SPLiT: Single Portrait Lighting Estimation via a Tetrad of Face Intrinsics

Fan Fei#, Yean Cheng#, Yongjie Zhu, Qian Zheng, Si Li, Gang Pan, *Senior Member, IEEE*, and Boxin Shi*, *Senior Member, IEEE*

Abstract—This paper proposes a novel pipeline to estimate a non-parametric environment map with high dynamic range from a single human face image. Lighting-independent and -dependent intrinsic images of the face are first estimated separately in a cascaded network. The influence of face geometry on the two lighting-dependent intrinsics, diffuse shading and specular reflection, are further eliminated by distributing the intrinsics pixel-wise onto spherical representations using the surface normal as indices. This results in two representations simulating images of a diffuse sphere and a glossy sphere under the input scene lighting. Taking into account the distinctive nature of light sources and ambient terms, we further introduce a two-stage lighting estimator to predict both accurate and realistic lighting from these two representations. Our model is trained supervisedly on a large-scale and high-quality synthetic face image dataset. We demonstrate that our method allows accurate and detailed lighting estimation and intrinsic decomposition, outperforming state-of-the-art methods both qualitatively and quantitatively on real face images.

Index Terms—Lighting Estimation, Intrinsic Image Decomposition, Face Modeling.

1 Introduction

Lighting estimation from a single image is a significant area of interest in computer vision [1], [2], [3], [4], [5], [6], [7], [8], as it enables applications such as realistic 3D object insertion. Estimating lighting from a single, ordinary photograph of a general scene presents a highly ill-posed problem, given that infinite combinations of scene illumination and underlying objects can produce the same captured image. Nonetheless, the ubiquity and prominence of faces in real-life photographs offer an opportunity to exploit explicit priors on face geometry or reflectance properties [9], [10], [11], as well as implicit priors derived from extensive human face datasets [12], [13], [14] to address this issue. By employing these rich priors, an essential alternative to the original problem involves estimating high dynamic range (HDR) scene lighting from a single, low dynamic range (LDR) face image (*i.e.*, a portrait) [15], [16], [17], [18], [19], [20], [21], [22], [23].

A task closely related to lighting estimation is intrinsic image decomposition [24], which decomposes an image into a set of images, each representing an intrinsic characteristic and aligned with the input. Classical intrinsic image decomposition task assumes a Lambertian world and separates the input image into diffuse albedo ($\mathbf{A}$, also known as reflectance) and diffuse shading ($\mathbf{D}$) [25], [26], [27]. More recent works additionally estimate surface normal ($\mathbf{N}$) and scene lighting ($\mathbf{L}$, which is not an intrinsic component) through inverse rendering of a scene [4], [28], [29], [30]. Specular reflection ($\mathbf{S}$) has also been considered an intrinsic component and isolated in highlight removal methods [23], [31], [32]. These four intrinsic components – $\{\mathbf{A}, \mathbf{N}, \mathbf{D}, \mathbf{S}\}$ – form a *tetrad*, and can be classified as either lighting-independent ($\mathbf{A}$ and $\mathbf{N}$) or lighting-dependent ($\mathbf{D}$ and $\mathbf{S}$). In this work, we estimate this *tetrad of face intrinsics* first because separating these components could reveal crucial cues for lighting estimation from a single image [4], [7], [33].

In the context of single portrait lighting estimation, where face properties (geometry and reflectance) can be approximately estimated using classical methods (*e.g.*, 3DMM [9]), a popular solution is the *reproduction-by-rendering* (also known as analysis-by-synthesis [9]) pipeline [15], [17], [18], [20], [21], [22]. This pipeline concurrently estimates scene components (including geometry, reflectance, and lighting) necessary for re-rendering the scene by a differentiable renderer, and constrains the re-rendered image to be close to the input image to find a plausible combination of scene components that well explains the input image. For example, SfSNet [20] estimates surface normal map $\mathbf{N}$ as geometry, the diffuse albedo map $\mathbf{A}$ as reflectance, and 2^{nd}-order spherical harmonics (SH) $\mathbf{L}$ as lighting, then try to reproduce the input image $\widetilde{\mathbf{I}} = \mathrm{R}(\mathbf{A}, \mathbf{N}, \mathbf{L})$ using a differentiable renderer R under Lambertian reflectance. This kind of method relies on the backpropagation of the gradient of the reconstruction error to estimate the scene components or to train the estimator network, thus the quality of estimation heavily depends on the accuracy of the involved forward rendering process that reconstructs the input image. However, it is difficult to in-

#Fan Fei and Yean Cheng contribute equally to this work.
**Boxin Shi is the corresponding author.*

- *Fan Fei, Yean Cheng, and Boxin Shi are with National Key Laboratory for Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing 100871, China. E-mails: {feifan_eecs, shiboxin}@pku.edu.cn, cya17@stu.pku.edu.cn.*
- *Yean Cheng and Boxin Shi are also with AI Innovation Center, School of Computer Science, Peking University, Beijing 100871, China.*
- *Yongjie Zhu and Si Li are with School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China. E-mails: {zhuyongjie, lisi}@bupt.edu.cn.*
- *Qian Zheng and Gang Pan are with The State Key Lab of Brain-Machine Intelligence and College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China. Emails: {qianzheng, gpan}@zju.edu.cn.*

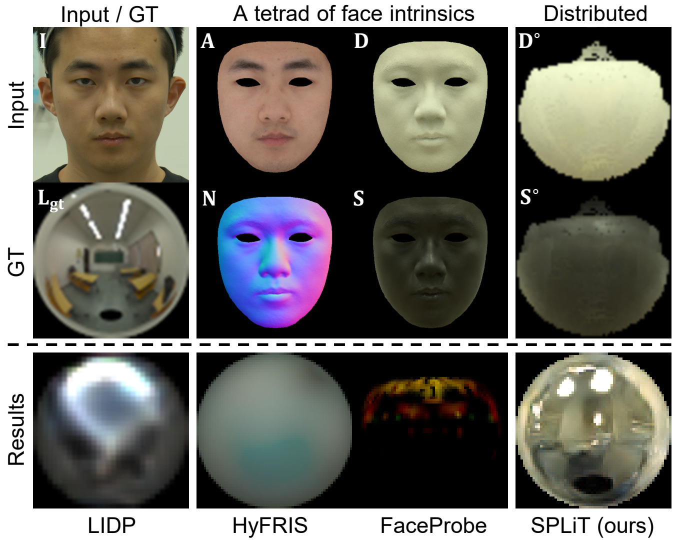


Fig. 1. SPLiT (ours) and state-of-the-art methods that use different sets of intrinsics in different manners for lighting estimation from a single portrait $\mathbf{I}$. SPLiT first estimates a tetrad of face intrinsics (diffuse albedo $\mathbf{A}$, surface normal $\mathbf{N}$, diffuse shading $\mathbf{D}$, and specular reflection $\mathbf{S}$), and uses spherically distributed $\mathbf{D}^\circ$ and $\mathbf{S}^\circ$ to estimate lighting. LIDP [34] does not involve any intrinsic. HyFRIS [22] employs $\mathbf{A}$ and $\mathbf{N}$ to estimate hybrid parametric lighting in a reproduction-by-rendering manner. FaceProbe [23] deconvolves $\mathbf{S}^\circ$ into an environment map. SPLiT produces the most accurate and realistic lighting estimation.

