Practical Wide-Angle Portraits Correction with Deep Structured Models

Jing Tan[1*] Shan Zhao[1*] Pengfei Xiong[2*] Jiangyu Liu[1] Haoqiang Fan[1] Shuaicheng Liu[3,1 †]

[1]Megvii Research [2]Tencent
[3]University of Electronic Science and Technology of China
{tanjing, zhaoshan, liujiangyu, fhq}@megvii.com
xiongpengfei2019@gmail.com, liushuaicheng@uestc.edu.cn
https://github.com/TanJing94/Deep_Portraits_Correction

Abstract

Wide-angle portraits often enjoy expanded views. However, they contain perspective distortions, especially noticeable when capturing group portrait photos, where the background is skewed and faces are stretched. This paper introduces the first deep learning based approach to remove such artifacts from freely-shot photos. Specifically, given a wide-angle portrait as input, we build a cascaded network consisting of a LineNet, a ShapeNet, and a transition module (TM), which corrects perspective distortions on the background, adapts to the stereographic projection on facial regions, and achieves smooth transitions between these two projections, accordingly. To train our network, we build the first perspective portrait dataset with a large diversity in identities, scenes and camera modules. For the quantitative evaluation, we introduce two novel metrics, line consistency and face congruence. Compared to the previous state-of-the-art approach, our method does not require camera distortion parameters. We demonstrate that our approach significantly outperforms the previous state-of-the-art approach both qualitatively and quantitatively.

1. Introduction

With the popularity of wide-angle cameras on smartphones, photographers can take pictures with broad vision. However, a wider field-of-view often introduces a stronger perspective distortion. All wide-angle cameras suffer from distortion artifacts that stretch and twist buildings, road ridges and faces, as shown in Fig. 1 (a).

There are relatively few works targeting on the perspective distortion correction in portrait photography[8, 7]. Previous methods apply perspective undistortion using camera calibrated distortion parameters [18, 23, 2, 6], which projects the image onto a plane for undistortion, as shown in Fig. 1 (b). Compared with Fig. 1 (a), the lines at the background become straight. Unfortunately, the faces are also projected as a plane, becoming unnaturally wider and asymmetric. It is then evident that background and faces require different types of corrections, to be separately handled with different strategies. As traditional calibration-based methods[1, 12, 19, 32] can only correct distortion in background regions, we need new ways to process faces.

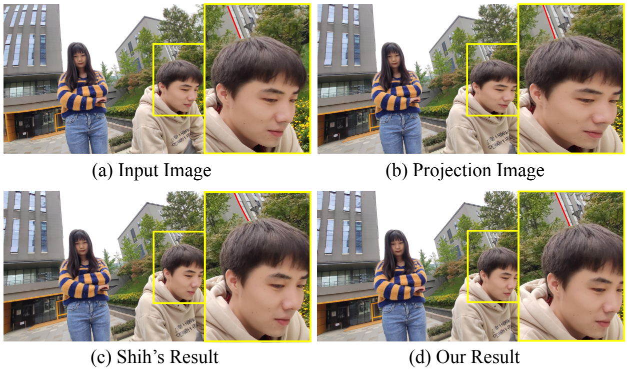

(a) Input Image
(b) Projection Image
(c) Shih's Result
(d) Our Result

Figure 1. Examples of distorted and corrected photographs. (a) the original distortion image with curved background and distorted faces. (b) projection image with straight lines. (c) result by Shih *et al.* [21], and (d) result of the proposed deep learning method. Both the background and faces are corrected in (d).

Recently, Shih *et al.* [21] proposes to deform a mesh which adapts to the stereographic projection [22] on facial regions, and applies perspective projection over the background, enabling different handling of the background and faces. However, a new problem arises, where the smooth transition between faces and background regions is nontrivial. In addition, the method [21] requires camera distortion parameters as well as the portrait segmentation mask as additional inputs. Fig. 1 (c) shows the result, where the face

[*]Equal contribution
[†]Corresponding author

in the corner has been over corrected and appear deformed.

In contrast, our approach does not rely on any prior calibrated parameters, thus being more flexible to various conditioned portraits. Compared to the mesh-based energy minimization [21], our deep solution works well in balancing the perspective projection on the background and stereographic projection on the faces, delivering smooth transitions between them. Fig. 1 (d) shows our result.

To this end, we propose a deep structured network to generate a content-aware warping flow field, which both straightens the background lines through perspective undistortion, and adapts to the stereographic projection on facial regions, notably achieving smooth transitions between these two projections. Our cascaded network includes a Line Correction Network (LineNet) and a Portrait Correction Network (ShapeNet). Specifically, given an input image, the LineNet is first applied to produce a flow field to undistort the perspective effects for line correction, where a Line Attention Module (LAM) is introduced to facilitate the localization of lines. Second, the projected image is fed into the ShapeNet for face correction, within which a Face Attention Module (FAM) is introduced for face localization. Furthermore, we design a Transition Module (TM) between LineNet and ShapeNet to ensure smooth transitions.

As there is no proper dataset readily available for training, we build a high-quality wide-angle portrait dataset. Specifically, we capture portrait photos by smartphones with various wide-angle lenses and then interactively correct them with a specially designed content-aware mesh warping tool, yielding $5,854$ pairs of input and output images for training. Moreover, for quantitative evaluations, we introduce two novel metrics, Line Straightness Metric (LineAcc) and Shape Congruence Metric (ShapeAcc) to evaluate the line straightness and face correctness accordingly. Previously, evaluation can only be made qualitatively.

Experimental results show that our approach can correct distortions in wide-angle portraits. Compared with calibration-based opponents, our method can rectify the faces faithfully without camera parameters. Compared with Shih's method [21], our method is calibration-free, and achieves good transitions between background and face regions. Both qualitative and quantitative evaluations are provided to validate the effectiveness of our method. Our main contributions are:

- The first deep learning based method to automatically remove distortions in wide-angle portraits from unconstrained photographs, without camera calibration and distortion parameters, delivering a better universality.
- We design a structured network to remove the distortion on background and portraits respectively, achieving smooth transitions between perspective-rectified background and stereographic-rectified faces.
- We provide a new perspective portrait dataset for image undistortion with a wide range of subject identities, scenes and camera modules. In addition, two universal metrics are designed for the quantitative evaluation.

