Learning Perspective Undistortion of Portraits

Yajie Zhao[1,*], Zeng Huang[1,2,*], Tianye Li[1,2], Weikai Chen[1], Chloe LeGendre[1,2], Xinglei Ren[1], Ari Shapiro[1], and Hao Li[1,2,3]

[1]USC Institute for Creative Technologies
[2]University of Southern California
[3]Pinscreen

We propose a learning-based method to remove perspective distortion. We show input photos (b) (e), our undistortion results (c) (f), and reference images (d) (g) captured simultaneously using a beam splitter rig (a). Our approach handles even extreme perspective distortions.

Abstract

Near-range portrait photographs often contain perspective distortion artifacts that bias human perception and challenge both facial recognition and reconstruction techniques. We present the first deep learning based approach to remove such artifacts from unconstrained portraits. In contrast to the previous state-of-the-art approach [23], our method handles even portraits with extreme perspective distortion, as we avoid the inaccurate and error-prone step of first fitting a 3D face model. Instead, we predict a distortion correction flow map that encodes a per-pixel displacement that removes distortion artifacts when applied to the input image. Our method also automatically infers missing facial features, i.e. occluded ears caused by strong perspective distortion, with coherent details. We demonstrate that our approach significantly outperforms the previous state-of-the-art [23] both qualitatively and quantitatively, particularly for portraits with extreme perspective distortion or facial expressions. We further show that our technique benefits a number of fundamental tasks, significantly improving the accuracy of both face recognition and 3D reconstruction and enables a novel camera calibration technique from a single portrait. Moreover, we also build the first perspective portrait database with a large diversity in identities, expression and poses.

1. Introduction

Perspective distortion artifacts are often observed in portrait photographs, in part due to the popularity of the "selfie" image captured at a near-range distance. The inset images, where a person is photographed from distances of 160cm and 25cm, demonstrate these artifacts.

When the object-to-camera distance is comparable to the size of a human head, as in the 25cm distance example, there is a large proportional difference between the camera-to-nose and camera-to-ear distances. This difference creates a face with unusual proportions, with nose and eyes appearing larger and ears vanishing all together [56].

Perspective distortion in portraits not only influences the way humans perceive one another [9], but also greatly impairs a number of computer vision-based tasks, such as face verification and landmark detection. Prior research [39, 40] has demonstrated that face recognition is strongly compromised by the perspective distortion of facial features. Additionally, 3D face reconstruction from such portraits is highly inaccurate, as geometry fitting starts from biased facial landmarks and distorted textures.

Correcting perspective distortion in portrait photography is a largely unstudied topic. Recently, Fried et al. [23] investigated a related problem, aiming to manipulate the relative pose and distance between a camera and a subject in a given portrait. Towards this goal, they fit a full perspective camera and parametric 3D head model to the input image and performed 2D warping according to the desired change in 3D. However, the technique relies on a potentially inaccurate 3D reconstruction from facial landmarks biased by perspective distortion. Furthermore, if the 3D face model fitting step failed, as it could for extreme perspective distortion or dramatic facial expressions, so would their broader pose manipulation method. In contrast, our approach does not rely on model fitting from distorted inputs and thus can handle even these challenging inputs. Our GAN-based synthesis approach also enables high-fidelity inference of any occluded features, not considered by Fried et al. [23].

In our approach, we propose a cascaded network that maps a near-range portrait with perspective distortion to its distortion-free counterpart at a canonical distance of $1.6m$ (although any distance between $1.4m \sim 2m$ could be used as the target distance for good portraits). Our cascaded network includes a distortion correction flow network and a completion network. Our distortion correction flow method encodes a per-pixel displacement, maintaining the original image's resolution and its high frequency details in the output. However, as near-range portraits often suffer from significant perspective occlusions, flowing individual pixels often does not yield a complete final image. Thus, the completion network inpaints any missing features. A final texture blending step combines the face from the completion network and the warped output from the distortion correction network. As the possible range of per-pixel flow values vary by camera distance, we first train a camera distance prediction network, and feed this prediction to the distortion correction network.

Training our proposed networks requires a large corpus of paired portraits with and without perspective distortion. However, to the best of our knowledge, no previously existing dataset is suited for this task. As such, we construct the first portrait dataset rendered from 3D head models with large variations in camera distance, head pose, subject identity, and illumination. To visually and numerically evaluate the effectiveness of our approach on real portrait photographs, we also design a beam-splitter photography system to capture portrait pairs of real subjects simultaneously on the same optical axis, eliminating differences in poses, expressions and lighting conditions.

Experimental results demonstrate that our approach removes a wide range of perspective distortion artifacts (e.g., increased nose size, squeezed face, *etc*), and even restores missing facial features like ears or the rim of the face. We show that our approach significantly outperforms Fried et al. [23] both qualitatively and quantitatively for a synthetic dataset, constrained portraits, and unconstrained portraits from the Internet. We also show that, when applied as a preprocessing step, our approach improves a wide range of fundamental tasks in computer vision and computer graphics, including face recognition/verification, landmark detection on near-range portraits (such as head mounted cameras in visual effects), and 3D face model reconstruction for avatar creation and 3D photos generation(Section 6.3). Additionally, our novel *camera distance prediction* provides accurate camera calibration from a single portrait.

