WildCap: Facial Albedo Capture in the Wild via Hybrid Inverse Rendering
Supplementary Material

Yuxuan Han[1] Xin Ming[1] Tianxiao Li[1] Zhuofan Shen[1] Qixuan Zhang[2,3] Lan Xu[2] Feng Xu[1]
[1]School of Software and BNRist, Tsinghua University [2]ShanghaiTech University [3]Deemos Technology

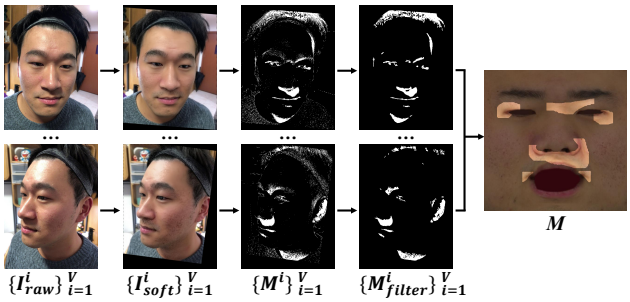


Figure 1. Pipeline of the proposed automatic method to obtain the shadow mask M.

Figure 2. Comparison of the manual and automatic shadow mask M. We visualize the shadow mask in the first row and show the reconstructed diffuse albedo map in the second row.

A. More Implementation Details

A.1. Shadow Detection

As mentioned in the main paper, our method only requires a coarse mask M to indicate the baking artifacts in I_{UV}. Thus, obtaining M is low-cost and easy. In the following, we propose two methods to obtain M, one is manual, the other is fully automatic.

Manual Method For the manual method, we open I_{UV} in Photoshop and use the Polygonal Lasso Tool to mark the facial regions containing baking artifacts. This step is easy and only requires a few mouse clicks.

Automatic Method We also develop a fully automatic method to obtain M as shown in Figure 1. Specifically, we detect shadow regions in the raw images $\{I^i_{raw}\}^V_{i=1}$, and then lift these image-space predictions into the UV space to obtain M. The rationale is to use shadow as a proxy to locate baking artifacts.

To detect shadow regions in $\{I^i_{raw}\}^V_{i=1}$, we adopt an existing work, *i.e.*, DiFaReli [9]. Following DiFaReli++ [10], we use DiFaReli to soften the shadows in $\{I^i_{raw}\}^V_{i=1}$. We denote the processed images as $\{I^i_{soft}\}^V_{i=1}$. Then, we compute the shadow mask M^i by thresholding the color difference between I^i_{raw} and I^i_{soft}. We further apply a median filter to M^i and remove connected areas smaller than a threshold; we denote the processed per-view shadow mask as M^i_{filter} Next, we lift $\{M^i_{filter}\}^V_{i=1}$ to the UV space to obtain M; we also dilate M to some extent to ensure it includes all the baking artifacts. All the hyperparameters, *e.g.*, thresholds and kernel sizes, are shared across different subjects.

Comparison of the Two Methods As shown in Figure 2, the automatic and manual methods reconstruct diffuse albedo maps in similar quality. Since the goal of our automatic method is to detect shadow regions as a proxy for baking artifacts, it also includes regions around eyes in M. However, we find that SwitchLight produces negligible baking artifacts around the eyes in the 2 cases shown in Figure 2, thus we do not mark them in the manual method. In addition, we notice that the automatic method fails to detect the baked ambient occlusion effects on the side nose, as shown in the rightmost column. To ensure the highest quality, we use the manual method by default. We leave training a portrait-delighting network with shadow removal confidence as our future work.

A.2. Light Stage Dataset

Our Light Stage dataset for training the diffusion prior is the same as that used in DoRA [3]. The dataset contains 6 Asians (2 males and 4 females), 9 African Americans (5 males and 4 females), and 33 Caucasians (17 males and 16 females). Please refer to DoRA for details on processing the dataset.

