Supplemental Materials for 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

1 Loss Functions

We now explain the loss functions that we apply to supervise the three steps of our method. Recall that we use a self-supervised approach to train our model involving synthetic shadow faces produced by synthesizing shadows on shadowless or near-shadowless FFHQ [3] images. We thus have both the input (synthetic shadow face) and the groundtruth (original FFHQ image) available during training.

Shadow removal loss With the paired shadow face and well-illuminated shadowless face, we enable a pixel level supervision on the recovery in grayscale. Specifically, we introduce a weighting map to encourage the loss to focus more on the shadow and shadow boundary, defined as

$$\mathcal{L}_{gs} = \left\| \hat{\mathbf{I}}_i^{b,gs} - \mathbf{I}_i^{b,gs} \right\|_1 \odot \frac{1 + \mathbf{B}_i + \mathbf{B}_i^{edge}}{\mathbf{R}}, \tag{10}$$

where $\hat{\mathbf{I}}_i^{b,gs}$ is the predicted grayscale deshadowed face, $\mathbf{I}_i^{b,gs}$ is the groundtruth grayscale deshadowed face, $\mathbf{B}^{edge}$ is the boundary of $\mathbf{B}$, $1 + \mathbf{B} + \mathbf{B}^{edge}$ is the weighting map, and $\mathbf{R}$ is the normalization term of the weighting map. A similar loss is also applied to the final RGB recovery and is denoted as L_{clr}.

Image gradient loss Human vision is very sensitive to high frequency artifacts, such as edges. To further suppress artifacts around shadow boundaries and recover high-frequency

* Equal Contribution.

details beneath shadows, we adopt an image gradient loss to encourage the image gradients between $\hat{\mathbf{I}}^b$ and $\mathbf{I}^b$ to be similar. This loss is denoted as:

$$\mathcal{L}_\nabla = \sum_k \|\nabla \lfloor \hat{\mathbf{I}}_i^b \rfloor_k - \nabla \lfloor \mathbf{I}_i^b \rfloor_k\|_1 \odot \frac{1+\mathbf{B}_i+\mathbf{B}_i^{edge}}{\mathbf{R}}, \tag{11}$$

where ∇ is the gradient operator and $\lfloor \cdot \rfloor_k$ denotes downsampling by the ratio $k = \{1,2,4,8\}$. Multiscale gradients help remove both sharp and blurry shadow boundaries.

Perceptual loss $\mathcal{L}_P$ To enhance the visual quality, we adopt the perceptual loss between the recovered face $\hat{\mathbf{I}}^b$ and $\mathbf{I}^b$. Similar to [5], our perceptual loss is defined as the VGG19 feature similarity of the deshadowed prediction and the groundtruth. We compute the VGG19 features using the first convolutional layer in each block and use the L_1 difference of the features, weighting each layer equally.

GAN loss Motivated by [4], we adopt a multiscale PatchGAN [7] at the scales 1, $1/2$, and $1/4$ of the original image's resolution. Each discriminator consists of 5 convolutional layers and 4 pooling layers, and outputs a 1-channel map in the range of $[0,1]$, where 0 denotes synthetic and 1 denotes real. We use the hinge loss in the GAN training:

$$\mathcal{L}_D = -\sum_{n=1}^{3} \min(0, D_n(\mathbf{I}_i^b) - 1) - \sum_{n=1}^{3} \min(0, -D_n(\hat{\mathbf{I}}_i^b) - 1),$$

$$\mathcal{L}_G = -\sum_{n=1}^{3} D_n(\hat{\mathbf{I}}_i^b), \tag{12}$$

where D_1, D_2 and D_3 are discriminators at 3 scales. $\mathcal{L}_D$ is the discriminator loss and $\mathcal{L}_G$ guides the shadow removal model to recover more realistic shadow-free faces.

Overall Loss The generator is thus supervised by the following loss:

$$\mathcal{L} = \alpha_1 \mathcal{L}_{gs} + \alpha_2 \mathcal{L}_{clr} + \alpha_3 \mathcal{L}_\nabla + \alpha_4 \mathcal{L}_P + \alpha_5 \mathcal{L}_G. \tag{13}$$

The discriminators are supervised with adversarial loss $\mathcal{L}_D$ to compete with the generator. We execute the generator step and the discriminator step in each mini-batch iteration.