Our main contributions can be summarized as follows:

- The first deep learning based method to automatically remove perspective distortion from an unconstrained near-range portrait, benefiting a wide range of fundamental tasks in computer vision and graphics.
- A novel and accurate camera calibration approach that only requires a single near-range portrait as input.
- A new perspective portrait database for face undistortion with a wide range of subject identities, head poses, camera distances, and lighting conditions.

2. Related Work

Face Modeling. We refer the reader to [43] for a comprehensive overview and introduction to the modeling of digital faces. With advances in 3D scanning and sensing technologies, sophisticated laboratory capture systems [5, 6, 8, 21, 24, 38, 41, 57] have been developed for high-quality face reconstruction. However, 3D face geometry reconstruction from a single unconstrained image remains challenging. The seminal work of Blanz and Vetter [7] proposed a PCA-based morphable model, which laid the foundation for modern image-based 3D face modeling and inspired numerous extensions including face modeling from internet pictures [33], multi-view stereo [2], and reconstruction based on shading cues [34]. To better capture a variety of identities and facial expressions, the multi-linear face models [55] and the FACS-based blendshapes [11] were later proposed. When reconstructing a 3D face from images, sparse 2D facial landmarks [18, 19, 49, 60] are widely used for a robust initialization. Shape regressions have been exploited in the state-of-the-art landmark detection approaches [12, 32, 45] to achieve impressive accuracy.

Due to the low dimensionality and effectiveness of morphable models in representing facial geometry, there have been significant recent advances in single-view face reconstruction [52, 46, 36, 51, 28]. However, for near-range portrait photos, the perspective distortion of facial features may lead to erroneous reconstructions even when using the state-of-the-art techniques. Therefore, portrait perspective undistortion must be considered as a part of a pipeline for accurately modeling facial geometry.

Face Normalization. Unconstrained photographs often include occlusions, non-frontal views, perspective distortion, and even extreme poses, which introduce a myriad of challenges for face recognition and reconstruction. However, many prior works [25, 62, 47, 29] only focused on normalizing head pose. Hassner et al. [25] "frontalized" a face from an input image by estimating the intrinsic camera matrix given a fixed 3D template model. Cole et al. [15] introduced a neural network that mapped an unconstrained facial image to a front-facing image with a neutral expression. Huang et al. [29] used a generative model to synthesize an identity-preserving frontal view from a profile. Bas et al. [4] proposed an approach for fitting a 3D morphable model to 2D landmarks or contours under either orthographic or perspective projection.

Psychological research suggests a direct connection between camera distance and human portrait perception. Bryan et al. [9] showed that there is an "optimal distance" at which portraits should be taken. Cooper et al. [17] showed that the 50mm lens is most suitable for photographing an undistorted facial image. Valente et al. [54] proposed to model perspective distortion as a one-parameter family of warping functions with known focal length. Most related to our work, Fried et al. [23] investigated the problem of editing the facial appearance by manipulating the distance between a virtual camera and a reconstructed head model. Though this technique corrected some mild perspective distortion, it was not designed to handle extreme distortions, as it relied on a 3D face fitting step.

Image-based Camera Calibration. Camera calibration is an essential prerequisite for extracting precise and reliable 3D metric information from images. We refer the reader to [44, 48, 42] for a survey of such techniques. The state-of-the-art calibration methods mainly require a physical target such as checkerboard pattern [53, 64] or circular control points [14, 16, 26, 31, 20], used for locating point correspondences. Flores et al. [22] proposed the first method to infer camera-to-subject distance from a single image with a calibrated camera. Burgos-Artizzu et al. [10] built the Caltech Multi-Distance Portraits Dataset (CMDP) of portraits of a variety of subjects captured from seven distances. Many recent works directly estimate camera parameters using deep neural networks. PoseNet [35] proposed an end-to-end solution for 6-DOF camera pose localization. Others [58, 59, 27] proposed to extract camera parameters using vanishing points from a single scene photos with horizontal lines. To the best of our knowledge, our method is the first to estimate camera parameters from a single portrait.

3. Overview

The overview of our system is shown in Fig. 1. We pre-process the input portraits with background segmentation, scaling, and spatial alignment (*see appendix*), and then feed them to a *camera distance prediction* network to estimates camera-to-subject distance. The estimated distance and the portrait are fed into our cascaded network including *FlowNet*, which predicts a distortion correction flow map, and *CompletionNet*, which inpaints any missing facial features caused by perspective distortion. Perspective undistortion is not a typical image-to-image translation problem, because the input and output pixels are not spatially corresponded. Thus, we factor this challenging problem into two sub tasks: first finding a per-pixel undistortion flow map, and then image completion via inpainting. In particular, the vectorized flow representation undistorts an input image at its original resolution, preserving its high frequency details, which would be challenging if using only generative image synthesis techniques. In our cascaded architecture (Fig. 1 middle), *CompletionNet* is fed the warping result of *FlowNet*. We provide details of *FlowNet* and *CompletionNet* in Sec. 4. Finally, we combined the results of the two cascaded networks using the Laplacian blending [1] to synthesize a complete undistorted image with inference of missing features while keeping the high-resolution details as shown in Fig. 2.