A.3. Lighting Regularization

As mentioned in the main paper, during optimization, we add a regularization term $\mathcal{L}_{reg}$ to our lighting model Γ_θ:

$$\mathcal{L}_{reg} = 0.1 \cdot \mathcal{L}_{TV} + \mathcal{L}_{neg} \tag{1}$$

We apply a total variation regularization $\mathcal{L}_{TV}$ to constrain the spatial smoothness of the actual lighting parameters γ:

$$\mathcal{L}_{TV} = \sum_{u,v} ||\gamma_{u,v} - \gamma_{u,v-1}||_2^2 + ||\gamma_{u,v} - \gamma_{u-1,v}||_2^2 \tag{2}$$

We apply a negative shading regularization $\mathcal{L}_{neg}$ to constrain the shading of γ^V to be negative:

$$\mathcal{L}_{neg} = \sum_{u,v} \max(0, s_{u,v}^V)^2 \tag{3}$$

Here, $s_{u,v}^V$ is the shading of γ^V at UV location (u, v). The rationale of $\mathcal{L}_{neg}$ is that we expect baking artifacts to be explained as a clean diffuse albedo map illuminated by local dark lights.

A.4. Super-Resolution Network

We adopt RCAN [15] as our super-resolution network $\mathcal{U}$ to upsample the 1K resolution reflectance maps into 4K. Similar to previous works [7], we train $\mathcal{U}$ at the patch level. At inference time, we directly send a 1K-resolution reflectance map to $\mathcal{U}$. During training, we cropped paired reflectance patches from the 1K and 4K versions of the scan. The patch size is set to 48×48, and $\mathcal{U}$ is trained to upsample it to 192×192. We also modify the number of input and output channels of RCAN to 7 to support upsampling the concatenated diffuse albedo, specular albedo, and detailed normal map simultaneously.

B. More Experiments

B.1. Evaluation on Skin Tone Control

Recall that in our method, we control the skin tone via initialization. Specifically, we set the sampling start point $x_{T_{init}}$ as the noised version of a Light Stage scan x_0^{ref} whose skin tone is similar to the provided one. We also modify the diffuse albedo component of x_0^{ref} using the color-matching transform to better align with the provided skin tone. At the same time, we initialize the lighting so that the shaded initial reflectance map x_0^{ref} has a consistent color tone as I_{UV}. As shown in Figure 3, our strategy can effectively control the skin tone of the solved diffuse albedo maps (2nd row) to match the initialization x_0^{ref} (3rd row).

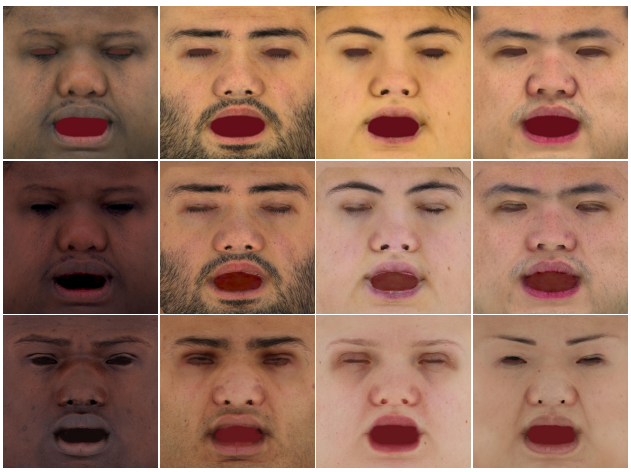


Figure 3. Evaluation on skin tone control. We show the texture map I_{UV}, the reconstructed diffuse albedo map A, and the initialization x_0^{ref} from the top row to the bottom row.

B.2. Baking Artifacts of SwitchLight

Since one of our core contributions is to explain SwitchLight's baking artifacts as lighting effects, a natural question is, when will SwitchLight produce these artifacts? In Figure 4, we comprehensively test our method on diverse in-the-wild cases, ranging from simple cases captured under low-frequency lighting to hard cases with apparent shadow and specularity appearing on the face.

