Blind Removal of Facial Foreign Shadows

Yaojie Liu*[1]
https://yaojieliu.github.io/
Andrew Hou*[1]
https://andrewhou1.github.io/
Xinyu Huang[2]
xinyu.huang@us.bosch.com
Liu Ren[2]
https://sites.google.com/site/liurenshomepage/
Xiaoming Liu[1]
http://www.cse.msu.edu/~liuxm/index2.html

[1] Michigan State University
East Lansing, MI
[2] Bosch Research North America
Sunnyvale, CA

Abstract

In-the-wild face photographs often suffer from undesirable foreign shadows cast by external objects (*e.g.* hands, phones, and trees). Removing facial foreign shadows not only improves image aesthetics but also mitigates the negative impacts on face-related tasks. This paper tackles the blind removal of facial foreign shadows for both single images and videos by making three contributions. First, we propose a novel two-stage shadow modeling algorithm that consists of gray-scale shadow removal and colorization. This decomposition provides an effective way to handle both color distortion and sub-surface scattering effects. Second, we propose a novel Temporal Sharing Module (TSM) to extract hierarchical features across multiple aligned video frames, which represent the shadow-free faces. Third, we collect a real face database with 280 videos captured under highly dynamic environments and annotate pixel-level shadow segmentation maps. Extensive experiments demonstrate that our approach outperforms state-of-the-art methods quantitatively and qualitatively. Code, our database, and pre-trained models are publicly available at https://github.com/andrewhou1/BlindShadowRemoval.

1 Introduction

In our daily activities, many external objects around us can cast shadows on faces, known as *facial foreign shadows*. For instance, when we take selfies outdoors, our hand or camera might block part of the sunlight and create a shadow on the face. Dynamic and scattered shadows may be produced by leaves when walking under trees. While driving, the driver may confront the high-contrast lighting caused by the direct sunlight and car pillars. One may want to remove these cast shadows for aesthetic purposes, such as in photoshop and face editing. In other cases, the cast shadows should be removed since they could negatively impact face-related tasks, such as face recognition, age estimation, and driver monitoring.

* Equal Contribution.

Figure 1: Results of our shadow removal model on images in our Shadow Face in the Wild (SFW) dataset (*top*) and UCB dataset [55] (*bottom*). From left to right are the input, output, and shadow matte.

Many works aim to handle self shadows and relight the face via quotient images [19, 48, 56], inverse rendering [30, 31, 32], or style transfer [16, 26]. Those methods focus more on the global lighting distribution and are limited in handling arbitrary high-frequency structures caused by harsh foreign shadows. There are also face completion works under structured occlusions, such as squares, circles, and lattices [27, 52, 54]. Compared with foreign shadow removal, face completion is easier as the shape is less complicated and the occlusion is often filled with a single color such as white. Some works study shadows on generic objects [25, 37]. While they excel at shadow detection, when applied to faces, observable artifacts can be detected on deshadowed results due to the lack of face priors.

The major problem with prior methods is they cannot handle the high-frequency structure caused by harsh shadows, as demonstrated in [41, 55]. Instead of predicting illumination, Zhang *et al.* [55] propose a single image approach using only perceptual and pixel intensity losses and train the network on a synthetic shadow dataset. These losses are effective in removing harsh shadows and recovering fine details. However, their model based only on perceptual and pixel intensity losses does not generalize well in practice, as it is hard to build a training dataset that covers the complex and diverse lighting conditions in the real world.

This work aims to detect and remove foreign shadows from in-the-wild faces, where we face three major challenges. First, the shadows of in-the-wild faces are arbitrary, varying in sizes, shapes, locations, colors, blurriness, and intensities. Prior works [25, 37, 55] model shadows directly in RGB. Given the high diversity, they have a hard time addressing all the discrepancies, leaving observable artifacts in deshadowed faces. Second, there are very few public databases for training and evaluation for this problem. To capture paired shadow and non-shadow faces, both the subject and photographer need to be perfectly still, which is rarely feasible. Third, sometimes the shadow removal is extended from single images to videos, which requires additional temporal consistency.

To address the aforementioned challenges, we propose a novel blind facial foreign shadow removal model. To handle the shadow diversity, we propose a simple yet effective approach to decompose the direct RGB shadow removal into grayscale shadow removal and colorization. We show that, without color, the shadow modeling becomes a much simpler task and the grayscale removal model more easily generalizes to unseen data. After detecting the shadow regions using grayscale shadow removal, the colorization step is an image inpainting process. Without seeing the biased color information from the shadow region, the colorization process also becomes more generalizable. To ensure temporal consistency, we also propose

a temporal sharing module (TSM) to aggregate information among multiple frames. TSM includes an efficient warping layer to handle frames with pose and expression variations.