corporate a both differentiable and accurate (*i.e.*, physically-based) renderer. Differentiable physically-based rendering (PBR) remains a challenging problem due to the biased estimation of gradients caused by discontinuities in ray visibility [35] and the high demand of memory and computation time [36]. To make things worse, using a single image as input leaves unobserved regions to be dealt with when using PBR, which requires the full 3D reconstruction of the entire scene, including its invisible parts [5], [7]. Facing these challenges, previous reproduction-by-rendering methods sacrifice accuracy and use over-simplified renderers to reconstruct the input image, often assuming Lambertian reflectance and ignoring effects such as cast shadows [15], [17], [18], [20], [21], [22]. As a consequence, they often only estimate face intrinsics of limited quality and low-frequency scene lighting (Fig. 1, HyFRIS [22]).

Recent works circumvent the usage of a differentiable renderer and the mentioned challenges coming with it by predicting high-frequency lighting from a single portrait in an *end-to-end* manner, training "black-box" deep lighting estimators on synthetic face images [37], [38] or Light Stage-captured One-Light-At-a-Time (OLAT) images [39] with lighting labels (Fig. 1, the method proposed by LeGendre *et al.* [34], abbreviated as LIDP). However, for the end-to-end methods, an estimator may face increasing difficulty due to the enlarged dimensionality of output, as it can be easily distracted by scene components independent of lighting (*i.e.*, $\mathbf{A}, \mathbf{N}$). Another way to avoid the usage of a renderer is to trace the extracted facial highlight back to the scene and deconvolve it, which is sensitive to error and unable to produce realistic LDR textures in the environment map (Fig. 1, the method proposed by Yi *et al.* [23], abbreviated as FaceProbe). But these preceding lighting estimation methods [22], [23], [34] do not fully explore the interactions among the tetrad of face intrinsics. In Fig. 1, our method produces the most accurate and realistic lighting estimation compared to these methods, thanks to the proposed pipeline utilizing the full tetrad of face intrinsics to facilitate lighting estimation, without using an in-network renderer.

In this paper, we propose the **S**ingle **P**ortrait **Li**ghting estimation via a **T**etrad of face intrinsics (**SPLiT**). Taking as input a single LDR face image with the face region mask, we "split" the input into a tetrad of face intrinsic components under a non-Lambertian reflectance model, which we then fully leverage to estimate an HDR environment map $\mathbf{L}$, represented as an image of a mirror sphere light probe. Given the mentioned challenges faced by reproduction-by-rendering or end-to-end methods, SPLiT informs the lighting estimator by first removing irrelevant and distracting lighting-independent scene components from the input, turning the human face into standard spheres with simpler reflective characteristics (*i.e.*, diffuse only / specular only) to facilitate the subsequent lighting estimation. To achieve this, we design a cascaded network that first estimates lighting-independent $\{\mathbf{A}, \mathbf{N}\}$ and subsequently estimates lighting-dependent $\{\mathbf{D}, \mathbf{S}\}$ by incorporating $\{\mathbf{A}, \mathbf{N}\}$ into the input. The inherently constrained $\{\mathbf{A}, \mathbf{N}\}$ restricts the plausible space of $\{\mathbf{D}, \mathbf{S}\}$, based on which we can eliminate the influence of complex face geometry by distributing $\mathbf{D}$ and $\mathbf{S}$ pixel-wise onto spherical representations using the surface normal (from $\mathbf{N}$) as indices. During this process, cast shadows are largely suppressed by retaining the maximum. The distributed spherical $\mathbf{D}^\circ$ and $\mathbf{S}^\circ$ are BRDF-convolved and degraded versions of $\mathbf{L}$ and are used as input by a lighting estimator to recover $\mathbf{L}$, which should approximately reproduce $\mathbf{D}^\circ$ and $\mathbf{S}^\circ$ if used to render a diffuse and a specular sphere. This visual affinity between $\{\mathbf{D}^\circ, \mathbf{S}^\circ\}$ and $\mathbf{L}$ reduces the difficulty faced by a deep lighting estimator (such as [34]). Taking into account the distinctive nature of light sources and the ambient environment, we further incorporate a two-stage lighting estimation module that first estimates high-dynamic-range light sources and then fills in realistic textures, inspired by recent works [6], [8], [40], to qualify our estimation for rendering mirror-like objects in the object insertion application. SPLiT does not impose an in-network differentiable rendering process, thereby avoiding potential issues arising from an over-simplified rendering model [22]. The entire network is trained under supervision using a large-scale, high-quality synthetic dataset rendered on a 3D face scan dataset [13] and real HDR panoramas [1], [3]. SPLiT outperforms state-of-the-art methods employing either a reproduction-by-rendering [22] or an end-to-end [34] pipeline.

In summary, we contribute a face-to-lighting pipeline, characterized by the discriminative usage of estimated and distributed lighting-(in)dependent face intrinsics for accurate and realistic lighting inference. We build a test dataset consisting of real images of human faces with environment lighting and conduct extensive experiments to demonstrate that our method outperforms previous single-portrait lighting estimation methods both qualitatively and quantitatively.

TABLE 1
Summary of methods for lighting estimation that apply to a single portrait. Notations: input face image ($\mathbf{I}$), output lighting ($\mathbf{L}$); diffuse albedo ($\mathbf{A}$), surface normal ($\mathbf{N}$), diffuse shading ($\mathbf{D}$), and specular reflection ($\mathbf{S}$), with variants in face-like forms ($*$) or spherical forms ($*^\circ$), inferred directly ($*$) or via reproduction-by-rendering with inferred components ($\widetilde{*}$), estimated by 3DMM [9] and constraint within its parameter space ($*^{\mathrm{MM}}$) or not ($*$). Lambertian reflectance model: Yes, No, or N/A (lighting estimation module trained in an end-to-end manner). Estimated lighting representation: general or parametric outdoor environment map, 2^{nd}-order spherical harmonics (SH), or SH + pre-defined distant lights, each annotated with sizes of output parameters or image.