2 SFW Video Shadow Removal

To demonstrate our video shadow removal performance and to show our Temporal Sharing Module's (TSM's) ability to improve temporal consistency, we include results for 4 subjects from our SFW database. We find that our proposed model with TSM is able to better maintain temporal consistency compared to our model with only grayscale shadow removal and colorization (GS+C). In particular, there is less flickering and other noticeable artifacts when transitioning between frames compared to the GS+C model. Our proposed model with TSM is also able to remove a wide range of foreign shadows (*e.g.* the hand shadow of the first subject and the phone shadow of the fourth subject) as well as strong self shadows (*e.g.* the eye shadows of the second subject and the nose shadow of the third subject).

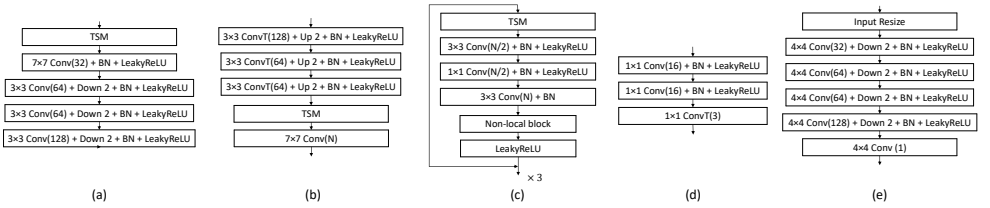

Figure 9: The network architecture of each component in our shadow removal network. (a) Encoder, (b) Decoder, (c) Non-local bottleneck module, (d) Color space transformation, and (e) Discriminator. Down 2 indicates down-sampling with a stride of 2 and Up 2 indicates up-sampling with a stride of 2. The number in each block indicates the number of output channels. TSM denotes our Temporal Sharing Module.

3 SFW Database Collection

We provide more details on the contents of the SFW database here, our large-scale, in-the-wild video database for foreign shadow removal and segmentation. SFW contains 280 videos from 20 subjects captured under highly dynamic environments as well as 440 annotated frames for evaluating shadow segmentation performance. Most videos are captured at $1,080$p resolution with various smartphone cameras. For each subject, videos are collected in five sessions: indoor, outdoor standing, outdoor walking, outdoor extreme, and driving. The indoor session collects videos in an indoor environment, where the lighting is relatively soft with no strong specular lights. For outdoor collection, the standing session requires the subject to hold a standing position with no ambient light variations, and the walking and extreme sessions require the subject to be moving, creating a changing ambient light. For the first three sessions, subjects use common objects to create shadows, including hands, phones, paper, and pens. For the outdoor extreme session, we strive to create more complex shadow patterns and require the subjects to walk under trees to create leaf-shape shadows. In the last session, the subjects record videos in a moving car, where shadows may be generated by the sun visor, the rear-view mirror, a pillar or bridge, and surrounding buildings.

4 Network Architecture

The architecture details of our network are shown in Fig. 9. It consists of five sub-networks: the encoder, the decoder (the same for both the grayscale shadow removal module and the colorization module), the non-local bottleneck module, and the discriminators. Each convolution layer is concatenated with batch normalization and a leaky ReLU layer (except the output layer). The bottleneck module (Fig. 9(c)) is repeated 3 times (no sharing) for both the grayscale shadow removal module and the colorization module, and the channel size is 256.

