Research Article | Vol. 29, No. 20 / 27 Sep 2021 / Optics Express 32256

Optics EXPRESS

Research on face specular removal and intrinsic decomposition based on polarization characteristics

BIN LIANG,[1] DONGDONG WENG,[1,2,*] ZIQI TU,[1] LE LUO,[1] AND JIE HAO[1]

[1]*Beijing Engineering Research Center of Mixed Reality and Advanced Display, Beijing Institute of Technology, Beijing, 100081, China*
[2]*AICFVE of Beijing Film Academy, Beijing, 100081, China*
[*]*crgj@bit.edu.cn*

Abstract: It is well known that the specular component in the face image destroys the true informantion of the original image and is detrimental to the feature extraction and subsequent processing. However, in many face image processing tasks based on Deep Learning methods, the lack of effective datasets and methods has led researchers to routinely neglect the specular removal process. To solve this problem, we formed the first high-resolution Asian Face Specular-Diffuse-Image-Material (FaceSDIM) dataset based on polarization characterisitics, which consists of real human face specular images, diffuse images, and various corresponding material maps. Secondly, we proposed a joint specular removal and intrinsic decomposition multi-task GAN to generate a de-specular image, normal map, albedo map, residue map and visibility map from a single face image, and also further verified that the prediected de-specular images have a positive enhancement effect on face intrinsic decomposition. Compared with the SOTA algorithm, our method achieves optimal performance both in corrected linear images and in uncorrected wild images of faces.

1. Introduction

The visual appearance of the facial skin is a complex phenomenon, and to describe it properly it is important to understand how the skin interacts with the light. Generally speaking, a beam of light striking the skin surface separates into reflected and refracted light. The reflected light is reflected directly and does not enter the skin, i.e., specular light. The refracted light enters the skin and exits the skin, i.e. diffuse light. Specular and diffuse light will exhibit different light propagation properties and present a different visual appearance [1]. Among them, too much specular light can lead to overexposure of the image or produce abnormal information that is detrimental to feature extraction and subsequent processing of the image.

In computer graphics, generating a highly realistic face image from the interaction between the face geometry, surface material and ambient illumination is a very intuitive physical rendering process [2,3]. In contrast, the process of using a single face image to estimate the shape, normals, reflectance, and illumination is called intrinsic decomposition [4,5].

During the face image intrinsic decomposition, two problems need to be faced: 1) the specular component in the image can obscure the original shape, color, texture, and other features of the face, which is not conducive to face intrinsic decomposition; 2) different combinations of shape, material, and illumination may produce similar appearance. Therefore, common methods attempt to solve the restricted subproblems by imposing specific priors on surface color, shape, and/or reflection properties [4–12]. In recent years, deep learning methods have shown excellent performance in face image decomposition and re-editing [13–20]. However, these methods either assume some simple and unphysical face reflection models (e.g., Lambert model) and/or

#440778 https://doi.org/10.1364/OE.440778
Journal © 2021 Received 18 Aug 2021; revised 12 Sep 2021; accepted 12 Sep 2021; published 22 Sep 2021

Optics EXPRESS

illumination models (one or more distant point light sources), or they only perform well on synthetic data and fail to achieve the expected results on real data. Moreover, considering the lack of relevant real face datasets, almost none of the current face intrinsic decomposition methods will include the specular removal as a pre-processing task.

In this article, we are committed to collecting real face images to build a high-resolution human Face Specular-Diffuse-Image-Material (FaceSDIM) database and propose a more reasonable and effective face intrinsic decomposition network based on this dataset. Therefore, our thought process is as follows:

(1) Through a self-built polarization illumination spherical acquisition system, we captured the real face specular-diffuse images under different illumination and modeled the diffuse images by the photometric stereo (PMS) algorithm [21], and finally constructed the first more physical high-resolution (2048×2048) Asian Face Specular-Diffuse-Image-Materials dataset based on polarization archarateristics across the network.

(2) Based on the FaceSDIM dataset constructed, we proposed a multi-task adversarial neural network with joint specular removal and intrinsic decomposition, which not only perfectly implements specular removal and intrinsic decomposition of linear face images, but also further verifies that the de-specular images have positive enhancement of intrinsic decomposition. Compared with the SOTA algorithm, our method achieved the best performance.

(3) Combining the Deep Inverse Tone Mapping (DITMnet) technique proposed in our previous work [22] and the method of this paper, we achieved specular removal and intrinsic decomposition of a single face wild image and verified the effectiveness of our method in practical applications.

The complete project has been posted on Github (FaceIntrinsicDecomposition).

This paper is organized as follows. Firstly, we briefly review the relevant work in Section 2 and present our acquisition device and data preprocessing in Section 3. Next, we propose our approach and architecture in Section 4, and design several experiments to compare the performance of our method with other approaches in Section 5. Finally, we conclude our work in Section 6.

2. Related work

2.1. Specular removal

In traditional algorithms, researchers set up different prior conditions to separate specular-diffuse reflections, such as piecewise constancy prior for surface color [10], diffuse reflection smoothness [11], or specular independent subspace [12], etc. In recent years, Shi et al. [19] used deep learning methods to achieve specular removal on synthetic datasets, but could not show good generalization ability on real datasets. Yi et al. [20] proposed a step-by-step training strategy, pre-training on the synthetic dataset first and then unsupervised fine-tuning on the real dataset, but the performance still did not improve significantly. In addition, Nayar et al. [23] used polarizers to achieve specular removal, while Ma et al. [24] constructed a light stage system to separate the specular component. Kampouris et al. [25] built a binary spherical gradient illumination device to achieve specular-diffuse separation. Inspired by this, we assembled a multifunctional practical polarization illumination system to achieve the specular-diffuse separation by the polarization characteristics of the light on the skin surface.

2.2. Intrinsic decomposition

At present, there is a lot of work on the problem of shape, albedo, and illumination decomposed from a single image. Barron and Malik [4] assumed a Lambert rendering model with low-frequency illumination to recover surface normal, albedo, and illumination from an image under unknown illumination. Mian [7] used a computer screen as a programmable extended light source to illuminate the face from different directions and acquire images, and proposed an

Optics EXPRESS

efficient algorithm for reconstructing the 3D face modes from three images under arbitrary illumination. Ma et al. [8] proposed a calibration method for non-isotropic point light sources, which is capable of calibrating the position and orientation of the point light source with a single image synchronously for photometric stereo reconstruction. Zhou et al. [9] proposed an adaptive scheme based on FBEEMD and detail feature fusion to decompose a face image into a sub-image with high frequency matching detail features and a sub-image with low-frequency corresponding to contour features. These methods either assume that the skin is a lambert object or strongly rely on facial geometric priors (such as blendshape models). Importantly, these methods cannot provide a high-quality, physical facial reflection estimation.