	Method	Lambert.	Estimated intrinsics	Usage of intrinsics in estimating $\mathbf{L}$	Lighting representation
(1)	MoFA [18]	Yes	**3:** $\mathbf{A}^{\mathrm{MM}}, \mathbf{N}^{\mathrm{MM}}, \widetilde{\mathbf{D}}(\mathbf{N}^{\mathrm{MM}}, \mathbf{L})$	Make $\widetilde{\mathbf{I}}(\mathbf{A}^{\mathrm{MM}}, \mathbf{N}^{\mathrm{MM}}, \mathbf{L})$ approach $\mathbf{I}$	2^{nd}-order SH (27)
(2)	NeuralFace [21]	Yes	**3:** $\mathbf{A}, \mathbf{N}, \widetilde{\mathbf{D}}(\mathbf{N}, \mathbf{L})$	Make $\widetilde{\mathbf{I}}(\mathbf{A}, \mathbf{N}, \mathbf{L})$ approach $\mathbf{I}$	2^{nd}-order SH (27)
(3)	MLFace [17]	Yes	**3:** $\mathbf{A}, \mathbf{N}, \widetilde{\mathbf{D}}(\mathbf{N}, \mathbf{L})$	Make $\widetilde{\mathbf{I}}(\mathbf{A}, \mathbf{N}, \mathbf{L})$ approach $\mathbf{I}$	2^{nd}-order SH (27)
(4)	SfSNet [20]	Yes	**3:** $\mathbf{A}, \mathbf{N}, \widetilde{\mathbf{D}}(\mathbf{N}, \mathbf{L})$	Make $\widetilde{\mathbf{I}}(\mathbf{A}, \mathbf{N}, \mathbf{L})$ approach $\mathbf{I}$	2^{nd}-order SH (27)
(5)	FML [19]	Yes	**3:** $\mathbf{A}, \mathbf{N}, \widetilde{\mathbf{D}}(\mathbf{N}, \mathbf{L})$	Make $\widetilde{\mathbf{I}}(\mathbf{A}, \mathbf{N}, \mathbf{L})$ approach $\mathbf{I}$	2^{nd}-order SH (27)
(6)	OA3DMM [16]	No	**2:** $\mathbf{A}^{\mathrm{MM}}, \mathbf{N}^{\mathrm{MM}}$	Make $\widetilde{\mathbf{I}}$ approach $\mathbf{I}$ on consensus points	2^{nd}-order SH (27)
(7)	FFOLP [15]	Yes	**3:** $\bar{\mathbf{A}}, \mathbf{N}^{\mathrm{MM}}, \widetilde{\mathbf{D}}(\mathbf{N}, \mathbf{L})$	Make $\widetilde{\mathbf{I}}(\bar{\mathbf{A}}, \mathbf{N}^{\mathrm{MM}}, \mathbf{L})$ approach $\mathbf{I}$	Env. map (65, outdoor)
(8)	HyFRIS [22]	No	**4:** $\mathbf{A}, \mathbf{N}, \{\widetilde{\mathbf{D}}, \widetilde{\mathbf{S}}\}(\mathbf{N}, \mathbf{L})$	Make $\widetilde{\mathbf{I}}(\mathbf{A}, \mathbf{N}, \mathbf{L})$ approach $\mathbf{I}$	SH + distant lights (27 + 66)
(9)	SIPR [41]	N/A	**0:** end-to-end inference	Not involved	Env. map (32×16)
(10)	HDRLE [37]	N/A	**0:** end-to-end inference	Not involved	Env. map (21, outdoor)
(11)	LIDP [34]	N/A	**0:** end-to-end inference	Not involved	Env. map (32×32)
(12)	FaceProbe [23]	No	**2:** $\mathbf{N}^{\mathrm{MM}}, \mathbf{S}$	Deconvolve $\mathbf{S}^\circ(\mathbf{N}^{\mathrm{MM}}, \mathbf{S})$ into $\mathbf{L}$	Env. map (unfixed)
(13)	Ours	No	**4:** $\mathbf{A}, \mathbf{N}, \mathbf{D}, \mathbf{S}$	Learn $\mathbf{L}$ from $\{\mathbf{D}^\circ, \mathbf{S}^\circ\}(\mathbf{N}, \{\mathbf{D}, \mathbf{S}\})$	Env. map (64×64)

2 Related work

Lighting estimation from general scenes. The lighting estimation problem aims at taking as input one or more LDR images with a limited field of view (LFoV) of a general scene and estimating the HDR scene lighting as an output. The estimated lighting can be in the form of a parametric representation (such as low-order spherical harmonics (SH) [42], [43], multi-lobe spherical Gaussians (SG) [4], [40], [44], sky lighting models [45], [46], and 3D parametric lights [2], [7]), non-parametric environment maps [6], [8], or coordinate-based multilayer perceptrons (MLPs) [47], [48]. We adopt as our lighting representation the non-parametric environment map for its capability of modeling general-purpose and high-frequency incoming lighting. Within the single-image input category, non-learning-based [45] and deep learning-based [3], [33], [46], [49], [50], [51] methods often rely on a low-dimensional parametric lighting model for predicting outdoor lighting. Estimating indoor scene lighting [2], [6], [8], [29], [30], [52] is generally more challenging due to diverse light sources and near-field illumination effects such as occlusions and inter-reflections from nearby objects. These effects require spatially-varying lighting representations [4], [5], [7], [53], [54]. After the introduction of Neural Radiance Fields (NeRF) [55], holistic inverse rendering of a 3D scene into geometry, reflectance, and lighting from multi-view images has also become prevalent [44], [47], [56], [57].

Our method takes a single image as input. Many recent single-image methods utilize decomposed intrinsic components (*e.g.*, surface normal, depth, albedo, roughness) of the input scene to facilitate lighting estimation for indoor [4], [5], [7], [29], [30], [54] and outdoor [30], [33] scenes. Our task is related to, but distinct from, these methods, as they focus on predicting lighting from single or multi-view images of a specific type of scene without a particular class of objects. In contrast, this paper focuses on predicting lighting from an image containing a human face that can be captured both indoors and outdoors.

Intrinsic image decomposition of face images. Since this paper focuses on face images as input, we only discuss the intrinsic image decomposition of faces here. For general intrinsic image decomposition, we refer readers to the comprehensive survey [58].

Face intrinsic components are widely used in lighting estimation [15], [17], [18], [19], [20], [21], [22], [59] and relighting [38], [41], [60], [61] methods taking as input an face image. The seminal work of 3DMM [9] enables the recovery of 3D shape and reflectance from a face image. Two intrinsic components, the diffuse albedo $\mathbf{A}$ and the surface normal $\mathbf{N}$, can be rendered from the recovered 3D mesh and pose. However, 3DMM-generated intrinsics are often inaccurate and have limited variations due to the model's capability, leading many methods [17], [20], [21], [22], [59] to refine these estimations using deep networks. Recently, the usage of 3DMM [9] has been circumvented by directly learning from labels obtained by photometric stereo (PS) methods on real data [60] or rendered alongside synthetic face images using 3D faces [38]. Many methods, usually assuming a Lambertian world, further estimate the diffuse shading $\mathbf{D}$ by predicting the scene lighting $\mathbf{L}$ and rendering shading $\widetilde{\mathbf{D}}(\mathbf{N}, \mathbf{L})$ as a function of geometry and lighting [15], [17], [18], [19], [20], [21]. Specular reflection $\mathbf{S}$ has also been separated by highlight removal approaches [23], [32] or reproduced as $\widetilde{\mathbf{S}}(\mathbf{N}, \mathbf{L})$ using the Blinn-Phong model [22]. Our method accounts for non-Lambertian reflectance and separates a tetrad of face intrinsics $\{\mathbf{A}, \mathbf{N}, \mathbf{D}, \mathbf{S}\}$ from the input image to facilitate lighting estimation without using 3DMM [9], achieved by incorporating implicit priors from large-scale and high-quality face images and intrinsic labels.