For training the model, we follow [55] and build a synthetic database that contains paired shadow and shadow-free faces. Foreign shadows are generated with randomized properties. We also collect a face database with 280 videos under highly dynamic environments for evaluation. The external objects casting shadows include hands, books, leaves, trees, window blinds, car pillars, and buildings. To quantitatively evaluate shadow segmentation performance, we provide detailed pixel-wise segmentation annotations for this database.

In summary, the main contributions of this work include:

◇ A novel approach to decompose RGB shadow removal into grayscale shadow removal and colorization.

◇ A temporal sharing module to ensure video consistency and face symmetry.

◇ A face shadow database captured under dynamic environments, along with pixel-wise segmentation annotations.

◇ State-of-the-art shadow removal and shadow segmentation results and photo-realistic deshadow quality.

2 Related Work

Face relighting Face relighting methods can be roughly divided into three categories: quotient image-based, style transfer, and inverse rendering. The color ratio (*i.e.* quotient images) is first proposed in [35] to transfer a frontal face from one lighting to another. This basic idea has been extended to handle different poses, use ratios of images or radiance environment maps, and to generate synthetic relighting datasets [19, 40, 48, 56]. Facial lighting also can be changed by style transfer [16, 26, 28, 36, 38]. Similar to most quotient image methods, style transfer methods need at least a reference image as the target style. Moreover, the face poses of input and reference images need to be very close. In the category of inverse rendering, a face image is decomposed into multiple components such as geometry, reflectance, and lighting [1, 13, 20, 30, 31, 32, 34, 39, 42, 43, 46]. In general, inverse rendering methods only model facial shadows in the illumination and do not model foreign shadows.

Face completion Face completion aims to fill in the missing or occluded face regions with semantically meaningful information. In [54], Zhang *et al.* propose a DemeshNet with two sub-networks to remove mesh-like lines or watermarks on faces. Li *et al.* propose a disentangling and fusing network containing discriminators in three domains: occluded faces, clean faces, and structured occlusions [27]. The face inpainting network in [52] consists of a landmark predicting subnet and an inpainting subnet. More recently, Dey *et al.* [9] propose an analysis-by-synthesis approach for face completion that focuses on inpainting the albedo to resynthesize a complete face. Different from shadow removal, the structured occlusions are either opaque or contain repeated patterns. The networks are mainly used to hallucinate the invisible face regions.

Generic shadow detection and removal Without much training data, early works in general-purpose shadow detection and removal mainly study shadow properties, especially around shadow edges [5, 6, 14, 23, 49, 51]. Wu *et al.* [49] apply graph-cut inference to detect shadow regions, and then use the shadow matting to generate soft shadow boundaries. Deep learning based methods have been proposed to detect and remove shadows [10, 15, 25, 29, 33, 44, 53, 57, 59]. Hu *et al.* [22] design the direction-aware spatial context module and apply a spatial RNN to detect shadows. Cun *et al.* [8] learn to hierarchically aggregate the dilated multi-contexts and attentions. The authors in [55] demonstrate that the general-purpose methods

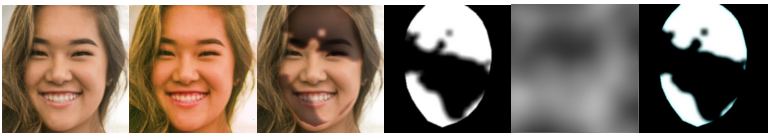

Figure 2: Illustration of data synthesis components.

such as [8, 22] cannot preserve the authenticity of the input faces. One reason is that these general-purpose networks are unable to capture specific face characteristics. For instance, human face skin is a highly scattering material that also has a complex absorption spectrum [11]. In this work, we propose a novel two-stage shadow modeling algorithm that can better handle both subsurface scattering effects and color distortion.

3 Proposed Method

3.1 Shadow synthesis and modeling

Shadows are produced by foreign objects that block light rays from arriving to the face. Let a matte $\mathbf{M}$ represent the shadow shape. The shadow formation can be modeled as a blending between the well-illuminated face $\mathbf{I}^b$ and the under-illuminated face $\mathbf{I}^d$:

$$\mathbf{I} = \mathbf{I}^b \odot (1-\mathbf{M}) + \mathbf{I}^d \odot \mathbf{M}, \tag{1}$$

where $\odot$ denotes element-wise multiplication. As real-world shadows vary in both shape and intensity, it's vital to have paired data $\{\mathbf{I}, \mathbf{I}^b\}$ with a large variety of $\mathbf{M}$ to train a generalizable shadow removal model. However, it's hardly feasible to collect a large-scale dataset with this type of paired data, as the subject needs to be perfectly static while capturing the pair. Therefore, creating a synthetic dataset becomes our go-to approach to tackle this problem.