2. Related Works

2.1. Image Distortions

Image distortions are often introduced when projecting a 3D scene to a 2D image plane through a limited Field-of-View (FOV) [35]. The perspective projection often distorts objects that are far away from the camera center [25]. Mesh-based methods have been attempted with user constraints, e.g., dominant straight lines and vanishing points, to cope with potential undesired mesh deformations [3, 13]. In this work, our method is calibration-free, which not only rectifies background perspective distortions, but also corrects the face distortions with a deep neural network.

2.2. Content-Aware Warping

Mesh-based content-aware warping have been widely applied in image and video manipulations, including image stitching [33, 5], video stitching [9, 15, 29], panorama rectangling [11], content-aware rotation [10], perspective manipulation [2], image retargeting [28], video retargeting [26, 27], stereoscopic editing [4], and video stabilization [16, 17, 34]. In this work, we propose deep structured models to produce content-aware flow fields for image warping.

2.3. Face Undistortion

Several methods are proposed to correct distorted faces.[24, 20, 30]. Fried *et al.* proposed to fit a 3D face model to modify the relative pose between the camera and the subject by manipulating the camera focal length, yielding photos with different perspectives [8]. A more related work is proposed by Shih *et al.* [21], which corrects wide-angle portraits by content-aware image warping. However, it requires the camera parameters and results in distortion either on the background or on the faces. In contrast, our method is calibration-free and achieves smooth transitions with a deep neural network.

3. Methodology

The proposed network contains two sub-networks, line correction network (LineNet) and portrait correction network (ShapeNet), Fig. 2 shows the overall architecture and the pipleline of the wide-angle image correction. As can be seen, the first LineNet generates a perspective projection flow from the given distortion image to project the image as flattened. Then the ShapeNet predicts the face correction flow from the flattened image. In order to make the two networks work on different deformations, we design

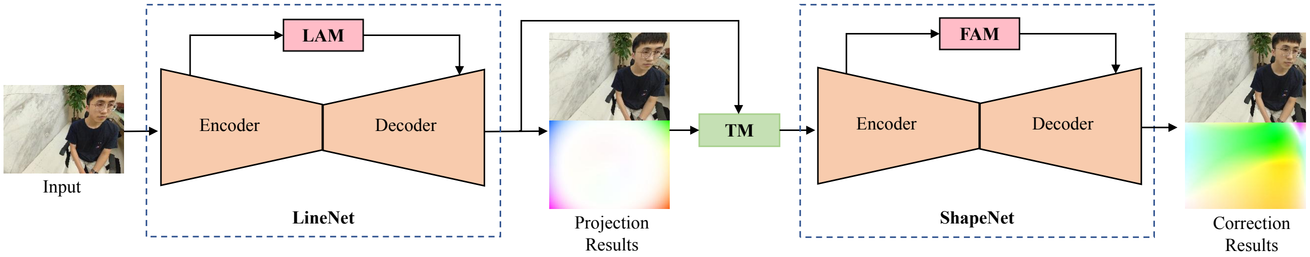


Figure 2. An overview of our network architecture.

self attention modules LAM and FAM to work on the two networks respectively. These two sub-networks are bridged by a transition module TM. Finally, the projection image is transformed to correction image with correction flows.

3.1. Line Correction Network

LineNet As shown in Fig. 3, a standard encoder decoder network is used to predict the corresponding deformed flow from a single image. It consists of two phases, namely, down-scale feature maps and up-scale feature maps. Given an original image I, we adopt *ResNeXt* [31] network as the backbone to extract feature maps. Then, the feature pyramid is input into up-scale decoders to generate the final flows. In each decoder, the feature maps are inferred by two 3×3 convolutional layers, an add operator with the previous encoded feature, and two additional 3×3 convolutional layers. Then, a deconvolution operator is adopted to up-scale the decoder feature map. After that, the projection image is obtained through the flow-based warping.

LAM Since the main purpose of the LineNet is to straighten the bended lines as shown in Fig. 2, we design a line attention module (LAM) to learn various lines. Different from other multi-branch methods, we use the intermediate results of encoder and decoder for prediction. LAM consists of two main blocks, channel attention block (CAB) and spatial attention block (SAB). As shown in Fig. 3, CAB is applied on high level features to generate the attention in channel perspective. It contains a global pooling operation and two convolution layers to generate a $C \times 1$ attention map from a given feature map $C \times H \times W$. All the high-level CAB outputs are concatenated into the following SAB. SAB combines the original encoder features and CAB features to output the final spatial attention feature maps.

In order to locate the lines in the image, we apply the two lowest level SAB outputs as additional line supervisions. Therefore, we use Sobel operator to extract the corresponding edge from the original image as the ground truth, and then calculate the loss between the edge map and the SAB output. LAM has two advantages. On one hand, it uses the information of high-level and low-level to strengthen the global and local features of the image. On the other hand, the edge constraint makes the encoder and decoder feature pay more attention to the edges. Without additional calculations, the model of line correction becomes feasible.

3.2. Portrait Correction Network

ShapeNet After obtaining the results of line correction, we need to further produce the flow of face correction. We use the same network architecture as LineNet. Portrait Correction Network also outputs a flow map, but it aims to correct the face areas while leaving the background unchanged.

FAM Different from the LAM of the first network, the main purpose of ShapeNet is to rectify the face area. Because the face is only a small part of the image, when the face is deformed, the boundary between the face and the background would inevitably be distorted. In order to describe the transition region more accurately, the most direct method is to segment the human image to obtain the accurate head boundary. However, depending on the energy transfer of the attention module, accurate segmentation is not the most necessary. Instead, we generate a heatmap of human face based on the results of face detection to adaptively learn the changes of face and transition region, as shown in Fig. 4 (d). In the same way, face heatmap is used as the supervisions of face areas.

