Supplementary on "Self-Supervised Robustifying Guidance for Monocular 3D Face Reconstruction"

Hitika Tiwari[1,4,*]
hitika@iitk.ac.in
Min-Hung Chen[2,**]
vitec6@gmail.com
Yi-Min Tsai[3]
yi-min.tsai@mediatek.com
Hsien-Kai Kuo[3]
hsienkai.kuo@mediatek.com
Hung-Jen Chen[3]
hj.chen@mediatek.com
Kevin Jou[3]
kevin.jou@mediatek.com
K. S. Venkatesh[1]
venkats@iitk.ac.in
Yong-Sheng Chen[4]
yschen@nycu.edu.tw

[1] Indian Institute of Technology Kanpur, India
[2] Microsoft AI R&D Center, Taiwan
[3] MediaTek Inc., Taiwan
[4] National Yang Ming Chiao Tung University, Taiwan

In the supplementary material, we would like to provide more technical details, experiments, and discussions about the limitations and societal impact.

1 More Technical Details

In this section, we detail the preliminaries of the proposed method (refer to Section 1.1). Further, in Section 1.2, we present the robustifying guidance using Guidance Pipeline, which aids the Robustification Pipeline in addressing the issue of occlusions and noise in the face images in a self-supervised manner.

1.1 Preliminaries: Monocular 3D Face Reconstruction

In this section, the preliminaries of the proposed approach are introduced, such as 3DMM [3], illumination assumptions, and 3D face projection, which are crucial to address the problem of 3D reconstruction from monocular face images.

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.
*Work partially done during the research internship at MediaTek Inc., Taiwan.
**Work partially done during MediaTek Inc., Taiwan.

3D Morphable Model (3DMM): In 3DMM, a set of geometry and texture coefficients lead to the formation of a 3D face. Thus the formulas of geometry vector $\mathbf{M}$ and texture vector $\mathbf{T}$ for the 3DMM model are stated as follows:

$$\mathbf{M} = \overline{\mathbf{M}} + \mathbf{s}\mathbf{B_s} + \mathbf{e}\mathbf{B_e}, \qquad \mathbf{T} = \overline{\mathbf{T}} + \mathbf{t}\mathbf{B_t}. \tag{1}$$

A linear combination of $\mathbf{B_s} \in \mathbb{R}^{3N\times 80}$ and $\mathbf{B_e} \in \mathbb{R}^{3N\times 64}$ (subsets of Principal Component Analysis basis for shape and expression) with the predicted shape parameter $\mathbf{s} = [s_1, \cdots s_{80}]$ and expression parameter $\mathbf{e} = [e_1, \cdots e_{64}]$ respectively morphs the mean 3D face geometry $\overline{\mathbf{M}} \in \mathbb{R}^{3N}$ (refer to Eq. 1). Similarly, texture morphing is facilitated by adding the mean texture $\overline{\mathbf{T}} \in \mathbb{R}^{3N}$ to the linear combination of texture basis vector $\mathbf{B_t} \in \mathbb{R}^{3N\times 80}$ and predicted texture parameter $\mathbf{t} = [t_1, \cdots t_{80}]$. The vectors $\mathbf{B_t}$, $\overline{\mathbf{T}}$, $\mathbf{B_s}$, and $\overline{\mathbf{M}}$ are obtained from the Basel Face Model [17] whereas we acquire $\mathbf{B_e}$ from the Facewarehouse model [4], following [13]. Note that we, as in [9], preclude the ear and neck regions; thus our mesh contains $N = 36$K vertices.

The 3D face illumination is represented using Spherical Harmonics by assuming a *Lambertian* surface reflectance [9]. To obtain the face image, the 3D face coordinates are mapped to the screen by assuming a pinhole camera under full perspective projection, as in [9]. We represent the 3D face illumination using Spherical Harmonics by assuming a *Lambertian* surface reflectance.

1.2 Guidance Pipeline

One of the main goals is learning reliable 3DMM coefficients with *the least* supervision and dependencies. Inspired by R-Net [9] which contains comparatively fewer dependencies, we propose the **Self-Supervised Guidance Pipeline** to learn the coefficients C_G by exploiting the cycle-consistency in a self-supervised manner (see main paper for details). All the components of the *Guidance Pipeline* are presented as follows:

Obtaining 3D Face Alignment: The first consistency we aim to maintain is the face alignment between the guiding image I_G and the reconstructed image $I_{G'}$, which is achieved by reducing the discrepancy between landmark coordinates of the faces. We represent the discrepancy using the *Landmark Loss* $\mathcal{L}_K$ as follows: Following R-Net, we obtain the alignment between the 2D non-occluded noise-free face image and the corresponding 3D face projection by reducing the discrepancies between the 68 landmark coordinates of the faces.

$$\mathcal{L}_K = ||L_G - L_{G'}||. \tag{2}$$

where L_G and $L_{G'}$ denote a set of 68 landmark coordinates of 2D face image and the rendered counterpart, respectively, and $||.||$ represents the L2 norm.

Obtaining Photometric Consistency: To reach the goal of fewer dependencies, we directly regress the pixels of the rendered 3D face ($I_{G'}$) over the corresponding guiding face image (I_G) and obtain the pixel-wise consistency between them by the *Guiding Photometric Loss* $\mathcal{L}_{GP}$ as follows:

$$\mathcal{L}_{GP} = ||I_G - I_{G'}||. \tag{3}$$

It is worth noting that unlike R-Net [9], we relax the requirement of skin masks as additional dependencies.

Obtaining Perceptual Loss: In addition to image-level information, reducing the feature-level discrepancy is critical to obtaining the perceptual accuracy of 3D faces. Thus, we adopt the *Perceptual Loss* $\mathcal{L}_P$ as follows:

$$\mathcal{L}_P = 1 - \frac{\langle \theta, \theta' \rangle}{||\theta||||\theta'||}. \tag{4}$$

where θ and θ' are the feature representations obtained from the pre-trained FaceNet model [20] for target image I_G and the corresponding rendered face $I_{G'}$, respectively.