4. Portrait Undistortion

4.1. Camera Distance Prediction Network

Rather than regress directly from the input image to the camera distance D, which is known to be challenging to train, we use a distance classifier. We check if the camera-to-subject distance of the input is larger than an query distance $d \in (17.4cm, 130cm)$[1]. Our strategy learns a continuous mapping from input images to the target distances. Given any query distance D and input image pair $(input, D)$, the output is a floating point number in the range of $0 \sim 1$, which indicates the probability that the distance of the input image is greater than the query distance D. As shown in Fig. 3, the vertical axis indicates the output of our distance prediction network while the horizontal axis is the query distance. To predict the distance, we locate the query distance with a network output of 0.5. With our network denoted as ϕ, our network holds the transitivity property that if $d1 > d2$, $\phi(input, d1) > \phi(input, d2)$.

To train the camera distance network, we append the value of $\log_2 d$ as an additional channel of the input image and extract features from the input images using the VGG-11 network [50], followed by a classic classifier consisting of fully connected layers. As training data, for each of the training image with ground truth distance d, we sample a set of $\log_2 d$ using normal distribution $\log_2 d \sim \mathcal{N}(\log_2 D, 0.5^2)$.

[1] $17.4cm$ and $130cm$ camera-to-subject distances correspond to $14mm$ and $105mm$, respectively, in $35mm$ equivalent focal length

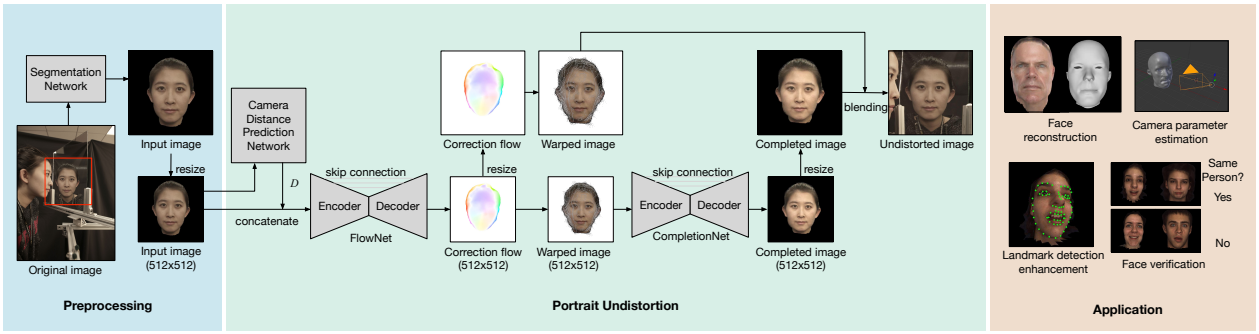


Figure 1: The pipeline workflow and applications of our approach. The input portrait is first segmented and scaled in the preprocessing stage and then fed to a network consisting of three cascaded components. The *FlowNet* rectifies the distorted artifacts in the visible regions of input by predicting a distortion correction flow map. The *CompletionNet* inpaints the missing facial features due to the strong perspective distortions and obtains the completed image. The outcomes of two networks are then scaled back to the original resolution and blended with high-fidelity mean texture to restore fine details.

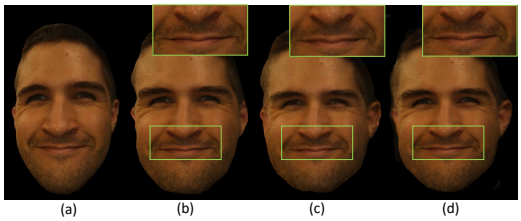


Figure 2: Illustration of blending. (a) Input; (b) is the result of *FlowNet* with interpolation; (c) is the result of *CompletionNet*; (d) is the final result with blending (b) and (c). ***Green box*** shows the close-up of facial details.

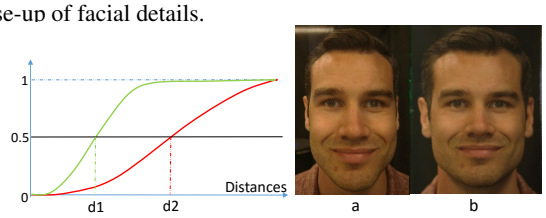


Figure 3: Illustration of Camera Distance Prediction Classifier. ***Green Curve*** and ***Red Curve*** are the response curves of input ***a*** and ***b***; ***d1*** and ***d2*** are the predicted distances of input ***a*** and input ***b***.

4.2. FlowNet

The FlowNet operates on the normalized input image (512×512) $\mathcal{A}$ and estimates a correction forward flow $\mathcal{F}$ that rectifies the facial distortions. However, due to the immense range of possible perspective distortions, the correction displacement for portraits taken at different distances will exhibit different distributions. Directly predicting such high-dimensional per-pixel displacement is highly under-constrained and often leads to inferior results (Fig. 11). To ensure more efficient learning, we propose to attach the estimated distance to the input of *FlowNet* in the similar way as in Section 4.1. Instead of directly attaching the predicted number, we propose to classify these distances into eight intervals[2] and use the class label as the input to *FlowNet*. The use of label will decrease the risk of accumulation error from camera distance prediction network, because the accuracy of predicting a label is higher than floating number.