In recent years, deep learning-based methods have shown excellent results in intrinsic decomposition problems, especially in joint shape-material-lighting untangling problems. SFSnet [15] used a deep network to decompose the face image into normal, albedo and illumination, learning low-frequency variations and high-frequency details of the face from virtual and real face data. This method assumed that the face was a Lambert object, and used a simplified spherical harmonic function to simulate the illumination, which couldn't produce the realistic shadow casting effect. Unsuper3D [17] used five sub-networks to decompose an input image into depth, albedo, viewpoint, and lighting, together with a pair of confidence maps by training to reconstruct the input without external supervision. These materials had low resolution (64×64), poor accuracy, and couldn't be applied to the fields of high-precision face rendering. AvatarMe [26] used GANfit networks, super-resolution networks, De-lighting networks and BRDF inference networks to generate four high-resolution surface normal and albedo maps from a wild face image, and finally re-render a high-fidelity 3D face avatar. Learning [18] used multiple U-net cascade networks to decompose the input image into components such as human face normal, albedo, and non-diffuse residue, and used these components to render and generate face images under new illuminations. This method neglected the effect of specular component without specular removal process in image acquisition phase, data preprocessing phase, and network inference phase, which resulted in poor results.

Unlike Leanring [18], we proposed a joint specular removal and intrinsic decomposition multi-task GAN approach to generate a de-specular image and several corresponding material maps from a single face image. Such a strategy can not only effectively reconstruct the information in the specular region of the face image, but also further improve the performance of the intrinsic decomposition with the joint input image and de-specualr image predicted. In addition, we generate the first high-resolution Asian FaceIM database, which can be widely used in areas including but not limited to face recognition, re-lighting, 3D virtual human, etc.

3. Data capture and preprocessing

This section describes the polarization illumination system used to capture the specular-diffuse image of faces, including the equipment configuration, acquisition process, and data preprocessing.

3.1. Equipment configuration

The system is 2.5 meters in diameter, and it contains a set of combinable metal brackets, 156 sets of programmable illuminations, and 40 DSLR cameras, as shown in Fig. 1. Among them, the brackets present a spherical structure of subdivided icosahedron, which is used to mount light sources, cameras, and corresponding controllers. All light sources are pointed at the center of the system, and each group of light sources consists of three sub-lights, which are no polarization lamp (No.0), horizontal polarization lamp (No.1) vertical polarization lamp (No.2). There are 40 DSLR cameras mounted around the system (36 of them were used to construct the geometric mesh of the human face), and in this paper only 4 Canon EOS-1D X Mark II digital cameras with vertical polarizers in the front area were selected to capture the high-precision skin appearance.

Optics EXPRESS

Fig. 1. Our polarization illumination spherical illumination system, including brackets, cameras, and illuminations.

To calibrate the system to obtain linear images with consistent accuracy, we proposed a DITMnet to obtain linear HDR images from a single wild image for unknown cameras in the previous papers [22,27]. In this paper for known cameras, we first used a 24-color standard color plate to calibrate them precisely; then a non-metal dielectric spherical reflector (a plastic ball) was used to correct the polarizers of all sources [28,29], resulting in the cross and parallel polarization modes; finally, a 24-color standard color plate (perfect uniform opaque Lambert object) was employed to calibrate the luminosity in different polarization modes (the image value of the lambert object remained consistent in different polarization modes).

3.2. Acquisition process

In this experiment, we recorded face specular-diffuse images by OLAT (one-light-at-a-time) mode [16] under 28 light sources visible to all cameras. All images are linear raw images, and the effective resolution is 2048×2048. We recruited 27 Asian participants and got a total of 27×4×28×2 raw images. All participants had fully read and understood the informed consent form, and signed a portrait rights authorization agreement. All data is only used for non-profit academic research areas.

3.3. Data preprocessing

For all the data, we first performed the specular-diffuse separation, then calculated the face material maps under different polarization modes, and finally selected the best material maps to bulid the FaceIM database according to the re-rendering effect. The specific processing is as follows:

Step 1: According to the principle of light propagation and polarization characteristics in the skin [24,28,29], we can obtain the face mix image with specular $I_m^k = I_{parallel}^k$ under single light L_k, diffuse image $I_d^k = I_{cross}^k$, non-diffuse residual image $I_{rs}^k = I_{parallel}^k - I_{cross}^k$ for each single light $L_k, (k = 0, 1, \ldots, 27)$, and a face mask map M_m for separating the front and back scenes of the face, as shown in Fig. 2. This part of the data can be used to derive the specular-diffuse separation task of the face image.

Step 2: The face albedo A_d, A_m and normal N_d, N_m in different polarization modes were obtained by the photometric stereo (PMS) method [21], as shown in Fig. 3 (left). Compared to the performance of the maps in parallel mode, the albedo map in cross mode is closer to the real skin color and texture, and the surface normal has less abnormal white light.

Step 3: Specifying a new illumination randomly, we rendered the face re-lighting images in two polarization modes, as shown in Fig. 3 (right). Comparing with the real image, it can

Optics EXPRESS

Fig. 2. The mix images (first four rows) and diffuse images (last four rows) of a subject under different lighting sources and viewpoints, the last column is the face mask map. The mix images (parallel mode) and diffuse images (cross mode) of a participant under different illuminations L and viewpoints V, the last column is the face mask map M_m.