Regularization: To ensure the plausible face geometry and texture of the reconstructed 3D face, we adopt the *Regularization* term $\mathcal{L}_R$, which enforces the coefficients to follow the (normal) distribution of 3DMM, as follows:

$$\mathcal{L}_R = w_s||\mathbf{s}|| + w_t||\mathbf{t}|| + w_e||\mathbf{e}||, \tag{5}$$

where w_s, w_t and w_e are the weights associated with shape $\mathbf{s}$, texture $\mathbf{t}$, and expression $\mathbf{s}$ coefficients, respectively.

Overall Loss for Guidance Pipeline: The overall loss function $\mathcal{L}_{guide}$ for the *Self-Supervised Guidance Pipeline* can be expressed below:

$$\mathcal{L}_{guide} = \alpha_K \mathcal{L}_K + \alpha_{GP}\mathcal{L}_{GP} + \alpha_P \mathcal{L}_P + \alpha_R \mathcal{L}_R. \tag{6}$$

where $\alpha_K, \alpha_{GP}, \alpha_P$, and α_R are the weights for landmark loss (Eq. (2)), guiding photometric loss (Eq. (3)), perceptual loss (Eq. (4)) and regularization term (Eq. (5)), respectively. For simplicity, the notation of the image index is ignored throughout the whole paper.

2 More Experimental Details

In this section, we detail the datasets and the corresponding variations deployed for training the proposed *Self-Supervised **RO**bustifying **GU**idanc**E** (**ROGUE**)* framework (refer to Sec. 2.1). In addition, we present the experimental details, such as network architecture, weights for the losses, etc., in Sec. 2.2. Finally, more ablation studies and experiments are shown in Sec. 2.4 and 2.3, respectively.

2.1 Datasets

Procuring 3D ground-truth data is difficult due to privacy concerns and monetary issues. To validate the robustness against occlusions and noise, we exploit the proposed 1) ReaChOcc dataset, 2) SynChOcc dataset, and build a variant of the test set of CelebA [16] 3) *noisy face* set. A sample of the proposed ReaChOcc and SynChOcc datasets is shown in Fig. 1. For quantitative evaluation, we use a total of 7 high-performing face recognition models (the results of 2 are shown in the main paper and the remaining in supplementary) to determine the perceptual accuracy between the input face image and the corresponding rendered face. Our occlusion-based testing (real-life) and training (occluded with random shapes) images belong to different domains. Therefore, our method is distinct from the conventional data augmentation-based techniques, where the domains of test and training data are needed to be the same.

Figure 1: A demonstration of the samples from the proposed ReaChOcc and SynChOcc datasets. Our ReaChOcc contains unpaired clean images, whereas SynChOcc provides paired clean images for comparison.

2.2 Implementation Details

The proposed *Self-Supervised **RO**bustifying **GU**idanc**E** (**ROGUE**)* framework contains a coefficient network and two discriminators to facilitate the overall learning of the model. The coefficient network exploits ResNet-50 as backbone architecture with a modified classification layer by 257 nodes. The face images in the dataset are cropped, aligned (using the method in [7]), and reshaped to size 224×224. These images serve as the input to our model. We opt for a batch of 5 for each case: clean images, occluded faces, and noisy face images. Thus, the proposed network is trained with a net batch size of 15. In addition, we exploit linear discriminators with 3 fully-connected layers containing 257, 124, and 2 nodes, respectively, in the Robustification Pipeline. Our model is initialized with ImageNet weights [19]. In addition, an Adam optimizer [14] is deployed for training the model with an initial learning rate of 10^{-4} for the coefficient network and 10^{-8} for the discriminators. The proposed model contains the *Guidance Pipeline* and the *Robustification Pipeline*, where the weights associated with the losses in Guidance Pipeline are $\alpha_K = 1.6 \times 10^{-3}, \alpha_{GP} = 1.92, \alpha_P = 0.2$ and $\alpha_R = 3 \times 10^{-4}$ (as in R-Net), and the weights for Robustification Pipeline are $\beta_O = 1.92, \beta_N = 1.92$ and $\beta_C = 10^{-3}$ (please refer to Sec. 2.4 for more ablation experimental results).

2.3 More Experimental Results

This section qualitatively compares our method with DECA, MoFA, and R-Net on real-world and synthetic occlusions, and noisy images. Moreover, we provide the quantitative evaluation of our model on 7 face recognition models: VGG-Face [5], FaceNet [20], FaceNet-512 [20], OpenFace [1], DeepFace [21], ArcFace [8] and SFace [23]. *It is worth noting that*

Methods	Learning	Require Real Occluded Input	External Aid	Reported Robustness		
				Real-world Occlusions	Synthetic Occlusions	Noise
MoFA (TPAMI'18)	N/A	✓	N/A	✓	×	×
3DMM (CVPR'05)	N/A	✓	N/A	✓	×	×
Flow (Cambridge Press'03)	N/A	✓	N/A	✓	×	×
3DFFA (CVPR'16)	N/A	✓	N/A	✓	×	×
Sela et al. (ICCV'17)	N/A	✓	N/A	✓	×	×
Tran et al. (CVPR'18)	Weakly-Supervised	✓	Bump Map	✓	×	×
R-Net (CVPRW'19)	Weakly-Supervised	✓	Attention Mask	✓	×	×
DECA (TOG'21)	Weakly-Supervised	✓	Attention Mask	✓	×	×
MICA (ECCV'22)	Fully-Supervised	✓	Pre-trained ArcFace	✓	×	×
ROGUE (Ours)	Self-Supervised	×	No Aid Required	✓	✓	✓

Table 1: A comparison of the proposed method with various state-of-the-art approaches based on the learning scheme, input constraints, the requirement of external aid, and the reported robustness. Unlike SoTA approaches, our method performs well for four challenging data variations, three variants of occlusions, and noise in the images, without posing any dependency.

we demonstrate the perceptual results from DECA to emphasize the necessity of occlusion robust 3D texture reconstruction. However, we understand that DECA is not designed for occlusion-robust texture reconstruction. We also provide the details on how our work differs from various other approaches in Table 1.