FlowNet takes $\mathcal{A}$ and distance label $\mathcal{L}$ as input, and it will predict a forward flow map $\mathcal{F}_{\mathcal{AB}}$, which can be used to obtain undistorted output $\mathcal{B}$ when applied to $\mathcal{A}$. For each pixel (x, y) of $\mathcal{A}$, $\mathcal{F}_{\mathcal{AB}}$ encodes the translation vector (Δ_x, Δ_y). Denote the correspondence of (x, y) on $\mathcal{B}$ as (x', y'), then $(x', y') = (x, y) + (\Delta_x, \Delta_y)$. In *FlowNet*, we denote generator and discriminator as G and D separately. Then L1 loss of flow is as below:

$$\mathcal{L}_G = \|\boldsymbol{y} - \mathcal{F}_{\mathcal{AB}}\|_1 \tag{1}$$

In which $\boldsymbol{y}$ is the ground truth flow. For the discriminator loss, as forward flow $\mathcal{F}_{\mathcal{AB}}$ is per-pixel correspondence to $\mathcal{A}$ but not $\mathcal{B}$, thus $\mathcal{B}$ will have holes, seams and discontinuities which is hard to used in discriminator. To make the problem more tractable, instead of applying discriminator on $\mathcal{B}$, we use the $\mathcal{F}_{\mathcal{AB}}$ to map $\mathcal{B}$ to $\mathcal{A}'$ and use $\mathcal{A}$ and $\mathcal{A}'$ as pairs for discriminator on the condition of $\mathcal{L}$.

$$\begin{aligned}\mathcal{L}_{\mathrm{D}} = \min_G \max_D \; & \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{data}}(\boldsymbol{x})}\big[\log D(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{L}})\big] + \\ & \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\big[\log\left(1 - D(\mathcal{A}', \boldsymbol{\mathcal{L}})\right)\big].\end{aligned} \tag{2}$$

where $p_{\mathrm{data}}(\boldsymbol{x})$ and $p_{\boldsymbol{z}}(\boldsymbol{z})$ represent the distributions of real data $\boldsymbol{x}$(input image $\mathcal{A}$ domain) and noise variables $\boldsymbol{z}$ in the domain of $\mathcal{A}$ respectively. The discriminator will penalize the joint configuration in the space of $\mathcal{A}$, which leads to shaper results.

4.3. CompletionNet

The distortion-free result will then be fed into the *CompletionNet*, which focuses on inpainting missing features and filling the holes. Note that as trained on a vast number of paired examples with large variations of camera distance, *CompletionNet* has learned an adaptive mapping regarding to varying distortion magnitude inputs.

[2]The eight distance intervals are $[23, 26)$, $[26, 30)$, $[30, 35)$, $[35, 43)$, $[43, 62)$, $[62, 105)$, $[105, 168)$ and $[168, +\infty)$, measured in centimeters.

4.4. Network Architecture

We employ a U-net structure with skip connections similar to [30] to both *FlowNet* and *CompletionNet*. There is not direct correspondence between each pixel in the input and those in the output. In the *FlowNet*, the L1 and GAN discriminator loss are only computed within the segmentation mask, leading the network to focus on correcting details that only will be used in the final output. In the *CompletionNet*, as the output tends to cover more pixels than the input, during training, we compute a new mask that denotes the novel region compared to input and assign higher L1 loss weight to this region. In implementation, we set the weight ratio of losses computed inside and outside the mask as $5:1$ for the *CompletionNet*. We found that this modifications leads to better inference performance while producing results with more details.

Implementation Details. We train *FlowNet* and *CompletionNet* separately using the Adam optimizer [37] with learning rate 0.0002. All training was performed on an NVIDIA GTX 1080 Ti graphics card. As shown in Fig. 1 middle, the generator in both networks uses the mirrored structure, in which both encoder and decoder use 8-layer convolutional network. ReLU activation and batch normalization are used in all layers except the input and output layers. The discriminator consists of four convolutional layers and one fully connected layer. For input image of 512×512, the run times are: $73ms$ (distance prediction), $330ms$ (*FlowNet*), and $340ms$ (*CompletionNet*).

5. Data Preparation

Training Data Acquisition and Rendering. As there is no database of paired portraits with *only* perspective changing. We therefore generate a novel training dataset where we can control the subject-to-camera distance, head pose, illumination and ensure that the image differences are only caused *only* by perspective distortion. Our synthetic training corpus is rendered from 3D head models acquired by two scanning system. The first is a light-stage[24, 41] scanning system which produces pore-level 3D head models for photo-real rendering. Limited by the post-processing time, cost and number of individuals that can be rapidly captured, we also employ a second capture system engineered for rapid throughput. In total, we captured 15 subjects with well defined expressions in the high-fidelity system, generating 307 individual meshes, and 200 additional subjects in the rapid-throughput system.

We rendered synthetic portraits using a variety of camera distances, head poses, and incident illumination conditions. We sample distances distributed from $23cm$ to $160cm$ with corresponding $35mm$ equivalent focal length from $18mm$ to $128mm$ to ensure the same framing. We also randomly sample candidate head poses in the range of -45° to +45° in pitch, yaw, and roll. We used 107 different image-based environment for global illumination combined with point lights. With the random combination of camera distances, head poses, and illuminations, we totally generate 35,654 pairs of distroted/ undistored portraits along with forward flow.