From Figure 4, we find that SwitchLight performs quite well in easy cases with low-frequency lighting, such as the first two rows. As the scene illumination becomes high-frequency, shadows and specularity appear on the face. We empirically find that SwitchLight works well in removing specularity, but struggles in shadows. However, shadows are ubiquitous in everyday captures. For example, both the sun and the roof light bulb would cast shadows on the face. This drawback prevents SwitchLight from becoming an ideal method for facial albedo capture in the wild. Fortunately, thanks to our model-based optimization, we successfully explain the shadow-baking artifacts as a clean diffuse albedo illuminated by a dark shading.

B.3. Deeper Analysis of Texel Grid Lighting

As shown in Figure 5 of the main paper, there is a trade-off controlled by the grid size: a larger grid (with smaller g) can better preserve details, while a smaller grid (with larger g) can produce a cleaner albedo due to increased representation capacity. Here, we test on a more challenging case to

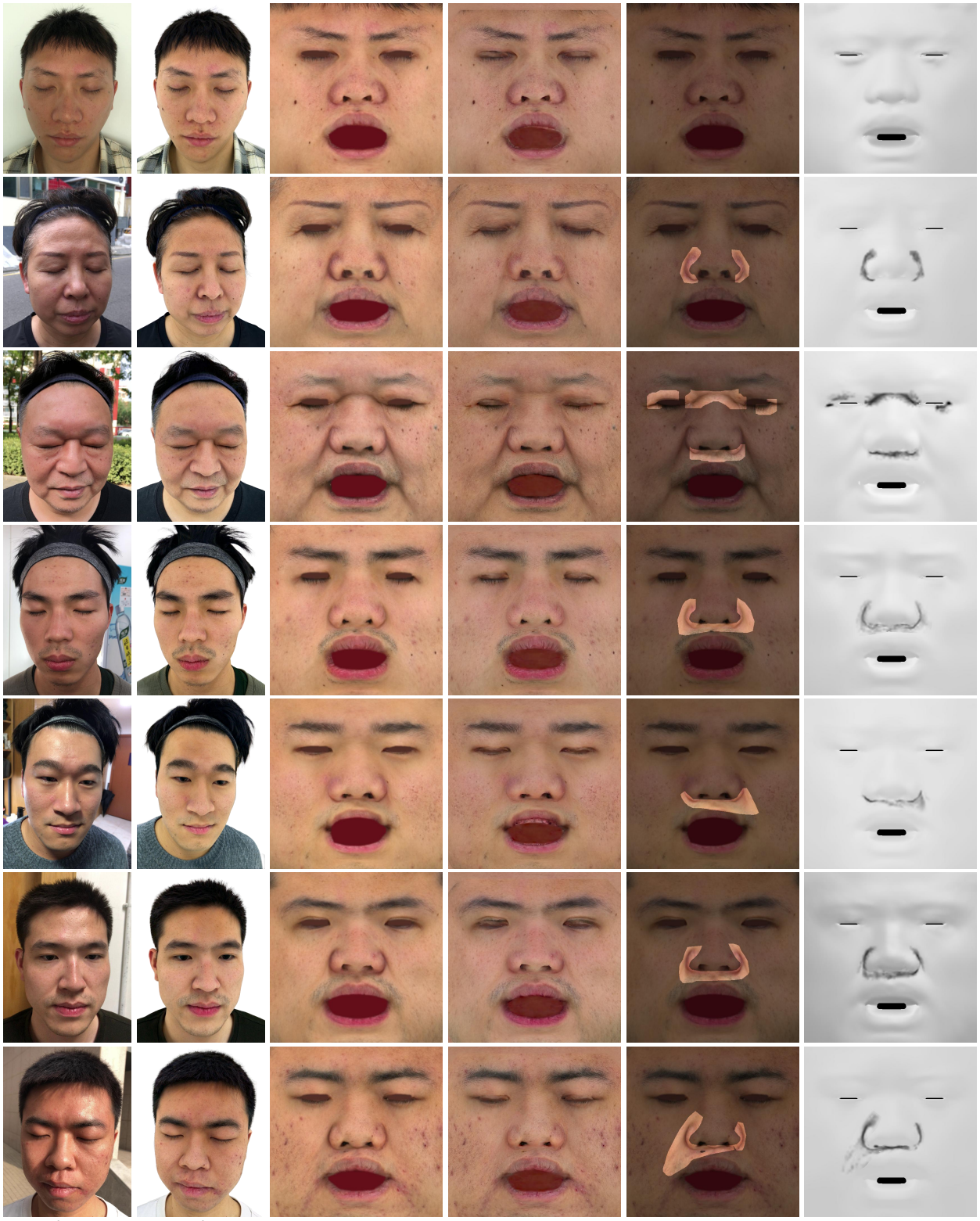
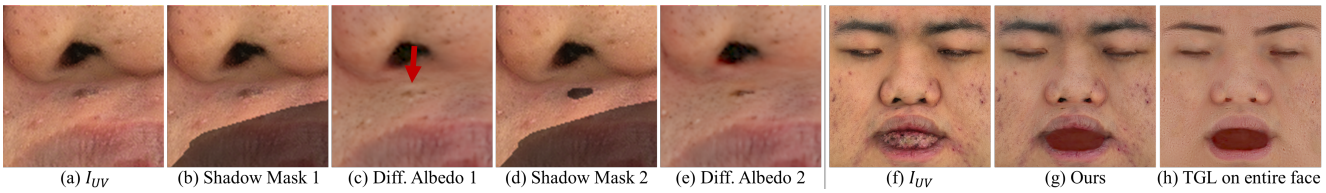