TM Considering that while ShapeNet is performing face correction, the perspective projection transformation from the LineNet should be maintained, so in order to make it easier for LineNet and ShapeNet to keep the consistency of the non-portrait area, we proposed a transition module (TM) to transfer the distortion from the LineNet to ShapeNet.

The TM has three parts, as shown in Fig. 3. The first part is the decoder feature maps of the penultimate layer, following with convolution and up sampling. The second part is the final flow map, and the third part is the projection image. Three feature maps are concatenated together as inputs to the ShapeNet. This module contains image features, location features, as well as hidden semantic information.

3.3. Training and Inference

Loss Function. Our model is learned in an end-to-end manner. For each sub-network, we adopt L2 loss between the generated and ground truth of flows and image respectively. Besides, we apply a boundary preservation L2 loss to enhance the edge accuracy. Sobel operator is adopted to generate the edge of ground truth and predication. Different

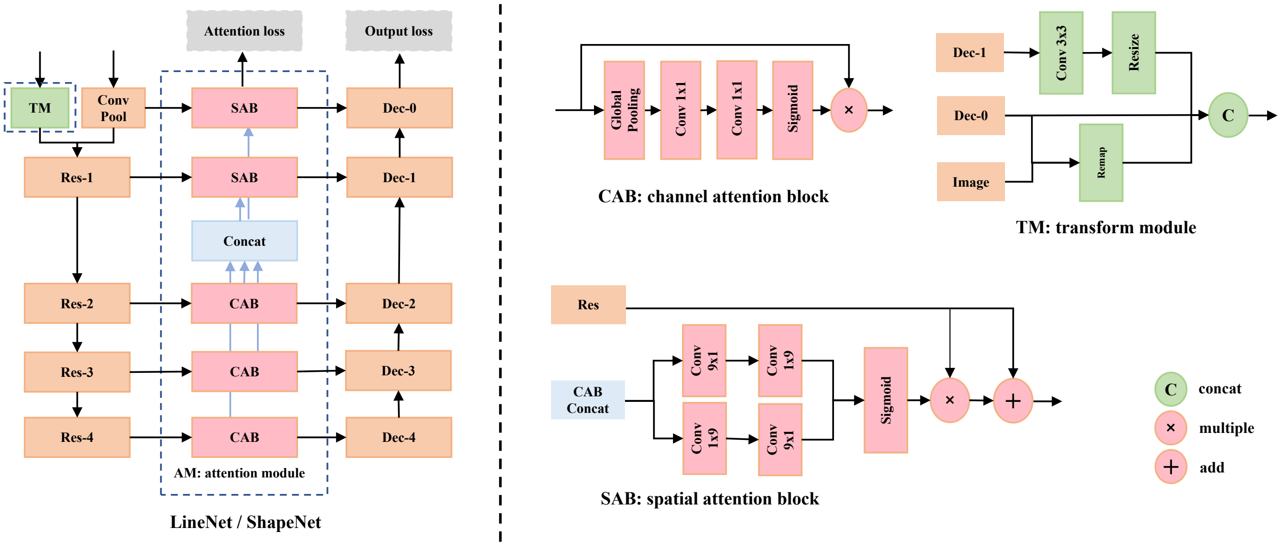


Figure 3. The overall structure of our model. **ShapeNet** and **LineNet** share the similar structure, with **TM** avaliable only for **ShapeNet**.

from boundary supervision used in several previous works, Sobel operator can also smooth flows to avoid aliasing in the undistorted images. The formulas are as follows:

$$L_{line} = ||F_{flow} - I_{flow}||_{2,s2} + ||F_{proj} - I_{proj}||_{2,s2} \quad (1)$$

$$L_{shape} = ||F_{flow} - I_{flow}|_{2,s2} + ||F_{out} - I_{out}||_{2,s2} \quad (2)$$

where L_{line}, L_{shape} are the losses of the output of LineNet and ShapeNet. Two kinds of losses, L2 and $sobel_L2$ are used on both flow and images. Similar loss is also applied to the attention modules.

$$L_{LAM} = ||F_{lam} - I_{edge}||_2 \quad (3)$$

$$L_{FAM} = ||F_{fam} - I_{face}||_2 \quad (4)$$

L_{LAM} is the difference between LAM output and the provided edge ground truth, while L_{FAM} is the difference between FAM output and the labeled face heatmap. The total loss function is the weighted sum of the above losses.

$$L = \lambda_1 L_{LAM} + \lambda_2 L_{FAM} + \lambda_3 L_{line} + \lambda_4 L_{shape} \quad (5)$$

The $\lambda_{1,2,3,4}$ is used to balance the importance among the reconstruction and attention losses. we set them to 5, 5, 1, 1 respectively in all experiments.

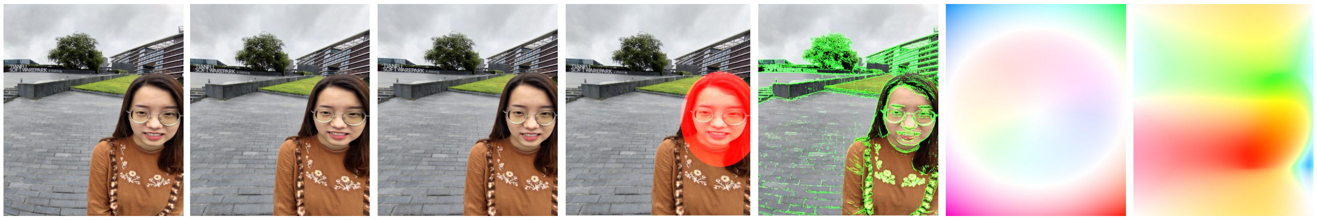
Inference. In the inference stage, the generated two flows can be combined into one flow to describe the offset directly from the original image to the final corrected. Given an image, it is first reduced to 256 × 384 to obtain two flow maps of 256 × 384. After fusion, the fused flow is resized to the original size to generate the correction image.