Shadow synthesis As indicated in [55], using Eqn. 1 to synthesize natural face shadows introduces multiple variations in shape, intensity, subsurface scattering and color. Let shape $\mathbf{B}$ be a binary mask that defines the whole region affected by foreign shadow. The shadow is often unevenly distributed, such as in mottled patterns or gradually changing patterns, depending on the relative distance between the object and the face as well as the environmental lighting. We use a gray-scale matte $\mathbf{M}_\mathrm{I}$ to represent the uneven intensity. In addition, the light outside the shadow region would penetrate beneath the skin, reach the vessels and reflect back, creating a red band around the shadow boundary. We represent this subsurface scattering effect by $\mathbf{M}_\mathrm{ss}$, which is computed by blurring $\mathbf{B}$ with a different kernel per RGB channel. Therefore, Eqn. 1 can be updated to:

$$\mathbf{I} = \mathbf{I}^b \odot (1-\mathbf{B} \odot \mathbf{M}_\mathrm{ss}) + \mathbf{I}^d \odot \mathbf{B} \odot \mathbf{M}_\mathrm{ss} \odot \mathbf{M}_\mathrm{I}. \tag{2}$$

$\mathbf{B}$, $\mathbf{M}_\mathrm{I}$, and $\mathbf{M}_\mathrm{ss}$ are illustrated in Fig. 2. Moreover, the shadow region may be under certain color distortion, because parts of the light are blocked. We formulate such color distortion by a 3×3 color transfer matrix $\mathbf{C}$:

$$\mathbf{I}^d = \mathbf{I}^b \mathbf{C}. \tag{3}$$

During the synthesis, given a well-illuminated face $\mathbf{I}^b$, we generate random parameters for each component to synthesize different shadow faces $\mathbf{I}$, which is detailed in Sec. 4.

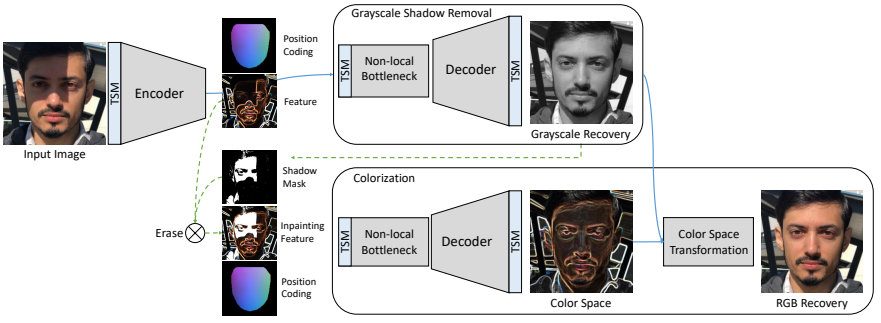

Figure 3: Our model begins with a grayscale shadow removal module that predicts the deshadowed face in grayscale. The shadow region is detected via thresholding of the grayscale difference between the deshadowed and input faces. We erase the face features within the shadow, and our colorization module is essentially an image inpainting step. The final RGB deshadowed image is estimated using the deshadowed grayscale and the learned color space. Our Temporal Sharing Module (TSM) ensures face symmetry using mirrored images, and can also be applied to video to ensure temporal consistency.

Shadow modeling With synthetic pairwise data, we can train a model $G(\cdot)$ to detect and remove foreign shadows ($\mathbf{I} \rightarrow \mathbf{I}^b$). Despite the complexity of the shadow synthesis process, prior works [25, 37, 55] opt to simplify the relation between $\mathbf{I}$ and $\mathbf{I}^b$ in $G(\cdot)$ as:

$$\mathbf{W}, \mathbf{N} \leftarrow G(\mathbf{I} \mid \omega), \tag{4}$$

$$\hat{\mathbf{I}}^b = \mathbf{I} \odot \mathbf{W} + \mathbf{N}, \tag{5}$$

where ω are the parameters of the shadow removal model, and both the scaling $\mathbf{W}$ and offset $\mathbf{N}$ are of the same size as $\mathbf{I}$. The motivation for this simplification is two-fold: 1) precisely estimating all shadow components (*i.e.* $\mathbf{B}$, $\mathbf{M}_{\mathrm{I}}$, $\mathbf{M}_{\mathrm{ss}}$, $\mathbf{C}$) can be very challenging, and 2) even with full supervision of all the components, reversing the shadow formation may raise a convergence issue. This is due to the ambiguity in the shadow parameterization, where the same shadow could be generated from different combinations of shadow components.

