utilizing coarse-to-fine strategies [57] as well as generative adversarial networks (GANs) combined with either cycle consistency [50] or vision-language priors [39] like CLIP [40]. More recently, multi-task approaches simultaneously solving unwarping and illumination correction have emerged [41, 49]. The latest unified document enhancement models tackle all forms of document degradation concurrently, such as deshadowing, deblurring and illumination correction, by leveraging advanced feature generators [59] or diffusion modules [63].

Datasets. Since illumination correction has received considerably less attention than geometric unwarping, dedicated training datasets are rare. Consequently, many deep learning-based illumination correction methods rely on the warped albedo or original document textures provided by unwarping datasets such as Doc3D [7] and DocProj [24]. While a few datasets target this specific task, they exhibit significant limitations. Doc3DShade [8] follows a capture procedure similar to Doc3D [7] but additionally captures real-world shading. However, it inherits Doc3D's scalability issues and contains very few distinct document textures and geometries, with the different samples varying mostly in the applied shading. RealDAE [57] is the only real-world dataset for this task. Its ground truth images were obtained by manually correcting the deformed images using Adobe Photoshop, making the collection process difficult and restricting its size to just 450 training samples. Unlike SyntheticDoc, none of these datasets provide explicit shading annotations, offering only the albedo (see Tab. 2). In addition, SyntheticDoc is created with over 500,000 distinct geometries and an equal number of unique document images, all rendered using a high-quality path tracer and diverse lighting configurations for enhanced realism.

Fig. 3: Meshes generated by our physics based simulation pipeline using the different scenarios.

3 SyntheticDoc dataset generation pipeline

The generation of the SyntheticDoc dataset consists of two main steps. First, the 3D geometries representing the deformed sheets of paper are simulated using ArcSim [35,36], a simulation engine specialized for sheets of deformable materials. Then, the resulting meshes are rendered using Blender [3] to produce the final photorealistic images along with their corresponding ground-truth annotations.

3.1 Warped paper meshes generation

The paper mesh, which represents the geometry of the document, is the most crucial component of each sample. It defines both the distortion in the document and its shading and cast shadows. To generate a dataset that is an order of magnitude larger than existing ones, we need a scalable approach to obtain these meshes. Prior methods based on 3D scanning [7] or depth capture [47] require manual work and are inherently unscalable.

To overcome this challenge, we rely on physical simulation. We utilize ArcSim [35, 36], a powerful adaptive simulator designed for sheets of deformable materials. While capable of simulating fabrics and plastics, it is particularly well-suited for paper. Paper exhibits a specific physical characteristic that makes it difficult to simulate: it does not stretch. ArcSim enforces this constraint, simulating paper while preventing unnatural stretching. In addition, regardless of its deformation, a sheet of paper mathematically forms a developable surface; ArcSim preserves this geometric property throughout the simulation.

To generate realistic meshes that accurately represent real-world document deformations, we design three distinct simulation scenarios. Each scenario is

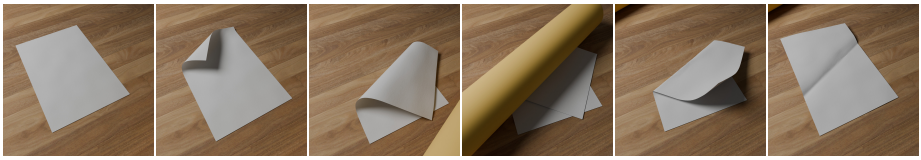

Fig. 4: The process of simulating a folded document. Starting from a flat mesh, a boundary vertex is pulled over another one. A roller simulates a hand flattening the crease, after which the mesh is unfolded. Note that the material textures shown here are for aesthetic purposes only.

initialized with a flat, rectangular A4-sized mesh, a common format for documents. Examples of the resulting deformed document meshes are presented in Fig. 3.

Pull scenario. The first scenario is straightforward and involves pulling on specific vertices of the original flat mesh. One or two vertices are randomly selected and displaced vertically to a random height. These control points can lie on the boundaries of the paper, mimicking a person picking up the document by its edge, or within its interior, which simulates pinching the paper. As only one or two vertices are displaced, this scenario predominantly produces smoothly curved documents, but it can still generate sharper features when lifting a central vertex. Meshes obtained with this scenario are presented in the first column of Fig. 3.

Fold scenario. The second scenario is designed to simulate folded documents by mimicking the physical folding process. To that end, two boundary vertices are selected, and one is pulled over the other. Then, a rigid cylinder is rolled over the resulting mesh to flatten it and form a crease. Finally, the paper is unfolded by returning the displaced vertex back to its original position. This process, illustrated in Fig. 4, creates convincing geometries with one fold. Examples of the resulting meshes are shown in the third column of Fig. 3.