2.3.1 Qualitative Evaluation

To validate the generalization ability of our method, we rigorously test our model on real and synthetic occlusion and noisy scenarios in the main paper. Following the course, in Fig. 2, Fig. 3, and Fig. 4, we present various examples to show the robustness of our model toward *unseen* real-life, synthetic occlusions and noisy challenging cases in the face images. It is worth noting that the improvement over MoFA and R-Net is more evident when the occlusion cover more than half of the face and the occlusion is not in skin color. Besides, DECA reproduces occlusion on the 3D faces, thus following a different line of research. We infer the same for the case of noisy face images. For a fair comparison, MoFA is re-trained to reconstruct 3D faces with the same number of face vertices ($N = 36K$), and only the trained CNN-based models (e.g., our coefficient network) are used at the inference stage.

2.3.2 Quantitative Evaluation

We investigate the quantitative performance for the *real-world challenging occlusions*, *synthetic challenging occlusions* and *noisy images*. As evident in the qualitative results in the main paper that only three methods, MoFA, R-Net, and DECA, reconstruct both 3D face shape and texture; thus, we compare the quantitative performance of our method with these three approaches.

1) Real-World Challenging Occlusions: Our quantitative results in Table 2 show better perceptual accuracy for the reconstructed 3D face than R-Net and MoFA. Our proposed method reduces perceptual errors between the vectors obtained from VGG-Face, FaceNet, FaceNet-512, OpenFace, DeepFace, ArcFace, and SFace by a large margin of **17.44**% (from 0.952 to 0.786), **21.55**% (from 1.202 to 0.943), **25.8**% (from 1.237 to 0.918), **11.3**% (from

Figure 2: A qualitative comparison for different methods on *real-life* occlusions.

Figure 3: *A qualitative comparison for different methods on synthetic occlusions.*

0.769 to 0.682), **8.1**% (from 0.656 to 0.603), **21.9**% (from 1.313 to 1.025), **13.4**% (from 1.215 to 1.052), respectively, compared to MoFA. In addition, the proposed method reduces **5.8**% (from 0.834 to 0.786), **8.0**% (from 1.025 to 0.943), **12.1**% (from 1.045 to 0.918), **11.3**% (from 0.706 to 0.682), **3.4**% (from 0.621 to 0.603), **13.7**% (from 1.188 to 1.025),

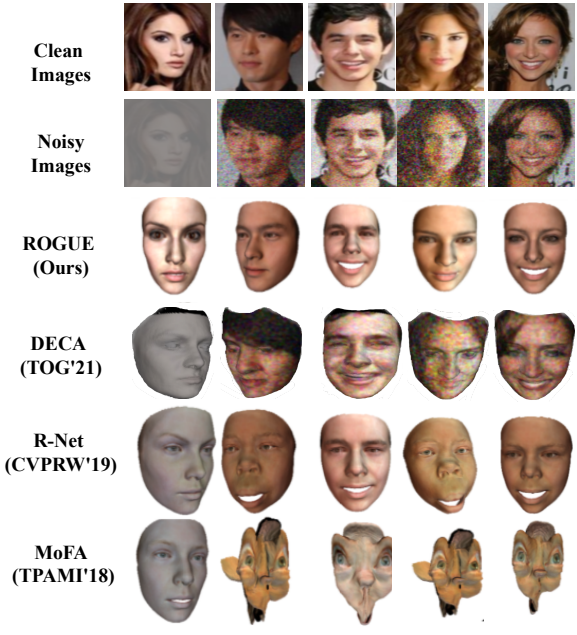

Figure 4: A qualitative comparison for different methods on *noisy images* (various noise types).

6.0% (from 1.119 to 1.052) of the perceptual distance between vectors obtained using VGG-Face, FaceNet, FaceNet-512, OpenFace, DeepFace, ArcFace, and SFace, respectively, from R-Net. Moreover, the proposed method reduces **19.4**% (from 0.975 to 0.786), **18.3**% (from 1.154 to 0.943), **16.3**% (from 1.097 to 0.918), **32.2**% (from 1.005 to 0.682), **38.3**% (from 0.977 to 0.603), **14.3**% (from 1.196 to 1.025), **22.4**% (from 1.356 to 1.052) of the perceptual distance between vectors obtained using VGG-Face, FaceNet, FaceNet-512, OpenFace, DeepFace, ArcFace, and SFace, respectively, from DECA. All the above results show that the proposed approach tackles real-world occlusions more effectively.

Methods	Perceptual Error (↓)						
	VGG-Face	FaceNet	FaceNet-512	OpenFace	DeepFace	ArcFace	SFace
MoFA (TPAMI'18)	0.952 ± 0.129	1.202 ± 0.146	1.237 ± 0.141	0.769 ± 0.189	0.656 ± 0.187	1.313 ± 0.114	1.215 ± 0.119
R-Net (CVPRW'19)	0.834 ± 0.143	1.025 ± 0.195	1.045 ± 0.173	0.706 ± 0.189	0.621 ± 0.180	1.188 ± 0.171	1.119 ± 0.143
DECA (TOG'21)	0.975 ± 0.131	1.154 ± 0.178	1.097 ± 0.176	1.005 ± 0.207	0.977 ± 0.111	1.196 ± 0.176	1.356 ± 0.105
ROGUE (Ours)	$\mathbf{0.786 \pm 0.138}$	$\mathbf{0.943 \pm 0.187}$	$\mathbf{0.918 \pm 0.171}$	$\mathbf{0.682 \pm 0.189}$	$\mathbf{0.603 \pm 0.176}$	$\mathbf{1.025 \pm 0.168}$	$\mathbf{1.052 \pm 0.140}$

Table 2: A quantitative comparison of the perceptual error metric with other approaches on the proposed *ReaChOcc dataset*, where the error numbers are the lower the better.

