

000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044

– Supplementary Material –

Learning an Isometric Surface Parameterization for Texture Unwrapping

Sagnik Das¹, Ke Ma², Zhixin Shu³, and Dimitris Samaras¹

¹ Stony Brook University, Stony Brook NY 11790, USA

{sadas}@cs.stonybrook.edu

² Snap Inc.

³ Adobe Research

Overview

In this supplementary material we include the following sections:

1. High-resolution results for arbitrary surface texture editing
2. High-resolution results for document texture editing
3. Video with additional results
4. More qualitative comparison with DewarpNet [4] on synthetic evaluation set
5. Qualitative comparison with [4] for different types of real documents
6. Qualitative comparison with [16] on their test-set
7. Usefulness of L_{uv}
8. Details of weighting function used in L_z
9. Training details of the UV prior network
10. Initializing S and F_z
11. Unwarping and texture editing details
12. Pre-processing details for the real scenes
13. Detailed ablation figure
14. Limitations
15. Example of a failure case
16. Training time

1 High-resolution results for arbitrary surface texture editing

In Fig. 1, 2 and 3 we show the examples of editing arbitrary surface texture. Furthermore, in Fig. 4, 5 we show examples of face [10] texture unwrapping and editing. These examples show that our learned F_{uv} prior and the proposed method works beyond documents as long as the isometry assumption is not strongly violated.

2 High-resolution results for document texture editing

In Fig. 6, 7 we show the examples of texture editing in higher resolution.



Fig. 1. Example of texture edited images rendered from different views. Note the perspective changes and deformation on the edited texture due to the surface. The input foreground mask is shown using dashed yellow polygon.

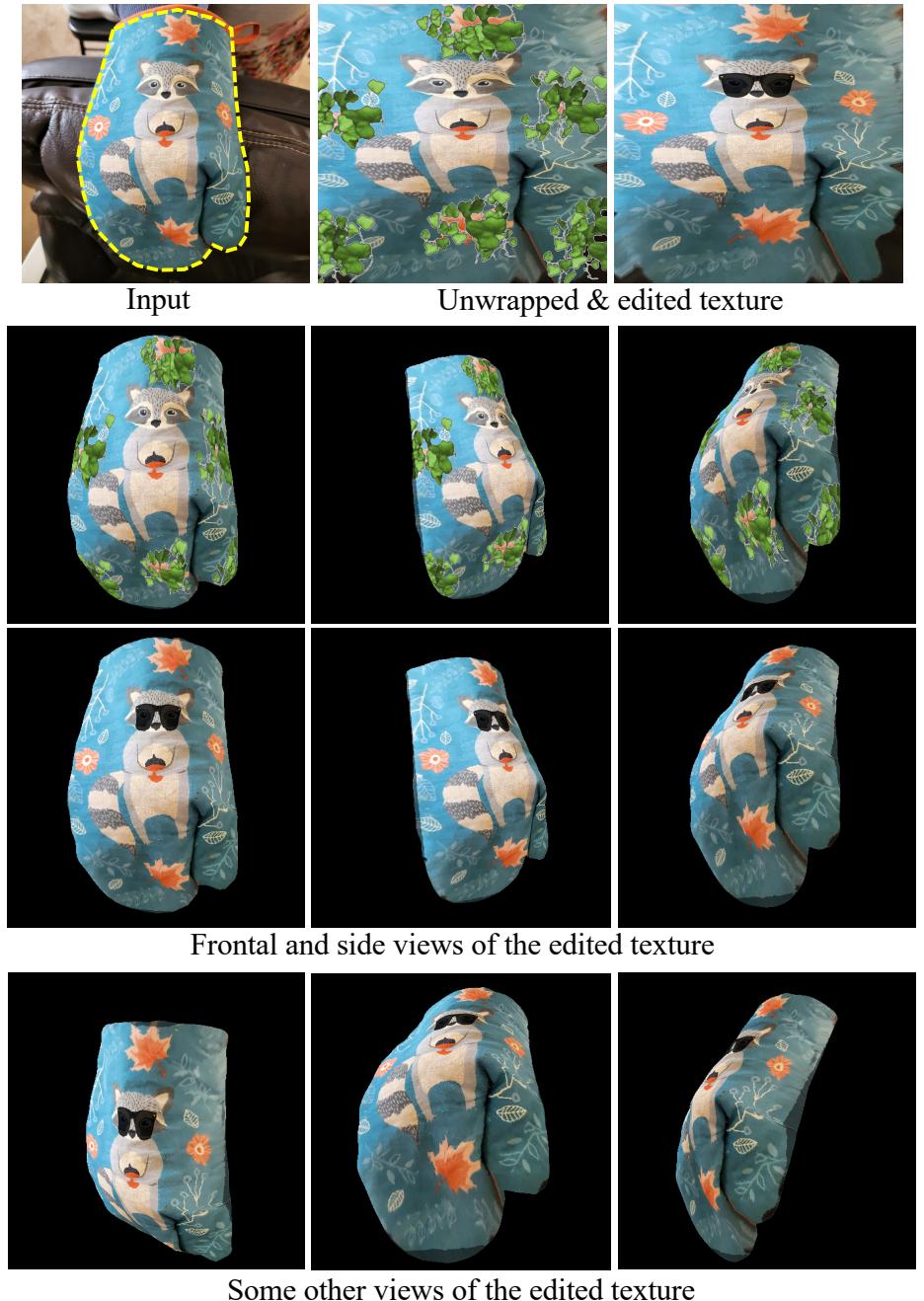


Fig. 2. Example of texture edited images rendered from different views. Note the perspective changes and deformation on the edited texture due to the surface. The input foreground mask is shown using dashed yellow polygon.

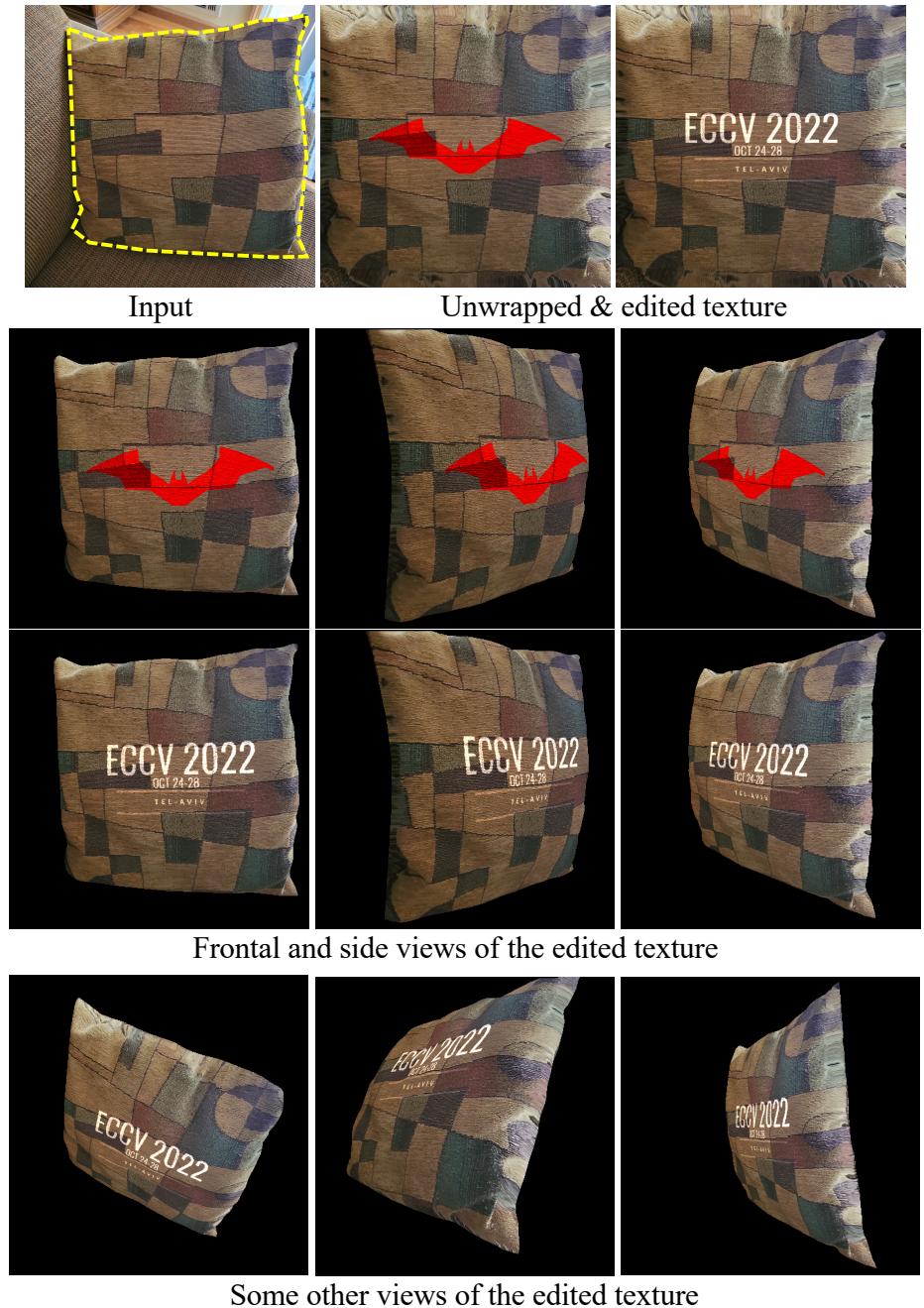


Fig. 3. Example of texture edited images rendered from different views. Note the perspective changes and deformation on the edited texture due to the surface. The input foreground mask is shown using dashed yellow polygon.