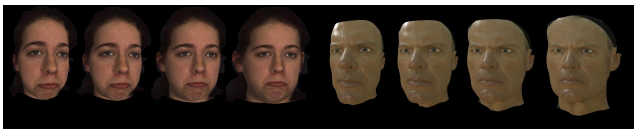


Figure 4: Training dataset. ***The left side triplet*** are the synthetic images generated from BU-4DFE dataset[61]. ***From the left to the right***, the camera-to-subject distances are: *24cm, 30cm, 52cm, 160cm*. ***The right side triplet*** are the synthetic images rendered from high-resolution 3D model. ***From left to the right***, the camera-to-subject distances are: *22cm, 30cm, 53cm, 160cm*.

17,000 additional portrait pairs warped from BU-4DFE dataset [61](38 females and 25 males subjects for training while 20 females and 17 males for testing) are used to expand diversity of identities.

Test Data Acquisition. To demonstrate that our system scales well to real-world portraits, we also devised a two-camera beam splitter capture system that would enable simultaneous photography of a subject at two different distances. As shown in Teaser left, we setup a beam-splitter (50% reflection, 50% transmission) at 45° along a metal rail, such that a first DSLR camera was set at the canonical fixed distance of $1.6m$ from the subject, with a $180mm$ prime lens to image the subject directly, while a second DSLR camera was set at a variable distance of $23cm$ - $1.37m$, to image the subject's reflection off the beam-splitter with a $28mm$ zoom lens. With carefully geometry and color calibration, the two hardware-synced cameras were able to capture nearly ground truth portraits pairs of real subjects both with and without perspective distortions. (*More details can be found in the appendix*).

6. Results and Experiments

6.1. Evaluations

Face Image Undistortion. In Fig. 5, Fig. 6 and Fig. 7 we show the undistortion results of ours compared to Fried et al.[23]. To better visualize the reconstruction error, we also show the error map compared to groundtruth or references in Fig. 5 and Fig. 6. We perform histogram equalization before computing the error maps. Our results are visually more accurate than Fried et al.[23] especially on the face boundaries. To numerically evaluate our undistortion accuracy, we compare with Fried et al.[23] the average error over 1000 synthetic pair from BU-4DFE dataset. With an average intensity error of 0.39 we significantly outperform Fried et al.[23] which has an average intensity error of 1.28. In Fig. 7, as we do not have references or ground truth, to

Figure 5: Undistortion results of beam splitter system compared to Fried et al. [23]. *From left to the right* : inputs, results of Fried et al. [23], error maps of Fried et al. [23] compared to references, ours results, error map of our results compared to references, refere

Figure 6: Undistortion results of synthetic data generated from BU-4DFE dataset compared to Fried et al. [23]. *From left to the right* : inputs, results of Fried et al. [23], error map of Fried et al. [23] compared to ground truth, ours results, error map of our results compared to ground truth, ground truth.

better visualize the motion of before-after undistortion, we replace the g channel of input with g channel of result image to amplify the difference.

Camera Parameter Estimation. Under the assumption of same framing(keeping the head size in the same scale for all the photos), the distance is equivalent to focal length by multiplying a scalar. The scalar of converting distances to focal length is $s = 0.8$, which means when taking photo at $160cm$ camera-to-subject distance, to achieve the desired framing in this paper, a $128mm$ focal length should be used. Thus, as long as we can predict accurate distances of the input photo, we can directly get the $35mm$ equivalent focal length of that photo. We numerically evaluate the accuracy of our *Camera Distance Prediction* network by testing with 1000 synthetic distorted portraits generated from BU-4DFE dataset. The mean error of distance prediction is 8.2% with a standard deviation of 0.083. We also evaluate the accuracy of labeling. As the intervals mentioned in Section 4.2

Figure 7: Evaluation and comparisons with Fried et al. [23] on a variety of datasets with in the wild database. (a). inputs; (b). Fried et al. [23]; (c). Ours; (d). The Mixture of (a) and (b) for better visualization of undistortion; (e). The Mixture of (a) and (c); Shaded portraits indicate the failure of Fried et al. [23].

are successive, some of the images may lie on the fence of two neighboring intervals. So we regard label prediction as correct within its one step neighbor. Under this assumption, the accuracy of labeling is 96.9% which insures input reliability for the cascade networks. Fig. 8 shows the distance prediction probability curve of three images. For each of them we densely sampled query distances along the whole distance range and and the classifier results are monotonically changing. We tested on 1000 images and found that

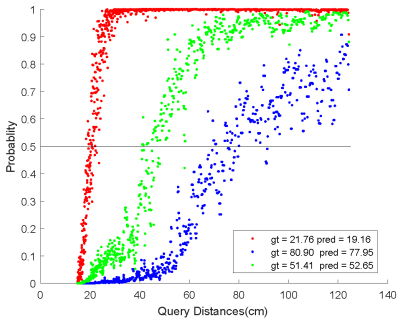


Figure 8: Distance prediction probability curve of three different input portraits with query distance sampled densely along the whole range.