2) Synthetic Challenging Occlusions: To further investigate the impact of occlusions, we build the *synthetic occlusion* set SynChOcc by overlaying natural occlusions at the specific spatial location on clean face images, leading to much more challenging occlusion conditions than the real-world ones. Synthetic occlusions are hard to be tackled. Thus, we present comprehensive comparisons of the proposed method with the other recent approaches in Table 3. Our quantitative results show better perceptual accuracy for the reconstructed 3D face than

R-Net and MoFA. Our proposed method reduces perceptual distance between vectors obtained from VGG-Face, FaceNet, FaceNet-512, OpenFace, DeepFace, ArcFace, and SFace by a large margin of **15**.**8**% (from 0.952 to 0.802), **23**.**0**% (from 1.157 to 0.891), **26**.**4**% (from 1.195 to 0.879), **2**.**6**% (from 0.894 to 0.871), **0**.**8**% (from 0.854 to 0.847), **23**.**4**% (from 1.284 to 0.983), **18**.**1**% (from 1.241 to 1.016), respectively, compared to MoFA. In addition, the proposed method reduces **2**.**7**% (from 0.824 to 0.802), **21**.**55**% (from 0.968 to 0.891), **8**.**0**% (from 0.955 to 0.879), **11**.**3**% (from 0.880 to 0.871), **1**.**0**% (from 0.860 to 0.847), **13**.**1**% (from 1.131 to 0.983), **12**.**6**% (from 1.162 to 1.016) of the perceptual distance between vectors obtained using VGG-Face, FaceNet, FaceNet-512, OpenFace, DeepFace, ArcFace, and SFace, respectively. Furthermore, our approach shows an improvement of **4**.**2**% (from 0.837 to 0.802), **11**.**1**% (from 1.002 to 0.891), **7**.**6**% (from 0.951 to 0.879), **3**.**9**% (from 0.906 to 0.871), **7**.**6**% (from 0.917 to 0.847), **7**.**4**% (from 1.061 to 0.983), **19**.**5**% (from 1.262 to 1.016) in the perceptual distance between vectors obtained using VGG-Face, FaceNet, FaceNet-512, OpenFace, DeepFace, ArcFace, and SFace, respectively than DECA.

	Perceptual Error (↓)						
Methods	VGG-Face	FaceNet	FaceNet-512	OpenFace	DeepFace	ArcFace	SFace
MoFA (TPAMI'18)	0.952 ± 0.130	1.157 ± 0.133	1.195 ± 0.126	0.894 ± 0.196	0.854 ± 0.165	1.284 ± 0.150	1.241 ± 0.129
R-Net (CVPRW'19)	0.824 ± 0.155	0.968 ± 0.186	0.955 ± 0.187	0.880 ± 0.195	0.860 ± 0.165	1.131 ± 0.194	1.162 ± 0.170
DECA (TOG'21)	0.837 ± 0.144	1.002 ± 0.197	0.951 ± 0.184	0.906 ± 0.201	0.917 ± 0.162	1.061 ± 0.210	1.262 ± 0.166
ROGUE (Ours)	$\mathbf{0.802 \pm 0.151}$	$\mathbf{0.891 \pm 0.179}$	$\mathbf{0.879 \pm 0.174}$	$\mathbf{0.871 \pm 0.195}$	$\mathbf{0.847 \pm 0.165}$	$\mathbf{0.983 \pm 0.186}$	$\mathbf{1.016 \pm 0.168}$

Table 3: A quantitative comparison of the perceptual distance using perceptual error with other approaches on the proposed *SynChOcc dataset*, where the error numbers are the lower the better.

3) Noisy Faces: Finally, we investigate the case of noisy face images by introducing various types of noise such as speckle, salt, pepper, Gaussian, etc. The quantitative results are shown in Table 4. Our proposed method reduces perceptual distance between vectors obtained from VGG-Face, FaceNet, FaceNet-512, OpenFace, DeepFace, ArcFace, and SFace by a large margin of **19**.**6**% (from 0.996 to 0.801), **22**.**9**% (from 1.265 to 0.976), **22**.**7**% (from 1.245 to 0.963), **6**.**8**% (from 0.923 to 0.860), **17**.**2**% (from 0.791 to 0.655), **18**.**6**% (from 1.250 to 1.017), and **10**.**1**% (from 1.260 to 1.133) respectively, compared to MoFA. Moreover, the proposed method reduces **15**.**3**% (from 0.946 to 0.801), **16**.**0**% (from 1.165 to 0.979), **17**.**1**% (from 1.161 to 0.963), **7**.**3**% (from 0.928 to 0.860), **20**.**7**% (from 0.826 to 0.655), **16**.**7**% (from 1.221 to 1.017), **9**.**7**% (from 1.255 to 1.133) of the perceptual distance between vectors obtained using VGG-Face, FaceNet, FaceNet-512, OpenFace, DeepFace, ArcFace, and SFace, respectively with regard to R-Net. Furthermore, our approach shows an improvement of **16**.**2**% (from 0.956 to 0.801), **17**.**6**% (from 1.188 to 0.979), **17**.**5**% (from 1.167 to 0.963), **21**.**0**% (from 1.089 to 0.860), **29**.**0**% (from 0.923 to 0.655), **13**.**1**% (from 1.170 to 1.017), **13**.**6**% (from 1.312 to 1.133) of the perceptual distance between vectors obtained using VGG-Face, FaceNet, FaceNet-512, OpenFace, DeepFace, ArcFace, and SFace, respectively with regard to DECA.