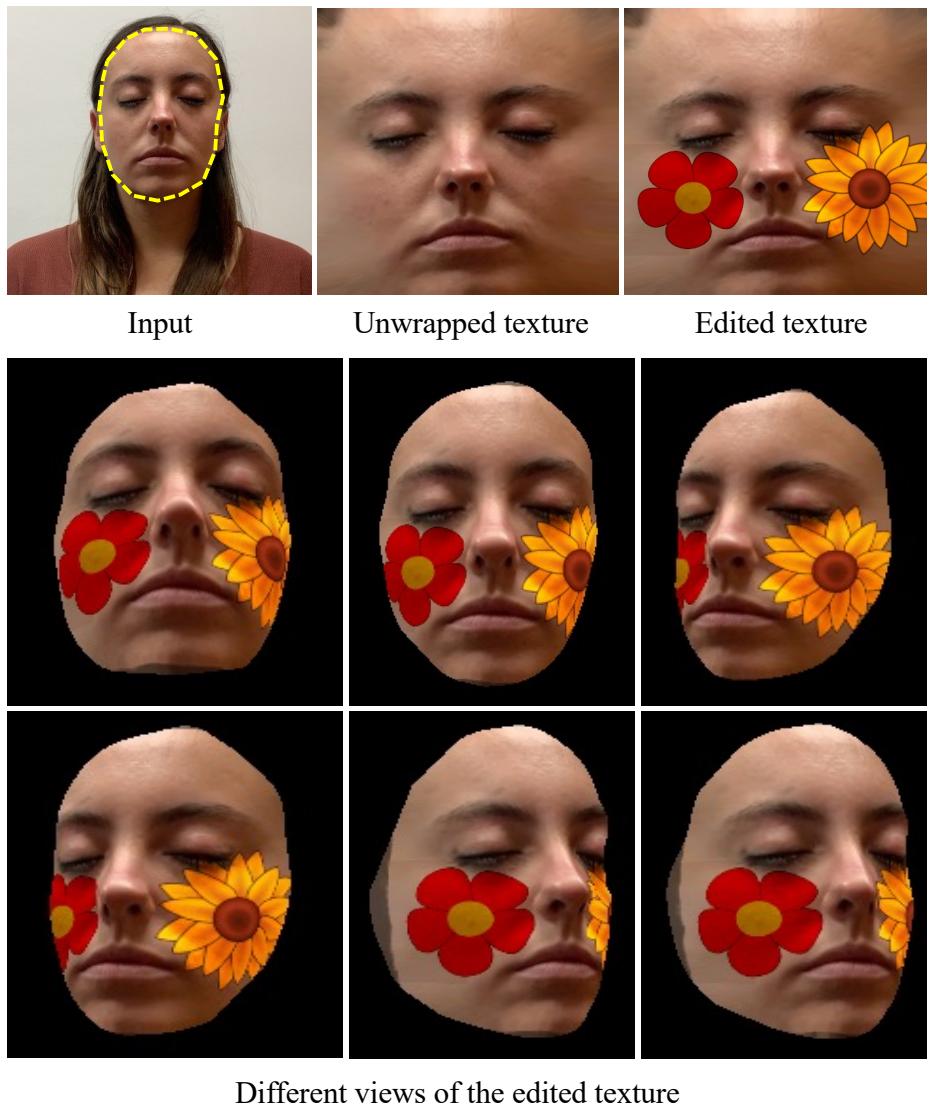


Fig. 4. Example of texture edited faces [10] rendered from different views. Note the perspective changes and deformation on the edited texture due to the surface. The input foreground mask is shown using dashed yellow polygon.

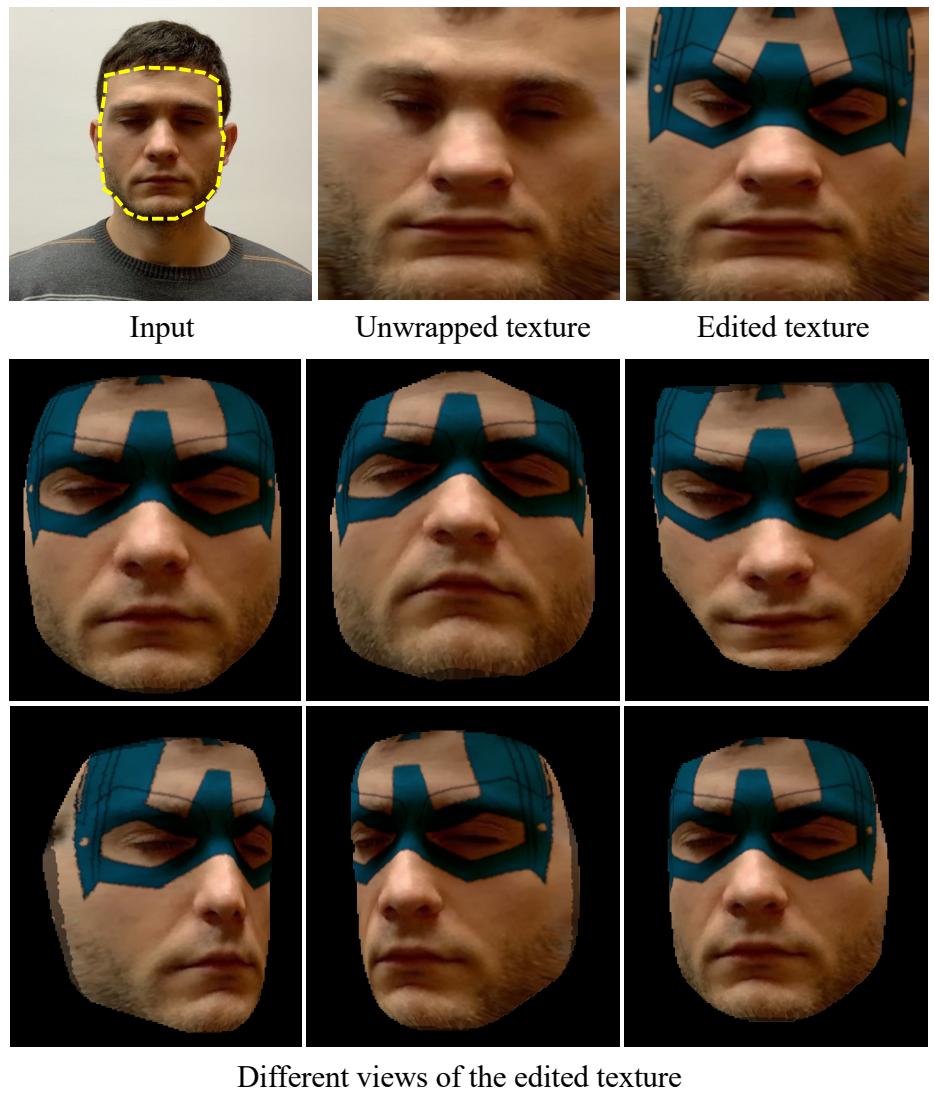


Fig. 5. Example of texture edited faces [10] rendered from different views. Note the perspective changes and deformation on the edited texture due to the surface. The input foreground mask is shown using dashed yellow polygon.

270
271
272
273
274
275
276
277
278
279
280
281
282
283
284

Input



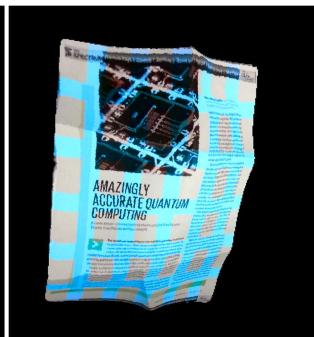
Unwarped



Edited Texture



Texture Edited Images



The image shows a book cover with a blue and white striped pattern. The title 'AMAZINGLY ACCURATE QUANTUM COMPUTING' is printed in large, bold, black capital letters. Below the title, there is a small green square icon containing a white atom symbol, followed by the text 'Quantum Computing for Everyone'. The book is shown from a three-quarter perspective, revealing its front cover and part of the spine.



Fig. 6. Example of texture edited images from different views. Note the perspective changes and deformation on the edited texture due to the complex shape of the paper

314



Fig. 7. Example of texture edited images from different views. Note the perspective changes and deformation on the edited texture due to the complex shape of the paper.

360 3 Video with additional results

361
362 We include a video (3394-supp.mp4) to demonstrate the quality of our texture
363 editing results. It includes continuous view of the edited textures from different
364 camera perspectives.

365 366 4 More qualitative comparison with DewarpNet [4] on 367 synthetic evaluation set

368
369 In Fig. 8, 9 we show more qualitative comparison with DewarpNet [4] on un-
370 warping frontal view of a document. For a better illustrative comparison we
371 also show qualitative results of the 4 best (lowest LD) unwarped views using [4]
372 in Fig.10, and 11. Clearly in all of the cases we achieve better or comparative
373 results. Furthermore, we can see that it is hard to predict which view will per-
374 form best for [4], and results vary significantly even if the views are reasonably
375 frontal. Comparatively, being a multi-view method, our approach produces more
376 consistent unwarping across all views.