Lighting estimation from faces. We summarize recent works on the problem of estimating lighting from a single portrait in Table 1. Previous works predominantly use 3DMM [9] to obtain an initial estimation of $\{\mathbf{A}, \mathbf{N}\}$, either maintaining it within the space of 3DMM [15], [16], [18], [23] or using it as a prior to constrain predictions of deep networks [17], [20], [21], [22]. With these estimations necessary for re-rendering the face image, the lighting estimation

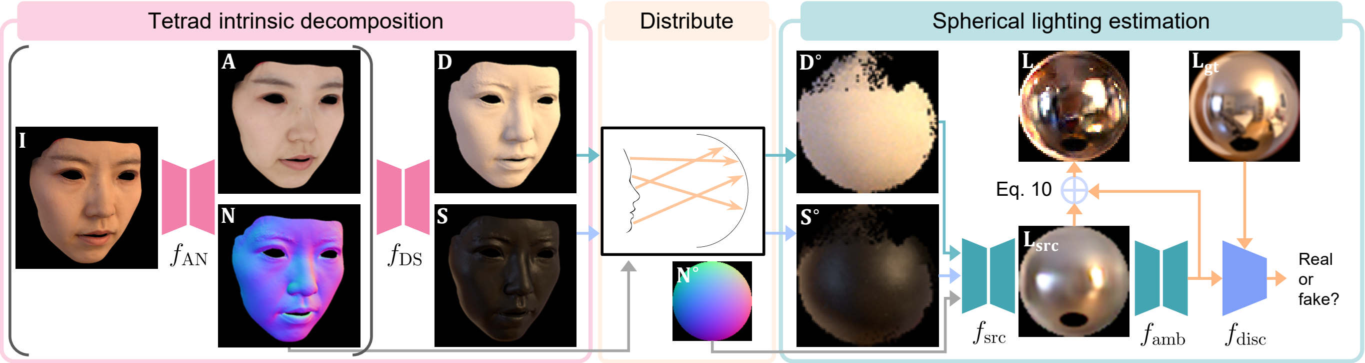


Fig. 2. The pipeline of SPLiT. The tetrad intrinsic decomposition module predicts the tetrad of face intrinsics $\{\mathbf{A}, \mathbf{N}, \mathbf{D}, \mathbf{S}\}$ from the input image $\mathbf{I}$ by cascaded networks, f_{AN} and f_{DS}. The lighting-dependent shading $\{\mathbf{D}, \mathbf{S}\}$ are distributed onto spherical representations $\{\mathbf{D}^\circ, \mathbf{S}^\circ\}$ using $\mathbf{N}$ as per-pixel indices. The spherical lighting estimation module takes $\{\mathbf{D}^\circ, \mathbf{S}^\circ\}$ and the constant normal map $\mathbf{N}^\circ$ as input. Taking into account the different nature of HDR light sources and ambient lights, it first estimates light sources and the ambient lights by the light source network f_{src}, then enriches realistic textures into the ambient environment by the ambient texture network f_{amb} and the discriminator f_{disc} via a generative adversarial network (GAN) framework.

problem could be simplified by assuming a Lambertian reflectance model with low-frequency lighting (*e.g.*, 2^{nd}-order SH) and solved in a reproduction-by-rendering framework by minimizing the difference between the rendered image and the input image [16], [17], [18], [20], [21]. However, the assumption of diffuse reflectance prevents these methods from predicting high-frequency lighting. More recent methods in this framework supplement the low-frequency lighting with a sparse basis of distant lights while adding a specular term into rendering [22], or replace the lighting with a parametric outdoor lighting model [15]. Nevertheless, the reproduction-by-rendering framework (rows 1-8) may have reached a bottleneck in dealing with more complex lighting with higher frequency due to the limited capability and accuracy of the rendering process. Efforts have been made to move away from this framework and to predict high-frequency lighting in an end-to-end approach (rows 9-11) by training networks on face images with lighting labels generated by rendering scanned face meshes [38] or relighting OLAT basis of human subjects [34], [41]. This approach achieves the state of the art [34] but risks losing interpretability. An alternative approach [23] (row 12) that retains high-frequency lighting information involves extracting facial highlights and tracing them back onto the environment map before performing color correction and deconvolution. Our SPLiT (row 13) gains advantages from both separated and simplified scene representations (*i.e.*, the spherically distributed intrinsics) and learned priors on scene components, lighting, and their correlations, thus outperforming the aforementioned approaches.

3 Proposed Method

Given an LDR input image containing a human face, the tightly cropped face image $\mathbf{I}$ and the face region mask $\mathbf{M}$ can be automatically generated by off-the-shelf face detectors [62] and semantic segmentation models [63] if not provided. As illustrated in Fig. 2, the proposed network takes $\mathbf{I}$ (whose image formation model is introduced in Sec. 3.1) and $\mathbf{M}$ as input, and consists of three parts: the tetrad intrinsic decomposition module (Sec. 3.2), the normal-indexed distributing onto spheres (Sec. 3.3), and the two-stage lighting estimation module (Sec. 3.4). Implementation details are elaborated in Sec. 3.5.

3.1 Image Formation Model

Assuming distant and uniform lighting, we adopt a non-Lambertian reflectance model and separate the reflection of light into surface reflection, which produces specular highlights, and body reflection, which produces the diffuse image. The intrinsic decomposition of the input can thus be formulated as

$$\mathbf{I} = \mathbf{A} \odot \mathbf{D} + \mathbf{S}, \tag{1}$$

where $\mathbf{I}$ is the input face image[1], $\mathbf{A}$ is the diffuse albedo that scales the diffuse shading $\mathbf{D}$ at every pixel according to the spatially-varying skin reflectance, and $\mathbf{S}$ is the specular reflection. We do not specify a BRDF model and associated parametrization here, only requiring the diffuse albedo and the division of diffuse and specular reflection to be present in the chosen BRDF, so the intrinsic components $\mathbf{A}, \mathbf{D}$, and $\mathbf{S}$ can be correctly generated by the renderer as labels used in training.