2.4 Ablation Studies

Impact of Occlusion Color and Size. A critical question also arises: *How do colors and sizes of occlusions affect the reconstructed 3D faces?* To answer this, we present a detailed study with various sizes and colors of facial occlusions across the fixed face image. In

Methods	Perceptual Error (↓)						
	VGG-Face	FaceNet	FaceNet-512	OpenFace	DeepFace	ArcFace	SFace
MoFA (TPAMI'18)	0.996 ± 0.150	1.265 ± 0.168	1.245 ± 0.171	0.923 ± 0.221	0.791 ± 0.175	1.250 ± 0.274	1.260 ± 0.112
R-Net (CVPRW'19)	0.946 ± 0.203	1.165 ± 0.250	1.161 ± 0.253	0.928 ± 0.263	0.826 ± 0.207	1.221 ± 0.217	1.255 ± 0.156
DECA (TOG'21)	0.956 ± 0.204	1.188 ± 0.262	1.167 ± 0.295	1.089 ± 0.223	0.923 ± 0.151	1.170 ± 0.298	1.312 ± 0.142
ROGUE (Ours)	**0.801 ± 0.171**	**0.979 ± 0.195**	**0.963 ± 0.185**	**0.860 ± 0.194**	**0.655 ± 0.181**	**1.017 ± 0.146**	**1.133 ± 0.121**

Table 4: A quantitative comparison of the perceptual error metric with other approaches on the proposed *noisy variant* of CelebA-test dataset, where the error numbers are the lower the better.

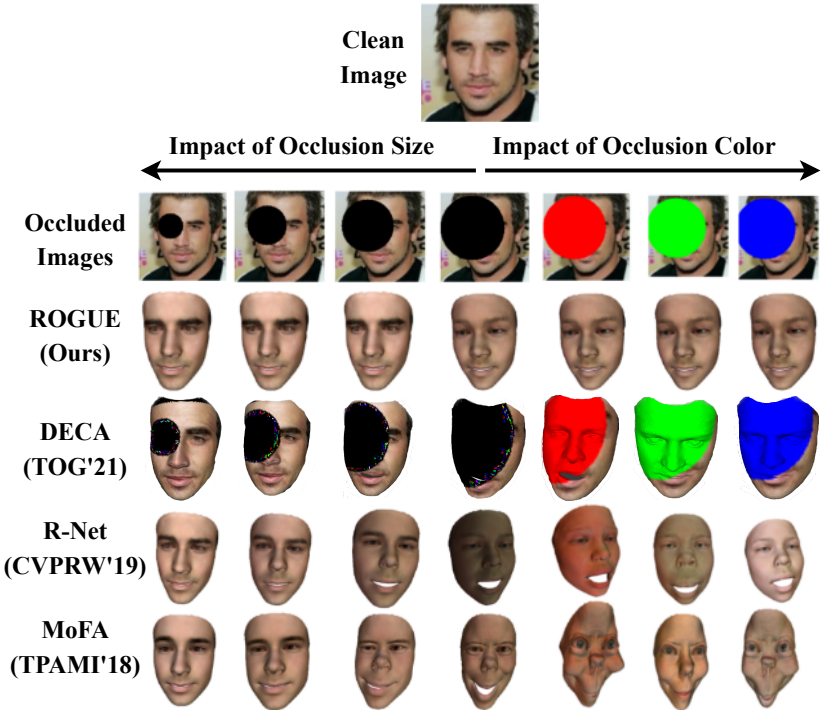

Figure 5: A qualitative comparison of the reconstructed 3D faces with other approaches for different *occlusion sizes and colors*. The results show that the proposed method carries better robustness against various colors and sizes of occlusions.

Fig. 5, it is observed that our method is highly robust to the sizes of occlusions. Moreover, the reconstructed 3D faces are barely affected by the occlusion colors, showing that the proposed model can handle heavily occluded facial regions with varying pixel values.

L2 vs. Adversarial Consistency: To show the effectiveness of our discriminator-based loss, we replace it with naive L2 loss, and the performance degrades, indicating that L2 loss is not enough to achieve consistency between clean and occluded/noisy images in terms of coefficients. The experiments are performed on three datasets: 1) ReaChOcc, 2) SynChOcc, and 3) noisy face set, as shown in Table 5, 6 and 7, respectively.

Choice of Weights for the Losses: To choose the weights associated with various losses for training the Self-Supervised Robustifying Guidance, we perform several experiments as

	Perceptual Error (↓)						
Losses	VGG-Face	FaceNet	FaceNet-512	OpenFace	DeepFace	ArcFace	SFace
L2	0.879 ± 0.132	1.135 ± 0.147	1.184 ± 0.149	0.751 ± 0.186	0.648 ± 0.185	1.283 ± 0.138	1.157 ± 0.162
Adversarial	**0.786 ± 0.138**	**0.943 ± 0.187**	**0.918 ± 0.171**	**0.682 ± 0.189**	**0.603 ± 0.176**	**1.025 ± 0.168**	**1.052 ± 0.140**

Table 5: A quantitative comparison of the perceptual error metric with other approaches on the proposed ReaChOcc dataset, where the error numbers are the lower the better.

	Perceptual Error (↓)						
Losses	VGG-Face	FaceNet	FaceNet-512	OpenFace	DeepFace	ArcFace	SFace
L2	0.895 ± 0.140	1.0753 ± 0.150	1.116 ± 0.141	0.880 ± 0.196	0.954 ± 0.159	1.227 ± 0.159	1.114 ± 0.152
Adversarial	**0.802 ± 0.151**	**0.891 ± 0.179**	**0.879 ± 0.174**	**0.871 ± 0.195**	**0.847 ± 0.165**	**0.983 ± 0.186**	**1.016 ± 0.168**

Table 6: A quantitative comparison of the perceptual error with other approaches on the proposed SynChOcc dataset, where the error numbers are the lower the better.