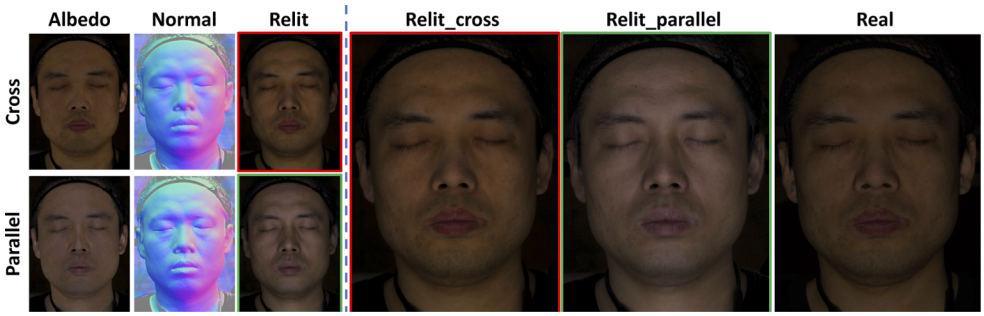

Fig. 3. The left part shows the albedo and normal map of the face in different polarization modes obtained by the PMS algorithm. The right part shows the comparison between the real face image and the relighting face images under different polarization modes.

be found that the material maps in cross mode are more effective than those in parallel mode. Learning [18] could only perform face material fitting in parallel mode because it ignored the specular separation operation and didn't match the physical properties of the skin. This part of the data is used for the task of deriving the face material maps.

In short, by building a multi-functional combined polarized spherical illumination system, we constructed a real FaceSDIM database of the form $\{L^k, V^k, I_m^k, I_d^k, M_{rs}^k, M_v^k, M_a, M_n\}$ under a single illumination. To the best of our knowledge, this is the first high-resolution Asian face dataset based on polarization characteristics, which contains richer and more physical real face specular-diffuse images and material maps.

4. Algorithm flow

This section has four subsections, which describe the rendering equations, the network architecture, the loss functions, and the FaceSDIM dataset, respectively.

Optics EXPRESS

4.1. Rendering equation

The rendering equation used is as follows in Eq. (1):

$$L_o(\vec{\omega}_o) = \int_{\vec{\omega}_i \in \Omega} f(\vec{\omega}_i, \vec{\omega}_o) L_i(\vec{\omega}_i)(\vec{n} \cdot \vec{\omega}_i) d\vec{\omega}_i,$$
$$f(\vec{\omega}_i, \vec{\omega}_o) = k_d \frac{c}{\pi} + (1 - k_d) f_{non-lambert}. \tag{1}$$

where, $L_i(\vec{\omega}_i)$ and $L_o(\vec{\omega}_o)$ represent the incident light intensity and the output light intensity, respectively. $\vec{\omega}_i$ and $\vec{\omega}_o$ represent the direction of incident and output at the surface patch x_i from the normal $\vec{n}$, $f(\vec{\omega}_i, \vec{\omega}_o)$ represents the bidirectional reflectance distribution function (BRDF) [30], k_d represents the ratio of the energy refracted in the incident light, and c represents the skin surface color.

In this article, the rendering equation of a single illumination only needs to be integrated at one point, so the equation is simplified as follows in Eq. (2):

$$L_o(\vec{\omega}_o) = \left\{ k_d \frac{c}{\pi} + (1 - k_d) f_{non-lambert} \right\} L_i(\vec{\omega}_i)(\vec{n} \cdot \vec{\omega}_i),$$
$$= \{ M_a * (M_n \cdot \vec{\omega}_i) + M_{rs}(\vec{\omega}_i, \vec{\omega}_o) \} * M_v(\vec{\omega}_i, \vec{\omega}_o) * I. \tag{2}$$

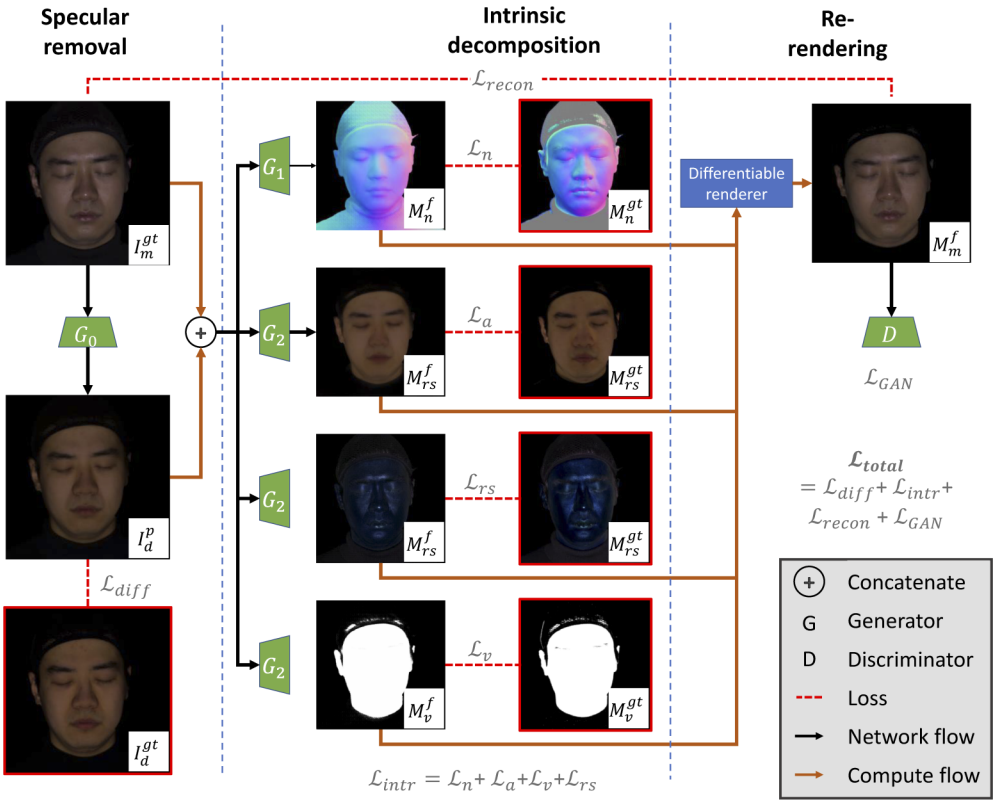

Fig. 4. The flow chart of the method of specular removal and intrinsic decomposition of the face image, including three stages of specular separation, intrinsic decomposition and re-rendering state.

Optics EXPRESS

where, M_a represents the albedo map, M_n represents the normal map, M_{rs} represents the residue map, M_v represents the visibility map, and I represents the illumination intensity.

4.2. Network architecture

Our proposed multi-task GAN method contains five U-net generators [31] and a Patch discriminator [32] to achieve specular removal and intrinsic decomposition of a single linear facial image, i.e., to obtain a specular-free diffuse image, normal map, albedo map, residue map, and visibility map, as shown in Fig. 4.