Figure 4. Evaluation of our method on various in-the-wild captures. From left to right, we show a raw captured image I^i_{raw}, the predicted diffuse albedo image I^i by SwitchLight, the texture I_{UV}, the reconstructed diffuse albedo map, the shadow mask used to modulate our texel grid lighting model, and the shading.



(a) I_{UV} (b) Shadow Mask 1 (c) Diff. Albedo 1 (d) Shadow Mask 2 (e) Diff. Albedo 2 (f) I_{UV} (g) Ours (h) TGL on entire face

Figure 5. A deeper analysis of the Texel Grid Lighting.

	PSNR ↑	SSIM [12] ↑	LPIPS [14] ↓
DeFace*	28.43	0.9791	0.0826
FLARE*	22.48	0.9742	0.0571
Ours	**28.71**	**0.9802**	**0.0388**

Table 1. Quantitative comparison on diffuse albedo reconstruction. We compare our method with DeFace* and FLARE* using the scan of Digital Emily. The metric is computed on the same cropped facial skin region.

highlight the potential loss of real texture details due to the high-capacity representation of TGL.

In an extreme case shown in Figure 5 where *the nevus is located in the shadow region and its size is coincidentally close to the grid size* (a), our method ($g = 96$) somehow absorbs it into the lighting (b)(c). This is because the small grid lacks global information to distinguish nevus from baking shadow. To address this while preserving the expressive power of the small grid, we can exclude the nevus from the shadow mask (d)(e). This aligns with our design: we apply TGL only to the shadow region (g), rather than the entire face (h).

B.4. Quantitative Comparison on Synthetic Data

We conduct a quantitative comparison of diffuse albedo reconstruction using the Digital Emily project [1]. We render multi-view images from the scan and then run each method (Ours, DeFace* [4], and FLARE* [2]). We compare the reconstructed diffuse albedo and GT in image space. As shown in Table 1, our method obtains the best metrics.

B.5. Results on Studio-Captured Dataset

Our method can also be applied to studio-captured multi-view face datasets, like NeRSemble [6] and Ava256 [8]. We show some results on NeRSemble in Figure 6. Compared to in-the-wild videos captured by a smartphone camera, these studio-captured datasets are less challenging. The reason is that the lighting conditions in these studio-captured datasets are low-frequency. For example, Ava256 uses uniform white light to capture the data, and the captured images are almost shadow-free. We believe using our method to create an open-sourced, large-scale Light Stage dataset from existing studio-captured datasets is a valuable future direction.

B.6. Position and Contribution of WildCap

Despite previous closed-source works, such as Xu et al. [13] and Rainer et al. [11], proposing to capture appearance from multi-view images, *we emphasize that we consider a more challenging and practical problem*.