377 378 5 Qualitative comparison with [4] for different types of 379 real documents

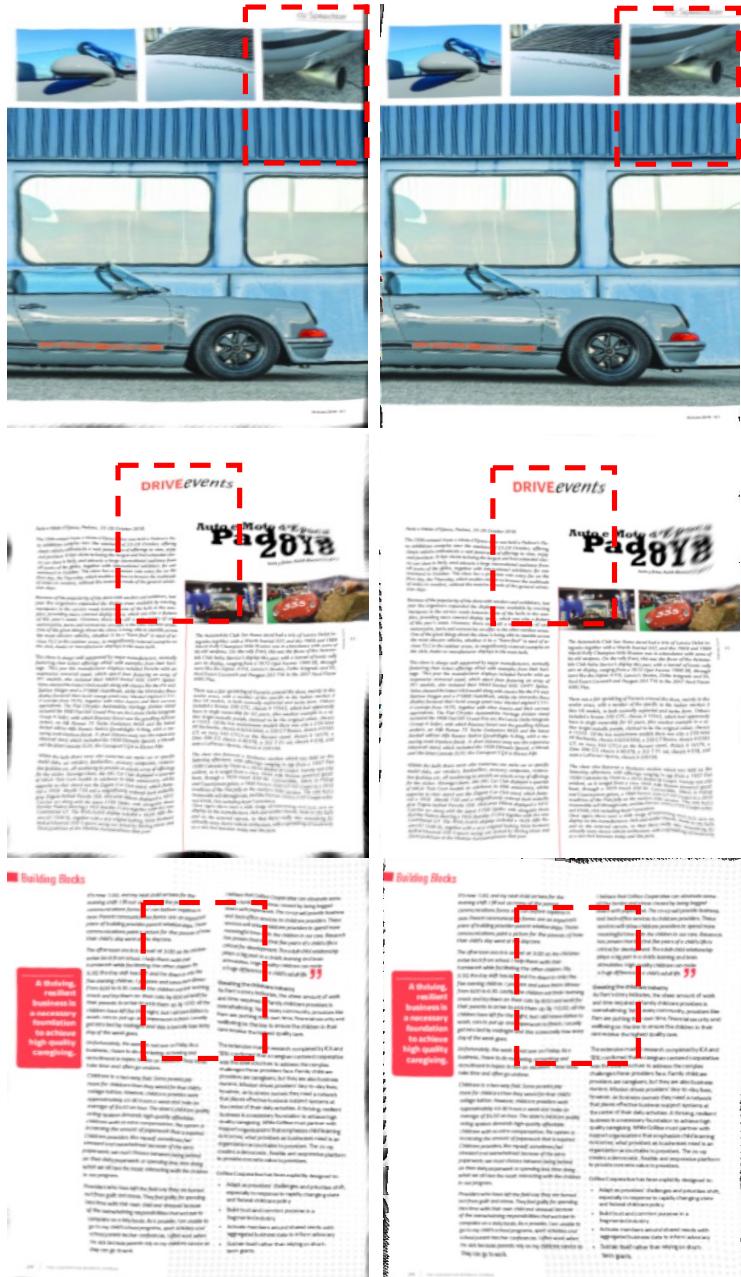
380
381 In Fig. 12, 13, 14, and 15, we show qualitative unwarping result for four different
382 type of documents, e.g. book, receipt, flyer, and magazine. In all the views our
383 method shows consistent and good quality unwarping results.

384 385 6 Qualitative comparison with [16]

386
387 In Fig. 16, 17 we provide a qualitative comparison with 5 publicly available
388 images from [16]. The results are competitive and often produce better unwarping.
389 Quantitative numbers couldn't be reported because the high-res/original
390 unwarped results are not publicly available.

391 392 7 Usefulness of L_{uv}

393
394 In section 3.3 of the main submission, we define L_{uv} (Eq. 8) to prevent non-
395 uniform mapping between the 3D and the UV domain. Specifically, we constrain
396 the output of F_{uv} to be $\sim \mathcal{U}(0, 1)$ using L_{uv} . Without L_{uv} , F_{uv} is prone to
397 produce a mapping $\sim \mathcal{U}(a, b)$ where $a > 0$ or $b < 1$. Consequently, F_z also learns
398 an incorrect mapping between the texture and the 3D domain. As a result, the
399 unwarped texture gets stretched or squeezed. We demonstrate two such examples
400 in Fig. 18.



DewarpNet

Proposed

Fig. 8. Comparison of frontal view unwarping: left is DewarpNet and right is our approach. Our results are clearly better with straighter lines. Discriminative regions are highlighted with red dashed rectangles.

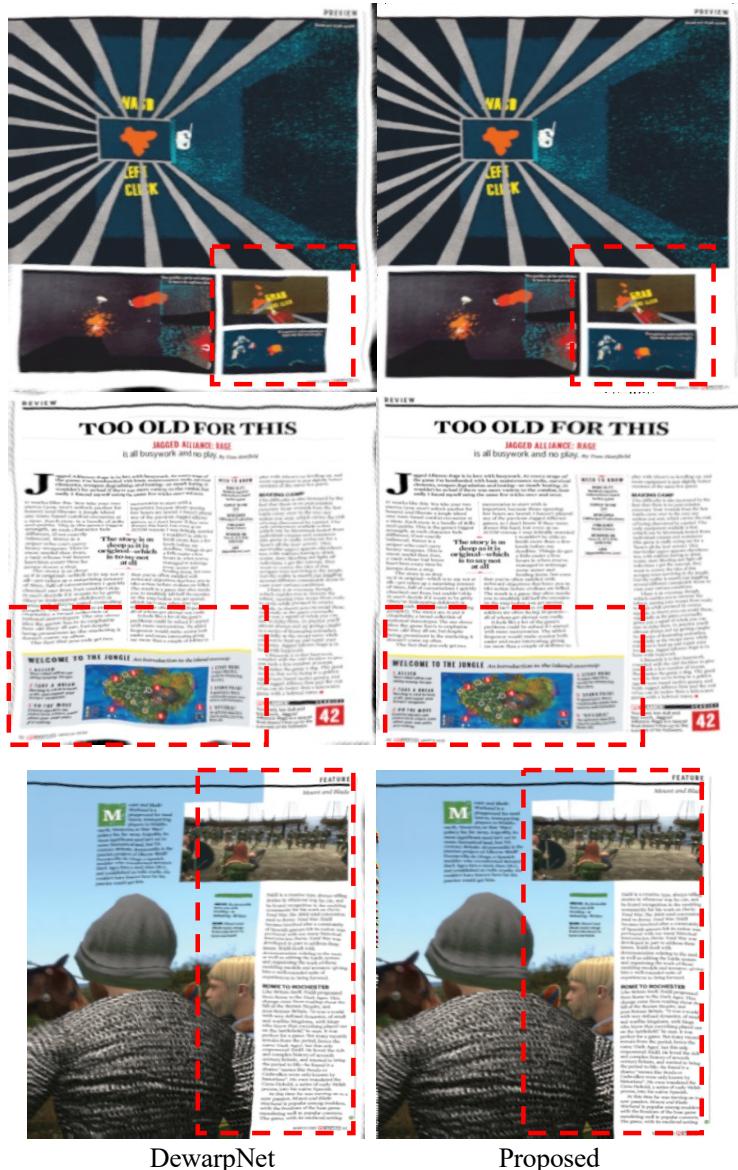


Fig. 9. Comparison of frontal view unwarping: left is DewarpNet and right is our approach. Our results are clearly better with straighter lines. Discriminative regions are highlighted with red dashed rectangles.



535 **Fig. 10.** 4 best results (sorted in ascending order from top to bottom according to
536 LD score [lower better]) of (b) DewarpNet compared to (c) proposed unwarping for a
537 specific scene. For all the views proposed unwarping shows better and consistent visual
538 results than DewarpNet. (a) is the input. Blue dashed boxes denote the discriminative
539 areas in the unwarped results.

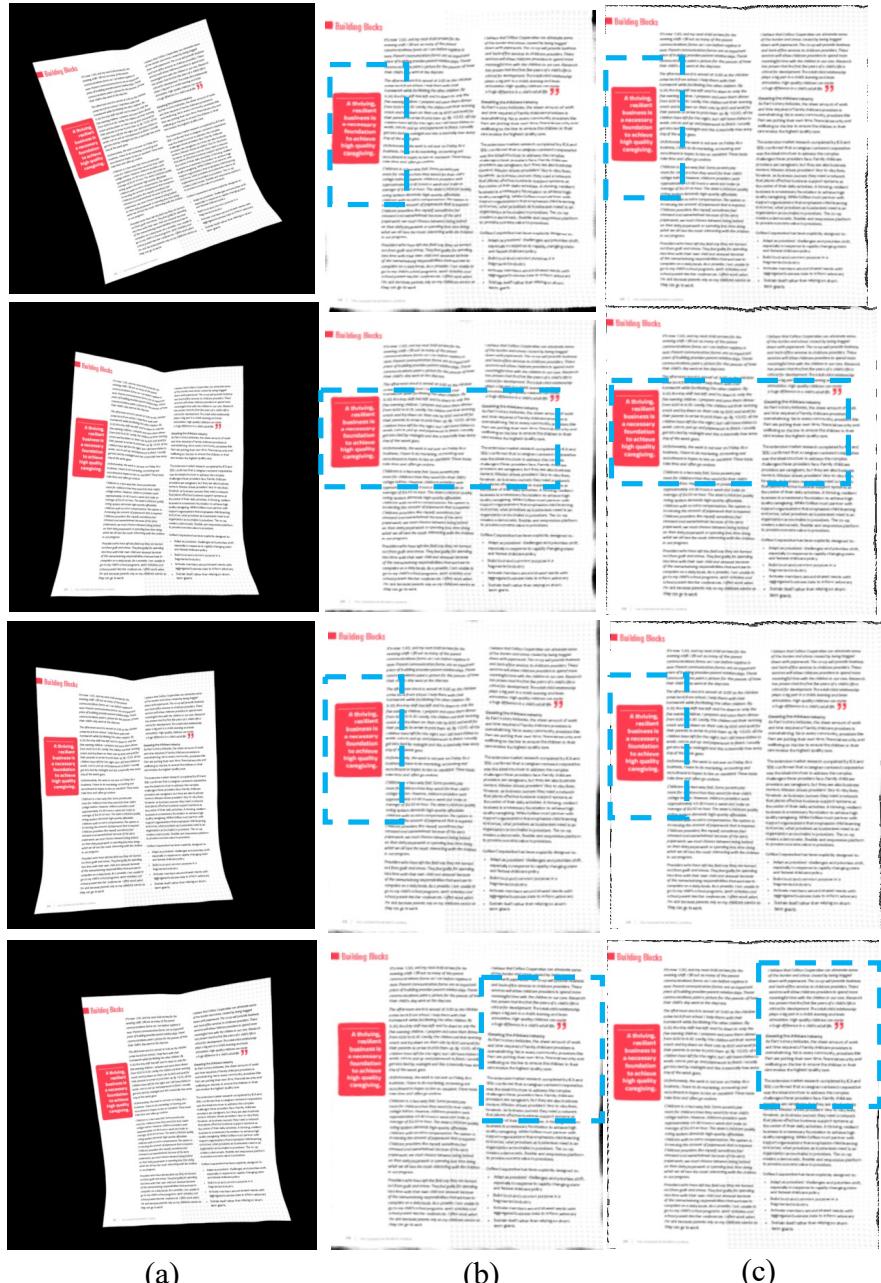


Fig. 11. 4 best results (sorted in ascending order from top to bottom according to LD score [lower better]) of (b) DewarpNet compared to (c) proposed unwarping for a specific scene. In all the views proposed unwarping shows better and consistent visual results than DewarpNet. (a) is the input. Blue dashed boxes denote the discriminative areas in the unwarped results.