However, prior works based on Eqn. 5 have a hard time addressing the discrepancies between shadow and non-shadow regions, leaving some observable artifacts in deshadowed faces. We observe that it is not straightforward to derive Eqn. 5 from Eqn. 1. Due to the existence of color transfer matrix $\mathbf{C}$, $\mathbf{W}$ and $\mathbf{N}$ themselves become a function of $\mathbf{I}^b$, instead of being independent of $\mathbf{I}^b$. Thus, the model learning becomes a *chicken-and-egg* problem, which may easily turn into a memorization mode, *e.g.* a type of learning that generalizes poorly [7]. To tackle this issue, we propose to decompose the color shadow removal into grayscale shadow removal and colorization. While dealing with shadow removal in grayscale, $\mathbf{C}$ in Eqn. 3 simply becomes a scalar, and hence both $\mathbf{C}$ and $\mathbf{M}_{\mathrm{ss}}$ can be integrated into $\mathbf{M}_{\mathrm{I}}$ as $\mathbf{M}'_{\mathrm{I}}$. We can then transfer the relation of Eqn. 1 into:

$$\begin{aligned} \hat{\mathbf{I}}^{b,gs} &= \mathbf{I}^{gs} \odot (1 - \mathbf{B}) + \mathbf{I}^{gs} \odot \mathbf{B} \oslash \mathbf{M}'_{\mathrm{I}} \\ &= \mathbf{I}^{gs} \odot (1 - \mathbf{B} + \mathbf{B} \oslash \mathbf{M}'_{\mathrm{I}}) \\ &= \mathbf{I}^{gs} \odot \mathbf{W}, \end{aligned} \tag{6}$$

where $\hat{\mathbf{I}}^{b,gs}$ and $\mathbf{I}^{gs}$ are grayscale versions of $\mathbf{I}^b$ and $\mathbf{I}$, $\oslash$ is element-wise division, and $\mathbf{W} = 1 - \mathbf{B} + \mathbf{B} \oslash \mathbf{M}'_{\mathrm{I}}$. It's clear that Eqn. 6 is in a closed form and well aligned with Eqn. 5. As

$\mathbf{W}$ and $\mathbf{N}$ are detached from $\mathbf{I}^b$, they are easier to learn. Next, we simply need to colorize the grayscale face to get the final RGB face recovery. With the knowledge provided by grayscale shadow removal, we turn the blind color recovery into a mask-guided image inpainting.

The overall pipeline is shown in Fig. 3. Our approach consists of three major steps: 1*) grayscale shadow removal* (Sec. 3.2), 2*) colorization* (Sec. 3.3), and 3*) temporal information sharing* (Sec. 3.4). Steps 1 and 2 are the key ingredients for single frame shadow removal, and step 3 is the key ingredient for smooth video shadow removal and maintaining symmetry in the deshadowed face.

3.2 Grayscale shadow removal

The grayscale shadow removal module takes a RGB face $\mathbf{I} \in \mathbb{R}^{N^2 \times 3}$ as input, and outputs the scaling map $\mathbf{W} \in \mathbb{R}^{N^2 \times 1}$ and offset map $\mathbf{N} \in \mathbb{R}^{N^2 \times 1}$ that can recover a well-illuminated grayscale face $\hat{\mathbf{I}}^{b,gs} \in \mathbb{R}^{N^2 \times 1}$ based on Eqn. 5. The module consists of an encoder, a stack of residual non-local blocks, and a decoder. The encoder extracts features $\mathbf{F}$ from input images for shadow removal. It contains 4 convolution layers and 3 downsampling layers. To encourage spatial consistency for facial lighting and albedo, we leverage the latest design of non-local blocks and visual transformers [4, 12, 45]. We stack 3 residual non-local blocks to process the encoder features with positional encoding. The decoder then upsamples the features from non-local blocks via 3 transposed convolution layers, and estimates $\mathbf{W}$ and $\mathbf{N}$. We adopt a short-cut connection at each feature scale to pass along high-frequency information.

For positional encoding, we adopt the projected normalized coordinate code (PNCC) [58] and concatenate it to the encoder feature. PNCC is the normalized mean shape of 3DMM [2], and is projected to fit a given face. It encodes the face semantics as each vertex (*e.g.*, eye corner) has its unique 3D coordinate between $[0,0,0]$ and $[1,1,1]$, regardless of the pose, expression, and identity. Compared with conventional positional encoding in [4, 12], PNCC provides better face semantics that helps to detect and remove shadows.

3.3 Colorization

Using the grayscale shadow removal module, we can locate the shadow region as

$$\hat{\mathbf{B}} = |\hat{\mathbf{I}}^{b,gs} - \mathbf{I}^{gs}| > \beta, \tag{7}$$

where $\hat{\mathbf{B}}$ is the shadow segmentation mask binarized with the threshold of β. With this knowledge, we can turn the blind color recovery process into an image inpainting process with a given inpainting region. In comparison, if no knowledge is provided to the colorization process, this two-step approach is nearly identical to direct RGB shadow removal applied in previous work [25, 37, 55], which may still suffer from the poor generalization issue.