As shown in Figure 7, the test subjects in Xu et al. [13] and Rainer et al. [11] have *little or moderate* cast shadow (1st row), which can already be well resolved by SwitchLight [5] (2nd row) or our baseline method *w/o TGL*. However, we test on subjects with *strong* cast shadow (*e.g.*, the last 3 rows in Figure 4), where SwitchLight leaves apparent baking artifacts. We emphasize that cast shadow from indoor lighting or the sun is ubiquitous in the real world. Although we process more challenging data, our diffuse albedo maps have fewer artifacts and more details than Figure 1 of Xu et al. [13] and Figure 5 of Rainer et al. [11].

Thus, in addition to technical novelty, we convey to the community that low-cost techniques can handle challenging cases with strong cast shadows, which is *a new effect and a large step* in this field.

B.7. Failure Case of WildCap

Since our lighting representation is continuous, our method does not perform well when sharp shadow boundaries appear in SwitchLight's prediction. As shown in Figure 8, we test on a challenging case where the video is captured at noon under the sun. Even after being delighted by SwitchLight, there are still very sharp shadow boundaries on the face. Although our method obtains significantly better results, it still cannot totally remove these sharp boundaries. To address this, we leave training an improved portrait-delighting network as our future work.

References

[1] Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. The digital emily project: photoreal facial modeling and animation. In *Acm siggraph 2009 courses*, pages 1–15. 2009. 4

[2] Shrisha Bharadwaj, Yufeng Zheng, Otmar Hilliges, Michael J Black, and Victoria Fernandez-Abrevaya. Flare:

Figure 6. Results of our method on the NeRSemble [6] dataset (4 of 16 captured images are shown above).

Figure 7. Switchlight's prediction (the second row) of test subjects in Xu et al. [13] and Rainer et al. [11] (the first row).

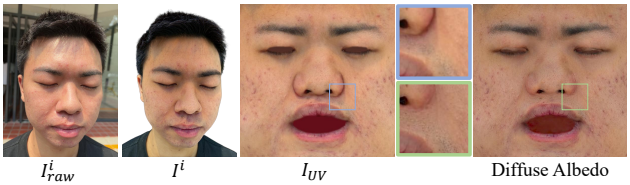


Figure 8. Failure case of our method. From left to right, we show the raw captured image I^i_{raw}, the predicted diffuse albedo image I^i by SwitchLight, the texture map I_{UV}, close-ups, and the reconstructed diffuse albedo map.

Fast learning of animatable and relightable mesh avatars. *arXiv preprint arXiv:2310.17519*, 2023. 4

[3] Yuxuan Han, Junfeng Lyu, Kuan Sheng, Minghao Que, Qixuan Zhang, Lan Xu, and Feng Xu. Facial appearance capture at home with patch-level reflectance prior. In *SIGGRAPH*, 2025. 2

[4] Tianxin Huang, Zhenyu Zhang, Ying Tai, and Gim Hee Lee. Learning to decouple the lights for 3d face texture modeling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 4

[5] Hoon Kim, Minje Jang, Wonjun Yoon, Jisoo Lee, Donghyun Na, and Sanghyun Woo. Switchlight: Co-design of physics-driven architecture and pre-training framework for human portrait relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25096–25106, 2024. 4

[6] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), 2023. 4, 5

[7] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme: Realistically renderable 3d facial reconstruction" in-the-wild". In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 760–769, 2020. 2

[8] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu,

Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. *NeurIPS Track on Datasets and Benchmarks*, 2024. 4

[9] Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. Difareli: Diffusion face relighting. 2023. 1

[10] Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. Difareli++: Diffusion face relighting with consistent cast shadows, 2025. 1

[11] Gilles Rainer, Lewis Bridgeman, and Abhijeet Ghosh. Neural shading fields for efficient facial inverse rendering. In *Computer Graphics Forum*, page e14943. Wiley Online Library, 2023. 4, 5

[12] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4

[13] Yingyan Xu, Kate Gadola, Prashanth Chandran, Sebastian Weiss, Markus Gross, Gaspard Zoss, and Derek Bradley. Monocular facial appearance capture in the wild. *ICCV*, 2025. 4, 5

[14] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 4

[15] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 2