In this paper, the two cases of known and unknown illumination were modeled separately as follows:

1) Specular removal: Input a face mixed image I_m^{gt} illuminated by the illumination L. Our goal was to learn a generator G_0 to obtain a pure diffuse face image I_d^p, as shown in Eq. (3):

$$G_0 : \begin{cases} G_0\left(I_m^{gt}, L, C\right) = I_d^p, known\ L, \\ G_0\left(I_m^{gt}, C\right) = I_d^p, unknown\ L. \end{cases} \tag{3}$$

Among them, C represents coordinates in pixel space [33].

2) Intrinsic decomposition: In combination with the above predicted diffuse image, our goal was to learn a series of generators to generate various material maps, as shown in Eq. (4):

$$\begin{aligned} G_1 &: \begin{cases} G_1\left(I_m^{gt}, I_d^p, L, C\right) = M_n^f, known\ L, \\ G_1\left(I_m^{gt}, I_d^p, C\right) = M_n^f, unknown\ L, \end{cases} \\ G_2 &: \begin{cases} G_2\left(I_m^{gt}, I_d^p, L, C\right) = M_a^f, known\ L, \\ G_2\left(I_m^{gt}, I_d^p, C\right) = M_a^f, unknown\ L, \end{cases} \\ G_3 &: \begin{cases} G_3\left(M_n^f, S^f, I_d^p, L, C\right) = M_v^f, known\ L, \\ G_3\left(M_n^f, I_d^p, C\right) = M_v^f, unknown\ L, \end{cases} \\ G_4 &: \begin{cases} G_4\left(M_n^f, M_a^f, S^f, I_d^f, I_{rs\_tmp}^f, I_d^p, L, C\right) = M_{rs}^f, known\ L, \\ G_4\left(M_n^f, M_a^f, I_d^p, C\right) = M_{rs}^f, unknown\ L. \end{cases} \end{aligned} \tag{4}$$

where, $S^f = M_n^f * L$, $I_d^f = M_a^f * S^f$, $I_{rs\_tmp}^f = I_m^{gt} - I_d^f$, $I_d^r = M_a^{gt} * (M_n^{gt} \cdot L)$.

3) Reconstruction rendering: According to the rendering equations, the face relighting image I_m^f under the illumination L can be obtained, as shown in Eq. (5):

$$I_m^f = \left(I_d^f + M_{rs}^f\right) * M_v^f. \tag{5}$$

4.3. Loss functions

The loss function L_{total} in this paper is composed of the specular separation loss function L_{diff}, the intrinsic decomposition loss function L_{intr}, the reconstruction rendering loss function L_{recon}, and the generative adversarial loss function L_{GAN}, which are defined as follows in Eq. (6) and Eq. (7):

$$\mathcal{L}_{total} = \mathcal{L}_{diff} + \mathcal{L}_{intr} + \mathcal{L}_{recon} + \mathcal{L}_{GAN}. \tag{6}$$

Optics EXPRESS

where,

$$\begin{aligned}
\mathcal{L}_{diff} &= \mathcal{L}\left(I_d^p, I_d^{gt}\right), \\
\mathcal{L}_{intr} &= \mathcal{L}\left(M_n^f, M_n^{gt}\right) + \mathcal{L}\left(M_a^f, M_a^{gt}\right) + \mathcal{L}\left(M_v^f, M_v^{gt}\right) + \lambda_{rs}\mathcal{L}\left(M_{rs}^f, M_{rs}^{gt}\right), \\
\mathcal{L}_{recon} &= \mathcal{L}\left(I_d^f, I_d^r\right) + \mathcal{L}\left(I_m^f, I_m^{gt}\right), \\
\mathcal{L}_{GAN}(G, D) &= \mathbb{E}_{I_m^{gt}, I_m^f}\left[logD\left(I_m^{gt}, I_m^f\right)\right] + \mathbb{E}_{I_m^{gt}, I_m^{gt}}\left[\log\left(1 - D\left(I_m^{gt}, G\left(I_m^{gt}, I_m^{gt}\right)\right)\right)\right].
\end{aligned} \tag{7}$$

Among them, L is composed of distance loss function L_1 and perceptual loss function L_p[17], as shown in Eq. (8):

$$\mathcal{L}\left(I_{ot}, I_{gt}\right) = \lambda_1\mathcal{L}_1 + \lambda_p\mathcal{L}_p = \lambda_1\left|I_{ot} - I_{gt}\right|_1 + \lambda_p\left|I_{ot} - I_{gt}\right|_p. \tag{8}$$

where, $\left|I_{ot} - I_{gt}\right|_1 = \left|I_{ot} - I_{gt}\right|$ represents the L_1 distance loss function between the output data I_{ot} and the real data I_{gt}, $\left|I_{ot} - I_{gt}\right|_p = \left|e^k\left(I_{ot}\right) - e^k\left(I_{gt}\right)\right|$ represents the perceptual loss function generated from the *k_th* layer of the off-the-shelf image encoder e. In this article, we use the features from only one layer relu3_3 of the encoder in VGG16 [34]. Here, $\lambda_1 = 20, \lambda_p = 0.5, \lambda_{rs} = 3$.

4.4. Training databases

We first resized the original data into 512×512, preserving the global information of the face, and then croped the original data with a window of size 512 and stride 256, preserving the local information of the face. We randomly selected 23 objects (85%) as the training dataset, 2 objects (7.5%) as the validation dataset, and 2 objects (7.5%) as the test dataset. Among them, the training dataset contains about 51k sets of image sets. In addition, the flip option (up, down, left, and right) can be turned on during the training process, and the training dataset can be ×4.

5. Experiments and analysis

We used PyTorch to implement the model and deploy it on a PC device with i7 CPU, 32GB RAM and NVIDIA GTX 1080Ti GPU. The network used an Adam optimizer with momentum parameters of 0.9 and 0.999, batch size of 8, the initial learning rate is 1e-4, and the number of training times is 20 epochs. In this section, we first introduce the evaluation metrics (Section 5.1), then present the performance of our method against SOTA algorithms with and without illumination (Section 5.2), respectively, and finally extend our method to face wild images (Section 5.3), where the above experiments verify the superiority of our method.