4. Data Preparation

There is no dataset of paired portraits. We therefore create a novel training dataset by ourselves which contains distorted and undistorted image pairs with various camera modules, scenes and identities. We used 5 different ultra wide-angle camera of smartphones, and photographed over 10 people in several scenes. The number of people in each photo ranges from 1 to 6. Overall, over 5,000 images were collected. Given a distorted portrait image, it is non-trivial to obtain its corresponding distortionless image. We propose to correct wide-angle portrait into a distortion-free image manually. To achieve this, we propose to first run an improved Shih's [21] algorithm iteratively and then further improve the results by our designed manual tool.

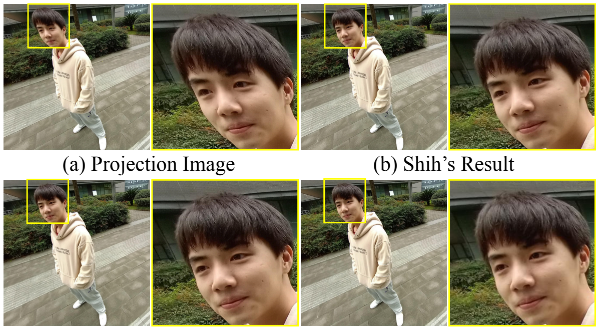
We notice that the edges near the faces are often distorted in [21]. We improve the results of [21] by adding explicit line constraints [15]. Fig 5 (a) is the input image. Fig 5 (b) is Shih's result [21], and Fig 5 (c) is our improved result. Notably, this improvement can alleviate the problem to some extent, but cannot be perfect. Specifically, we run our improved method iteratively in the following steps: 1) run the algorithm and obtain the initial results; 2) for results that look unnatural, we re-run the algorithm by adjusting the hyper-parameters. 3) Repeat step 2 until the correction results converge. In the experiment, about half of the images can be satisfactory after the first optimization, and the remaining half need 5 more iterations to complete. Finally, the image which cannot get good results is discarded.

Now, we obtain an initial dataset. To further improve the quality, we correct some unsatisfactory parts by our designed manual tool. Our manual tool is mesh-based, with as-rigid-as-possible quads constraints [16] and line-preserving constraints [15]. Users can drag the image content by a mouse to drive the mesh deformation interactively. As shown in Fig. 5 (d), we further correct the result of (c) for improvements. Fig. 4 shows an example. The flow motions, Fig. 4 (f) and (g), can be obtained during our manual correction, which are used as the guidance to the network LineNet and ShapeNet, accordingly. Face mask and lines,



(a) Input (b) Projection Image (c) Correction Image (d) Face Mask (e) Lines (f) Projection Map (g) Correction Map

Figure 4. Training example. (a)input image. (b)projection image with straight lines. (c)corrected image by enhanced Shih's method [21] and our manual tool. (d)face mask created by face detection used in FAM. (e)lines created by line detection used in LAM. (f)projection flow from (a) to (b), which is the guidance of LineNet. (g)correction flow from (b) to (c), which is the guidance of ShapeNet.



(a) Projection Image (b) Shih's Result

(c) Shih's Improved Result (d) Manually corrected

Figure 5. Example of our dataset creation. We built a tool that can adjust the warping of meshes for ground-truth generation. We use the results from Shih *et al.* [21] as input, and correct their problematic regions manually. (a) input image and the cropped region. (b) result by Shih *et al.* [21]. Please notice the bending of lines and unnatural shape of the face. (c)Shih *et al.* [21]'s improved result. (d) our manually corrected results of (c).

Fig. 4 (d) and (e), are used for LAM and FAM, accordingly.

5. Metrics

In this section, we introduce two novel evaluation metrics: Line Straightness Metric (LineAcc) and Shape Congruence Metric (ShapeAcc). As far as we know, there is no suitable quantitative metric in the field of distortion correction. The accuracy of quantitative calculation needs a corrected image as a reference, where our manually corrected images are used.

Line Straightness Metric: The salient lines should keep straight after correction. We mark salient lines in the test dataset and then calculate the curvature variation of the marked lines. For each line L, we uniformly sample n points p_0, p_1, ..., p_n. Then, the slope variation of the line can be calculated as:

$$S_g = \frac{y_{g_0} - y_{g_n}}{x_{g_0} - x_{g_n}} \tag{6}$$

$$LS = 1 - (\frac{1}{n} \sum_{i=0,..,n-1} [\frac{y_{d_i} - y_{d_{i-1}}}{x_{d_i} - x_{d_{i-1}}} - S_g]) \tag{7}$$

where $P_{g_i} = [x_{g_i}, y_{g_i}]$, $P_{d_i} = [x_{d_i}, y_{d_i}]$ depicts the location of corresponding point in reference and distortion images.

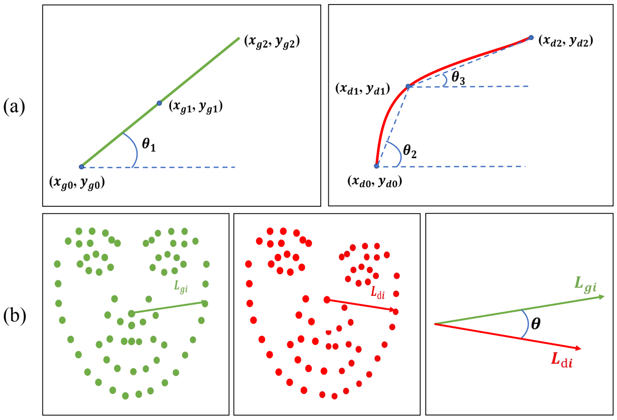
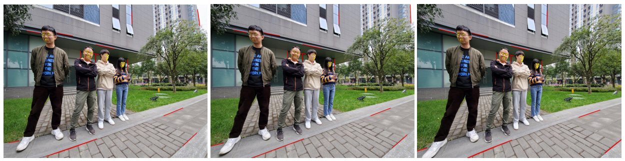


Figure 6. (a) Line Straightness metric (LineAcc) and (b) Shape Congruence Metric (ShapeAcc).