We represent lighting $\mathbf{L}(x, y)$ $(x^2 + y^2 \leq 1)$ as a mirror ball environment map, *i.e.*, an HDR orthographic capture of a mirror ball, where each position (x, y) stands for the incident radiance from the direction of the reflection vector $\mathbf{r} = \mathbf{v} - 2(\mathbf{v} \cdot \mathbf{n})\mathbf{n}$ of view vector $\mathbf{v} = (0, 0, 1)^\top$ regarding the surface normal $\mathbf{n} = (x, y, \sqrt{1 - x^2 - y^2})^\top$ of the ball. Among the two types of projection of environment maps commonly used in lighting estimation methods, the spherical projection [34], [52], [64] and the equirectangular projection [1], [3], [8], we choose the former in favor of its perceptual friendliness for qualitative evaluation [34], [52], [65], [66] and that its discontinuous edge is *in front of* the camera when projected back to 3D (in contrast to the equirectangular mapping whose vertical edge is usually placed *behind* the camera [6], [8], [40], [65], [67]), thus there

[1] We assume all input images are radiometrically calibrated (intensity linearly relates to scene radiance) throughout this paper. All images for display purposes, except normal maps, are gamma-corrected with $\gamma = 2.2$.

will not be visible vertical edges reflected by a mirror-like object in the virtual object relighting application. For a more elaborate discussion, please check the supplementary material.

3.2 Tetrad Intrinsic Decomposition Module

Since properly separated and physically meaningful representations of the input image can facilitate lighting estimation [22], [23] and enhance the interpretability of a method compared to end-to-end methods [34], we first perform intrinsic image decomposition using a cascaded module, inspired by prior works [38], [60], [68] using similar architectures for better decomposition. This module consists of two deep networks designed to estimate the tetrad of intrinsic components $\{\mathbf{A}, \mathbf{N}, \mathbf{D}, \mathbf{S}\}$ (Fig. 2, Tetrad intrinsic decomposition). The first network in the cascade, f_{AN}, takes concatenated $\mathbf{I}$ and $\mathbf{M}$, and simultaneously estimates the lighting-independent component $\{\mathbf{A}, \mathbf{N}\}$, where $\mathbf{N}$ is the unit-length surface normal map:

$$\{\mathbf{A}, \mathbf{N}\} = f_{\mathrm{AN}}(\mathbf{I}, \mathbf{M}). \tag{2}$$

The second network, f_{DS}, attempts to leverage the estimated $\{\mathbf{A}, \mathbf{N}\}$ to split apart $\{\mathbf{D}, \mathbf{S}\}$ from $\mathbf{I}$, where the intrinsic components $\mathbf{A}, \mathbf{D}$, and $\mathbf{S}$ are entangled together:

$$\{\mathbf{D}, \mathbf{S}\} = f_{\mathrm{DS}}(\mathbf{I}, \mathbf{M}, f_{\mathrm{AN}}(\mathbf{I}, \mathbf{M})). \tag{3}$$

We restrict f_{DS} to predict direct shading (both $\mathbf{D}$ and $\mathbf{S}$) since our goal of decomposition is to obtain clean representations independent of face properties for lighting estimation, rather than re-rendering realistic portraits that include effects like inter-reflection or subsurface scattering. This is achieved by generating labels of $\{\mathbf{D}, \mathbf{S}\}$ that are free of indirect reflections in our synthetic training dataset (see Sec. 4.1) and using them as targets in supervised training. Combining these two networks in a cascade, this module estimates the tetrad of intrinsics $\{\mathbf{A}, \mathbf{N}, \mathbf{D}, \mathbf{S}\}$.

Both f_{AN} and f_{DS} share a modified U-Net [69] architecture and are trained in a supervised manner. They minimize the following supervised loss terms:

$$\begin{aligned}\mathcal{L}_{\text{intrinsic}} = {} & \lambda_{\mathrm{A}}\mathcal{L}_1(\mathbf{A}, \mathbf{A}_{\mathrm{gt}}) + \lambda_{\mathrm{A_{vgg}}}\mathcal{L}_2(\mathrm{vgg}(\mathbf{A}), \mathrm{vgg}(\mathbf{A}_{\mathrm{gt}})) \\ & + \lambda_{\mathrm{N}}\mathcal{L}_1(\mathbf{N}, \mathbf{N}_{\mathrm{gt}}) + \lambda_{\mathrm{D}}\mathcal{L}_1(\mathbf{D}, \mathbf{D}_{\mathrm{gt}}) + \lambda_{\mathrm{S}}\mathcal{L}_1(\mathbf{S}, \mathbf{S}_{\mathrm{gt}}) \\ & + \lambda_{\mathrm{D_{vgg}}}\mathcal{L}_2(\mathrm{vgg}(\mathbf{D}), \mathrm{vgg}(\mathbf{D}_{\mathrm{gt}})) + \lambda_{\mathrm{rec}}\mathcal{L}_1(\mathbf{I}, \mathbf{A} \odot \mathbf{D} + \mathbf{S}),\end{aligned} \tag{4}$$

where we follow previous work [60] to compute $\mathcal{L}_2$ loss on the extracted features $\mathrm{vgg}(*)$ by a VGG-net [70] pre-trained on the ImageNet [71], and apply $\mathcal{L}_1$ supervision to the tetrad $\{\mathbf{A}, \mathbf{N}, \mathbf{D}, \mathbf{S}\}$ and the reconstructed image based on Eq. (1).

The insight behind this cascaded design is that lighting-dependent and -independent components should be treated separately. Although $\mathbf{D}$ and $\mathbf{S}$ depend on the unknown environment lighting, which can be quite arbitrary, the plausible space of reflectance $\mathbf{A}$ and geometry $\mathbf{N}$ of a human face is inherently constrained and can be learned in a data-driven manner. The predicted $\{\mathbf{A}, \mathbf{N}\}$ in turn restrict $\{\mathbf{D}, \mathbf{S}\}$ since they should approximately reproduce the input. In this manner, the color ambiguity between $\mathbf{A}$ and $\mathbf{D}$ is alleviated, although it can not be fully resolved. We further experimentally show (see Sec. 5.2) that the proposed one performs the best among decomposition pipelines with different individual networks and combinations, and can be further improved if the oracle of $\mathbf{A}$ is given which resolves the color ambiguity.

3.3 Normal-Indexed Distributing onto Sphere

The reflective properties of the face are simplified in the previous step since the estimated $\{\mathbf{D}, \mathbf{S}\}$ can be seen as captured images of an object with the same shape as the input face but with more spatially-uniform and simpler materials. However, their appearances are still determined by the interaction between the face's shape and the incident lighting. More specifically, in the absence of cast shadows, the value of a pixel in $\mathbf{D}$ equals the integral of the dot product of the surface normal and incoming lighting over the space of the visible hemisphere. That value in $\mathbf{S}$, though its precise number depends on the unknown underlying BRDF, is largely influenced by the proximity between incoming lighting direction and the reflection vector of the view vector with respect to the surface normal. The cast shadows in the input image are also fully preserved in the decomposed $\{\mathbf{D}, \mathbf{S}\}$. Thus, the dependency of $\{\mathbf{D}, \mathbf{S}\}$ on shape may complicate and impair subsequent lighting estimation while leading to data redundancy (*e.g.*, pixels with similar surface normal are very likely to have similar intensities in $\mathbf{D}$ or $\mathbf{S}$).