5.1. Evaluation metrics

In this paper, seven evaluation metrics are used to evaluate the performance of the prediction results, namely Mean Square Error (MSE), Mean Absolute Error (MAE), Learned Perceptual Image Patch Similarity (LPIPS), Structural Similarity (SSIM), Multi-scale Structural Similarity (MSSSIM), Peak Signal-to-Noise Ratio (PSNR), and Mean Degree Error (MDE). Among them, MDE is only used for normal map. The lower the scores of MDE, MSE, MAE and LPIPS, the higher the scores of SSIM, MSSSIM and PSNR, the better the performance of the algorithm.

In addition, in the subsequent pictures, **D**, **N**, **A**, **Rs**, **V**, **R** respectively represent predicted diffuse image, normal map, albedo map, residue map, visibility map, and relighting mix image. **[18]** denotes the current SOTA algorithm [18] (Learning), **Our** denotes our proposed method, and **Our/S** denotes the ablation method without specular removal stage. Similar to our method, Learning [18] also recovered the normal map and albedo map from a single face linear image and generated a face relighting image under a new illumination. In particular, if the new illumination

Optics EXPRESS

is the same as the original one, then the two methods have the same goal. However, the advantages of this paper are: 1) through the self-built polarization illumination spherical acquisition system, we captured the first real human Face Specular-Diffuse Image dataset; 2) based on the material modeling of diffuse images of human faces, we constructed the first high-resolution Asian Face Specular-Diffuse Image-Material (FaceSDIM) dataset that better matches the physical properties of skin; 3) based on the FaceSDIM dataset, we proposed a joint specular removal and intrinsic decomposition multi-task GAN method, which not only perfectly implements the specular removal and intrinsic decomposition tasks from a single linear face image , but also further explored the effect of de-specular images on intrinsic decomposition.

5.2. Comparisons with SOTA

First, we independently trained three models with illumination information, Learning [18] (SOTA algorithm), Our/S (ablation experiment without specular removal), and Our, to verify the effectiveness of our method. Because of copyright issues, the authors [18] provided their codes without their datasets, so it can only retrain on our dataset. Figure 5 (left) and Table 1 (left) represent the qualitative and quantitative results of the three models with known illumination, respectively.

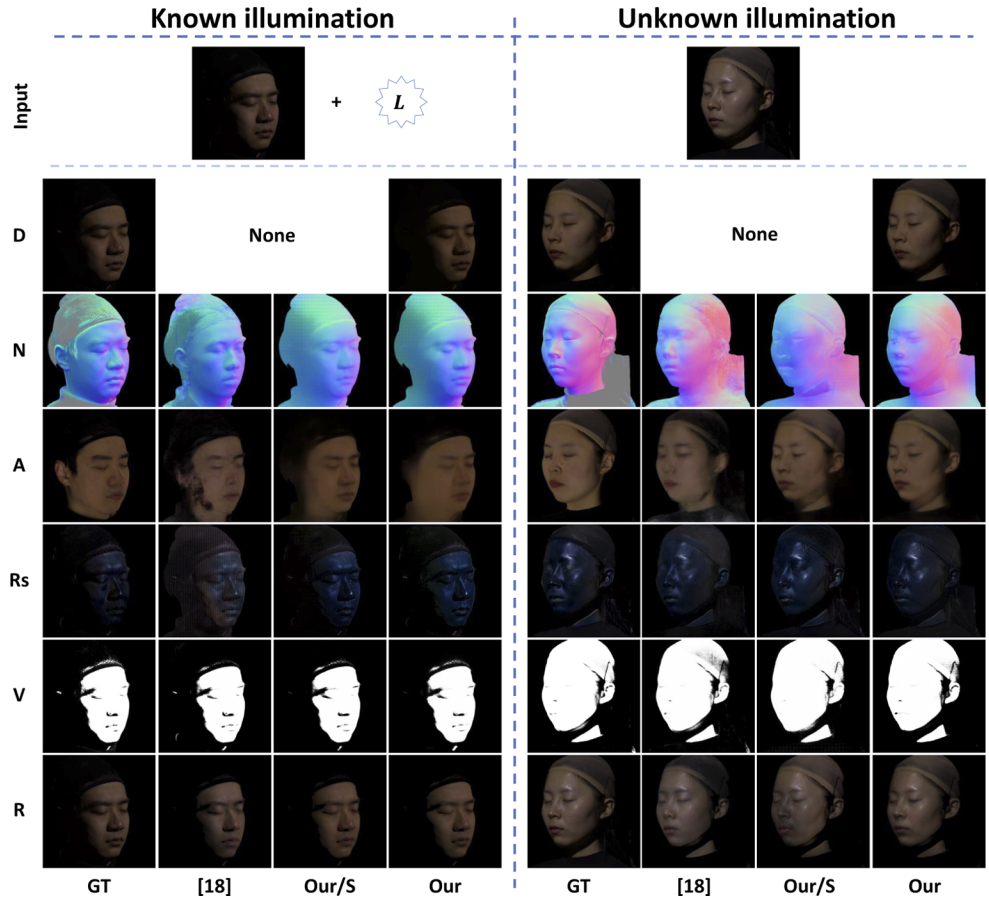

Fig. 5. The qualitative comparison of the prediction results of the three models with known (left) and unknown (right) illumination situations.

Optics EXPRESS

Table 1. The quantitative comparison of the three models, the left side indicates the results in the case of a known illumination and the right side indicates ones of an unknown illumination.