Our colorization module breaks down into 3 steps: 1) erasing, 2) inpainting, and 3) color space transformation. Structurally, the colorization module is similar to the grayscale shadow removal. It consists of 3 residual non-local blocks and a decoder. First, based on the shadow mask $\hat{\mathbf{B}}$, we set the shadow region of $\mathbf{F}$ to be 0 to circumvent any potential disturbance, and denote it as the inpainting feature. Second, the inpainting feature $\mathbf{F} \odot (1 - \hat{\mathbf{B}})$ is concatenated with $\hat{\mathbf{B}}$ and the PNCC encoding, and fed to the module. The non-local blocks aim to fill in the missing region in $\mathbf{F}$, and the decoder is designed to produce a M-channel color space $\mathbf{C} \in \mathbb{R}^{N^2 \times M}$. In the end, we use three 1×1 convolution layers to transfer the grayscale face $\hat{\mathbf{I}}^{b,gs}$ with the color space $\mathbf{C}$ back to the RGB face $\hat{\mathbf{I}}^b$. During training, no gradients from the colorization module will be sent back to the grayscale shadow removal module via $\hat{\mathbf{B}}$.

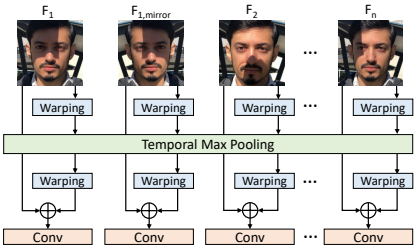

Figure 4: Illustration of the Temporal Sharing Module (TSM). It can be applied to temporal frames to improve the temporal consistency of video shadow removal as well as mirrored input to improve face symmetry.

3.4 Temporal information sharing

We can extend our network for single-frame processing to leverage temporal information via a Temporal Sharing Module (TSM). Similar to other video-based image restoration problems, such as video deblurring, shadows can be arbitrary in shape and movement across different frames. Thus, the order of the frames might not carry useful cues for deshadowing. As a result, we propose to adopt a temporal-wise max pooling to aggregate the illumination information among different frames, shown in Fig. 4.

Let $\mathbf{F}_1, \mathbf{F}_2, ..., \mathbf{F}_k$ be the features to be shared among k frames. Before computing the temporal-wise max pooling, we first apply a warping layer to register features based on the face shape. After the temporal-wise max pooling, we apply an inverse warping to re-align the shared feature back to each frame feature, and concatenate with the original feature $\mathbf{F}_i$ for the next stage's computation. The TSM is a plug-in design for features at all scales. TSM can be used not only to share the temporal information, but also to enforce the prior knowledge of face symmetry, which has been used in other tasks [50]. To achieve this, we treat the mirrored face as a different frame, and send it to TSM for information sharing. In case there is only a single frame available, TSM simply concatenates with the original feature.

The warping layer leverages the pre-computed 68 facial landmarks via [3]. Given the landmarks for the neutral face $\mathbf{s}_0$ and face $\mathbf{s}_i$ at frame i, a sparse offset can be computed as $\Delta\mathbf{s}_{i\to 0} = \mathbf{s}_0 - \mathbf{s}_i \in \mathbb{R}^{68\times 2}$ to indicate where each pixel in the landmark position should be moved to. To obtain a dense offset map $\Delta\mathbf{S}_{i\to 0} \in \mathbb{R}^{N^2\times 2}$ indicating where each pixel in the entire feature map should be moved to, we apply a triangulation interpolation,

$$\Delta\mathbf{S}_{i\to 0} \leftarrow \mathrm{Tri}(\mathbf{s}_i, \Delta\mathbf{s}_{i\to 0}, N), \tag{8}$$

where $\mathrm{Tri}(\cdot)$ is Delaunay triangulation-based interpolation. The registration operation of feature $\mathbf{F}$ is denoted as:

$$\mathbf{F}_{i\to 0} = \mathbf{F}_i(\mathbf{S}^0 + \Delta\mathbf{S}_{i\to 0}), \tag{9}$$

where $\mathbf{S}_0 = \{(0,0),(0,1),...,(N,N)\} \in \mathbb{R}^{N^2\times 2}$ enumerates pixel locations in $\mathbf{F}_i$. Similarly, when we get the shared feature $\mathbf{F}_{\max}$, we can use $\Delta\mathbf{S}_{0\to i}$ to warp it back.

3.5 Training

We use synthetic shadow faces for training (Sec. 3.1). We apply multiple losses to supervise all three steps in the model, which are explained in detail in the Supplemental Materials.