To exclude this undesirable dependency and to condense lighting information, inspired by the reflectance map [72], we employ the surface normal $\mathbf{N}$ to distribute lighting-dependent $\{\mathbf{D}, \mathbf{S}\}$ onto a spherical representation (Fig. 2, Sph. distribute). This process is implemented via a pixel-wise operation on the face image coordinate system. For a pixel (p, q) with surface normal $\mathbf{N}(p, q) = (x, y, z)$, the normalized position (p°, q°) after distribution onto sphere coordinate system can be computed as $(p^\circ, q^\circ) = \left(\frac{1}{2}(1-y), \frac{1}{2}(1+x)\right)$. This process is performed on discretized pixels in practice. We assign the pixel intensities $\mathbf{D}(p, q)$ and $\mathbf{S}(p, q)$ to $\mathbf{D}^\circ(p^\circ, q^\circ)$ and $\mathbf{S}^\circ(p^\circ, q^\circ)$, respectively, while simultaneously generating the occupancy mask $\mathbf{M}^\circ$ that indicates the per-pixel state of being assigned. If multiple values are assigned to the same pixel, we retain the maximum value assuming it has a higher chance to belong to a face point free of cast shadows. This operation effectively eliminates cast shadows on $\{\mathbf{D}, \mathbf{S}\}$ as with high probability at least one of all occurrences of an orientation is unaffected by cast shadows. In this way, values in the estimated $\mathbf{N}$ are used as per-pixel indices to distribute $\{\mathbf{D}, \mathbf{S}\}$. This process can be summarized as:

$$\{\mathbf{D}^\circ, \mathbf{S}^\circ, \mathbf{M}^\circ\} = \mathrm{distribute}(\mathbf{D}, \mathbf{S}, \mathbf{N}, \mathbf{M}). \tag{5}$$

By performing this normal-indexed distributing, the condensed versions of $\{\mathbf{D}, \mathbf{S}\}$ (in our setting of 512×512 faces and 64×64 spheres, the latter is 64 times smaller in size) in a spherical form $\{\mathbf{D}^\circ, \mathbf{S}^\circ\}$ are obtained, each imitating an image of a standard ball with simple reflection lit by the scene lighting $\mathbf{L}$. Thus, the difficulty of subsequent lighting estimation has been reduced to almost the same level as inferring lighting from two light probe images [52].

3.4 Spherical Lighting Estimation Module

With the intrinsic module and the distribute operation, we obtain the occupancy mask $\mathbf{M}^\circ$ and two spherical representations $\{\mathbf{D}^\circ, \mathbf{S}^\circ\}$. These representations simulate light probe images, but the probes are diffuse and glossy balls instead of the perfect mirror. Moreover, these representations may contain errors, such as skin color bleeding into diffuse shading due to imperfect decomposition ($\mathbf{A} \rightarrow \mathbf{D} \rightarrow \mathbf{D}^{\circ}$ in Fig. 2) or deviated highlight directions resulting from inaccurate surface normal estimation ($\mathbf{N} \rightarrow \mathbf{S} \rightarrow \mathbf{S}^{\circ}$). Furthermore, many pixel values are missing due to the incomplete surface normal coverage of human faces.

To predict scene lighting from these degraded versions, we introduce a two-stage lighting estimation module by first estimating light sources, then hallucinating realistic textures, which is required if our estimated lighting is eventually employed to render highly glossy objects [67]. The rationale behind this two-stage design is that light sources with high dynamic ranges and ambient lights differ significantly: Light sources, or emitters, such as sun and sky in outdoor scenes or lamps and windows in indoor scenes, provide most of the scene's energy, but often lack visible textures. On the other hand, lights from other parts of the environment are often modeled as a whole as ambient lights [2], [6]. They contribute less to the appearance of illuminated objects but contain many details that can be reflected and directly seen when rendering mirror-like objects, in which case a realistic texture is required. This ambient term is conceptually closer to a daily LDR photograph rather than to an HDR record of lighting intensity and direction. As a result, some recent lighting estimation methods [6], [40], [65] treat these two types of incident radiance in distinct ways by using a generative adversarial network (GAN) [8], [14], [73], fed with a first estimated parametric lighting, to imagine a highly detailed environment map [6], [40], [65]. Inspired by this, we employ a two-stage network to estimate the entire environment map: a light source network that estimates HDR light sources with the overall intensity of the ambient term, and an ambient texture network that enriches realistic textures in the ambient environment.

Light source network. The first network, f_{src}, aims to reason about spatial distributions and intensities of the HDR light sources in the scene and accurately recover these properties by integrating information revealed by $\mathbf{D}^{\circ}$ and $\mathbf{S}^{\circ}$) while correcting errors introduced in previous steps (potentially by using learned priors of these aberrations):

$$\mathbf{L}_{\mathrm{src}} = f_{\mathrm{src}}(\mathbf{D}^{\circ}, \mathbf{S}^{\circ}, \mathbf{N}^{\circ}, \mathbf{M}^{\circ}), \tag{6}$$

where, in addition to the spherical representations $\{\mathbf{D}^{\circ}, \mathbf{S}^{\circ}, \mathbf{M}^{\circ}\}$, we also add the constant spherical surface normal map $\mathbf{N}^{\circ}$ into the input to inform the network about the relationship between pixel position and surface orientation. The network employs an encoder-decoder architecture and is trained in a supervised manner by minimizing several photometric losses computed on rendered images under the predicted lighting $\mathbf{L}_{\mathrm{src}}$ and the GT lighting $\mathbf{L}_{\mathrm{gt}}$. We achieve this by rendering images $\mathbf{L}^{\{\mathrm{d,s}\}}$ of two spheres [52], each with reflectance properties similar to the diffuse ($\mathbf{L}^{\mathrm{d}}$) or specular ($\mathbf{L}^{\mathrm{s}}$) part of human faces. We use image-based rendering [39] by linearly combining images in the 4D reflectance field $\mathrm{R}^{\{\mathrm{d,s}\}}(\theta, \phi, x, y)$ using either lighting as weights:

$$\mathbf{L}^{\{\mathrm{d,s}\}}_{\{\mathrm{src,gt}\}}(x, y) = \sum_{\theta,\phi} \mathrm{R}^{\{\mathrm{d,s}\}}(\theta, \phi, x, y)\mathbf{L}_{\{\mathrm{src,gt}\}}(\theta, \phi). \tag{7}$$

The loss function is

$$\mathcal{L}_{\mathrm{src}} = \lambda_{\mathrm{m}}\mathcal{L}_1(\mathbf{L}_{\mathrm{src}}, \mathbf{L}_{\mathrm{gt}}) + \lambda_{\mathrm{d}}\mathcal{L}_1(\mathbf{L}^{\mathrm{d}}_{\mathrm{src}}, \mathbf{L}^{\mathrm{d}}_{\mathrm{gt}}) + \lambda_{\mathrm{s}}\mathcal{L}_1(\mathbf{L}^{\mathrm{s}}_{\mathrm{src}}, \mathbf{L}^{\mathrm{s}}_{\mathrm{gt}}). \tag{8}$$

The loss terms encourage the module to produce a lighting estimation close to $\mathbf{L}_{\mathrm{gt}}$ in the sense that their rendered light probe images are close. In practice, we find that the light source network tends to output several high-intensity light sources with a fairly uniform ambient environment. This environment map, functionally similar to a parametric representation, can produce the same facial appearance as the GT environment map because ambient details do not manifest on faces that are not as highly specular as mirrors.