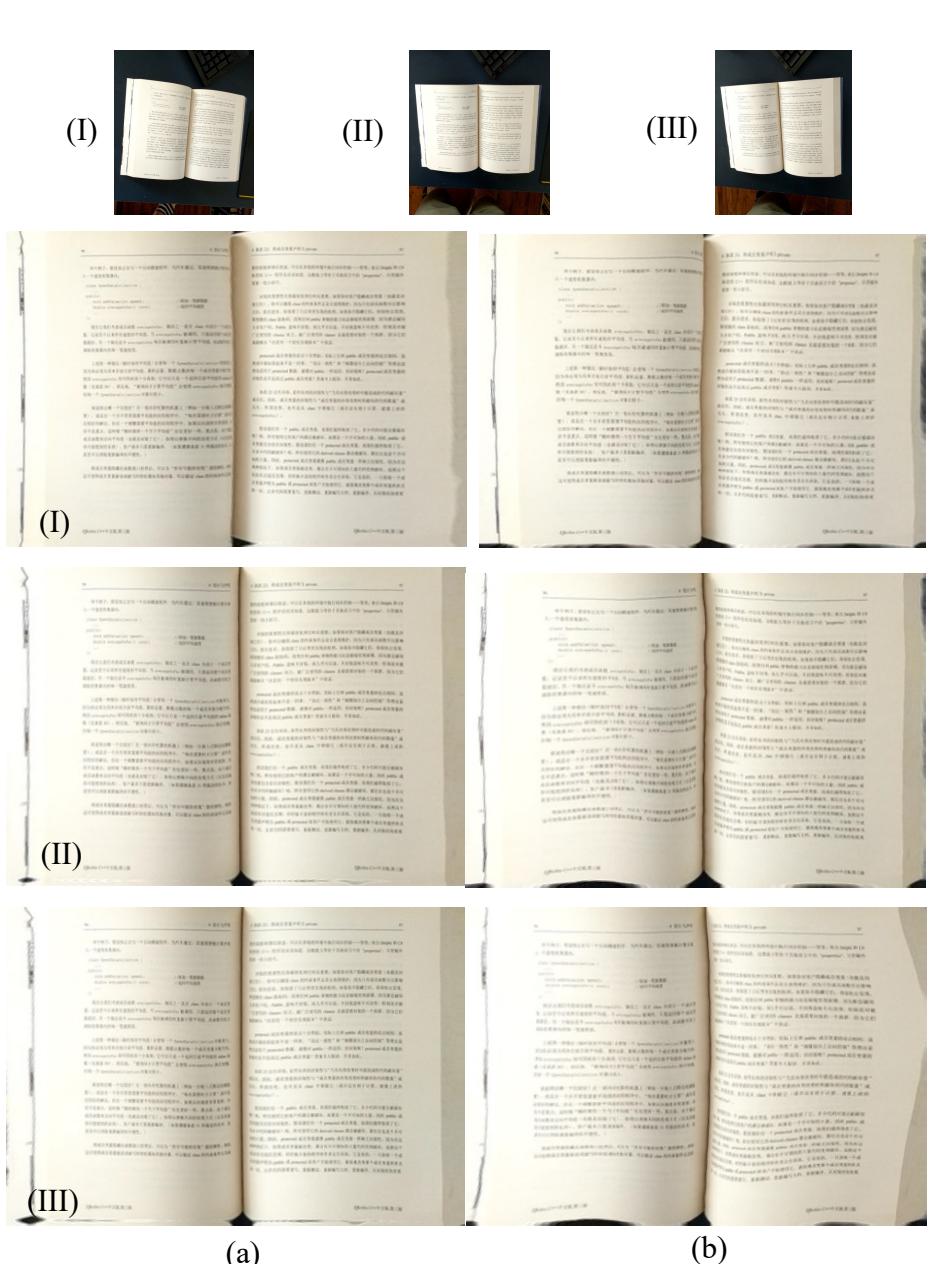


Fig. 12. Unwarping results on different views of a book. Top row shows the inputs. (a) Proposed, (b) DewarpNet. Our method generates good quality unwarping results with straighter text-lines.

630
631
632
633
634
635
636
637

(I)



(II)



(III)



(a)



(b)



Fig. 13. Unwarping results on different views of a receipt. Top row shows the inputs. (a) Proposed, (b) DewarpNet. Our method generates good quality unwarping results with straighter text-lines.

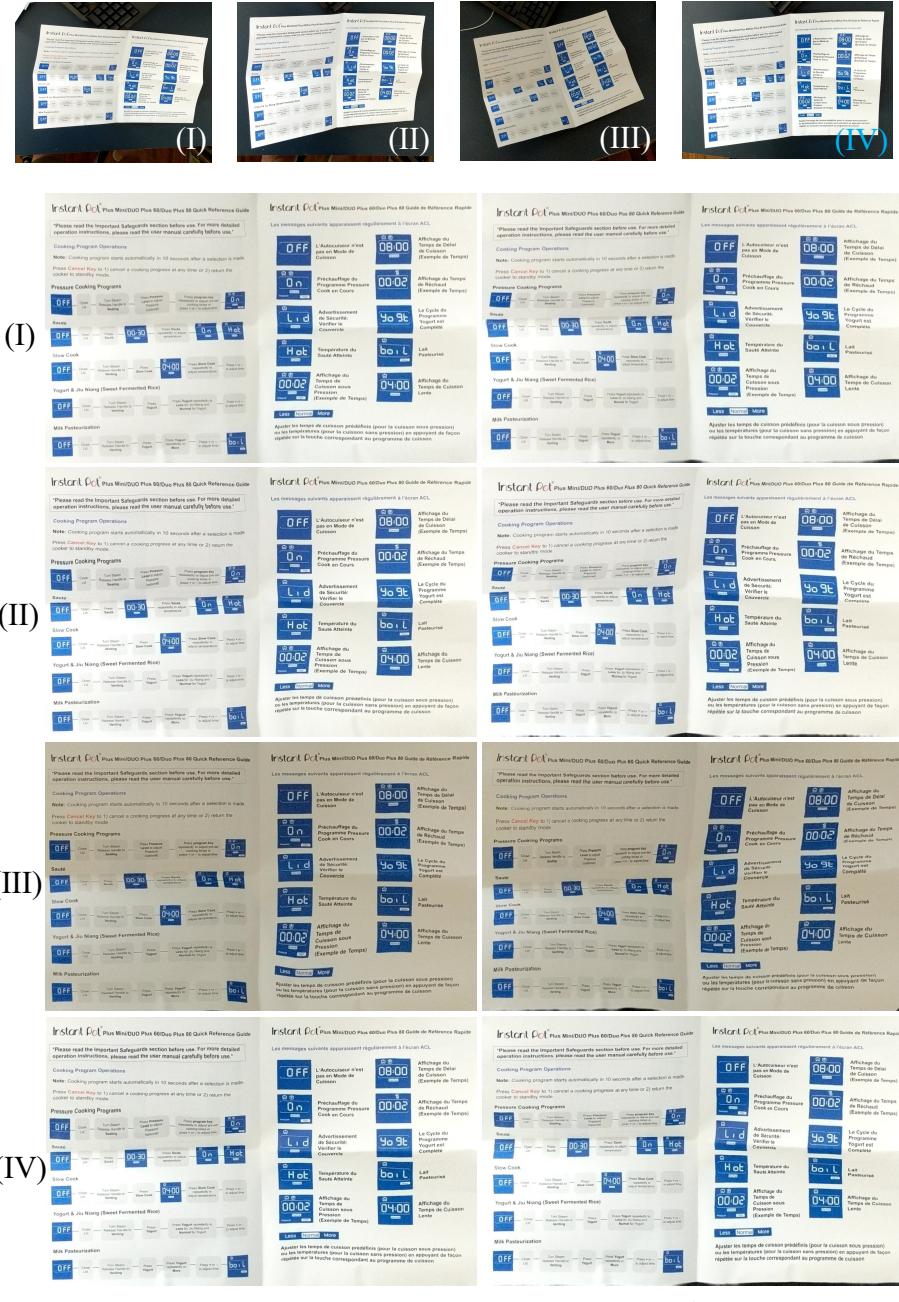


Fig. 14. Unwarping results on different views of a flyer. Top row shows the inputs. (a) Proposed, (b) DewarpNet. Our method generates good quality unwarping results with straighter text-lines.