		Known illumination			Unknown illumination		
		[18]	Our/S	Our	[18]	Our/S	Our
diffuse	**MSE**	-	-	**0.0001**	-	-	**0.0001**
	MAE	-	-	**0.0046**	-	-	**0.0044**
	LPIPS	-	-	**0.0112**	-	-	**0.0101**
	SSIM	-	-	**0.9500**	-	-	**0.9753**
	MSSSIM	-	-	**0.9914**	-	-	**0.9912**
	PSNR	-	-	**41.51**	-	-	**41.19**
normal	**Degress**	27.08	**27.04**	27.07	27.09	27.23	**27.08**
	MSE	0.0133	0.0129	**0.0128**	0.0140	0.0249	**0.0129**
	MAE	0.0596	0.0599	**0.0580**	0.0621	0.0980	**0.0593**
	LPIPS	**0.1868**	0.1931	0.1891	0.1978	0.2331	**0.1891**
	SSIM	0.7495	0.7576	**0.7656**	0.7532	0.7035	**0.7560**
	MSSSIM	0.7802	0.7846	**0.7974**	0.7693	0.6457	**0.7867**
	PSNR	18.78	18.93	**18.95**	18.56	16.10	**18.90**
albedo	**MSE**	0.0028	0.0014	**0.0014**	0.0026	0.0022	**0.0021**
	MAE	0.0271	0.0177	**0.0161**	0.0274	0.0211	**0.0189**
	LPIPS	0.1150	0.1173	**0.1115**	**0.1118**	0.1315	0.1206
	SSIM	0.8082	0.8303	**0.8637**	0.7997	0.8524	**0.8604**
	MSSSIM	0.8705	0.9084	**0.9093**	0.8722	0.8663	**0.8740**
	PSNR	26.20	29.69	**29.90**	26.03	27.88	**28.20**
residue	**MSE**	0.0001	0.0001	**0.0001**	0.0001	0.0001	**0.0001**
	MAE	0.0068	0.0046	**0.0042**	0.0049	0.0049	**0.0048**
	LPIPS	0.0649	0.0372	**0.0324**	0.0481	**0.0389**	0.0410
	SSIM	0.8347	0.9301	**0.9392**	**0.9235**	0.9168	0.9218
	MSSSIM	0.9487	0.9714	**0.9743**	**0.9688**	0.9624	0.9683
	PSNR	38.96	41.79	**42.21**	**41.64**	40.81	41.53
visibility	**MSE**	**0.0029**	0.0031	0.0032	**0.0027**	0.0060	0.0048
	MAE	**0.0108**	0.0111	0.0117	**0.0112**	0.0161	0.0133
	LPIPS	0.0697	**0.0625**	0.0648	**0.0646**	0.0892	0.0726
	SSIM	**0.9366**	0.9361	0.9327	**0.9325**	0.9077	0.9281
	MSSSIM	0.9733	**0.9741**	0.9734	**0.9728**	0.9558	0.9600
	PSNR	**26.14**	25.76	25.24	**26.36**	22.71	24.15
relit	**MSE**	0.0005	0.0002	**0.0002**	0.0007	0.0009	**0.0003**
	MAE	0.0113	0.0069	**0.0068**	0.0131	0.0145	**0.0093**
	LPIPS	0.0482	0.0283	**0.0271**	0.0564	0.0581	**0.0332**
	SSIM	0.8677	**0.8877**	0.8868	0.8610	0.8618	**0.8777**
	MSSSIM	0.9612	0.9856	**0.9861**	0.9465	0.9410	**0.9754**
	PSNR	33.47	37.84	**38.08**	32.00	31.02	**35.40**

Optics EXPRESS

As shown in Fig. 5 (left), first comparing Learning [18] (second column) and Our/S (third column) shows that Learning [18] has poor generalization ability for unknown ID data, especially for normal and albedo map, and even very obvious and undesired artifacts appear in the albedo map under extreme illumination. However, Our/S ameliorates this drawback well and fully illustrates the advantages of our face dataset based on polarization-guided. Second, comparing Learning [18] and Our (fourth column) shows that our method not only solves the generalization ability of unknown ID data well, but also achieves a better inpainting effect for predicted maps under extreme illumination. This proves that the effect of our model is optimal. Finally, comparing Our/S and Our shows that the addition of the specular removal constraint can better achieve the intrinsic decomposition of face images, and the quantitative comparison in Table 1 (left) also shows this point more intuitively. This also fully validates that our proposed multi-task network is better than the single-task one, and further explores the influence of specular removal on the intrinsic decomposition task.

Then, we retrained the above three models without light source information. Figure 5 (right) and Table 1 (right) represent the qualitative and quantitative results of the three models, respectively. As can be seen from Fig. 5 (right), the prediction results of the Learning [18] are generally blurred, especially the albedo map and the re-lighting image contain obvious specular components. Compared to Learning [18], the albedo map of the Our/S method is better, but a larger range of anomalous information appears in the normal map, which leads to poorer results in the re-lighting image as well. Compared with others, our method not only perfectly achieves the specular removal of face images, but also better predicts the face material maps, so the effect of the re-lighting image is also closer to that of the real image. The statistical comparison results in Table 1 (right) also better validate the superiority of our dataset and multi-task network.

Finally, from the predicted albedo mapping in Fig. 5 we found that our method showed perfect results with high-frequency information in the case of good illumination (small shadow regions in the original image in the right side of Fig. 5), but showed reasonable results with only low-frequency information in the case of extreme illumination (large shadow regions in the original image in the left side of Fig. 5). Although other methods had worse results, this

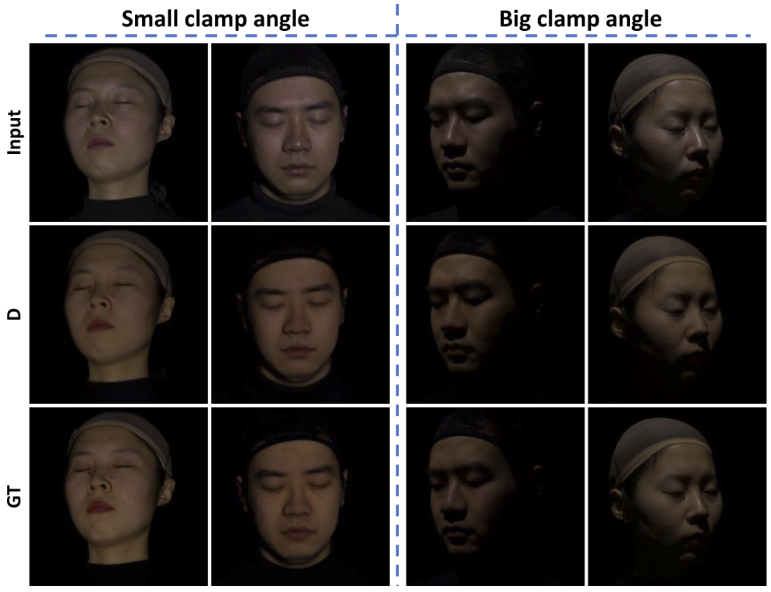

Fig. 6. The qualitative comparison of the effects of our method on specular removal in the case of small clamp angle (left) and large clamp angle (right).

Optics EXPRESS

results also illustrated the particularly strong influence of extreme illumination in face intrinsic decomposition, which is a challenge for the current study.