5 Transition from synthesis to modeling

In Sec. 3.1, we mention that while dealing with shadows in grayscale, $\mathbf{C}$ in Eqn. 3 becomes a scalar, and thus it is feasible to obtain a closed-form transition from synthesis (Eqn. 1) to

modeling (Eqn. 6). We show a step-by-step process here:

$$\begin{aligned}\mathbf{I} &= \mathbf{I}^b \odot (1-\mathbf{B}\odot\mathbf{M}_{\text{ss}}) + \mathbf{I}^d \odot \mathbf{B} \odot \mathbf{M}_{\text{ss}} \odot \mathbf{M}_{\text{I}} \\ &= \mathbf{I}^b \odot (1-\mathbf{B}\odot\mathbf{M}_{\text{ss}}) + \mathbf{I}^b\mathbf{C} \odot \mathbf{B} \odot \mathbf{M}_{\text{ss}} \odot \mathbf{M}_{\text{I}},\end{aligned} \tag{14}$$

$$\begin{aligned}\mathbf{I}\odot\mathbf{B} &= \mathbf{I}^b \odot (1-\mathbf{B}\odot\mathbf{M}_{\text{ss}}) \odot \mathbf{B} + \\ &\qquad \mathbf{I}^b\mathbf{C} \odot \mathbf{B} \odot \mathbf{M}_{\text{ss}} \odot \mathbf{M}_{\text{I}} \odot \mathbf{B} \\ &= \mathbf{I}^b \odot \mathbf{B} \odot (1-\mathbf{M}_{\text{ss}}) + \mathbf{I}^b\mathbf{C} \odot \mathbf{B} \odot \mathbf{M}_{\text{ss}} \odot \mathbf{M}_{\text{I}} \\ &= \mathbf{I}^b \odot \mathbf{B} \odot (1-\mathbf{M}_{\text{ss}} + \mathbf{C}\mathbf{M}_{\text{ss}} \odot \mathbf{M}_{\text{I}}),\end{aligned} \tag{15}$$

$$\begin{aligned}\mathbf{I}\odot(1-\mathbf{B}) &= \mathbf{I}^b \odot (1-\mathbf{B}\odot\mathbf{M}_{\text{ss}}) \odot (1-\mathbf{B}) + \\ &\qquad \mathbf{I}^b\mathbf{C} \odot \mathbf{B} \odot \mathbf{M}_{\text{ss}} \odot \mathbf{M}_{\text{I}} \odot (1-\mathbf{B}) \\ &= \mathbf{I}^b \odot (1-\mathbf{B}) \odot (1-\mathbf{B}\odot\mathbf{M}_{\text{ss}}) \\ &= \mathbf{I}^b \odot (1-\mathbf{B}).\end{aligned} \tag{16}$$

From Eqn. 15, we can derive $\mathbf{I}^b \odot \mathbf{B} = \mathbf{I}\odot\mathbf{B} \oslash (1-\mathbf{M}_{\text{ss}} + \mathbf{C}\mathbf{M}_{\text{ss}} \odot \mathbf{M}_{\text{I}})$, and then we derive:

$$\begin{aligned}\mathbf{I}^b &= \mathbf{I}^b \odot \mathbf{B} + \mathbf{I}^b \odot (1-\mathbf{B}) \\ &= \mathbf{I}\odot(1-\mathbf{B}) + \mathbf{I}\odot\mathbf{B} \oslash (1-\mathbf{M}_{\text{ss}} + \mathbf{C}\mathbf{M}_{\text{ss}} \odot \mathbf{M}_{\text{I}}) \\ &= \mathbf{I}\odot(1-\mathbf{B} + \mathbf{B} \oslash (1-\mathbf{M}_{\text{ss}} + \mathbf{C}\mathbf{M}_{\text{ss}} \odot \mathbf{M}_{\text{I}})) \\ &= \mathbf{I}\odot(1-\mathbf{B} + \mathbf{B} \oslash \mathbf{M}'_{\text{I}}),\end{aligned} \tag{17}$$

where $\mathbf{M}'_{\text{I}} = 1-\mathbf{M}_{\text{ss}} + \mathbf{C}\mathbf{M}_{\text{ss}} \odot \mathbf{M}_{\text{I}}$.

6 Implementation details

Our shadow removal network is implemented in Tensorflow with an initial learning rate of $1e$-4. We train for $186,000$ iterations in total with a batch size of 10, and decrease the learning rate by a factor of 0.9 every $20,000$ iterations. We initialize the weights with the normal distribution of $[0, 0.02]$. $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \beta\}$ are set to be $\{400, 400, 2, 0.005, 1, 0.1\}$. We use [1] to crop the face and provide 68 facial landmarks.

References

[1] 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.

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

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

[4] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018.

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