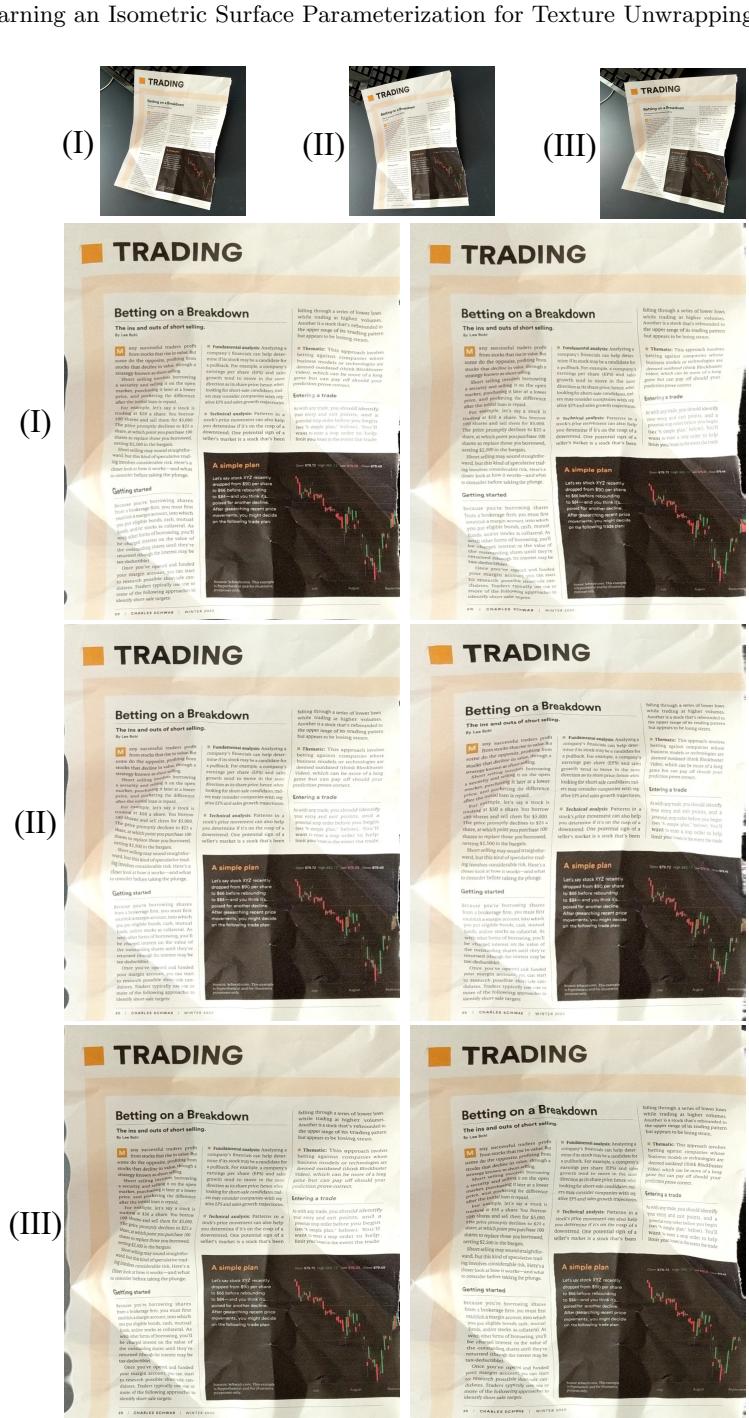
720
721
722
723
724725
726
727728
729
730731
732
733734
735
736737
738
739740
741
742743
744
745746
747
748749
750
751752
753
754755
756
757758
759
760761
762
763720
721
722
723
724725
726
727728
729
730731
732
733734
735
736737
738
739740
741
742743
744
745746
747
748749
750
751752
753
754755
756
757758
759
760761
762
763

Fig. 15. Unwarping results on different views of a magazine page. Top row shows the inputs. (a) Proposed, (b) DewarpNet. Our method generates good quality unwarping results with straighter text-lines.

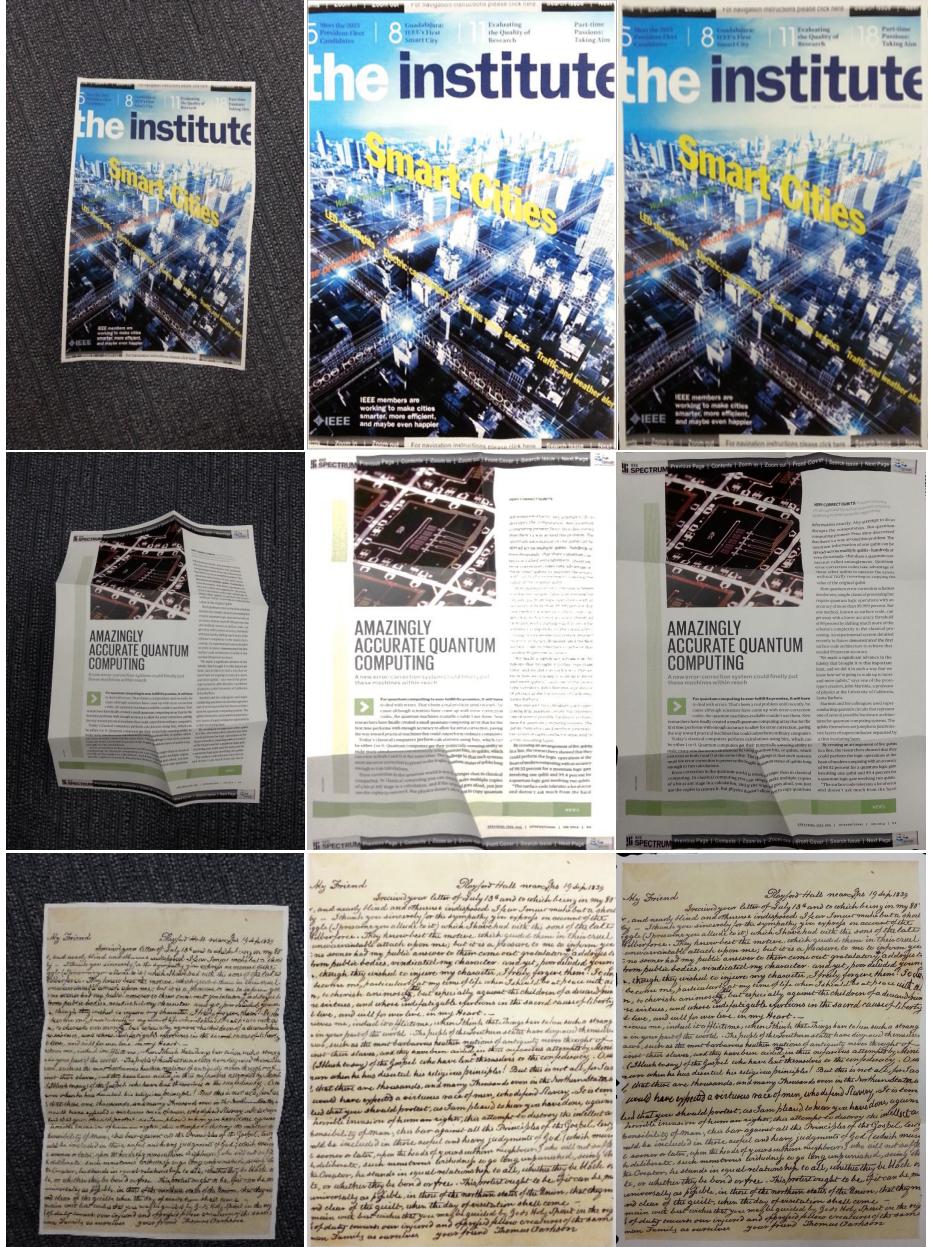


Fig. 16. Comparison with [16]: We show competitive unwarping results compared to a prior multi-view unwarping approach. A quantitative comparison could not be performed because high-res/original unwarped results are not publicly available.

810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854



Active Flattening of Curved Document Images via Two Structured Beams

Geling Meng*, Ying Wang*, Yiqian Xiong, Youming Xiang, Chuanqi Pan,
Institute of American Culture, Academy of Sciences (CAS),
Beijing, China, 100080, China
(Corresponding author: youming.xiang@iaoc.ac.cn)

Abstract: Document images captured by a digital camera often have non-planar surface distortions, the primary reason is that the camera is usually held in hand. In this paper, we propose a novel method to flatten curved document images. Our method is based on the active flattening framework, which is a well-known framework for dealing with non-rigid surface deformation. It is a two-step process: first, we estimate the initial shape of the document image; second, we deform the document image to a flat plane by using a series of soft-affine transformations. The proposed method can handle various curved document images with different contents. Experimental results on several curved document images show that our method is effective and efficient in the proposed model.

*Equal contribution. Correspondence to: Youming Xiang, Institute of American Culture, Academy of Sciences (CAS), Beijing, China, 100080, China (Corresponding author: youming.xiang@iaoc.ac.cn).

In a competitive situation, digital cameras often just advantages against the digital cameras in the market who offer low resolution and poor performance. The main reason is that the digital cameras are usually held in hand and the camera is usually held in hand. In this paper, we propose a novel method to flatten curved document images. This is especially true when we consider the non-planar surface distortion of the document images. There are several ways to handle such non-planar surface distortion. One way is to use a camera with a fixed lens, another way is to use a camera with a zoom lens, and the third way is to use a camera with a wide-angle lens. The last way is the most common way because it is the easiest way to handle the non-planar surface distortion.

For the second way, geometric distortion reduction of document images has received great attention and many methods have been proposed [1-10]. For example, [1] proposed a method for removing the curved document images [11-13], [14-16], [17-19], [20-22], [23-25]. These methods are mainly based on the multi-view geometry [1-10]. The methods mentioned above are mainly based on the multi-view geometry [1-10]. The methods mentioned above are mainly based on the multi-view geometry [1-10].

To simplify and make the extraction process easier, we propose a novel method to flatten curved document images. For it [21-25] derive a new frame-shifting, which is a very good way to handle the curved document images. They are a cylindrical surface to represent



Part 6 → Curv 6 End.