Fall scenario. The last scenario, while the simplest, is capable of generating the largest variety of outputs. In this setup, we simulate the document falling onto various 3D primitives under the influence of gravity. The outputs vary widely depending on the shape of the collider and the gravitational acceleration. We use three types of colliders: spheres, cylinders and gables (see Fig. 5). Dropping the paper onto a cylinder produces a smoothly curved surface (Fig. 3, second column). When falling onto a gable, the mesh contains much sharper, fold-like deformations (Fig. 3, fourth column). Draping the document over a sphere forces the generation of complex wrinkles, as the sphere's non-zero Gaussian curvature geometrically conflicts

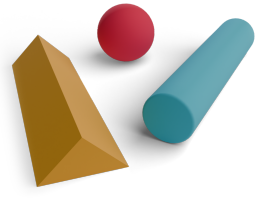

Fig. 5: Primitives used for the fall scenario.

with the developable nature of the paper (Fig. 3, fifth column). Lastly, letting the paper fall onto multiple small spheres creates many creases, simulating a heavily crumpled document (Fig. 3, last column). In all these simulations, the scale, position and orientation of the primitives are randomized. In the last two cases, we apply highly randomized gravitational acceleration to create various crease patterns.

Together, these simulation scenarios enable the generation of curved (*Pull*, *Fall on cylinder*), folded (*Fold*, *Fall on gable*) and crumpled (*Fall on sphere*, *Fall on multiple spheres*) documents. For each of these six configurations, we generate nearly 50,000 unique meshes. Each mesh can be flipped to simulate viewing from the opposite side of the paper sheet. This simple operation produces vastly different appearance, ultimately resulting in a total of 584,708 distinct deformed document geometries.

3.2 Rendering

To generate high-quality samples for the final dataset, we import the warped paper meshes from the previous step into Blender and set up a unique scene configuration for each sample. The scene setup involves positioning the virtual camera, placing explicit light sources, defining a background surface for the paper to lie on and applying a physically based paper material paired with a document texture to the paper mesh. This scene is then rendered using Cycles, Blender's path-tracing engine [3]. We opt for the more computationally expensive path tracing over rasterization to produce physically realistic images, which is especially important to accurately capture the self-shadowing and specular highlights of warped documents. We render the images and their corresponding ground truths at a resolution of 1024×1440 pixels with 128 render samples per pixel as a tradeoff between scalability and visual quality.

Camera. For each sample, we select a camera position from a set of valid viewing angles. An angle is considered valid if the document is fully contained within the camera's field of view and all mesh face normals are consistently front-facing. This ensures that the entire document is visible and completely free of self-occlusion. The candidate angles are constrained to a moderate inclination range around the top-down view of the paper, closely reflecting real-world capture conditions for document images.

Lighting. We use four predefined light setups, with randomized parameters, to cover a wide range of real-world conditions. Each setup can be globally rotated around the mesh, and the color temperature of the lights can be adjusted. The four setups range from controlled studio environments to natural daylight (see Fig. 6): *3-point lighting* mimics a classic studio setup with key-, fill- and backlight, producing neutral illumination with soft, controlled shadows; *Softbox lighting* utilizes several large area lights that cast very soft shadows, as is common in photography or indoor settings; *Rim lighting* employs strong backlights on either

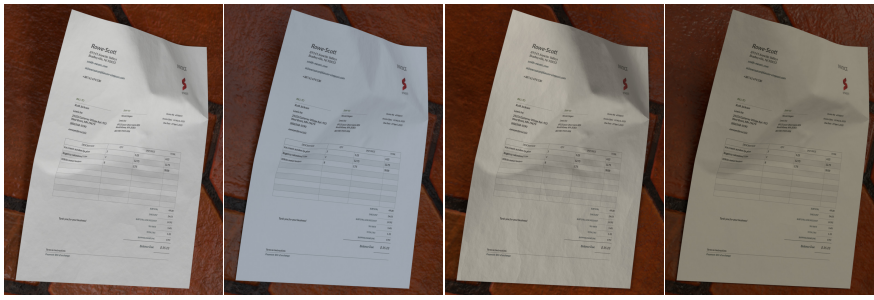

Fig. 6: Various lighting setups. From left to right: *neutral 3-point lighting*, *cool softbox lighting*, *warm rim lighting* and *warm natural lighting*.

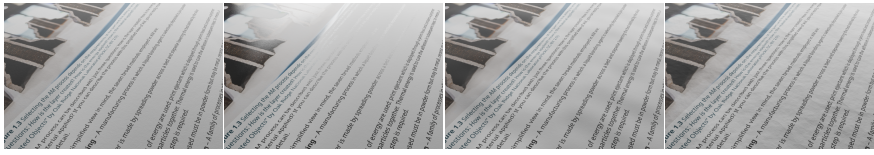

Fig. 7: Variation in the paper texture (zoomed view to highlight the material texture). From left to right: basic grainy paper texture, paper texture with glossy ink, slightly wavy paper texture, creased paper texture.

side of the document, creating sharp edge highlights and hard shadows; *Natural lighting* uses strong directional sunlight to simulate daylight, replicating window-lit or outdoor environments. We use these explicit light sources rather than image-based lighting (i.e., illumination derived from HDR environment maps) to retain more control over the scene and to generate more pronounced specular highlights and shadows on the paper surface.