Ambient texture network. The environment map estimated in the previous step suffices for the majority of situations. However, when there is a need for rendering mirror-like objects, high-frequency and realistic texture are desired instead of a uniform ambient environment [67]. We then use the second network f_{amb}, which is a conditional GAN, to enrich the detailed texture into the ambient part of the previously estimated $\mathbf{L}_{\mathrm{src}}$:

$$\mathbf{T} = f_{\mathrm{amb}}(\mathrm{tonemap}(\mathbf{L}_{\mathrm{src}})). \tag{9}$$

Here, $\mathrm{tonemap}(\cdot)$ transforms the environment map into a gamma-corrected LDR image, as seen in daily photographs. The output texture $\mathbf{T}$ is also an LDR image with light source positions aligned with the input $\mathbf{L}_{\mathrm{src}}$. This enables the network to operate on a scale close to human perception, making it easier to produce realistic images. We then replace the ambient part of the $\mathbf{L}_{\mathrm{src}}$ with the texture $\mathbf{T}$ without changing its overall intensity:

$$\mathbf{L} = \mathbf{M}_{\mathrm{src}} \odot \mathbf{L}_{\mathrm{src}} + \overline{\mathbf{M}_{\mathrm{src}}} \odot \mathrm{adjust}(\mathrm{tonemap}^{-1}(\mathbf{T})). \tag{10}$$

Here, $\mathbf{M}_{\mathrm{src}}$ is the mask of light sources (*i.e.*, pixels with values above a threshold) and $\overline{\cdot}$ is logical negation. We first inversely tonemap the texture $\mathbf{T}$ back to the linear space, then retain the overall ambient light intensity predicted by f_{src} by adjusting the colored mean intensity of $\mathrm{tonemap}^{-1}(\mathbf{T}) \odot (1 - \mathbf{M}_{\mathrm{src}})$ to the same level as $\mathbf{L}_{\mathrm{src}} \odot (1 - \mathbf{M}_{\mathrm{src}})$ using per-channel scaling coefficients.

The input of the discriminator f_{disc} is the concatenated $\mathrm{tonemap}(\mathbf{L}_{\mathrm{src}})$ and fake texture $\mathbf{T}$ (or real texture $\mathrm{tonemap}(\mathbf{L}_{\mathrm{gt}})$) to discern the consistency of spatial distribution of light sources. The generator f_{amb} and the discriminator f_{disc} are trained using losses in pix2pixHD [74]. f_{amb} tries to minimize the following losses:

$$\begin{aligned}\mathcal{L}_{\mathrm{amb}} =& \lambda_{\mathrm{adv}}\mathcal{L}_2(f_{\mathrm{disc}}(\mathrm{tonemap}(\mathbf{L}_{\mathrm{src}}), \mathbf{T}), 1) \\ &+ \lambda_{\mathrm{feat}}\mathcal{L}_1(\mathrm{feature}(f_{\mathrm{disc}}, \mathbf{T}), \mathrm{feature}(f_{\mathrm{disc}}, \mathrm{tonemap}(\mathbf{L}_{\mathrm{gt}})) \\ &+ \lambda_{\mathrm{vgg}}\mathcal{L}_2(\mathrm{vgg}(\mathbf{T}), \mathrm{vgg}(\mathrm{tonemap}(\mathbf{L}_{\mathrm{gt}}))),\end{aligned} \tag{11}$$

Here, $\mathrm{feature}(f_{\mathrm{disc}}, \cdot)$ represents intermediate features of the input texture, extracted by f_{disc}. The loss function of f_{disc} is

$$\begin{aligned}\mathcal{L}_{\mathrm{disc}} =& \mathcal{L}_2(f_{\mathrm{disc}}(\mathrm{tonemap}(\mathbf{L}_{\mathrm{src}}), \mathbf{T}), 0) \\ &+ \mathcal{L}_2(f_{\mathrm{disc}}(\mathrm{tonemap}(\mathbf{L}_{\mathrm{src}}), \mathrm{tonemap}(\mathbf{L}_{\mathrm{gt}})), 1).\end{aligned} \tag{12}$$

3.5 Implementation Details

The two sub-networks in our tetrad intrinsic decomposition module, f_{AN} and f_{DS}, share a modified U-Net [69] architecture. The light source network f_{src} is implemented as a

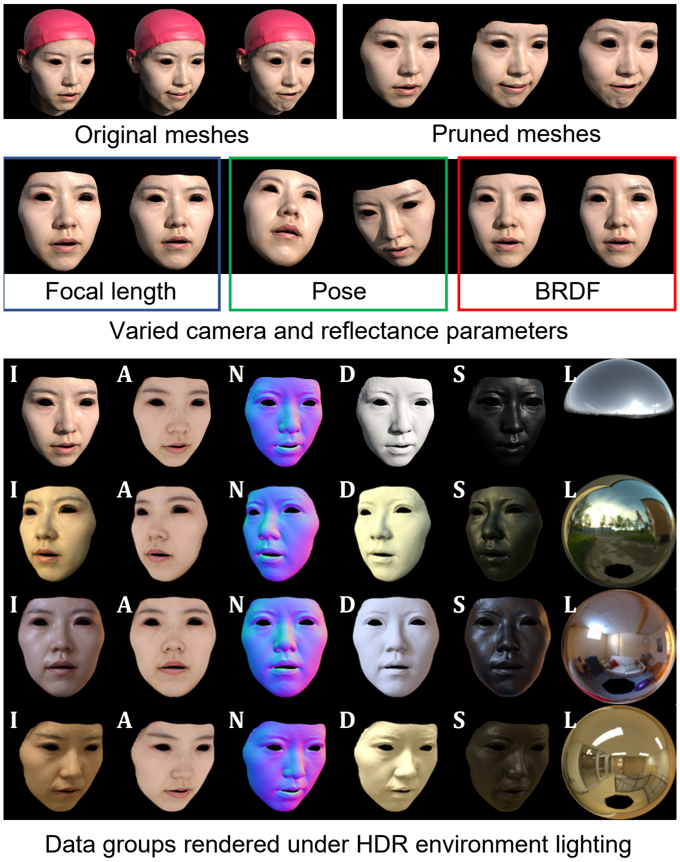

Original meshes
Pruned meshes
Focal length
Pose
BRDF
Varied camera and reflectance parameters
I
A
N
D
S
L
Data groups rendered under HDR environment lighting

Fig. 3. The generation process of our synthetic face image dataset. we first clip the face meshes in FaceScape [13] to exclude bath caps on the head and ears. The face images are then rendered using varying head poses, camera, and BRDFs. This results in our synthetic dataset with full labels of face region masks, lighting, and the tetrad of intrinsics.