Active Flattening of Curved Document Images via Two Structured Beams

Geling Meng*, Ying Wang*, Yiqian Xiong, Youming Xiang, Chuanqi Pan,
Institute of American Culture, Academy of Sciences (CAS),
Beijing, China, 100080, China
(Corresponding author: youming.xiang@iaoc.ac.cn)

Abstract: Document images captured by a digital camera often have non-planar surface distortions, the primary reason is that the camera is usually held in hand. In this paper, we propose a novel method to flatten curved document images. Our method is based on the active flattening framework, which is a well-known framework for dealing with non-rigid surface deformation. It is a two-step process: first, we estimate the initial shape of the document image; second, we deform the document image to a flat plane by using a series of soft-affine transformations. The proposed method can handle various curved document images with different contents. Experimental results on several curved document images show that our method is effective and efficient in the proposed model.

*Equal contribution. Correspondence to: Youming Xiang, Institute of American Culture, Academy of Sciences (CAS), Beijing, China, 100080, China (Corresponding author: youming.xiang@iaoc.ac.cn).

In a competitive situation, digital cameras often just advantages against the digital cameras in the market who offer low resolution and poor performance. The main reason is that the digital cameras are usually held in hand and the camera is usually held in hand. In this paper, we propose a novel method to flatten curved document images. This is especially true when we consider the non-planar surface distortion of the document images. There are several ways to handle such non-planar surface distortion. One way is to use a camera with a fixed lens, another way is to use a camera with a zoom lens, and the third way is to use a camera with a wide-angle lens. The last way is the most common way because it is the easiest way to handle the non-planar surface distortion.

For the second way, geometric distortion reduction of document images has received great attention and many methods have been proposed [1-10]. For example, [1] proposed a method for removing the curved document images [11-13], [14-16], [17-19], [20-22], [23-25]. These methods are mainly based on the multi-view geometry [1-10]. The methods mentioned above are mainly based on the multi-view geometry [1-10].

To simplify and make the extraction process easier, we propose a novel method to flatten curved document images. For it [21-25] derive a new frame-shifting, which is a very good way to handle the curved document images. They are a cylindrical surface to represent

the page shape of an opened book. By doing so, they can easily extract the document images from the page. In this paper, we propose a novel method to flatten curved document images. Our method is based on the active flattening framework, which is a well-known framework for dealing with non-rigid surface deformation. It is a two-step process: first, we estimate the initial shape of the document image; second, we deform the document image to a flat plane by using a series of soft-affine transformations. The proposed method can handle various curved document images with different contents. Experimental results on several curved document images show that our method is effective and efficient in the proposed model.

*Equal contribution. Correspondence to: Youming Xiang, Institute of American Culture, Academy of Sciences (CAS), Beijing, China, 100080, China (Corresponding author: youming.xiang@iaoc.ac.cn).

In a competitive situation, digital cameras often just advantages against the digital cameras in the market who offer low resolution and poor performance. The main reason is that the digital cameras are usually held in hand and the camera is usually held in hand. In this paper, we propose a novel method to flatten curved document images. This is especially true when we consider the non-planar surface distortion of the document images. There are several ways to handle such non-planar surface distortion. One way is to use a camera with a fixed lens, another way is to use a camera with a zoom lens, and the third way is to use a camera with a wide-angle lens. The last way is the most common way because it is the easiest way to handle the non-planar surface distortion.

For the second way, geometric distortion reduction of document images has received great attention and many methods have been proposed [1-10]. For example, [1] proposed a method for removing the curved document images [11-13], [14-16], [17-19], [20-22], [23-25]. These methods are mainly based on the multi-view geometry [1-10]. The methods mentioned above are mainly based on the multi-view geometry [1-10].

To simplify and make the extraction process easier, we propose a novel method to flatten curved document images. For it [21-25] derive a new frame-shifting, which is a very good way to handle the curved document images. They are a cylindrical surface to represent

the page shape of an opened book. By doing so, they can easily extract the document images from the page. In this paper, we propose a novel method to flatten curved document images. Our method is based on the active flattening framework, which is a well-known framework for dealing with non-rigid surface deformation. It is a two-step process: first, we estimate the initial shape of the document image; second, we deform the document image to a flat plane by using a series of soft-affine transformations. The proposed method can handle various curved document images with different contents. Experimental results on several curved document images show that our method is effective and efficient in the proposed model.

*Equal contribution. Correspondence to: Youming Xiang, Institute of American Culture, Academy of Sciences (CAS), Beijing, China, 100080, China (Corresponding author: youming.xiang@iaoc.ac.cn).

In a competitive situation, digital cameras often just advantages against the digital cameras in the market who offer low resolution and poor performance. The main reason is that the digital cameras are usually held in hand and the camera is usually held in hand. In this paper, we propose a novel method to flatten curved document images. This is especially true when we consider the non-planar surface distortion of the document images. There are several ways to handle such non-planar surface distortion. One way is to use a camera with a fixed lens, another way is to use a camera with a zoom lens, and the third way is to use a camera with a wide-angle lens. The last way is the most common way because it is the easiest way to handle the non-planar surface distortion.

For the second way, geometric distortion reduction of document images has received great attention and many methods have been proposed [1-10]. For example, [1] proposed a method for removing the curved document images [11-13], [14-16], [17-19], [20-22], [23-25]. These methods are mainly based on the multi-view geometry [1-10]. The methods mentioned above are mainly based on the multi-view geometry [1-10].

To simplify and make the extraction process easier, we propose a novel method to flatten curved document images. For it [21-25] derive a new frame-shifting, which is a very good way to handle the curved document images. They are a cylindrical surface to represent

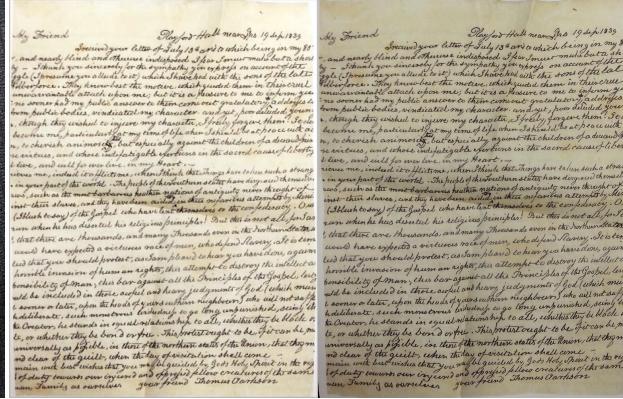
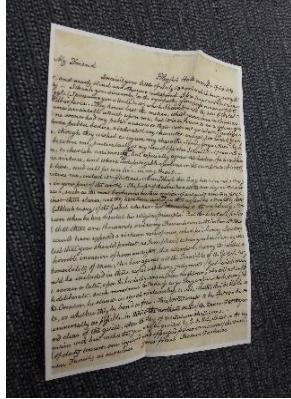
the page shape of an opened book. By doing so, they can easily extract the document images from the page. In this paper, we propose a novel method to flatten curved document images. Our method is based on the active flattening framework, which is a well-known framework for dealing with non-rigid surface deformation. It is a two-step process: first, we estimate the initial shape of the document image; second, we deform the document image to a flat plane by using a series of soft-affine transformations. The proposed method can handle various curved document images with different contents. Experimental results on several curved document images show that our method is effective and efficient in the proposed model.

*Equal contribution. Correspondence to: Youming Xiang, Institute of American Culture, Academy of Sciences (CAS), Beijing, China, 100080, China (Corresponding author: youming.xiang@iaoc.ac.cn).

In a competitive situation, digital cameras often just advantages against the digital cameras in the market who offer low resolution and poor performance. The main reason is that the digital cameras are usually held in hand and the camera is usually held in hand. In this paper, we propose a novel method to flatten curved document images. This is especially true when we consider the non-planar surface distortion of the document images. There are several ways to handle such non-planar surface distortion. One way is to use a camera with a fixed lens, another way is to use a camera with a zoom lens, and the third way is to use a camera with a wide-angle lens. The last way is the most common way because it is the easiest way to handle the non-planar surface distortion.

For the second way, geometric distortion reduction of document images has received great attention and many methods have been proposed [1-10]. For example, [1] proposed a method for removing the curved document images [11-13], [14-16], [17-19], [20-22], [23-25]. These methods are mainly based on the multi-view geometry [1-10]. The methods mentioned above are mainly based on the multi-view geometry [1-10].

To simplify and make the extraction process easier, we propose a novel method to flatten curved document images. For it [21-25] derive a new frame-shifting, which is a very good way to handle the curved document images. They are a cylindrical surface to represent