	Perceptual Error (↓)						
Losses	VGG-Face	FaceNet	FaceNet-512	OpenFace	DeepFace	ArcFace	SFace
L2	0.943 ± 0.157	1.197 ± 0.184	1.198 ± 0.161	0.774 ± 0.185	0.676 ± 0.180	1.239 ± 0.116	1.238 ± 0.106
Adversarial	**0.798 ± 0.173**	**0.976 ± 0.198**	**0.959 ± 0.187**	**0.706 ± 0.195**	**0.611 ± 0.181**	**1.014 ± 0.148**	**1.127 ± 0.123**

Table 7: A quantitative comparison of the perceptual distance using perceptual error metric with other approaches on the proposed noisy variant of CelebA-test dataset, where the error numbers are the lower the better.

shown in Table 8 and 9. In Table 8, we vary the weights β_O and β_N associated with occlusion- and noise-resistive photometric losses by fixing $\beta_C = 0.001$ corresponding to consistency loss. Besides, to obtain the best performance of the model against the consistency loss, we fix $\beta_O = \beta_N = 1.92$ and vary the weight β_C in Table 9. Based on the results in the tables, we conclude that the finest performance of the proposed model is obtained at $\beta_O = \beta_N = 1.92$ and $\beta_C = 0.001$.

		Perceptual Error (↓)		
Weights		Real	Synthetic	Noisy
0.48	0.48	1.092 ± 0.135	1.097 ± 0.162	1.159 ± 0.122
0.48	1.92	1.093 ± 0.135	1.097 ± 0.162	1.143 ± 0.122
1.62	1.62	1.071 ± 0.138	1.043 ± 0.166	1.135 ± 0.123
1.92	1.92	**1.052 ± 0.140**	**1.016 ± 0.168**	**1.127 ± 0.123**
2.22	2.22	1.097 ± 0.134	1.033 ± 0.167	1.153 ± 0.122
2.84	2.84	1.153 ± 0.164	1.45 ± 0.155	1.201 ± 0.118

Table 8: Impact of weights associated with occlusion- and noise-resistive photometric losses on our results. For this purpose, we fix $\beta_C = 10^{-3}$. The results are obtained using SFace.

Impact of Various Losses: To study the impact of losses for the occlusions, we perform a quantitative analysis on the losses for the various cases: *real-world occlusions*, *synthetic occlusions* and *noisy face* images. To answer this question, we train models with various combinations of the proposed losses and qualitatively evaluate them. We show these quantitative results in Table 10 and 11, $\mathcal{L}_O$ is insufficient to tackle real-world and synthetic occlusions. In addition, $\mathcal{L}_N$ is not effective for addressing the issue of occlusions. The combination of $\mathcal{L}_O$ and $\mathcal{L}_N$ shows no significant improvement in performance. However, the combination of

Weights	Perceptual Error (↓)		
	Real	Synthetic	Noisy
0.0001	0.148 ± 0.163	1.420 ± 0.0154	0.196 ± 0.118
0.001	**1.052 ± 0.140**	**1.016 ± 0.168**	**1.127 ± 0.123**
0.01	1.157 ± 0.161	1.125 ± 0.149	1.206 ± 0.117

Table 9: Impact of weights associated with discriminator loss on our results. For this purpose, we fix $\beta_O = \beta_N = 1.92$. The results are obtained using SFace.

the consistency loss $\mathcal{L}_C$ and $\mathcal{L}_O$ significantly improves the 3D vertex accuracy for delusional occlusions. Finally, by exploiting the losses $\mathcal{L}_O$, $\mathcal{L}_N$, and $\mathcal{L}_C$ altogether, we obtain the best performance of the proposed model. For the noisy image, Table 12 shows that the loss $\mathcal{L}_O$ does not contribute towards the improvement in the model performance. Besides, $\mathcal{L}_N$ alone is insufficient to tackle the face image noise. An improvement is observed in exploiting $\mathcal{L}_C$ for training the proposed model. The results demonstrate that the best model performance for noisy inputs can be obtained by either deploying a combination of $\mathcal{L}_C$ and $\mathcal{L}_N$ or using all three losses. We conjecture that the loss $\mathcal{L}_O$ is dedicated to addressing the issue of occlusions. Thus, the improvement is not significant when all three losses are used as compared to the grouping of $\mathcal{L}_N$ and $\mathcal{L}_C$. However, their cumulative usage is crucial for real-life scenarios as the face images contain both occlusions and noise.

Losses			Perceptual Error (↓)						
$\mathcal{L}_O$	$\mathcal{L}_N$	$\mathcal{L}_C$	VGG-Face	FaceNet	FaceNet-512	OpenFace	DeepFace	ArcFace	SFace
			0.952 ± 0.129	1.202 ± 0.146	1.237 ± 0.141	0.769 ± 0.189	0.656 ± 0.187	1.313 ± 0.114	1.215 ± 0.119
✓			0.874 ± 0.134	1.116 ± 0.153	1.177 ± 0.151	0.746 ± 0.189	0.641 ± 0.187	1.278 ± 0.129	1.151 ± 0.126
	✓		0.952 ± 0.129	1.202 ± 0.146	1.237 ± 0.141	0.769 ± 0.189	0.656 ± 0.187	1.313 ± 0.114	1.215 ± 0.119
		✓	0.813 ± 0.136	1.034 ± 0.178	1.015 ± 0.163	0.703 ± 0.189	0.624 ± 0.181	1.243 ± 0.159	1.097 ± 0.134
✓	✓		0.874 ± 0.134	1.116 ± 0.153	1.177 ± 0.151	0.746 ± 0.189	0.641 ± 0.187	1.278 ± 0.129	1.151 ± 0.126
✓		✓	**0.786 ± 0.138**	**0.943 ± 0.187**	**0.918 ± 0.171**	**0.682 ± 0.189**	**0.603 ± 0.176**	**1.025 ± 0.168**	**1.052 ± 0.140**
	✓	✓	0.813 ± 0.136	1.034 ± 0.178	1.015 ± 0.163	0.703 ± 0.189	0.624 ± 0.181	1.243 ± 0.159	1.097 ± 0.134
✓	✓	✓	**0.786 ± 0.138**	**0.943 ± 0.187**	**0.918 ± 0.171**	**0.682 ± 0.189**	**0.603 ± 0.176**	**1.025 ± 0.168**	**1.052 ± 0.140**