Background material. The background of the rendered image for each sample is a simple plane with a PBR (physically based rendering) material. The textures are sourced from the MatSynth dataset [46], which contains 5,789 high-quality PBR materials, including those captured by Deschaintre et al. [10], all under CC0 or CC-BY licenses. To maximize the diversity within the backgrounds of the SyntheticDoc dataset, we uniformly sample materials and randomize their rotation, scale and spatial offset parameters. Furthermore, to guarantee that the document rests naturally on the background plane and casts physically accurate shadows, we run a simple rigid-body simulation before rendering.

Procedural paper material. To make the rendered images as realistic as possible and mitigate the synthetic-to-real domain gap, we develop a highly parameterized procedural paper material using Blender's shader node system [3]. The surface imperfections are driven by a composite height map, which is subsequently converted into a normal map. This height map is created by combining three procedural components: a high-frequency Perlin noise [38] to simulate the

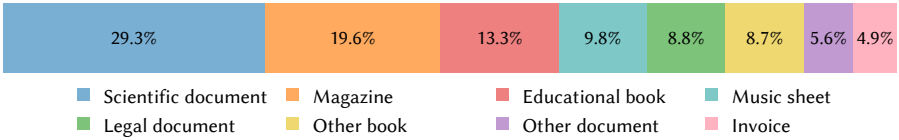

Fig. 8: Proportion of each type of document in our SyntheticDoc dataset.

microscopic fibers of the paper with a grainy texture, a low-frequency Perlin noise to model the macroscopic waving of the paper and a Worley noise [51] to generate micro-creases. In addition, we simulate glossy ink by modulating the surface roughness based on the document image texture, which also provides the albedo. The scale and strength of each component can be adjusted, making the paper material highly versatile. Examples of the resulting paper textures are presented in Fig. 7.

Documents. To make the SyntheticDoc dataset as comprehensive as possible, we render samples using a highly diverse set of document textures. In total, we collect 510,903 portrait-format document pages distributed across eight categories (see Fig. 8). We gather documents from various sources, all under permissive licenses to guarantee broad usability. Scientific papers are drawn from various arXiv [1] categories and are filtered for CC-BY licensing. Educational materials are sourced from OpenStax [37], while general books spanning multiple languages are obtained from the Directory of Open Access Books (DOAB) [11]. We also incorporate sheet music from IMSLP [18], enterprise documents from RealKIE [43] and procedurally generated invoices from Inv3D [17]. Additional documents, including math problems, diagrams and charts, are extracted from the CoSyn-400K dataset [55]. Finally, due to the scarcity of freely available magazine pages, we follow the approach of Verhoeven et al. [47] and use a text-to-image model to generate these. Specifically, we first create a templated prompt that we fill with magazine-specific attributes, refine it with a large language model (Gemini 3 Flash [16]) and generate the final image using FLUX.2-klein-9B [2]. While the text in these synthetic images may be illegible, the layout and structure of the documents closely match those of real magazines.

Additional renders. Alongside the main rendered image, we generate a comprehensive set of ground-truth annotations for each sample. To cleanly isolate the document, we first assign a black emissive material to the background plane. We capture the shading map by rendering the scene with identical parameters but removing the document's texture. The remaining annotations are generated by applying an unlit, emissive shader to the paper mesh and successively rendering the unshaded document albedo, surface normals, 3D world coordinates and UV maps. A sample from the dataset with all its annotations is presented in Fig. 1. Modern networks are commonly trained to predict a dense backward mapping from the distorted input to the flat document. This mapping can be easily obtained by inverting the provided UV map.

Fig. 9: Samples from our SyntheticDoc dataset. SyntheticDoc is composed of various document types (from left to right: book pages, invoices, legal documents, magazines, music sheets and scientific papers), with various deformations and lighting setups.

Each of the 584,708 generated meshes is rendered twice. Some meshes are discarded because they create self-occlusions across all sampled viewpoints. Each document texture is reused no more than three times. Through this highly scalable pipeline, we generate the SyntheticDoc dataset, comprising 1,000,000 training samples, 100,000 validation samples and over 38,000 test samples. A diverse selection of training images is presented in Fig. 9.

The simulation and rendering scripts, all assets used to create SyntheticDoc and the dataset itself will be made publicly available upon publication, allowing anyone to easily use and extend our work.