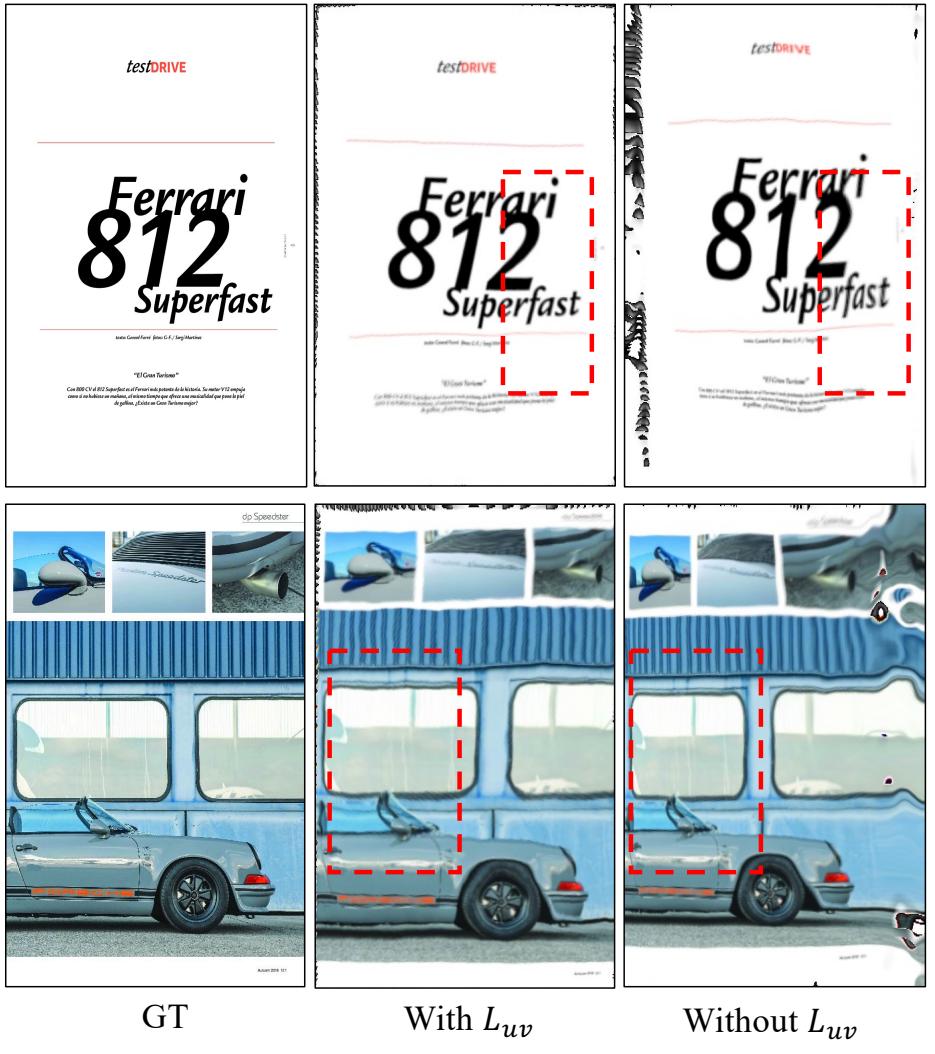


Input

You et al.

Proposed

Fig. 17. Comparison with [16]: We show competitive unwarping results compared to a prior multi-view unwarping approach. A quantitative comparison could not be performed because high-res/original unwarped results are not publicly available. The example with the dashed outline shows a failure case of our method: 'Real 6' (see figure 21).



888 **Fig. 18.** Usefulness of L_{uv} : Examples trained **without** L_{uv} show undesired stretches
889 and squeezes in the unwarped texture.

891 8 Details of weighting function used in L_z

893 We define L_z in Eq. 9 of the main submission:

$$895 \quad L_z = \frac{1}{|P_{in}|} \sum_{p \in P_{in}} w_p (\hat{z}_p - \hat{z}'_p)^2 \quad (1)$$

898 where $P \in P_{in}$ are the pixels for which ray-surface intersection is found and
899 $M_p = 1$. M_p , denote the pixel in the document mask M . M is a binary image,

855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899

where $M_p = 1$ denotes the pixel p is within the document region. w_p is a pre-calculated per-pixel weight based on the document mask (M). \hat{z}_p is the ray-surface intersection point obtained by sphere tracing, and \hat{z}'_p is the ray-surface intersection point predicted by F_z . To derive the 2D texture map of a 3D surface, constraint optimization-based techniques use user-defined keypoints [14]. The keypoints allow to constrain the 2D to 3D mapping estimation. For documents, we can consider the set of boundary points as the keypoints. From the application perspective, it helps to accurately map the texture boundary to the learned surface boundary (see Fig. 20(d) vs. (e) vs. (f)). Therefore, we employ a weighting function, which assigns a higher weight to the 3D surface points at the boundary. To implement $W(p)$ we use a Euclidean distance transform [3] on the document mask M , a binary image. Each pixel p , in the distance transformed image, D encodes the distance to the nearest non-zero pixel. We first normalize and invert the distance transformed image:

$$D^{norm} = \frac{D - \min(D)}{\max(D) - \min(D)}$$

$$D^{inv} = 1 - D^{norm}$$

Here $\max(\cdot)$ and $\min(\cdot)$ denote the maximum and minimum value of D_p over all the pixels. We assign the weights w_p as follows:

$$w_p = \begin{cases} 10.0, & \text{if } D_p^{inv} > 0.8 \\ 0.3, & \text{otherwise} \end{cases} \quad (2)$$

9 Training details of the UV prior network (\hat{F}_{uv})

We use an 8 layer MLP with a hidden layer of 512 units to learn the 3D to UV mapping prior for document shapes. Each hidden layer has a sine [12] activation function. The final layer uses a HardTanh activation function. To train \hat{F}_{uv} we utilize 10K UV mapped document meshes available in the Doc3D dataset. Each mesh is first registered with a $[-1, 1]$ uniform grid using a rigid transformation. Then the meshes are rendered in Blender [1] to obtain the projected geometry image (G) and the UV image (U). In G , each pixel p encodes the (X,Y,Z) coordinates. In U , p encodes the corresponding UV coordinates. During training, we randomly sample 10K pixels from each G as input to \hat{F}_{uv} and use the corresponding pixels in U as the ground-truth. We optimize the L1 loss for 150 epochs between the predicted and the ground-truth UV coordinates using the Adam optimizer with an initial learning rate of 10^{-5} . The learning rate is halved every 50 epochs. Following NeRF [9], we use a high dimensional Fourier mapping ($\chi_k : \mathbb{R} \rightarrow \mathbb{R}^{2k}$) to learn high-frequency details in the shape and the UV space. We empirically set the number of Fourier bands, $k = 10$.

945 10 Initializing S and F_z

946
947 We can start optimizing S from the standard IDR initialization (SDF of a
948 sphere). However, we notice that a better initialization can significantly improve
949 the training time as well as the quality of the shape reconstruction. For object-
950 specific applications like document unwarping, we found that initializing S with
951 a similar object can significantly reduce the training time and converge in half
952 the number of iterations (from 400K to 200K). Furthermore, we also found that
953 initializing F_z to produce a planar point cloud can further reduce our training
954 convergence time to ~ 6 hours (80K-100K iterations). To this end, we pre-train
955 F_z to produce a plane.

956 **Pre-training of F_z .** To initialize F_z such that it produces a planar point cloud,
957 we pre-train it by inputting points sampled from the UV space and predict the
958 point cloud with $Z = 0$. We employ Chamfer distance as a loss function between
959 ground truth and predicted 3D points. The ground truth points are sampled
960 from a plane. Additionally, we also apply the conformality constraints (defined
961 in section 3.2 of the main submission) for this pre-training step. The predicted
962 plane is bounded in $[-0.5, 0.5]$ in our implementation. This training step is quite
963 straightforward and converges in a few epochs.

964 11 Unwarping and texture editing details

965 To unwarped an input image, we determine a pixel at $p = (x, y)$ in the input
966 image should be projected to (u, v) in the unwarped image. Here the unwarped
967 image refers to the texture space. The coordinates (u, v) and p are associated by
968 F_z and τ : For a (u, v) coordinate, its corresponding point in 3D is obtained by
969 $\hat{z}'_p = F_z(u, v)$. Given the camera parameter τ , \hat{z}'_p is projected to p in the input
970 image. Thus, we can find its corresponding pixel in the input image for each pixel
971 in the unwarped image, which is all we need for unwarping. More specifically,
972 we use standard image projection and bilinear sampling [7] to implement the
973 unwarping step (see Fig. 19). The unwarping process can be realized as a grid
974 sampling step from the warped document image to a 2D rectangular uniform
975 grid. We can perform this sampling operation with a grid $G \in \mathbb{R}^{(H \times W \times 2)}$ and a
976 bi-linear sampler. Here H and W denote the height and the width of the grid.
977 Each location in G encodes a pixel coordinate \hat{p} of the input image.

978 At test time we sample F_z in a uniform grid and project using the known
979 camera pose (τ) to obtain the pixel coordinates. More specifically, sampling F_z
980 in a uniform grid $\in [0, 1]$ yields a uniform 2D grid $R_z \in \mathbb{R}^{(H \times W \times 3)}$. Each (u, v)
981 in R_z encodes a 3D coordinate of the document surface. The R_z representation
982 of the 3D shape is analogous to geometry images [6]. We obtain G from R_z with
983 a standard projection:

$$984 \hat{p} = K [R|T] \bar{\mathbf{z}} \quad (3)$$

985 Here, $\bar{\mathbf{z}}$ is the homogeneous coordinate representation of \mathbf{z} . $K \in \mathbb{R}^{3 \times 3}$, $[R|T] \in$
986 $\mathbb{R}^{4 \times 4}$, denote the intrinsic and extrinsic parameters of the camera.

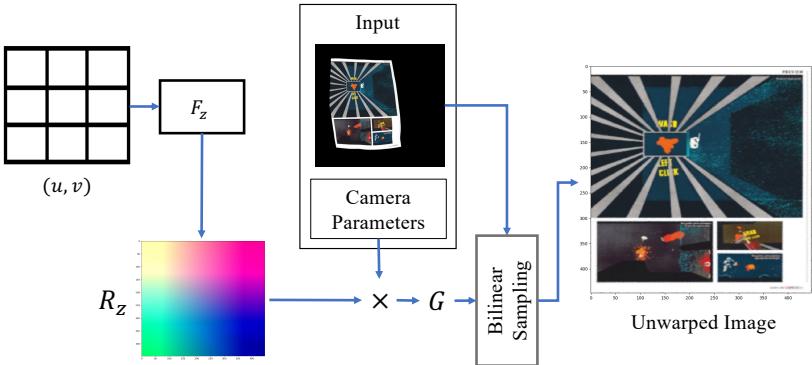


Fig. 19. Unwarping steps at test time: R_z denotes the flattened geometry in the texture space. Using the camera projection matrix for each view, we can obtain the unwarping grid G . \times denotes matrix multiplication. G can be used to sample [7] the input image to get the unwarped image.