4 Training and Evaluation Data

Training data To synthesize our training data based on Eqn. 1-3, we manually select 15,000 images from FFHQ [24] with no foreign and strong self shadows. The raw binary

Figure 5: Illustration of the SFW database. The first row shows shadow faces collected under highly dynamic settings (*e.g.* varying shadows and head poses from walking and driving). The second row shows pixel-level annotations for shadow segmentation. Zoom in to view the quality of our annotations.

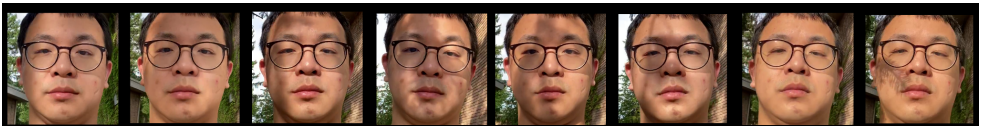

Figure 6: Samples from the SFW evaluation dataset. Here, we show the samples from one SFW video included in our 440 frame evaluation dataset. We selected frames with shadow patterns that are as diverse as possible and avoided redundant frames.

shadow shape $\mathbf{B}$ comes from 100 pre-defined silhouette shapes and the Perlin noise function. The raw shapes are randomly augmented with different scales, rotations, and boundary blurriness. Intensity map $\mathbf{M}_I$ is generated by a random Perlin noise function at two octaves.

Evaluation data To our knowledge, there is no large video database of real-world human faces with foreign shadows. One existing database, UCB [55], includes a very limited number of 100 face images. More importantly, this database contains only single images so that consistent image reconstruction on videos cannot be evaluated. In response to the need for a large video database, with the IRB approval, we collect a database that we call Shadow Faces in the Wild (SFW) for the evaluation of real-world facial shadow removal. In total, SFW includes 280 videos from 20 subjects. Some examples are shown in Fig. 5. Most videos are captured at $1,080$p resolution with various smartphone cameras. More details on the SFW database are provided in the Supplemental Materials.

For evaluation purposes, we annotated the pixel-wise shadow segmentation maps of key frames selected from the video set. We labeled 440 frames, where each frame was annotated by 2 people and a third person performed quality assurance. When selecting frames from each video to include in the SFW evaluation dataset, we ensured that the shadow patterns were as diverse as possible and avoided selecting redundant frames. Examples of frames we selected from one SFW video are shown in Fig. 6.

5 Experiments

5.1 Shadow removal and segmentation

To evaluate shadow removal, we use the UCB dataset and compute the PSNR and SSIM [47] between our deshadowed images and the groundtruth. We also evaluate shadow segmentation using SFW, which has groundtruth segmentations. We compute the area under curve (AUC) of ROC curve and accuracy based on the predicted (Eqn. 7) and groundtruth shadow masks. The accuracy is computed as $\frac{TP+TN}{N_p+N_n}$ where TP, TN, N_p, and N_n are true positives, true negatives, number of shadow pixels, and number of non-shadow pixels respectively. We binarize the shadow matte $\mathbf{M}$ into a shadow mask with a threshold of 0.1.

We first compare results on UCB. The primary baseline is [55], which also includes the performance of several previous works [8, 17, 22]. Our novel grayscale shadow removal and colorization model (GS+C) achieves state-of-the-art PSNR and SSIM, and outperforms the

Figure 7: Shadow removal quality on UCB database. From top to bottom, we show the input images, the groundtruth deshadowed images, the shadow removal results provided by [55], our network with naive RGB shadow modeling, and our single-frame network with grayscale shadow removal and colorization (GS+C). Blue arrows point to artifacts in [55] (*i.e.* unnatural gray patches and yellow traces).

Removal Model	PSNR	SSIM
Input Image	19.671	0.766
Guo *et al.*[17]	15.939	0.593
Hu *et al.*[22]	18.956	0.699
Cun *et al.*[8]	19.386	0.722
Zhang *et al.*[55]	23.816	0.782
RGB (Ours)	23.005	0.854
GS+C (Ours)	**23.829**	**0.866**

Table 1: Comparison of shadow removal performance on the UCB dataset. Our model outperforms all baselines in both PSNR and SSIM. Our GS+C model also outperforms naive RGB shadow modeling (RGB).

reported performance of [55] significantly on SSIM (See Tab. 1). For this experiment, we do not utilize TSM since the UCB test set consists of only single images and our model thus operates on single frames. A qualitative comparison on UCB is shown in Fig. 7. Our GS+C model is able to qualitatively improve over [55] by avoiding gray artifacts in the image (*i.e.* columns 1 and 2) as well as erroneous yellow patches (*i.e.* around the eye in column 3).