Table 10: A quantitative comparison of the perceptual error with other approaches on the proposed ReaChOcc dataset, where the error numbers are the lower the better.

Losses			Perceptual Error (↓)						
$\mathcal{L}_O$	$\mathcal{L}_N$	$\mathcal{L}_C$	VGG-Face	FaceNet	FaceNet-512	OpenFace	DeepFace	ArcFace	SFace
			0.952 ± 0.130	1.157 ± 0.133	1.195 ± 0.126	0.894 ± 0.196	0.854 ± 0.165	1.284 ± 0.150	1.241 ± 0.129
✓			0.894 ± 0.140	1.065 ± 0.152	1.056 ± 0.147	0.880 ± 0.196	0.851 ± 0.165	1.167 ± 0.164	1.143 ± 0.154
	✓		0.952 ± 0.130	1.157 ± 0.133	1.195 ± 0.126	0.894 ± 0.196	0.854 ± 0.165	1.284 ± 0.150	1.241 ± 0.129
		✓	0.823 ± 0.146	0.936 ± 0.161	0.978 ± 0.159	0.879 ± 0.196	0.849 ± 0.165	1.093 ± 0.179	1.098 ± 0.161
✓	✓		0.894 ± 0.140	1.065 ± 0.152	1.056 ± 0.147	0.880 ± 0.196	0.851 ± 0.165	1.167 ± 0.164	1.143 ± 0.154
✓		✓	**0.802 ± 0.151**	**0.891 ± 0.179**	**0.879 ± 0.174**	**0.871 ± 0.195**	**0.847 ± 0.165**	**0.983 ± 0.186**	**1.016 ± 0.168**
	✓	✓	0.823 ± 0.146	0.936 ± 0.161	0.978 ± 0.159	0.879 ± 0.196	0.849 ± 0.165	1.093 ± 0.179	1.098 ± 0.161
✓	✓	✓	**0.802 ± 0.151**	**0.891 ± 0.179**	**0.879 ± 0.174**	**0.871 ± 0.195**	**0.847 ± 0.165**	**0.983 ± 0.186**	**1.016 ± 0.168**

Table 11: A quantitative comparison of the perceptual error with other approaches on the proposed SynChOcc dataset, where the error numbers are the lower the better.

Losses			Perceptual Error (↓)						
$\mathcal{L}_O$	$\mathcal{L}_N$	$\mathcal{L}_C$	VGG-Face	FaceNet	FaceNet-512	OpenFace	DeepFace	ArcFace	SFace
			0.934±0.151	1.209±0.187	1.203±0.159	0.799±0.183	0.689±0.178	1.283±0.107	1.256±0.115
✓			0.934±0.151	1.209±0.187	1.203±0.159	0.799±0.183	0.689±0.178	1.283±0.107	1.256±0.115
	✓		0.883±0.162	1.195±0.188	1.140±0.175	0.768±0.188	0.671±0.180	1.191±0.123	1.198±0.118
		✓	0.809±0.168	1.096±0.194	1.074±0.183	0.741±0.191	0.637±0.180	1.085±0.136	1.153±0.121
✓	✓		0.883±0.162	1.195±0.188	1.140±0.175	0.768±0.188	0.671±0.180	1.191±1.123	1.198±0.118
✓		✓	0.809±0.168	1.096±0.194	1.074±0.183	0.741±0.191	0.637±0.180	1.085±0.136	1.153±0.121
	✓	✓	**0.798±0.173**	**0.976±0.198**	**0.959±0.187**	**0.706±0.195**	**0.611±0.181**	**1.014±0.148**	**1.127±0.123**
✓	✓	✓	**0.798±0.173**	**0.976±0.198**	**0.959±0.187**	**0.706±0.195**	**0.611±0.181**	**1.014±0.148**	**1.127±0.123**

Table 12: A quantitative comparison of the perceptual error metric with other approaches on the proposed noisy variant of CelebA-test dataset, where the error numbers are the lower the better.

3 More Discussions

3.1 R-Net and MoFA on Our Variant Datasets

One question may arise: *What if R-Net and MoFA are also trained using our occluded and noisy images?* Unfortunately, the original R-Net [9] is unsuitable to get trained on the variant datasets mainly due to the *unreliable skin masks*. Occlusions and image noise would distort the estimated skin masks, and the model may adapt to the distortions as the facial features. To make the model suitable to work with our occluded and noisy images, we exploit *clean skin masks* instead of regressing the pixel values of projected 3D faces directly on input images as in the original R-Net framework. Note that R-Net relies on the estimated skin masks to tackle minor occlusions, whereas our method can tackle heavy occlusion and noise issues without the mask dependency and tweaking tricks as mentioned above, indicating a wider usage of our ROGUE framework. The results for R-Net, re-trained on our variant datasets with clean skin masks, are much worse than ours, indicating that the proposed Robustification Pipeline can obtain better robustness than the skin mask technique, as shown in the second row of Table 13 and 14. Finally, MoFA [22] is also not designed to address heavy occlusions and noise. We re-trained MoFA with our variant datasets, and the reconstruction errors are much larger than ours, as shown in the first row of Table 13 and 14. All the results show that the superior performance of our method is not solely because of the variant data. Compared with MoFA and R-Net, our **ROGUE** framework can exploit all the variant data well, obtaining a more robust model for 3D face reconstruction.