Line acc: 56.310 Face acc: 95.950 (a) Input Img

Line acc: 100 Face acc: 91.475 (b) Projection Img

Line acc: 73.095 Face acc: 96.152 (c) Our Result

Figure 7. (a) input image with LineAcc and ShapeAcc. (b) Perspective undistorted image, where LineAcc increases and ShapeAcc decreases. (c) Our result with two scores balanced, both of which are increased compared with (a).

LS is the similarity between slope of these two lines. The line is more curved, the value of LineAcc is smaller. Fig. 6 (a) shows an illustration. We mark the salient lines near the image boundary and around the portraits.

Shape Congruence Metric: Given a face, we label the landmarks on the result and the reference image, and then calculate the similarity between the two groups of landmarks. Fig. 6 (b) shows an example, where the vectors are produced from the nose landmark to other landmarks.

$$FC = \frac{1}{n} \sum_{i=0,..,n-1} [cos(L_{g_i}, L_{d_i})] \tag{8}$$

$$cos(L_{g_i}, L_{d_i}) = \|L_{g_i}\| \|L_{d_i}\| cos\theta \tag{9}$$

where L_g and L_d depict the corresponding landmarks in the reference image and the result image. Fig. 7 shows an example of the two metrics. Fig. 7 (a) is the input with original LineAcc and ShapeAcc. Fig. 7 (b) is the line corrected by perspective undistortion, where the LineAcc increases, and the ShapeAcc decreases. Fig. 7 (c) is our result. We achieve a balanced scores regarding the two metrics. Note that, the dimensions of the two metrics are not the same.

6. Experiments

In this section, we first analyze the impact of each module by the ablation study, and then based on the best experimental module configuration, we conduct another two data ablation experiments on the training set of different phone modules to verify the generalization of the network. Finally, we make a quantitative and qualitative comparison with the related works, and also compare with some smartphones with wide-angle correction.

Implementation details All experiments are carried out on our training set. We use the standard data augmentation procedure to expand the diversity of training samples, including horizontal flip, random crop and scaling. Besides, we sample uniformly according to the number of people in the image, and increase the proportion of people in corner to solve the problem of imbalance between the portraits at the center and corner. We use the ADAM optimizer [14] with an initial learning rate of 5e-3 , which is divided by 10 at the 150-th epoch, and the total training epoch is 200.

Test set We construct a test set that contains both face landmarks and lines near the faces and at the corners of the image to appreciably evaluate the wide-angle portrait correction. It contains a total of 129 original wide-angle images of 5 different phone modules and the corresponding calibrated images according to camera parameters. Experiments are evaluated on our test set and Shih's [21] test set.

6.1. Ablation Studies

In this subsection, we step-wise decompose our approach to reveal the effect of each component. The basic LineNet is built on the straightforward encoder-decoder structure, which takes the original distorted image as input, and predicts the perspective projection flow map. Based on this, we verify the function of LAM, ShapeNet, TM and FAM module respectively. The LineAcc and ShapeAcc are adopted for the evaluation. LineAcc is evaluated on results of both LineNet and ShapeNet in Table. 1. As seen, all the proposed modules contribute to the performance.

LineNet The basic LineNet improves the Projection LineAcc from 66.064 to 66.745, which is obviously beyond Shih's [21] approach.

LAM Compare 1) and 2) in Table 1, after integrating the Line Attention Module with the basic LineNet, the LineAcc is improved from 66.745 to 66.856, because the Line Attention Module facilitates the learning of line-awareness features. As shown in Fig. 8 (a), the straight line in the corner is obviously straighter, with the constraint of LAM.

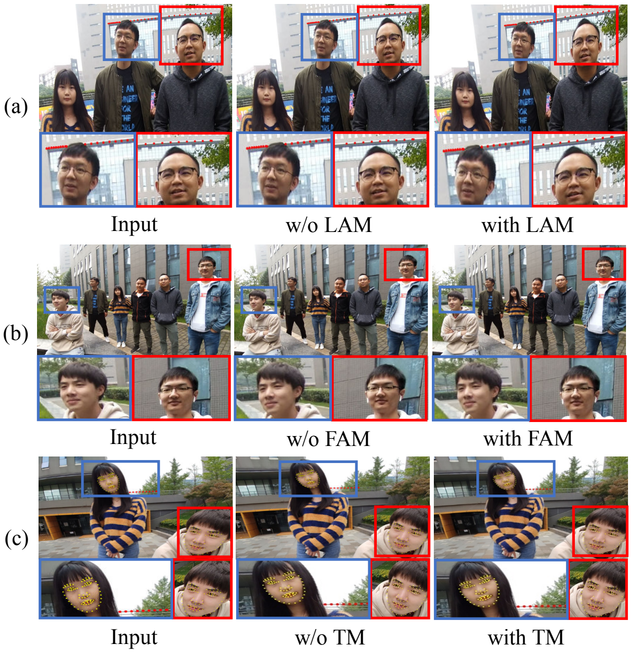


Figure 8. Visualization of ablation study. (a), (b), (c) represent the different performance of the network with or without LAM, FAM and TM, and prove their effects respectively.

ShapeNet Next, we integrate the Basic ShapeNet with LineNet. The evaluation result can be seen in 3) of Table. 1. ShapeNet affects the line accuracy of LineNet to a certain extent, but obviously significantly improves the accuracy of face correction. Furthermore, LAM shows strong robustness and improves the final line calibration accuracy in the end-to-end network by comparing 3) and 4).