4 Experiments

To showcase the practical utility of SyntheticDoc, we train a lightweight baseline model that jointly tackles document unwarping and illumination correction. The network is adapted from the one presented in UVDoc [47]. We employ the same architecture, simply swapping the 3D grid prediction head with a shading prediction module. The shading head uses a U-Net style decoder that concatenates encoder features via skip connections at each resolution stage to obtain higher precision results, especially along document boundaries. As a result, our network outputs both a coarse unwarping map and a shading image, which can be applied

Table 3: Quantitative comparison of the unwarping performance of various methods on the DocUNet benchmark. The first row reports the metrics for the original warped documents. The last row presents our lightweight baseline model trained exclusively on our SyntheticDoc dataset. Gold, silver and bronze backgrounds indicate the best, second-best and third-best scores, respectively.

Method	MS-SSIM ↑	LD ‘ ↓	AD ↓	CER ↓	ED ↓
Warped image	0.247	20.53	1.006	0.517	2029
DewarpNet [7]	0.472	8.38	0.395	0.216	828
DisplacementFlow [52]	0.432	7.62	0.395	0.291	1207
DDControlPoints [53]	0.473	8.93	0.423	0.272	1102
DocTr [13]	0.509	7.78	0.366	0.180	713
PieceWise [9]	0.490	8.65	0.430	0.247	977
FDRNet [54]	0.543	8.08	0.396	0.215	876
RDGR [19]	0.495	8.50	0.432	0.170	725
Marior [58]	0.476	7.37	0.404	0.198	788
PaperEdge [30]	0.472	7.98	0.367	0.189	751
DocGeoNet [15]	0.504	7.70	0.378	0.182	704
UVDoc [47]	0.544	6.83	0.315	0.171	704
DocRES [59]	0.464	9.40	0.470	0.231	890
DocScanner [14]	0.518	7.41	0.333	0.165	633
DvD [62]	0.548	6.60	0.280	0.171	643
AADD [48]	0.542	6.26	0.277	0.166	642
Ours	0.558	5.98	0.252	0.161	628

independently to perform document unwarping and illumination correction. We use this architecture to serve as a baseline and evaluate the benefits of training a model on our SyntheticDoc dataset over a combination of Doc3D and UVDoc. In addition, the lightweight architecture ensures that the model is fast at inference time for both document unwarping and illumination correction.

Training details. Similar to UVDoc [47], the predicted backward mapping is a coarse 45×31 unwarping grid. The input of the model is a 512×720 image of a warped document, tightly cropped around the document boundaries. We optimize the network using AdamW [20, 29] with a batch size of 16. The learning rate is initially set to 0.0001 with a weight decay of 0.0001, and evolves according to a cosine scheduler. We apply an L_1 loss to the coarse backward mapping, the shading map and the reconstructed image. The weight of the backward mapping loss is set to double that of the reconstruction and shading losses. We train the model for 38 epochs on SyntheticDoc alone, using eight NVIDIA GeForce RTX 4090 GPUs, with each epoch taking approximately 1.5 hours.

Evaluation. We assess our model's performance on the standard DocUNet benchmark [31], comparing it against current state-of-the-art approaches. We evaluate these methods across multiple metrics. Image similarity is measured using multi-scale structural similarity (MS-SSIM), local distortion (LD) and aligned distortion (AD), while optical character recognition (OCR) accuracy is evaluated via the

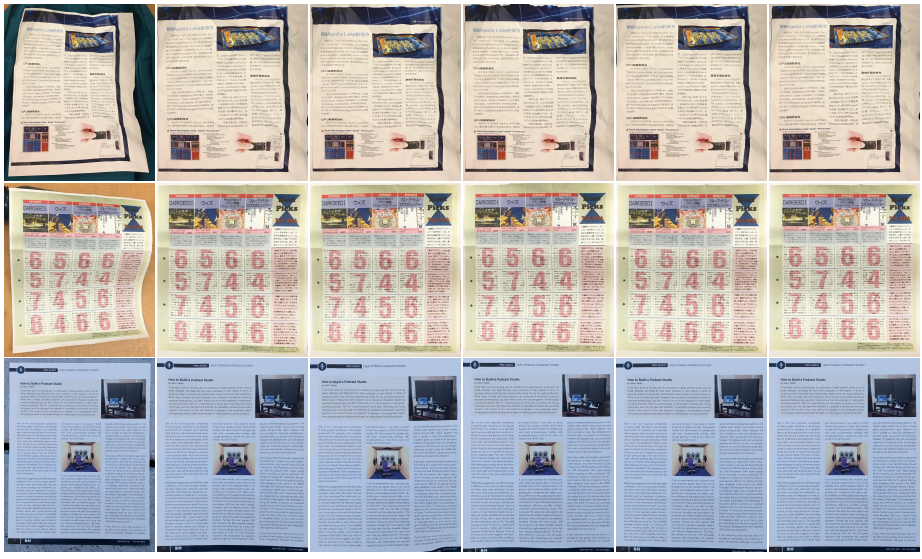

Fig. 10: Qualitative comparisons of unwarping-only results produced by various document unwarping methods on the DocUNet benchmark.

character error rate (CER) and edit distance (ED) metrics. Details about these metrics are provided in the supplementary material.