Limitations with DECA: DECA [10] reproduces occlusions on the 3D faces instead of removing them. Therefore, re-training of DECA on our dataset holds minimal significance.

Methods	MICC		
	Cooperative	Indoor	Outdoor
MoFA	2.29±0.57	2.29±0.54	2.37±0.62
R-Net + clean mask	1.96±0.54	1.95±0.51	2.03±0.58
ROGUE (Ours)	**1.69±0.51**	**1.69±0.47**	**1.73±0.56**

Table 13: A quantitative analysis on the *Synthetic Occlusion* variant of the standard MICC dataset [2].

Methods	MICC		
	Cooperative	Indoor	Outdoor
MoFA	2.49 ± 0.65	2.49 ± 0.62	2.66 ± 0.69
R-Net + clean mask	2.09 ± 0.59	2.09 ± 0.57	2.21 ± 0.61
ROGUE (Ours)	$\mathbf{1.85 \pm 0.53}$	$\mathbf{1.85 \pm 0.50}$	$\mathbf{1.91 \pm 0.58}$

Table 14: A quantitative analysis on the *Noisy* variant of the standard MICC dataset [2].

3.2 Comparison with High-Fidelity Face Reconstruction Methods

Recent GAN-based methods [6, 11, 12, 15, 18] show good performance of reconstructing facial details for 3D faces, such as wrinkles, pores, folds, etc. These methods focus on producing detailed high-fidelity facial textures, whereas ROGUE focuses on addressing the issues of large-scale occlusions and noise for 3D face reconstruction without posing additional dependencies, which are different aspects. Moreover, the ideas of these high-fidelity reconstruction methods are complementary to our method and can be integrated into our Guidance Pipeline to improve the texture details (without costing the robustness performance because of the Robustification Pipeline), which could be our future work.

3.3 Potential Negative Societal Impact

The proposed method estimates the closest possible 3D face from the occluded and noisy face images. The technology is highly efficient and effective but is not 100% perfect. This may lead to cumbersome situations, mainly when used with face recognition systems for identifying masked criminals at the crime scene. The situation might end up in capturing an innocent person. Along with the impact mentioned above, invasion of privacy is also a concern with such a technology. The estimation of 3D facial data from occluded images might raise concern among those who do not want to reveal their identity in certain situations. The negative impacts may be compensated by boosting the model's accuracy by deploying multiple copies of the same face image occluded with different patterns enabling the model to learn the significance of the visible region for reconstructing 3D face geometry and texture.

3.4 Limitations

The proposed Self-Supervised Robustifying Guidance network addresses the issues of occlusions and noise in monocular face images for reconstructing a 3D face. However, the method requires pre-processing of the images before serving them as the input to Guidance Pipeline and Robustification Pipeline. This increases the net time required for training the proposed model. Further, the automated pre-processing may fail for several reasons, for example, the inability of the face detection model to detect the face in the image, thus posing a challenge to the proposed approach. This issue may be addressed by deploying domain adaptation or generalization techniques in which the model learns the 3D face without requiring pre-processed face data. In addition, our method currently addresses occlusion and noise issues, which cover most but not all the distortion types for faces. This issue can be addressed by extending the ROGUE framework with more distortion types, such as blur or image compression. We plan to investigate these directions in our future work.

References

[1] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6 (2):20, 2016.

[2] Andrew D. Bagdanov, Alberto Del Bimbo, and Iacopo Masi. The florence 2d/3d hybrid face dataset. In *Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding*, J-HGBU '11, page 79–80, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0998-1. doi: 10.1145/2072572.2072597. URL http://doi.acm.org/10.1145/2072572.2072597.

[3] 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, 1999.

[4] 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, 2013.

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[6] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9429–9439, 2019.

[7] Dong Chen, Gang Hua, Fang Wen, and Jian Sun. Supervised transformer network for efficient face detection. In *European Conference on Computer Vision*, pages 122–138. Springer, 2016.

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[9] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[10] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021.

[11] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1155–1164, 2019.

[12] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. *arXiv preprint arXiv:2105.07474*, 2021.

[13] Yudong Guo, Jianfei Cai, Boyi Jiang, Jianmin Zheng, et al. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1294–1307, 2018.

[14] Diederik P Kingma and Jimmy Ba. Adam: A methodfor stochastic optimization. In *International Conference onLearning Representations (ICLR)*, 2015.

[15] Jiangke Lin, Yi Yuan, Tianjia Shao, and Kun Zhou. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5891–5900, 2020.

[16] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738. IEEE Computer Society, 2015. ISBN 978-1-4673-8391-2. URL http://dblp.uni-trier.de/db/conf/iccv/iccv2015.html#LiuLWT15.

[17] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009.

[18] Jingtan Piao, Keqiang Sun, Quan Wang, Kwan-Yee Lin, and Hongsheng Li. Inverting generative adversarial renderer for face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15628, 2021.

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[20] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[21] Sefik Ilkin Serengil. tensorflow-101. https://github.com/serengil/tensorflow-101, 2021.

[22] 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 *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1274–1283, 2017.

[23] Yaoyao Zhong, Weihong Deng, Jiani Hu, Dongyue Zhao, Xian Li, and Dongchao Wen. Sface: Sigmoid-constrained hypersphere loss for robust face recognition. *IEEE Transactions on Image Processing*, 30:2587–2598, 2021.