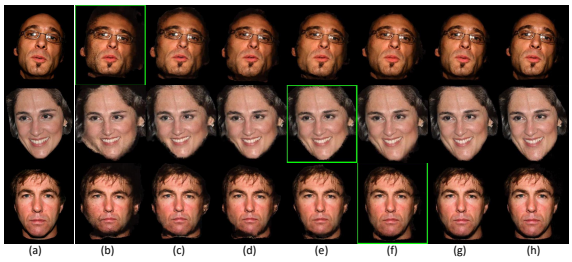


Figure 9: Results of different labels. (a) Inputs; (b) $\sim$ (h): Network outputs given labels from 1 to 7. ***Green box*** is the label predicted by the network.

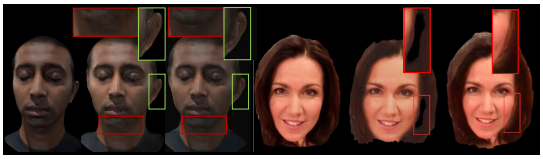


Figure 10: Ablation analysis on cascade network. In each triplet, *from left to the right*: inputs, results of single image-to-image network similar to [30], results of using cascade network including *FlowNet* and *CompletionNet*.

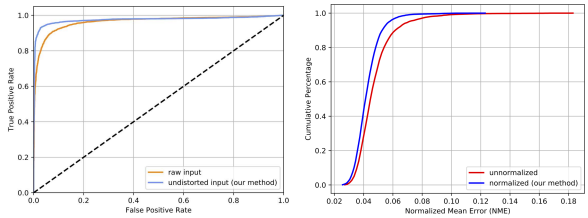
on average the transitivity holds 98.83%.

Evaluation on Interval Selection. In the *Camera Distance Prediction* network, some of the images might lie on the fence of neighboring intervals. To demonstrate the stability of our pipeline, we show several undistortion results for the same inputs using different distance labels in Fig. 9, with the network predicted label outlined in green. Visually, each outlined image and its left and right neighbors appear very similar, suggesting the results are not highly sensitive to adjacent interval categorization.

6.2. Ablation Study

In Fig. 10, we compare the single network and proposed cascade network. The results show that with a *FlowNet* as prior, the recovered texture will be sharper especially on boundaries. Large holes and missing textures are more likely to be inpainted properly. Fig. 11 demonstrates the effectiveness of our label channel introduced in *FlowNet*. The results without label channel are more distorted compared

Figure 11: Ablation analysis on attach label as a input to *FlowNet*. In each triplet, *from left to the right*: inputs, results from the network without label channel, results of our proposed network.



(a) (b)

Figure 12: (a). Receiver operating characteristic (ROC) curve for face verification performance on raw input and undistorted input using our method. Raw input (A,B) compares to undistorted input $(N(A),B)$; (b). Cumulative error curve for facial landmark detection performance given unnormalized image and image normalized by our method. Metric is measured in normalized mean error (NME).

to the ones with label as inputs, especially the proportions of noses, mouth regions and the face rims.

6.3. Applications

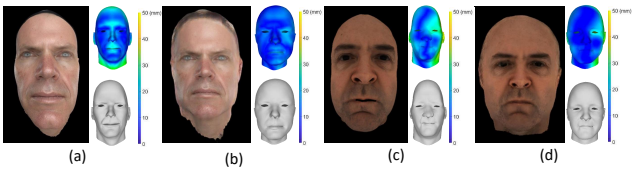


Figure 13: Comparing 3D face reconstruction from portraits, without and with our undistortion technique. (a) and (c) are heavily distorted portraits and the 3D mesh fitted using the landmarks of the portraits. (b) and (d) are undistorted results of (a) and (c) with 3D reconstruction based on them. Gray meshes show reconstructed facial geometry and color-coded meshes show reconstruction error.

Face Verification. Our facial undistortion technique can improve the performance of face verification, which we test using the common face recognition system OpenFace [3]. We synthesized 6,976 positive (same identity) and 6,980 negative (different identity) pairs from BU-4DFE dataset [61] as test data. We rendered one image A in a pair of

Input	mean	std
Raw input (A, B)	0.9137	0.0090
Undistorted input $(N(A), B)$	0.9473	0.0067

Table 1: Comparison of face verification accuracy for images with and without our undistortion as pre-processing. Accuracy is reported on random 10-folds of test data with mean and standard deviation.

images (A,B) as a near-distance portrait with perspective distortion; while we rendered B at the canonical distance of $1.6m$ to minimize the distortion. This is the setting of most face verification security system, which retrieves the database for the nearest neighbor. We evaluated face verification performance on raw data (A, B) and data $(N(A), B)$ and $(N(A), N(B))$ in which perspective distortion was removed using our method. Verification accuracy and receiver operating characteristic (ROC) comparisons are shown in Table 1 and Fig. 12a.

Landmark Detection Enhancement. We use the state-of-the-art facial landmark tracker OpenPose [13] on 6,539 renderings from the BU-4DFE dataset [61] as previously described, where each input image is rendered at a short camera-to-object distance with significant perspective distortion. We either directly apply the landmark detection to the raw image, or the undistorted image using our network and then locate the landmark on the raw input using the flow from our network. Landmark detection gives a 100% performance based on our pre-processed images, on which domain alignments are applied while fails on 9.0% original perspective-distorted portraits. For quantitative comparisons, we evaluate the landmark accuracy using a standard metric, Normalized Mean Error (NME) [63]. Given the ground truth 3D facial geometry, we can find the ground truth 2D landmark locations of the inputs. For images with successful detection for both the raw and undistorted portraits, our method produces lower landmark error, with mean NME = 4.4% (undistorted images), compared to 4.9% (raw images). Fig. 12b shows the cumulative error curves, showing an obvious improvement for facial landmark detection for portraits undistorted using our approach.