For the texture editing task, we first unwarp the image, then edit the texture, and finally warp each edited pixel p back to the original position using the predicted texture coordinates (t_p) . We can utilize the same bilinear sampling operation as the unwarping step.

12 Pre-processing details for the real scenes

To train our proposed approach on the real scenes, we first obtain the camera poses using COLMAP [11]. Each scene in the real data has 5-25 views. We pre-process the camera poses to a spherical domain following [8]. Since all the training meshes used to train F_{uv} are aligned with a $[-1, 1]$ uniform grid, we apply a fixed pre-computed rigid-transformation on the estimated 3D shape during the joint training of S , F_{uv} , and F_z . Specifically, we use a 6D rigid transformation, with two parameters for rotation (axis-angle representation), three for translation, and one for scale. We first train a vanilla IDR [15] for 10K iterations. Then we obtain a 3D point cloud representation of the surface by sphere-tracing the IDR estimated SDF. Each point in the point cloud is a ray-surface intersection point. Note that we do not need a very accurate geometry at this step. Therefore it is not required to optimize the SDF until convergence. Now we obtain the desired rigid transformation by optimizing the Chamfer distance between the obtained surface point cloud and 10K points sampled from a 2D uniform regular grid $\in [-1, 1]$. We use SGD [13] with a learning rate of 0.001 and momentum 0.9 and optimize for 10K iterations. Later, At every iteration during the joint training, we apply the estimated rigid transformation on the sphere traced surface points (\hat{z}_p) and use the transformed points as an input to the F_{uv} .

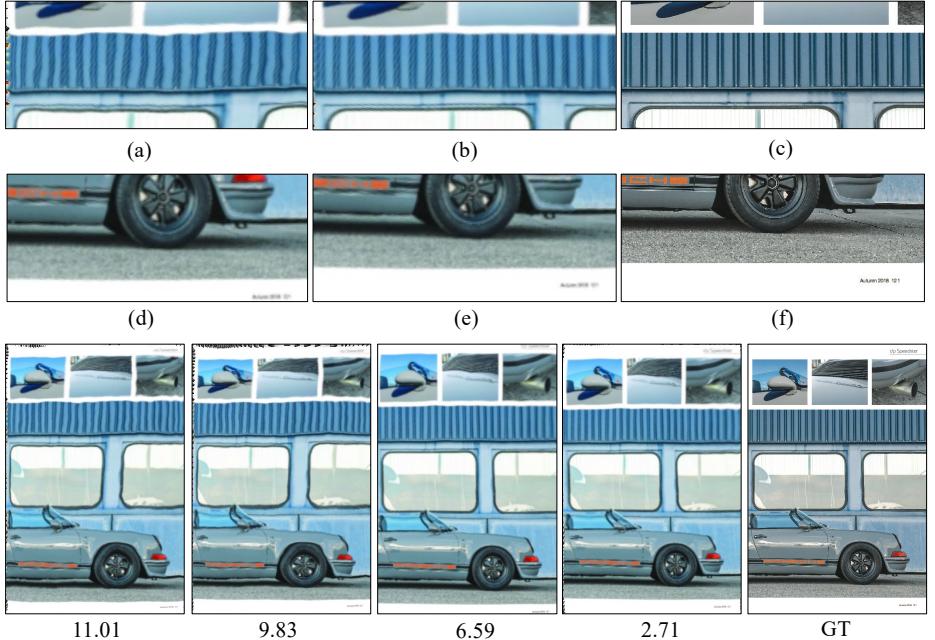


Fig. 20. Weighted L_z , and conformality effects. Top and middle row: (a) without conformality constraints, (b) with conformality constraints, (d) $w_p = 1$ in weighted L_z , (e) weighted L_z with w_p calculated using Eq. 2, (c,f) ground-truth; bottom-row (left-to-right): without conformality constraints and weighted L_z ; only with weighted L_z ; only with conformality constraints and weighted L_z ; ground-truth scan. Numbers in bottom denote the respective LD values.

13 Detailed ablation figure

We show a more detailed example of Fig. 8 of the main submission in Fig. 20, with zoomed-in regions to demonstrate the effect of the different components of L_T (Eq. 10 in the main paper).

14 Limitations

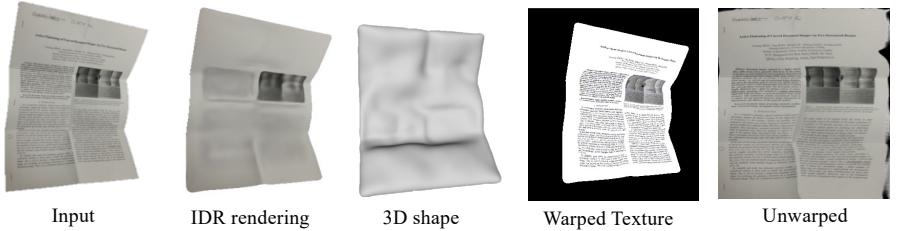
In the following, we discuss few potential limitations of our method:

- **3D reconstruction:** The main limitation of our method follows from IDR [15]. Inadequate number of images of a scene with large texture-less regions lead to inferior 3D reconstruction which affects our unwarping result (see section 15).
- **Training time:** The current approach takes ~ 6 hours to train a model and separate models must be trained for every scene which makes it unsuitable for real time applications. Runtime improvement will be addressed as a future work (see section 16).

- 1080 – **Need for masks:** We assume masks are available for every image. Although
 1081 masks are currently provided as manual inputs, we believe it's fairly straight-
 1082 forward to train a foreground-background segmentation models to automate
 1083 the task.

1084 15 Example of a failure case

1085 Our method might fail due to imperfect 3D reconstruction. We show one such
 1086 case for a scene from [16]. Mainly, there are two reasons for failure cases: first,
 1087 fewer views (only 5), and second, insufficient textured documents. IDR has in-
 1088 sufficient information to reconstruct the 3D shape. As a result of the poor 3D
 1089 shape, our texture parameterization network produces an inferior unwarping re-
 1090 sult. For illustration, we show the reconstructed 3D shape, warped texture, and
 1091 unwarped texture in Fig. 21.



1092 **Fig. 21.** Shows a failure case of our method due to inferior 3D reconstruction. This
 1093 happened due to fewer available views for the scene and insufficient texture.

1109 16 Training Time

1110 Our proposed method for a scene can be trained in approximately 6 hours for
 1111 448×448 resolution images using a single Titan Xp GPU. The current training
 1112 time per scene is high compared to DewarpNet's inference time which makes
 1113 it unsuitable for real-time applications. However, we would like to note that in
 1114 the current implementation sphere-tracing takes almost 50-60% of the running
 1115 time. With a faster version of the sphere-tracing we can readily achieve a faster
 1116 framework. Moreover, neural rendering is an active research field and there are
 1117 multiple other works that are focusing on improving the speed and generalization
 1118 abilities [5,2]. Therefore, a faster training can be achieved following any newer
 1119 or faster alternatives of IDR.

1122 1123 References

- 1124 1. Blender - a 3D modelling and rendering package

- 1125 2. Bergman, A.W., Kellnhofer, P., Wetzstein, G.: Fast training of neural lumigraph
1126 representations using meta learning (2021)
- 1127 3. Borgefors, G.: Distance transformations in digital images. Computer vision, graph-
1128 ics, and image processing **34**(3), 344–371 (1986)
- 1129 4. Das, S., Ma, K., Shu, Z., Samaras, D., Shilkrot, R.: DewarpNet: Single-image doc-
1130 ument unwarping with stacked 3D and 2D regression networks. In: Proceedings of
the International Conference on Computer Vision (2019)
- 1131 5. Garbin, S.J., Kowalski, M., Johnson, M., Shotton, J., Valentin, J.: Fastnerf: High-
1132 fidelity neural rendering at 200fps (2021)
- 1133 6. Gu, X., Gortler, S.J., Hoppe, H.: Geometry images. ACM Transactions on Graphics
(TOG) **21**(3), 355–361 (2002)
- 1134 7. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer
1135 networks. In: Advances in Neural Information Processing Systems (2015)
- 1136 8. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi,
1137 R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescrip-
1138 tive sampling guidelines. ACM Transactions on Graphics (TOG) (2019)
- 1139 9. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng,
1140 R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In:
1141 Proceedings of the European Conference on Computer Vision (2020)
- 1142 10. Ramon, E., Triginer, G., Escur, J., Pumarola, A., Garcia, J., Giro-i Nieto, X.,
1143 Moreno-Noguer, F.: H3d-net: Few-shot high-fidelity 3d head reconstruction. In:
1144 Proceedings of the IEEE/CVF International Conference on Computer Vision. pp.
1145 5620–5629 (2021)
- 1146 11. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection
1147 for unstructured multi-view stereo. In: European Conference on Computer Vision
(ECCV) (2016)
- 1148 12. Sitzmann, V., Martel, J.N., Bergman, A.W., Lindell, D.B., Wetzstein, G.: Implicit
1149 neural representations with periodic activation functions. In: arXiv (2020)
- 1150 13. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization
1151 and momentum in deep learning. In: International conference on machine learning.
1152 pp. 1139–1147. PMLR (2013)
- 1153 14. Tzur, Y., Tal, A.: FlexiStickers: Photogrammetric texture mapping using casual
1154 images. In: Proceedings of the ACM SIGGRAPH Conference on Computer Graph-
1155 ics. Association for Computing Machinery (2009)
- 1156 15. Yariv, L., Kasten, Y., Moran, D., Galun, M., Atzmon, M., Ronen, B., Lipman, Y.:
1157 Multiview neural surface reconstruction by disentangling geometry and appear-
1158 ance. Advances in Neural Information Processing Systems **33** (2020)
- 1159 16. You, S., Matsushita, Y., Sinha, S., Bou, Y., Ikeuchi, K.: Multiview Rectification
1160 of Folded Documents. IEEE Transactions on Pattern Analysis and Machine Intel-
1161 ligence (2017)