Second, we evaluate the models on the SFW database, which is more challenging due to its highly dynamic environments. We conduct a quantitative comparison on the performance of shadow segmentation (See Tab. 2). As no pre-trained models, training data, and training scripts for [8, 17, 22, 55] are available, it is not possible to reproduce the exact models reported in those papers. Therefore, we compare with them on the UCB dataset using the reported performance from [55] but not on the SFW dataset. Our method outperforms all others in AUC and accuracy. Fig. 8 visually demonstrates that our method is better at removal and segmentation of facial foreign shadows compared to the baselines. As our datasets are highly diverse, we find that [25] and [18] cannot generalize well.

5.2 Ablation studies

To ablate GS+C, our baseline is our method with direct RGB shadow modeling. For a fair comparison, we merge the computation resources of the GS+C model (*i.e.* doubling the bottleneck depth and the decoder channels). GS+C outperforms the RGB model and achieves the best PSNR and SSIM (see Tab. 1) thanks to the effectiveness of our novel shadow mod-

Figure 8: Qualitative shadow removal results on the SFW database. We show the input, shadow removal results from [25] and [18], our single-frame model, our temporal model, groundtruth shadow segmentations (in bright purple), and our predicted shadow masks (before thresholding).

Table 2: Comparison of shadow segmentation performance on the SFW database. Our model outperforms all baselines in AUC and Accuracy. Our Temporal GS+C model (with TSM) leads to improved shadow segmentation over GS+C.

Segmentation Model	AUC	Accuracy
Le and Samaras [25]	0.603	0.683
Hu *et al.*[21]	0.540	0.604
He *et al.*[18]	0.725	0.858
GS+C (Ours)	0.824	0.888
Temporal GS+C (Ours)	**0.836**	**0.890**

eling and decomposition, as well as better visual quality especially in the deshadowed region (see Fig. 7). GS+C leaves behind less noticeable artifacts compared to the RGB model (*e.g.* columns 3, 4, 7, 8, and 9). GS+C is also more comprehensive in removing the entire foreign shadow (*e.g.* column 7). We also ablate TSM on the SFW database, which is a video dataset. For the shadow segmentation experiment, we apply TSM to each input image by treating the horizontally flipped image as a second frame. TSM achieves the best performance in AUC and accuracy (see Tab. 2). As seen in Fig. 8, adding TSM also suppresses artifacts in the deshadowed region (*e.g.* columns 1, 5, and 6) due to enforcing face symmetry.

6 Conclusion

We introduce a new problem: blind removal of facial foreign shadows, and propose an effective shadow modeling algorithm to improve generalizability. We decompose conventional RGB shadow modeling into grayscale shadow modeling and colorization and propose a temporal sharing module (TSM) that can be integrated into other methods to impose temporal consistency and face symmetry. Our method produces photo-realistic deshadowed faces with SoTA PSNR and SSIM. Our SFW video database collected under highly dynamic environments is another major contribution that can benefit face-related research and applications.

References

[1] Mallikarjun B R, Ayush Tewari, Tae-Hyun Oh, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, and Christian Theobalt. Monocular reconstruction of neural face reflectance fields. In *CVPR*, 2021.

[2] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3D morphable model. *PAMI*, 2003.

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem?(and a dataset of 230,000 3D facial landmarks). In *ICCV*, 2017.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[5] Zipei Chen, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Canet: A context-aware network for shadow removal. In *ICCV*, 2021.

[6] Yung-Yu Chuang, Dan B Goldman, Brian Curless, David H Salesin, and Richard Szeliski. Shadow matting and compositing. In *SIGGRAPH*, 2003.

[7] Ciprian A Corneanu, Meysam Madadi, Sergio Escalera, and Aleix M Martinez. What does it mean to learn in deep networks? and, how does one detect adversarial attacks? In *CVPR*, 2019.

[8] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting GAN. In *AAAI*, 2020.

[9] Rahul Dey and Vishnu Boddeti. 3DFaceFill: An analysis-by-synthesis approach to face completion. In *WACV*, 2022.

[10] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao. ARGAN: Attentive recurrent generative adversarial network for shadow detection and removal. In *ICCV*, 2019.

[11] Craig Donner and Henrik Wann Jensen. A spectral BSSRDF for shading human skin. In *EGSR*, 2006.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[13] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3D morphable models and an illumination prior for face image analysis. *IJCV*, 2018.

[14] Graham D Finlayson, Steven D Hordley, and Mark S Drew. Removing shadows from images. In *ECCV*, 2002.

[15] Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang. Auto-exposure fusion for single-image shadow removal. In *CVPR*, 2021.

[16] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. Mask-guided portrait editing with conditional GANs. In *CVPR*, 2019.

[17] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. *PAMI*, 2012.

[18] Yingqing He, Yazhou Xing, Tianjia Zhang, and Qifeng Chen. Unsupervised portrait shadow removal via generative priors. In *MM*, 2021.

[19] Andrew Hou, Ze Zhang, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Towards high fidelity face relighting with realistic shadows. In *CVPR*, 2021.