In addition, it can be seen from Fig. 6 that when the clamp angle between the light source direction and the viewing direction is small (left side of Fig. 6), the predicted diffuse image of our method is excellent and extremely close to the real image; when the clamp angle is large (right side of Fig. 6), the polarizers cannot form a perfect vertical cross state, and a slight specular omission phenomenon (especially the parts of the nose tip and forehead) will appear in the diffuse image of the face, while our method can better eliminate this phenomenon and enhance the robustness to polarization correction errors. This also validates the superiority of our method in real face specular removal.

In summary, with or without illumination information, the effect of our method exhibits optimal performance compared to other SOTA algorithms, and the ablation experiment Our/S validates the superiority of our dataset and multi-task strategy.

5.3. *Extended experiments based on a single face wild image*

The methods in this paper are all trained on corrected linear images, but in real scenes, the intensity values of the face image are not linearly correlated with the radiance of the scene due to camera sensors, gamma correction, and other factors. The HDR linear image can be first obtained from a single LDR wild image by using the previously proposed DITMnet method [22],

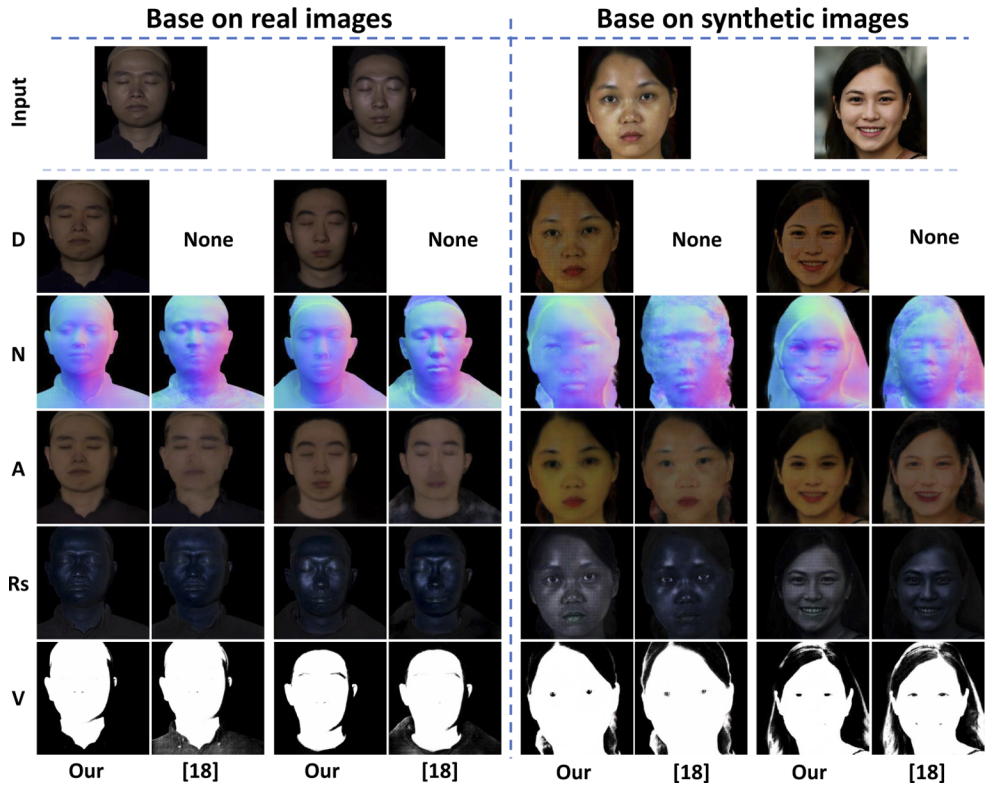

Fig. 7. Comparison of the extended application of our method with other SOTA algorithms for face wild images shotted by DSLR camera shots (left) and synthesized by the StyleGAN algorithm (right). Among them, the first row represents the face linear image prediected from the wild image by the DITMnet method in our previous paper.

Optics EXPRESS

and then the specular removal and intrinsic decomposition tasks of a single face wild image can be achieved using our method in this paper.

Figure 7 shows the performance of our method and other SOTA algorithm [18] on face wild images. In this subsection, the data on the left side are from real face images taken by DSLR cameras, and the data on the right side are synthesized by the StyleGAN algorithm [35]. The normal map and albedo map predicted by Learning [18] exhibit very obvious and undesired artifacts, especially in the face synthetic images. In contrast, our method eliminates this phenomenon very well and shows better and more credible performance. In addition, the performance of all methods in real face images is better than that of synthetic images, which is because all methods are trained in real face data and are slightly weaker for synthetic images.

Considered together, the performance of our method is optimal, both in our acquired dataset and in the wild images of the actual scenes, obtaining the best level so far.

6. Conclusion

In this paper, we first obtained the real face specular-diffuse images based on polarization-guided through a self-built multifunctional practical polarization illumination spherical acquisition system, next constructed a more physical high-resolution Asian FaceSDIM database, then propose a joint specular removal and intrinsic decomposition multi-task GAN method to generate a de-specular image, normal map, albedo map, residue map and visibility map from a single face image, and finally further extended to face wild images in practical applications. The effectiveness of our method is verified by comparing with ablation experiments and SOTA algorithms.

Limitation. First, the poor performance of our method in some extreme illumination conditions (large shadows in face images), which is a limitation of all current methods; second, limited by the performance of the equipment and the capacity of the model, our high-resolution database cannot be fully utilized; finally, limited by the size of the subjects, the accuracy of the normal map predicted is relatively poor.

Future work. On the one hand, we try to improve the network to take full advantage of our database. Secondly, the generalization and robustness of our algorithm can be imporved by adding new participants and/or suitable synthetic datasets. In addition, we will try to improve the rendering model by modeling the residue map (non-diffuse part) with a suitable specular reflection model to achieve a more physical and realistic face relighting image under a custom illumination.

Funding. the Key-Area Research and Development Program of Guangdong Province (No.2019B010149001); National Natural Science Foundation of China (No.62072036); the 111 Project (B18005).

Acknowledgments. This work was supported by the Key-Area Research and Development Program of Guangdong Province (No.2019B010149001) and the National Natural Science Foundation of China (No.62072036) and the 111 Project (B18005)

Disclosures. The authors declare that there are no conflicts of interest related to this article.