FAM Furthermore, FAM is applied onto ShapeNet. As shown in 4) and 5) of Table. 1, compared to the standard structure, FAM improves the ShapeAcc Metric from 97.472 to 97.479, which is due to the improved confidence of the face areas. The same conclusion can be found in Fig. 8 (b).

TM The purpose of TM is to further balance the deformation of salient straight lines and faces. As depicted in 7) in Table. 1, the correction LineAcc and ShapeAcc are improved from 66.484 to 66.575, and from 97.479 to 97.485, respectively. As shown in Fig. 8 (c), the transition area between head and background is more natural while TM transfers the line projection features to ShapeNet.

Lmk loss Finally, we verify the effectiveness of the landmark loss. Experimental results show that landmark loss does not bring a particularly large improvement in the accuracy of face correction, because FAM has improved the accuracy of face correction with a relatively large margin.

Table 1. Ablation study of the proposed network. ”LineNet”, ”LAM”, ”ShapeNet”, ”TM” and ”FAM” refer to the basic LineNet, Line Attention Module, ShapeNet, Transtion Module, Face Attention Module, respectively. ”Lmk Loss” refers to the adaption of landmark loss onto ShapeNet. ”Proj LineAcc”, ”Corr LineAcc”, and ”ShapeAcc” refer the Projection and Correction LineAcc Metric of two networks and the final Shape Correction Accuracy. In addition, results of three methods are adopted for comparisons. They are Input Image, projection result with calibrated params, and Shih’s [21] result. The best is marked in red and the second best is marked in blue.

No.	LAM	ShapeNet	TM	FAM	Lmk Loss	Proj LineAcc	Corr LineAcc	ShapeAcc
1) LineNet						66.745	\	97.380
2)	✓					66.856	\	97.391
3)		✓				66.707	66.439	97.458
4)	✓	✓				66.873	66.472	97.472
5)	✓	✓		✓		66.938	66.484	97.479
6)	✓	✓			✓	66.985	66.541	97.473
7)	✓	✓	✓	✓		67.069	66.575	97.485
8)	✓	✓	✓	✓	✓	67.135	66.784	97.490
9) Input						66.064	66.064	97.455
10) Proj Img						\	\	96.876
11) Shih [21]						66.143	66.143	97.253

Table 2. Quantitative comparisons of ours and Shih [21] on different test sets. The first three rows in the table indicate the models training without note data, without vivo data and with the total training set, respectively. And correspondingly test on note, vivo, ours whole test and google test. The best two scores are shown in red and blue.

No.	note testset		vivo testset		all testset		google	
	LineAcc	ShapeAcc	LineAcc	ShapeAcc	LineAcc	ShapeAcc	LineAcc	ShapeAcc
1) ours wot note	67.605	97.061	64.997	98.341	66.464	97.464	\	\
2) ours wot vivo	68.299	97.109	63.418	98.361	66.324	97.483	\	\
3) ours with all	68.683	97.115	65.148	98.363	66.784	97.490	64.650	97.499
4) google(Shih [21])	66.886	97.267	63.087	98.238	66.143	97.253	61.551	97.464

But on the other hand, it improves the accuracy of line correction. This is mainly due to more accurate face edge constraints, which smooths the transition at face boundary regions. Based on the superposition of the above modules, the accuracy of the proposed model is much higher than that of the Shih’s [21], in terms of both LineAcc and ShapeAcc.

Generalization verification Based on the model 8) in Table. 1, we select two smartphone modules from the training and testing set respectively for cross testing, and the full test set and test set of [21] are also used for the final evaluation to verify the robustness and generalization of the network. As shown in Table. 2, comparing 1) 3) and 2) 3), adding data of different models can improve the robustness. At the same time, except for the ShapeAcc on note testset, our methods perform better than Shih [21] on almost all 4 test sets, which indicates the good performance of generalization.

6.2. Comparison with Other Methods

Fig. 9 shows the visual comparisons with Shih’s results [21]. The projection image can correct lines but cause serious distortion on face regions, while the stereographic projection can maintain the faces but suffer from structure bending. Shih’s results [21] can seek a balance between the faces and the background. However, some structures are still bended at the background, and some faces still suffer from a bit distortion, unbalanced with the body. In contrast, our results are more natural in the correction of the head, and the transition area between the portrait and the background is smoother. The line in the background is also closer to the result of perspective projection, and the faces look more natural. More notably, in the second row of Fig. 9, our results can correct the rightmost face while keeping the architectural lines above it still straight, while the lines above the face in Shih’s [21] are obviously deformed. Metrics in Table. 1 and Table. 2 also confirm the conclusion, since the accuracy in terms of both LineAcc and ShapeAcc has been greatly improved from Shih’s [21].

6.3. Comparison with Other Phones

Furthermore, we compare the results with some smartphones with wide-angle portrait correction. Two flagship phones of Xiaomi and iPhone are applied as comparisons. As shown in Fig. 10, there is serious stretching of portraits and some bending of background in the result of iPhone 12. The result of Xiaomi 10 is close to the result of perspective projection, and there is still slight deformation of the face. Our results are significantly better than others, as the face is undistorted while correcting the background lines.