Face Reconstruction. One difficulty of reconstructing highly distorted faces is that the boundaries can be severely self-occluded (e.g., disappearing ears or occlusion by the hair), which is a common problem in 3D face reconstruction methods regardless if the method is based on 2D landmarks or texture. Fig. 13 shows that processing a near-range portrait input using our method in advance can significantly improves 3D face reconstruction. The 3D facial geometry is obtained by fitting a morphable model (FaceWarehouse [11]) to 2D facial landmarks. Using the original perspective-distorted image as input, the reconstructed geometry appears distortion, while applying our technique as a pre-processing step retains both identity and geometric details. We show error map of 3D geometry compared to ground truth, demonstrating that our method applied as a pre-processing step improves reconstruction accuracy, compared with the baseline approach without any perspective-distortion correction.

7. Conclusion

We have presented the first automatic approach that corrects the perspective distortion of unconstrained near-range portraits. Our approach even handles extreme distortions. We proposed a novel cascade network including camera parameter prediction network, forward flow prediction network and feature inpainting network. We also built the first database of perspective portraits pairs with a large variations on identities, expressions, illuminations and head poses. Furthermore, we designed a novel duo-camera system to capture testing images pairs of real human. Our approach significantly outperforms the state-of-the-art approach [23] on the task of perspective undistortion with an accurate camera parameter prediction. Our approach also boosts the performance of fundamental tasks like face verification, landmark detection and 3D face reconstruction.

Limitations and Future Work. One limitation of our work is that the proposed approach does not generalize to lens distortions, as our synthetic training dataset rendered with an ideal perspective camera does not include this artifact. Similarly, our current method is not explicitly trained to handle portraits with large occlusions and head poses. We plan to resolve both of the limitations in future work by augmenting training examples with lens distortions, large facial occlusions and head poses. Another future avenue is to investigate end-to-end training of the cascaded network, which could further boost the performance of our approach, but would require fully-differentiable image warping.

8. Acknowledgments

Hao Li is affiliated with the University of Southern California, the USC Institute for Creative Technologies, and Pinscreen. This research was conducted at USC and was funded by in part by the ONR YIP grant N00014-17-SFO14, the CONIX Research Center, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA, the Andrew and Erna Viterbi Early Career Chair, the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005, Adobe, and Sony. This project was not funded by Pinscreen, nor has it been conducted at Pinscreen or by anyone else affiliated with Pinscreen. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

[1] Edward H Adelson, Charles H Anderson, James R Bergen, Peter J Burt, and Joan M Ogden. Pyramid methods in image processing. *RCA engineer*, 29(6):33–41, 1984. 3

[2] Brian Amberg, Andrew Blake, Andrew Fitzgibbon, Sami Romdhani, and Thomas Vetter. Reconstructing high quality face-surfaces using model based stereo. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007. 2

[3] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 2016. 7

[4] Anil Bas and William AP Smith. What does 2d geometric information really tell us about 3d face shape? *arXiv preprint arXiv:1708.06703*, 2017. 3

[5] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. High-quality single-shot capture of facial geometry. In *ACM Transactions on Graphics (ToG)*, volume 29, page 40. ACM, 2010. 2

[6] Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W Sumner, and Markus Gross. High-quality passive facial performance capture using anchor frames. In *ACM Transactions on Graphics (TOG)*, volume 30, page 75. ACM, 2011. 2

[7] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. 2

[8] Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer. High resolution passive facial performance capture. In *ACM transactions on graphics (TOG)*, volume 29, page 41. ACM, 2010. 2

[9] Ronnie Bryan, Pietro Perona, and Ralph Adolphs. Perspective distortion from interpersonal distance is an implicit visual cue for social judgments of faces. *PloS one*, 7(9):e45301, 2012. 1, 3

[10] Xavier P Burgos-Artizzu, Matteo Ruggero Ronchi, and Pietro Perona. Distance estimation of an unknown person from a portrait. In *European Conference on Computer Vision*, pages 313–327. Springer, 2014. 3

[11] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014. 2, 8

[12] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014. 2

[13] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018. 8

[14] Qian Chen, Haiyuan Wu, and Toshikazu Wada. Camera calibration with two arbitrary coplanar circles. In *European Conference on Computer Vision*, pages 521–532. Springer, 2004. 3

[15] Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T Freeman. Synthesizing normalized faces from facial identity features. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3386–3395. IEEE, 2017. 3

[16] Carlo Colombo, Dario Comanducci, and Alberto Del Bimbo. Camera calibration with two arbitrary coaxial circles. In *European Conference on Computer Vision*, pages 265–276. Springer, 2006. 3

[17] Emily A Cooper, Elise A Piazza, and Martin S Banks. The perceptual basis of common photographic practice. *Journal of vision*, 12(5):8–8, 2012. 3

[18] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001. 2

[19] David Cristinacce and Tim Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067, 2008. 2

[20] Ankur Datta, Jun-Sik Kim, and Takeo Kanade. Accurate camera calibration using iterative refinement of control points. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1201–1208. IEEE, 2009. 3