Results. Quantitative results on the DocUNet benchmark are presented in Tab. 3. Note that the results of other SOTA methods were computed based on the results provided by their authors, or using the code they made available. Works that released neither their code nor their results on the DocUNet benchmark, such as Uni-DocDiff, are therefore not included in the comparison. As shown, our simple baseline network trained solely on our SyntheticDoc dataset outperforms all state-of-the-art methods across the visual (MS-SSIM, LD and AD) and OCR (ED and CER) metrics, with improvements ranging from 1.8% to 9%. We provide qualitative visual comparisons of the unwarping against current state-of-the-art models in Fig. 10 and in the supplementary material.

We also compare the results of our method after both document unwarping and illumination correction on the DocUNet benchmark to approaches capable of performing both tasks [13, 59]. Quantitative results are presented in Tab. 4. Our simple baseline network outperforms all competing methods on the visual metrics (MS-SSIM and PSNR). While it does not achieve the best performance on the OCR metrics, it produces the most visually convincing results, better preserving the original colors of the document, as visible in the qualitative visual comparisons presented in the supplementary material.

Table 4: Quantitative comparison on both the unwarping and illumination correction tasks of various methods on the DocUNet benchmark. The first row reports the metrics for the original warped documents. The last row presents our lightweight baseline model trained exclusively on our SyntheticDoc dataset. Gold and silver backgrounds indicate the best and second-best scores, respectively.

Method	MS-SSIM ↑	PSNR ↑	CER ↓	ED ↓
Warped Image	0.247	7.92	0.517	2029
DocTr [13]	0.496	10.79	0.126	491
DocRES [59]	0.473	11.68	0.173	637
Ours	0.584	12.68	0.149	601

Table 5: Quantitative comparison of the unwarping performance of variants of our methods on the DocUNet benchmark. The gray row is our original experiment.

Method			MS-SSIM ↑	LD ↓	AD ↓	CER ↓	ED ↓
#samples	Input	Output					
100k	512×720	31×45	0.532	6.89	0.323	0.194	704
500k	512×720	31×45	0.546	6.52	0.278	0.168	654
1M	512×720	31×45	0.558	5.98	0.252	0.161	628
1M	1024×1440	121×177	0.570	5.51	0.219	0.164	651

4.1 Ablation studies

Training with fewer samples. We trained our baseline model on subsets of 100K and 500K samples from our SyntheticDoc dataset (see Tab. 5). These models were trained for fewer epochs due to time constraints. Even when trained with fewer samples, our simple model achieves similar performance to most SOTA methods, demonstrating the high quality of the dataset. Our model trained on the full dataset still performs best, highlighting the value of SyntheticDoc's scale.

Training with higher resolution. We intentionally use a lightweight baseline model to prove that our SyntheticDoc dataset yields SOTA results without architectural tricks. To demonstrate that our high-resolution, large-scale data can push performance further, we modified this simple model by increasing the input/output resolution. As shown in Tab. 5 (last row), this increased capacity yields massive gains in visual metrics, with the AD score improving by more than 20% over SOTA.

5 Discussion

We presented SyntheticDoc, a large-scale synthetic dataset for document unwarping and illumination correction. Comprising 1,000,000 training samples,

SyntheticDoc represents the first dataset of this size for these tasks. High-quality samples are obtained through the combination of physically simulated geometries and high-resolution path-traced rendering, mitigating the sim-to-real gap. We demonstrate a practical application of this dataset by training a lightweight baseline model for document unwarping and illumination correction. Our dataset is available at https://igl.ethz.ch/projects/SyntheticDoc/ and the code used to generate it at https://github.com/tanguymagne/SyntheticDoc.

Limitations. Due to the scale and high-quality rendering, the final dataset is very large in size. The training set alone, containing only the rendered images, albedos, shadow maps and UV maps, is over 5 TB. This makes model training computationally demanding, causing us to opt for a relatively lightweight architecture. We believe that with less constrained hardware capacities, a very large model could truly benefit from the size of this dataset to push performance even further. Furthermore, since we focused on the most common real-world capture scenarios, the current dataset does not cover extreme viewpoints or diverse camera types. However, our setup can be easily extended to cover such cases.

Future work. Built entirely upon highly scalable steps, our generation pipeline can be readily reproduced to further expand the dataset with more document textures or new physical simulation scenarios for the deformed document meshes, including more complex fold scenarios (explicit Z-fold or tri-fold simulations) or other paper formats. In addition, the procedural nature of our approach makes it relatively straightforward to extract additional ground-truth annotations such as OCR transcripts or document layouts, extending the utility of SyntheticDoc to other downstream document analysis tasks. Furthermore, our generation pipeline could be used to create a multi-view dataset for document processing.