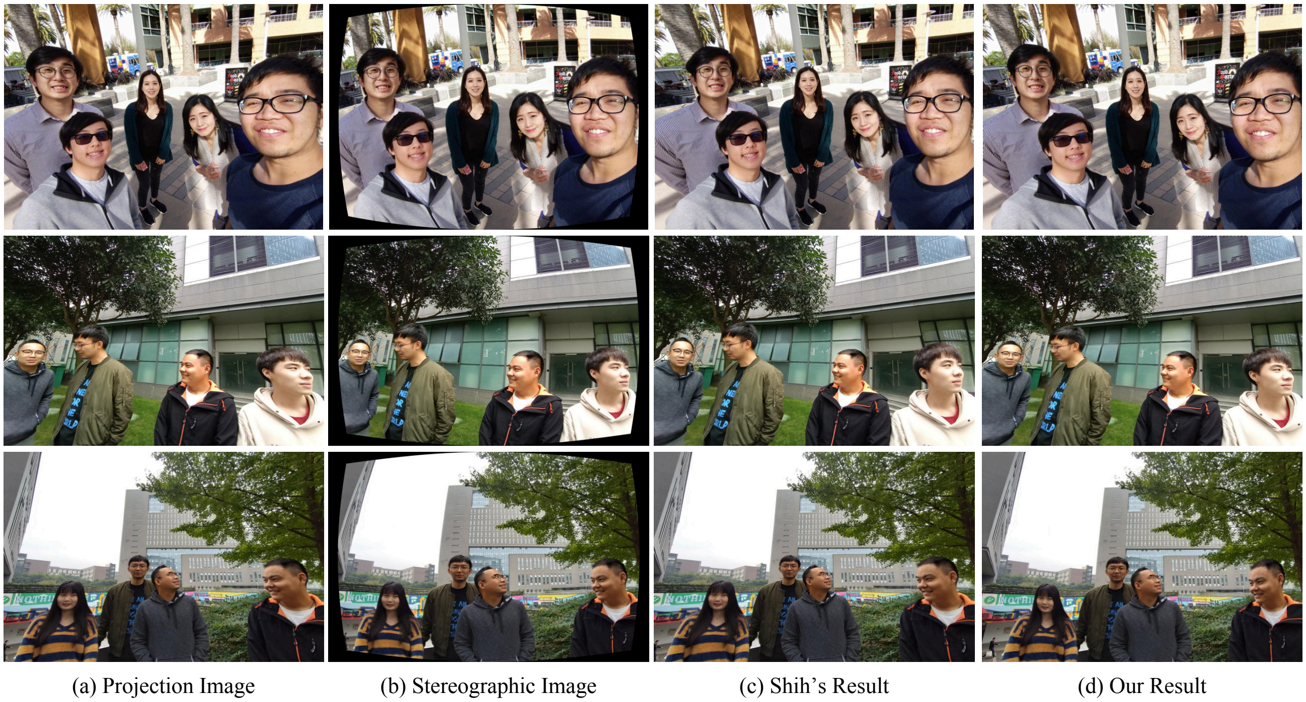


Figure 9. Qualitative evaluation of undistortion methods. Notice the coordination of face area and line curvatures in the transition area.

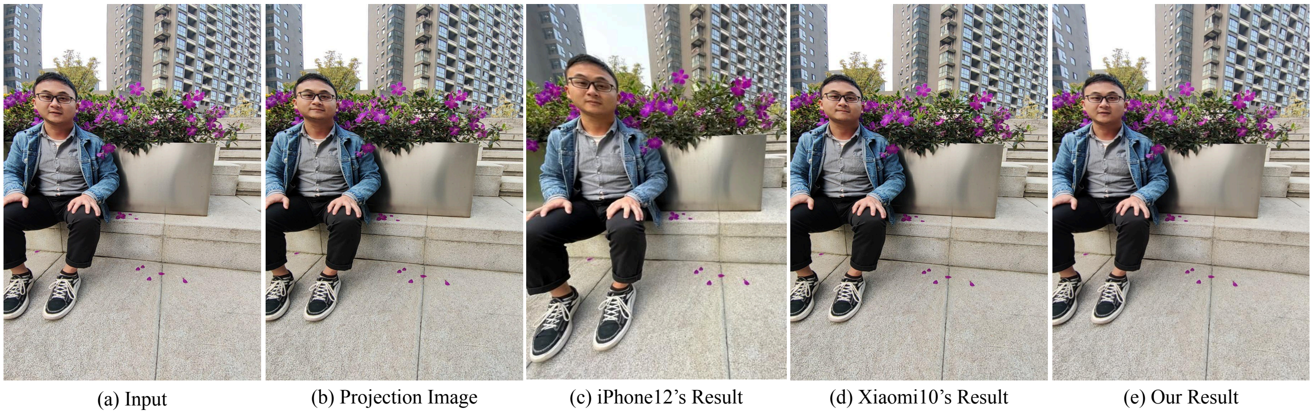


Figure 10. Qualitative comparison between our method and some phones with wide-angle portrait correction.

More comparisons are shown in our project page, including corrections for photos from the Internet, as well as some failure cases.

7. Conclusion

This paper proposes a deep structured network to automatically correct distorted portraits in wide-angle photos, which applies perspective projection to background and stereographic projection to portraits, and achieves a smooth transition between them. It does not rely on any prior calibrated parameters, thus being more flexible to various conditioned portraits. Besides, we construct a high-quality wide-angle portrait dataset and design two metrics for quantitative evaluations. Considerable experiments verify the robustness and generalization of our method. We believe this work is of great practical value.

Acknowledgement

This research was supported in part by National Natural Science Foundation of China (NSFC, No.61872067, No.61720106004), in part by Research Programs of Science and Technology in Sichuan Province (No.2019YFH0016).

References

[1] Jun-Sik Kim Ankur Datta and Takeo Kanade. Accurate camera calibration using iterative refinement of control points. In *Proc. ICCVW*, pages 1201–1208, 2009.

[2] Robert Carroll, Aseem Agarwala, and Maneesh Agrawala. Image warps for artistic perspective manipulation. *ACM Trans. Graphics*, 29(4):1–9, 2010.

[3] Robert Carroll, Maneesh Agrawala, and Aseem Agarwala. Optimizing content-preserving projections for wide-angle images. *ACM Trans. Graphics*, 28(3):43, 2009.

[4] Che-Han Chang, Chia-Kai Liang, and Yung-Yu Chuang. Content-aware display adaptation and interactive editing for stereoscopic images. *IEEE Trans. on Multimedia*, 13(4):589–601, 2011.

[5] Che-Han Chang, Yoichi Sato, and Yung-Yu Chuang. Shape-preserving half-projective warps for image stitching. In *Proc. CVPR*, pages 3254–3261, 2014.