[21] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156. ACM Press/Addison-Wesley Publishing Co., 2000. 2

[22] Arturo Flores, Eric Christiansen, David Kriegman, and Serge Belongie. Camera distance from face images. In *International Symposium on Visual Computing*, pages 513–522. Springer, 2013. 3

[23] Ohad Fried, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. Perspective-aware manipulation of portrait photos. *ACM Transactions on Graphics (TOG)*, 35(4):128, 2016. 1, 2, 3, 5, 6, 8

[24] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. In *ACM Transactions on Graphics (TOG)*, volume 30, page 129. ACM, 2011. 2, 5

[25] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4295–4304, 2015. 3

[26] Janne Heikkila. Geometric camera calibration using circular control points. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1066–1077, 2000. 3

[27] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Jonathan Eisenmann, Matthew Fisher, Emiliano Gambaretto, Sunil Hadap, and Jean-François Lalonde. A perceptual measure for deep single image camera calibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2354–2363, 2018. 3

[28] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics (TOG)*, 36(6):195, 2017. 2

[29] Rui Huang, Shu Zhang, Tianyu Li, Ran He, et al. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. *arXiv preprint arXiv:1704.04086*, 2017. 3

[30] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017. 5, 7

[31] Guang Jiang and Long Quan. Detection of concentric circles for camera calibration. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 333–340. IEEE, 2005. 3

[32] Vahid Kazemi and Sullivan Josephine. One millisecond face alignment with an ensemble of regression trees. In *27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, United States, 23 June 2014 through 28 June 2014*, pages 1867–1874. IEEE Computer Society, 2014. 2

[33] Ira Kemelmacher-Shlizerman. Internet based morphable model. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3256–3263. IEEE, 2013. 2

[34] Ira Kemelmacher-Shlizerman and Ronen Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE transactions on pattern analysis and machine intelligence*, 33(2):394–405, 2011. 2

[35] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 3

[36] Hyeongwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt, and Christian Theobalt. Inversefacenet: Deep single-shot inverse face rendering from a single image. *arXiv preprint arXiv:1703.10956*, 2017. 2

[37] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[38] Hao Li, Bart Adams, Leonidas J Guibas, and Mark Pauly. Robust single-view geometry and motion reconstruction. In *ACM Transactions on Graphics (TOG)*, volume 28, page 175. ACM, 2009. 2

[39] Chang Hong Liu and Avi Chaudhuri. Face recognition with perspective transformation. *Vision Research*, 43(23):2393–2402, 2003. 1

[40] Chang Hong Liu and James Ward. Face recognition in pictures is affected by perspective transformation but not by the centre of projection. *Perception*, 35(12):1637–1650, 2006. 1

[41] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, and Paul Debevec. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 183–194. Eurographics Association, 2007. 2, 5

[42] Gerard Medioni and Sing Bing Kang. *Emerging topics in computer vision*. Prentice Hall PTR, 2004. 3

[43] Frederic I Parke and Keith Waters. *Computer facial animation*. CRC Press, 2008. 2

[44] Fabio Remondino and Clive Fraser. Digital camera calibration methods: considerations and comparisons. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 36(5):266–272, 2006. 3

[45] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014. 2

[46] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5553–5562. IEEE, 2017. 2

[47] Christos Sagonas, Yannis Panagakis, Stefanos Zafeiriou, and Maja Pantic. Robust statistical face frontalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3871–3879, 2015. 3

[48] Joaquim Salvi, Xavier Armangué, and Joan Batlle. A comparative review of camera calibrating methods with accuracy evaluation. *Pattern recognition*, 35(7):1617–1635, 2002. 3

[49] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011. 2

[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[51] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, 2017. 2

[52] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2387–2395. IEEE, 2016. 2

[53] Roger Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal on Robotics and Automation*, 3(4):323–344, 1987. 3

[54] Joachim Valente and Stefano Soatto. Perspective distortion modeling, learning and compensation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2015 IEEE Conference on*, pages 9–16. IEEE, 2015. 3

[55] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. *ACM transactions on graphics (TOG)*, 24(3):426–433, 2005. 2

[56] Brittany Ward, Max Ward, Ohad Fried, and Boris Paskhover. Nasal distortion in short-distance photographs: The selfie effect. *JAMA facial plastic surgery*, 2018. 1

[57] Thibaut Weise, Hao Li, Luc Van Gool, and Mark Pauly. Face/off: Live facial puppetry. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer animation*, pages 7–16. ACM, 2009. 2

[58] Scott Workman, Connor Greenwell, Menghua Zhai, Ryan Baltenberger, and Nathan Jacobs. Deepfocal: a method for direct focal length estimation. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 1369–1373. IEEE, 2015. 3

[59] Scott Workman, Menghua Zhai, and Nathan Jacobs. Horizon lines in the wild. In *British Machine Vision Conference (BMVC)*, 2016. 3

[60] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE, 2013. 2

[61] Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. A high-resolution 3d dynamic facial expression database. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008. 5, 7, 8

[62] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *Proc. ICCV*, pages 1–10, 2017. 3

[63] Stefanos Zafeiriou, George Trigeorgis, Grigorios Chrysos, Jiankang Deng, and Jie Shen. The menpo facial landmark localisation challenge: A step towards the solution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, volume 1, page 2, 2017. 8

[64] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000. 3