Ethical considerations. When gathering the assets required to produce our dataset, we ensured that they were all available under permissive licenses. While utilizing text-to-image models to generate magazine pages can raise concerns regarding data provenance and copyright infringement, we mitigate this issue by employing strictly generic prompts that do not attempt to replicate specific real-world magazines.

Our data generation pipeline is computationally efficient. Simulating a mesh takes at most 10 minutes on a CPU-only machine, while rendering each sample takes approximately 5 seconds on a machine equipped with a consumer-grade RTX 4090 GPU. In addition, the cluster we use to generate our dataset and train our model is carbon neutral and powered entirely by renewable energy, minimizing the environmental footprint of our work.

Acknowledgements

We thank the anonymous reviewers for their insightful feedback and constructive suggestions. We are also grateful to Danielle Luterbacher for her help in managing the hardware required to create and store a dataset of this size.

References

1. arXiv: arXiv (2026), https://arxiv.org/ 10
2. Black Forest Labs: Flux.2 klein 9b (2026), https://huggingface.co/black-forest-labs/FLUX.2-klein-9B 10
3. Blender Foundation: Blender 4.5 (2026), https://www.blender.org/ 2, 6, 8, 9
4. Brown, M., Seales, W.: Document restoration using 3d shape: a general deskewing algorithm for arbitrarily warped documents. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001. vol. 2, pp. 367–374 vol.2 (2001). https://doi.org/10.1109/ICCV.2001.937649 3
5. Brown, M., Seales, W.: Image restoration of arbitrarily warped documents. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(10), 1295–1306 (2004). https://doi.org/10.1109/TPAMI.2004.87 3
6. Brown, M., Tsoi, Y.C.: Geometric and shading correction for images of printed materials using boundary. IEEE Transactions on Image Processing **15**(6), 1544–1554 (2006). https://doi.org/10.1109/TIP.2006.871082 5
7. Das, S., Ma, K., Shu, Z., Samaras, D., Shilkrot, R.: Dewarpnet: Single-image document unwarping with stacked 3d and 2d regression networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019) 2, 3, 4, 5, 6, 12
8. Das, S., Sial, H., Ma, K., Baldrich, R., Vanrell, M., Samaras, D.: Intrinsic decomposition of document images in-the-wild. In: Proceedings of the 31st British Machine Vision Conference (BMVC) (2020). https://doi.org/10.5244/C.34.188 5
9. Das, S., Singh, K.Y., Wu, J., Bas, E., Mahadevan, V., Bhotika, R., Samaras, D.: End-to-end piece-wise unwarping of document images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4268–4277 (October 2021) 3, 12
10. Deschaintre, V., Aittala, M., Durand, F., Drettakis, G., Bousseau, A.: Single-image svbrdf capture with a rendering-aware deep network. ACM Trans. Graph. **37**(4) (Jul 2018). https://doi.org/10.1145/3197517.3201378 9
11. DOAB: Directory of open access books (2026), https://www.doabooks.org/ 10
12. Feng, H., Liu, S., Deng, J., Zhou, W., Li, H.: Deep unrestricted document image rectification. IEEE Transactions on Multimedia **26**, 6142–6154 (2024). https://doi.org/10.1109/TMM.2023.3347094 4
13. Feng, H., Wang, Y., Zhou, W., Deng, J., Li, H.: Doctr: Document image transformer for geometric unwarping and illumination correction. In: Proceedings of the 29th ACM International Conference on Multimedia. p. 273–281. MM '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3474085.3475388 3, 5, 12, 13, 14
14. Feng, H., Zhou, W., Deng, J., Tian, Q., Li, H.: Docscanner: Robust document image rectification with progressive learning. Int. J. Comput. Vision **133**(8), 5343–5362 (May 2025). https://doi.org/10.1007/s11263-025-02431-5, https://doi.org/10.1007/s11263-025-02431-5 12, 13
15. Feng, H., Zhou, W., Deng, J., Wang, Y., Li, H.: Geometric representation learning for document image rectification. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 475–492. Springer Nature Switzerland, Cham (2022) 3, 4, 12
16. Google: Gemini 3 flash preview (2026), https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview 10