[20] Andrew Hou, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Face relighting with geometrically consistent shadows. In *CVPR*, 2022.

[21] X Hu, CW Fu, L Zhu, J Qin, and PA Heng. Direction-aware spatial context features for shadow detection and removal. *PAMI*, 2019.

[22] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *CVPR*, 2018.

[23] Yeying Jin, Aashish Sharma, and Robby T. Tan. DC-ShadowNet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network. In *ICCV*, 2021.

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[25] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *ICCV*, 2019.

[26] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020.

[27] Zhihang Li, Yibo Hu, Ran He, and Zhenan Sun. Learning disentangling and fusing networks for face completion under structured occlusions. *Pattern Recognition*, 2020.

[28] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088*, 2017.

[29] Zhihao Liu, Hui Yin, Xinyi Wu, Zhenyao Wu, Yang Mi, and Song Wang. From shadow generation to shadow removal. In *CVPR*, 2021.

[30] Koki Nagano, Huiwen Luo, Zejian Wang, Jaewoo Seo, Jun Xing, Liwen Hu, Lingyu Wei, and Hao Li. Deep face normalization. *TOG*, 2019.

[31] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In *CVPR*, 2020.

[32] Rohit Pandey, Sergio Orts-Escolano, Chloe LeGendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: Learning to relight portraits for background replacement. In *SIGGRAPH*, 2021.

[33] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *CVPR*, 2017.

[34] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. SfSNet: Learning shape, reflectance and illuminance of facesin the wild'. In *CVPR*, 2018.

[35] Amnon Shashua and Tammy Riklin-Raviv. The quotient image: Class-based rerendering and recognition with varying illuminations. *PAMI*, 2001.

[36] YiChang Shih, Sylvain Paris, Connelly Barnes, William T Freeman, and Frédo Durand. Style transfer for headshot portraits. In *SIGGRAPH*, 2014.

[37] Yael Shor and Dani Lischinski. The shadow meets the mask: Pyramid-based shadow removal. *CGF*, 2008.

[38] Zhixin Shu, Sunil Hadap, Eli Shechtman, Kalyan Sunkavalli, Sylvain Paris, and Dimitris Samaras. Portrait lighting transfer using a mass transport approach. *TOG*, 2017.

[39] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, 2017.

[40] Arne Stoschek. Image-based re-rendering of faces for continuous pose and illumination directions. In *CVPR*, 2000.

[41] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul E Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *TOG*, 2019.

[42] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCVW*, 2017.

[43] Luan Tran and Xiaoming Liu. Nonlinear 3D face morphable model. In *CVPR*, 2018.

[44] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *CVPR*, 2018.

[45] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[46] Yang Wang, Lei Zhang, Zicheng Liu, Gang Hua, Zhen Wen, Zhengyou Zhang, and Dimitris Samaras. Face relighting from a single image under arbitrary unknown lighting conditions. *PAMI*, 2008.

[47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004.

[48] Zhen Wen, Zicheng Liu, and Thomas S Huang. Face relighting with radiance environment maps. In *CVPR*, 2003.

[49] Qi Wu, Wende Zhang, and BVK Vijaya Kumar. Strong shadow removal via patch-based shadow edge detection. In *ICRA*, 2012.

[50] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3D objects from images in the wild. In *CVPR*, 2020.

[51] Tai-Pang Wu and Chi-Keung Tang. A bayesian approach for shadow extraction from a single image. In *ICCV*, 2005.

[52] Yang Yang, Xiaojie Guo, Jiayi Ma, Lin Ma, and Haibin Ling. Lafin: Generative landmark guided face inpainting. *arXiv preprint arXiv:1911.11394*, 2019.

[53] Edward Zhang, Ricardo Martin-Brualla, Janne Kontkanen, and Brian Curless. No shadow left behind: Removing objects and their shadows using approximate lighting and geometry. In *CVPR*, 2021.

[54] Shu Zhang, Ran He, Zhenan Sun, and Tieniu Tan. DeMeshNet: Blind face inpainting for deep meshface verification. *TIFS*, 2017.

[55] Xuaner Zhang, Jonathan T. Barron, Yun-Ta Tsai, Rohit Pandey, Xiuming Zhang, Ren Ng, and David E. Jacobs. Portrait shadow manipulation. *TOG*, 2020.

[56] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *ICCV*, 2019.

[57] Lei Zhu, Ke Xu, Zhanghan Ke, and Rynson W.H. Lau. Mitigating intensity bias in shadow detection via feature decomposition and reweighting. In *ICCV*, 2021.

[58] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3D total solution. *PAMI*, 2017.

[59] Yurui Zhu, Jie Huang, Xueyang Fu, Feng Zhao, Qibin Sun, and Zheng-Jun Zha. Bijective mapping network for shadow removal. In *CVPR*, 2022.