[6] Song-Pei Du, Shi-Min Hu, and Ralph R Martin. Changing perspective in stereoscopic images. *IEEE Trans. on Visualization and Computer Graphics*, 19(8):1288–1297, 2013.

[7] Elise A Piazza Emily A Cooper and Martin S Banks. The perceptual basis of common photographic practice. In *Journal of vision*, pages 8–12, 2012.

[8] Ohad Fried, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. Perspective-aware manipulation of portrait photos. *ACM Trans. Graphics*, 35(4):1–10, 2016.

[9] Heng Guo, Shuaicheng Liu, Tong He, Shuyuan Zhu, Bing Zeng, and Moncef Gabbouj. Joint video stitching and stabilization from moving cameras. *IEEE Trans. on Image Processing*, 25(11):5491–5503, 2016.

[10] Kaiming He, Huiwen Chang, and Jian Sun. Content-aware rotation. In *Proc. CVPR*, pages 553–560, 2013.

[11] Kaiming He, Huiwen Chang, and Jian Sun. Rectangling panoramic images via warping. *ACM Trans. Graphics*, 32(4):1–10, 2013.

[12] Janne Heikkila. Geometric camera calibration using circular control points. In *IEEE Trans. on pattern analysis and machine intelligence*, page 1066–1077, 2000.

[13] Yoshihiro Kanamori, Nguyen Huu Cuong, and Tomoyuki Nishita. Local optimization of distortions in wide-angle images using moving least-squares. In *Conference on Computer Graphics*, pages 51–56, 2011.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *arXiv preprint arXiv:1412.6980*, 2014.

[15] Kaimo Lin, Shuaicheng Liu, Loong-Fah Cheong, and Bing Zeng. Seamless video stitching from hand-held camera inputs. In *Computer Graphics Forum*, volume 35, pages 479–487, 2016.

[16] Feng Liu, Michael Gleicher, Hailin Jin, and Aseem Agarwala. Content-preserving warps for 3d video stabilization. *ACM Trans. Graphics*, 28(3):44, 2009.

[17] Shuaicheng Liu, Lu Yuan, Ping Tan, and Jian Sun. Bundled camera paths for video stabilization. *ACM Trans. Graphics*, 32(4):1–10, 2013.

[18] Darko Pavić, Volker Schönefeld, and Leif Kobbelt. Interactive image completion with perspective correction. *The Visual Computer*, 22(9-11):671–681, 2006.

[19] Haiyuan Wu Qian Chen and Toshikazu Wada. Camera calibration with two arbitrary coplanar circles. In *Proc. ECCV*, page 521–532, 2004.

[20] Pietro Perona Ronnie Bryan and Ralph Adolphs. Perspective distortion from interpersonal distance is an implicit visual cue for social judgments of faces. *PloS one*, 7(9):e45301, 2012.

[21] YiChang Shih, Wei-Sheng Lai, and Chia-Kai Liang. Distortion-free wide-angle portraits on camera phones. *ACM Trans. Graphics*, 38(4):1–12, 2019.

[22] Benny Stale Svardal, Kjell Einar Olsen, and Odd Ragnar Andersen. Stereographic projection system, Apr. 15 2003. US Patent 6,547,396.

[23] Mahdi Abbaspour Tehrani, Aditi Majumder, and M Gopi. Correcting perceived perspective distortions using object specific planar transformations. In *Proc. ICCP*, pages 1–10, 2016.

[24] Paul Beardsley Bob Sumner Thabo Beeler, Bernd Bickel and Markus Gross. High-quality single-shot capture of facial geometry. *ACM Trans. Graphics*, 29(4):1–9, 2010.

[25] Dhanraj Vishwanath, Ahna R Girshick, and Martin S Banks. Why pictures look right when viewed from the wrong place. *Nature neuroscience*, 8(10):1401–1410, 2005.

[26] Yu-Shuen Wang, Hongbo Fu, Olga Sorkine, Tong-Yee Lee, and Hans-Peter Seidel. Motion-aware temporal coherence for video resizing. *ACM Trans. Graphics*, 28(5):1–10, 2009.

[27] Yu-Shuen Wang, Hui-Chih Lin, Olga Sorkine, and Tong-Yee Lee. Motion-based video retargeting with optimized crop-and-warp. *ACM Trans. Graphics*, 29(4):1–9, 2010.

[28] Yu-Shuen Wang, Chiew-Lan Tai, Olga Sorkine, and Tong-Yee Lee. Optimized scale-and-stretch for image resizing. *ACM Trans. Graphics*, 27(5):1–8, 2008.

[29] Wei Jiang and Jinwei Gu. Video stitching with spatial-temporal content-preserving warping. In *Proc. CVPRW*, pages 42–48, 2015.

[30] Matteo Ruggero Ronchi Xavier P Burgos-Artizzu and Pietro Perona. Distance estimation of an unknown person from a portrait. In *Proc. ECCV*, page 313–327, 2014.

[31] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proc. CVPR*, pages 1492–1500, 2017.

[32] Jonathan Eisenmann Matthew Fisher Emiliano Gambaretto Sunil Hadap Yannick Hold-Geoffroy, Kalyan Sunkavalli and Jean-Franc¸ois Lalonde. A perceptual measure for deep single image camera calibration. In *Proc. CVPR*, page 2354–2363, 2018.

[33] Julio Zaragoza, Tat-Jun Chin, Michael S Brown, and David Suter. As-projective-as-possible image stitching with moving dlt. In *Proc. CVPR*, pages 2339–2346, 2013.

[34] Fang-Lue Zhang, Xian Wu, Hao-Tian Zhang, Jue Wang, and Shi-Min Hu. Robust background identification for dynamic video editing. *ACM Trans. Graphics*, 35(6):1–12, 2016.

[35] Denis Zorin and Alan H Barr. Correction of geometric perceptual distortions in pictures. In *Computer graphics and interactive techniques*, pages 257–264, 1995.