17. Hertlein, F., Naumann, A., Philipp, P.: Inv3d: a high-resolution 3d invoice dataset for template-guided single-image document unwarping. Int. J. Doc. Anal. Recognit. **26**(3), 175–186 (Apr 2023). https://doi.org/10.1007/s10032-023-00434-x 3, 4, 10
18. IMSLP: International music score library project (imslp) / petrucci music library (2026), https://imslp.org/ 10
19. Jiang, X., Long, R., Xue, N., Yang, Z., Yao, C., Xia, G.S.: Revisiting document image dewarping by grid regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4543–4552 (June 2022) 3, 12
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980 12
21. Koo, H.I., Kim, J., Cho, N.I.: Composition of a dewarped and enhanced document image from two view images. IEEE Transactions on Image Processing **18**(7), 1551–1562 (2009). https://doi.org/10.1109/TIP.2009.2019301 3
22. Kumari, P., Das, S.: Document image rectification using stable diffusion transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 3426–3435 (June 2025) 3
23. Li, H., Wu, X., Chen, Q., Xiang, Q.: Foreground and text-lines aware document image rectification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19574–19583 (October 2023) 3
24. Li, X., Zhang, B., Liao, J., Sander, P.V.: Document rectification and illumination correction using a patch-based cnn. ACM Trans. Graph. **38**(6) (Nov 2019). https://doi.org/10.1145/3355089.3356563 3, 4, 5
25. Li, Z., Chen, X., Pun, C.M., Cun, X.: High-resolution document shadow removal via a large-scale real-world dataset and a frequency-aware shadow erasing net. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12449–12458 (October 2023) 5
26. Liang, J., DeMenthon, D., Doermann, D.: Geometric rectification of camera-captured document images. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(4), 591–605 (2008). https://doi.org/10.1109/TPAMI.2007.70724 3
27. Lin, Y.H., Chen, W.C., Chuang, Y.Y.: Bedsr-net: A deep shadow removal network from a single document image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 5
28. Liu, S., Feng, H., Luan, B., Hou, M., Deng, J., Zhou, W.: Booknet: Book image rectification via cross-page attention network (2026), https://arxiv.org/abs/2601.21938 3, 4
29. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7 12
30. Ma, K., Das, S., Shu, Z., Samaras, D.: Learning from documents in the wild to improve document unwarping. In: ACM SIGGRAPH 2022 Conference Proceedings. SIGGRAPH '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3528233.3530756 3, 4, 12
31. Ma, K., Shu, Z., Bai, X., Wang, J., Samaras, D.: Docunet: Document image unwarping via a stacked u-net. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018) 3, 12

32. Markovitz, A., Lavi, I., Perel, O., Mazor, S., Litman, R.: Can you read me now? content aware rectification using angle supervision. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII. p. 208–223. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-58610-2_13 4
33. Meng, G., Su, Y., Wu, Y., Xiang, S., Pan, C.: Exploiting vector fields for geometric rectification of distorted document images. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018) 3
34. Meng, G., Wang, Y., Qu, S., Xiang, S., Pan, C.: Active flattening of curved document images via two structured beams. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3890–3897 (2014). https://doi.org/10.1109/CVPR.2014.497 3
35. Narain, R., Pfaff, T., O'Brien, J.F.: Folding and crumpling adaptive sheets. ACM Trans. Graph. **32**(4) (Jul 2013). https://doi.org/10.1145/2461912.2462010 2, 6
36. Narain, R., Samii, A., O'Brien, J.F.: Adaptive anisotropic remeshing for cloth simulation. ACM Trans. Graph. **31**(6) (Nov 2012). https://doi.org/10.1145/2366145.2366171 2, 6
37. OpenStax: OpenStax (2026), https://openstax.org/ 10
38. Perlin, K.: An image synthesizer. In: Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques. p. 287–296. SIGGRAPH '85, Association for Computing Machinery, New York, NY, USA (1985). https://doi.org/10.1145/325334.325247 9
39. Quan, J., Wang, H., Wu, C., Cao, G.: Dle: Document illumination correction with dynamic light estimation. In: 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 3701–3707 (2024). https://doi.org/10.1109/SMC54092.2024.10831684 5
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/radford21a.html 5
41. Tang, H., Guo, J., Wang, T., Yu, Y., Wang, C.: Efficient joint rectification of photometric and geometric distortions in document images. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3690–3694 (2024). https://doi.org/10.1109/ICASSP48485.2024.10447446 3, 5
42. Tian, Y., Narasimhan, S.G.: Rectification and 3d reconstruction of curved document images. In: CVPR 2011. pp. 377–384 (2011). https://doi.org/10.1109/CVPR.2011.5995540 3
43. Townsend, B., May, M., Mackowiak, K., Wells, C.: Realkie: Five novel datasets for enterprise key information extraction (2025), https://arxiv.org/abs/2403.20101 10
44. Tsoi, Y.C., Brown, M.S.: Multi-view document rectification using boundary. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (2007). https://doi.org/10.1109/CVPR.2007.383251 3
45. Ulges, A., Lampert, C.H., Breuel, T.: Document capture using stereo vision. In: Proceedings of the 2004 ACM Symposium on Document Engineering. p. 198–200. DocEng '04, Association for Computing Machinery, New York, NY, USA (2004). https://doi.org/10.1145/1030397.1030434 3