Data availability. Data underlying the results presented in this paper are not fully publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. S. A. Shafer, "Using color to separate reflection components," Color Res. Appl. **10**(4), 210–218 (1985).
2. M. Pharr, W. Jakob, and G. Humphreys, *Physically based rendering: From theory to implementation* (Morgan Kaufmann, 2016).
3. T. Igarashi, K. Nishino, and S. K. Nayar, *The appearance of human skin: A survey* (Now Publishers Inc, 2007).
4. J. T. Barron and J. Malik, "Shape, illumination, and reflectance from shading," IEEE Trans. Pattern Anal. Mach. Intell. **37**(8), 1670–1687 (2015).
5. H. Barrow, J. Tenenbaum, A. Hanson, and E. Riseman, "Recovering intrinsic scene characteristics," Comput. Vis. Syst. **2**, 2 (1978).
6. G. Oxholm and K. Nishino, "Shape and reflectance estimation in the wild," IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 376–389 (2016).

Optics EXPRESS

7. A. Mian, "Illumination invariant recognition and 3d reconstruction of faces using desktop optics," Opt. Express **19**(8), 7491–7506 (2011).
8. L. Ma, J. Liu, X. Pei, Y. Hu, and F. Sun, "Calibration of position and orientation for point light source synchronously with single image in photometric stereo," Opt. Express **27**(4), 4024–4033 (2019).
9. Y. Zhou, S.-T. Zhou, Z.-Y. Zhong, and H.-G. Li, "A de-illumination scheme for face recognition based on fast decomposition and detail feature fusion," Opt. Express **21**(9), 11294–11308 (2013).
10. G. J. Klinker, S. A. Shafer, and T. Kanade, "The measurement of highlights in color images," Int. J. Comput. Vis. **2**(1), 7–32 (1988).
11. L. Quan and H.-Y. Shum, "Highlight removal by illumination-constrained inpainting," in *Proceedings ninth ieee international conference on computer vision*, (IEEE, 2003), pp. 164–169.
12. T. Su, Y. Zhou, Y. Yu, X. Cao, and S. Du, "Illumination separation of non-lambertian scenes from a single hyperspectral image," Opt. Express **26**(20), 26167–26178 (2018).
13. Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras, "Neural face editing with intrinsic image disentangling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2017), pp. 5541–5550.
14. C. Innamorati, T. Ritschel, T. Weyrich, and N. J. Mitra, "Decomposing single images for layered photo retouching," in *Computer Graphics Forum*, vol. 36 (Wiley Online Library, 2017), pp. 15–25.
15. S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs, "Sfsnet: Learning shape, reflectance and illuminance of facesin the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2018), pp. 6296–6305.
16. T. Sun, J. T. Barron, Y.-T. Tsai, Z. Xu, X. Yu, G. Fyffe, C. Rhemann, J. Busch, P. E. Debevec, and R. Ramamoorthi, "Single image portrait relighting," ACM Trans. Graph. **38**(4), 1–12 (2019).
17. S. Wu, C. Rupprecht, and A. Vedaldi, "Unsupervised learning of probably symmetric deformable 3d objects from images in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), pp. 1–10.
18. T. Nestmeyer, J.-F. Lalonde, I. Matthews, and A. Lehrmann, "Learning physics-guided face relighting under directional light," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), pp. 5124–5133.
19. J. Shi, Y. Dong, H. Su, and S. X. Yu, "Learning non-lambertian object intrinsics across shapenet categories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), pp. 1685–1694.
20. R. Yi, C. Zhu, P. Tan, and S. Lin, "Faces as lighting probes via unsupervised deep highlight extraction," in *Proceedings of the European Conference on computer vision (ECCV)*, (2018), pp. 317–333.
21. Y. Xiong, A. Chakrabarti, R. Basri, S. J. Gortler, D. W. Jacobs, and T. Zickler, "From shading to local shape," IEEE Trans. Pattern Anal. Mach. Intell. **37**(1), 67–79 (2015).
22. B. Liang, D. Weng, Y. Bao, Z. Tu, and L. Luo, "Method for reconstructing a high dynamic range image based on a single-shot filtered low dynamic range image," Opt. Express **28**(21), 31057–31075 (2020).
23. S. K. Nayar, X.-S. Fang, and T. Boult, "Separation of reflection components using color and polarization," Int. J. Comput. Vis. **21**(3), 163–186 (1997).
24. W.-C. Ma, T. Hawkins, P. Peers, C.-F. Chabert, M. Weiss, and P. E. Debevec, "Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination," Render. Tech. **2007**, 10 (2007).
25. C. Kampouris, S. Zafeiriou, and A. Ghosh, "Diffuse-specular separation using binary spherical gradient illumination," in *EGSR (EI&I)*, (2018), pp. 1–10.
26. A. Lattas, S. Moschoglou, B. Gecer, S. Ploumpis, V. Triantafyllou, A. Ghosh, and S. Zafeiriou, "Avatarme: Realistically renderable 3d facial reconstruction in-the-wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), pp. 760–769.
27. B. Liang, D. Weng, Y. Bao, Z. Tu, and L. Luo, "Reconstructing hdr image from a single filtered ldr image base on a deep hdr merger network," in *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, (IEEE, 2019), pp. 257–258.
28. A. Ghosh, T. Hawkins, P. Peers, S. Frederiksen, and P. Debevec, "Practical modeling and acquisition of layered facial reflectance," in *ACM SIGGRAPH Asia 2008 papers*, (2008), pp. 1–10.
29. A. Ghosh, G. Fyffe, B. Tunwattanapong, J. Busch, X. Yu, and P. Debevec, "Multiview face capture using polarized spherical gradient illumination," in *Proceedings of the 2011 SIGGRAPH Asia Conference*, (2011), pp. 1–10.
30. I. G. Renhorn and G. D. Boreman, "Developing a generalized brdf model from experimental data," Opt. Express **26**(13), 17099–17114 (2018).
31. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, (Springer, 2015), pp. 234–241.
32. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2017), pp. 1125–1134.
33. R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, and J. Yosinski, "An intriguing failing of convolutional neural networks and the coordconv solution," arXiv preprint arXiv:1807.03247 (2018).
34. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F.-F. Li, "Imagenet large scale visual recognition challenge," Int. J. Comput. Vis. **115**(3), 211–252 (2015).

Optics EXPRESS

35. T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), pp. 8110–8119.