46. Vecchio, G., Deschaintre, V.: Matsynth: A modern pbr materials dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22109–22118 (June 2024) 9
47. Verhoeven, F., Magne, T., Sorkine-Hornung, O.: Uvdoc: Neural grid-based document unwarping. In: SIGGRAPH Asia 2023 Conference Papers. SA '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3610548.3618174 2, 3, 4, 6, 10, 11, 12, 13
48. Wang, C., Shen, I.C., Igarashi, T., Jiang, C.: Axis-aligned document dewarping (2025), https://arxiv.org/abs/2507.15000 12, 13
49. Wang, R., Xue, Y., Jin, L.: Docnlc: A document image enhancement framework with normalized and latent contrastive representation for multiple degradations. Proceedings of the AAAI Conference on Artificial Intelligence **38**(6), 5563–5571 (Mar 2024). https://doi.org/10.1609/aaai.v38i6.28366, https://ojs.aaai.org/index.php/AAAI/article/view/28366 5
50. Wang, Y., Zhou, W., Lu, Z., Li, H.: Udoc-gan: Unpaired document illumination correction with background light prior. In: Proceedings of the 30th ACM International Conference on Multimedia. p. 5074–5082. MM '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3503161.3547916 5
51. Worley, S.: A cellular texture basis function. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques. p. 291–294. SIGGRAPH '96, Association for Computing Machinery, New York, NY, USA (1996). https://doi.org/10.1145/237170.237267 10
52. Xie, G.W., Yin, F., Zhang, X.Y., Liu, C.L.: Dewarping document image by displacement flow estimation with fully convolutional network. In: Bai, X., Karatzas, D., Lopresti, D. (eds.) Document Analysis Systems. pp. 131–144. Springer International Publishing, Cham (2020) 4, 12
53. Xie, G.W., Yin, F., Zhang, X.Y., Liu, C.L.: Document dewarping with control points. In: Lladós, J., Lopresti, D., Uchida, S. (eds.) Document Analysis and Recognition – ICDAR 2021. pp. 466–480. Springer International Publishing, Cham (2021) 3, 4, 12
54. Xue, C., Tian, Z., Zhan, F., Lu, S., Bai, S.: Fourier document restoration for robust document dewarping and recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4573–4582 (June 2022) 3, 12
55. Yang, Y., Patel, A., Deitke, M., Gupta, T., Weihs, L., Head, A., Yatskar, M., Callison-Burch, C., Krishna, R., Kembhavi, A., Clark, C.: Scaling text-rich image understanding via code-guided synthetic multimodal data generation. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 17486–17505. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.acl-long.855 10
56. You, S., Matsushita, Y., Sinha, S., Bou, Y., Ikeuchi, K.: Multiview rectification of folded documents. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(2), 505–511 (2018). https://doi.org/10.1109/TPAMI.2017.2675980 3
57. Zhang, J., Liang, L., Ding, K., Guo, F., Jin, L.: Appearance enhancement for camera-captured document images in the wild. IEEE Transactions on Artificial Intelligence **5**(5), 2319–2330 (2024). https://doi.org/10.1109/TAI.2023.3321257 5
58. Zhang, J., Luo, C., Jin, L., Guo, F., Ding, K.: Marior: Margin removal and iterative content rectification for document dewarping in the wild. In: Proceedings of the 30th

ACM International Conference on Multimedia. p. 2805–2815. MM '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3503161.3548214 3, 12
59. Zhang, J., Peng, D., Liu, C., Zhang, P., Jin, L.: Docres: A generalist model toward unifying document image restoration tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15654–15664 (June 2024) 3, 5, 12, 13, 14
60. Zhang, L., Yip, A.M., Tan, C.L.: Photometric and geometric restoration of document images using inpainting and shape-from-shading. In: Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2. p. 1121–1126. AAAI'07, AAAI Press (2007) 5
61. Zhang, L., Zhang, Y., Tan, C.: An improved physically-based method for geometric restoration of distorted document images. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(4), 728–734 (2008). https://doi.org/10.1109/TPAMI.2007.70831 3
62. Zhang, W., Lu, H., Ning, M., Huang, X., Wang, W., Huang, K., Wang, Q.: Dvd: Unleashing a generative paradigm for document dewarping via coordinates-based diffusion model. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. SA Conference Papers '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3757377.3763913 3, 12, 13
63. Zhao, F., Zeng, W., Li, Z., Yang, D., Li, B., Bi, X., Zhou, Y.: Uni-docdiff: A unified document restoration model based on diffusion. In: Proceedings of the 33rd ACM International Conference on Multimedia. p. 8204–8213. MM '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3746027.3755362 3, 5
64. Zhou, X., Li, G., Jiang, N., Wang, D.H., Zhang, X.Y., Zhu, S.: Dochformer: Document image dewarping via harmonized modeling of hierarchical priors. In: Antonacopoulos, A., Chaudhuri, S., Chellappa, R., Liu, C.L., Bhattacharya, S., Pal, U. (eds.) Pattern Recognition. pp. 29–44. Springer Nature Switzerland, Cham (2025) 3