T-network [37], [75] consisting of two parts: an autoencoder and an estimator. For the implementation details including weights of loss terms, network architecture diagrams, and training schemes, please check the supplementary material.

4 Dataset

In this section, we introduce our synthetic face image dataset (Sec. 4.1), our captured real face image dataset (Sec. 4.2), and an in-the-wild face image dataset (Sec. 4.3) used for training and evaluation.

4.1 Synthetic Dataset Generation

To train and validate our network for estimating face intrinsics and scene lighting without resorting to a reproduction-by-rendering approach, a face image dataset with full labels of intrinsics and lighting is needed. However, real image datasets generated from OLAT sequences captured by the expensive Light Stage [39] can hardly meet these conditions [34], [60], [76] due to difficulties in obtaining high-quality labels of diffuse shading and specular reflections. Existing synthetic datasets satisfying the requirements are either rendered using 3DMM-generated human head meshes without rich details and variations [20], [23] or using scanned head meshes but are not available in large quantities [38]. Therefore, we construct a large-scale, high-quality synthetic dataset for training (SynTrain[2]) and validation (SynValid), as illustrated in Fig. 3.

Base head mesh pre-processing. Our dataset contains rendered images of 3D face scans from the publicly available FaceScape [13] 3D face dataset[3], each consisting of a base mesh, displacement map, and texture map. We select face scans of the first 359 subjects without mosaic in the eye region (otherwise deliberately added for privacy protection of subjects as they required), using 320 for training and the remaining 39 for validation. Each subject is scanned under 20 different expressions, resulting in 6.3K meshes for SynTrain and 775 for SynValid. We clip the face meshes to exclude bath caps on the head and ears, leveraging dense correspondence between meshes for automation.

Environment map pre-processing. We use the Laval Indoor, Outdoor, and Sky HDR database [1], [3] and select 2.2K indoor and 1.2K outdoor HDR panoramas as our environment lighting. We adjust their mean intensities to the same level and spin them horizontally by random angles for each rendered image group. We use $5/6$ of these panoramas for SynTrain and the remaining for SynValid and generate the same amount of image groups under indoor and outdoor lighting.

Physically-based rendering of face images. Our synthetic dataset contains 253K groups of face images with face region masks, labels of lighting, and the tetrad of intrinsics for SynTrain and 23.2K groups in SynValid. Images are rendered using the Blender Cycles ray-tracing engine [77] and a physically-based microfacet BRDF [78] with varying specular intensity, and roughness. We randomly set the pose, specular intensity, and roughness of the face and camera focal length to improve network robustness and generalization.

Auxiliary dataset for broader skin tone. To address the limited skin tone in SynTrain, we create an additional dataset for skin tone augmentation. Although SynTrain is large-scale, high-quality, and contains GT of all intrinsics, its skin tone range is restricted due to FaceScape [13] dataset containing predominantly light to medium skin tones. This limitation could hinder our method's applicability to diverse populations. To make our method suitable for a diverse population, we generate a face image dataset using OLAT images of subjects with a broader range of skin tones, obtained from DataTang[4]. Importantly, this dataset includes a significant proportion of images featuring individuals with darker skin tones, effectively complementing the previously limited scope of light to medium skin tones. This generated dataset, comprising 30K face images with inferred pseudo-labels of $\{\mathbf{A}, \mathbf{N}\}$, is utilized to train the intrinsic decomposition module alongside SynTrain, with supervision on $\mathbf{D}$ and $\mathbf{S}$ disabled. This dataset is not used in ablation studies.

Data augmentation. During training, we augment the training sets through various transformations, including random flipping, resized cropping, random exposure and white balance adjustments, and random Gaussian noise. We ensure

[2] All dataset names are noted using small capitals throughout this paper.

[3] In this paper, we only show publishable data specified in the FaceScape [13] dataset license agreement.

[4] https://www.datatang.ai/datasets/4

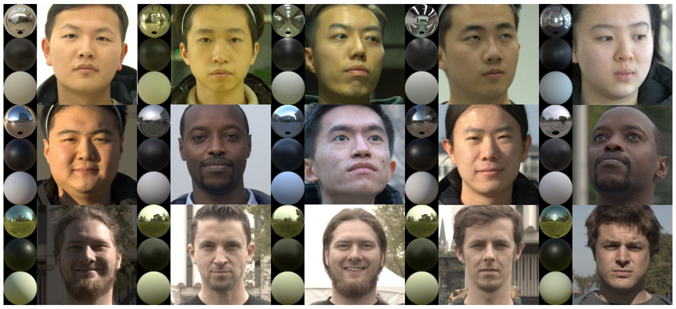


Fig. 4. Examples from our LIGHTTEST dataset. GT lighting $\mathbf{L}$ and the rendered $\mathbf{L}^{\mathrm{s}}$, $\mathbf{L}^{\mathrm{d}}$ (Eq. (7)) are shown to the left of the face images. Images are captured in indoor scenes (the first row), outdoor scenes (the second row), or directly from the LAVAL FACE&LIGHTING HDR Database [15] (the third row).

that the correctness of intrinsic labels is maintained, *e.g.*, applying the same exposure change to $\mathbf{D}$ and $\mathbf{S}$ when adjusting $\mathbf{I}$, and negating the X-axis of $\mathbf{N}$ when flipping $\mathbf{I}$ and the intrinsics.

4.2 Real Test Dataset Capture

We capture real images of human subjects with diverse skin tones and HDR environment lighting in various indoor and outdoor scenes to quantitatively evaluate different methods for lighting estimation from a single face image. These captured image-lighting pairs, combined with pairs from the LAVAL FACE&LIGHTING Database [15], form our LIGHTTEST dataset. We use a Ricoh Z1 camera with dual fish eyes to capture HDR environment maps, while the face images are captured with a Sony α7R III camera. Following the guidelines in [79], we capture the full dynamic range of the scenes using sequences of 7 LDR images taken with different exposures, each at most 3 stops apart, and a 3.0 neutral density (ND) filter to additionally capture the sun without saturation. To eliminate the color difference between the two cameras, we calibrate a 3×3 color correction matrix on their raw responses of a color panel under the same lighting. The color change introduced by the ND filter is also eliminated in a similar way. By using the color-corrected raw image and the same ISP pipeline, we obtain face-lighting pairs with consistent colors. Our captured data contains 66 pairs of 9 human subjects in 11 outdoor scenes and 84 pairs of 10 human subjects in 9 indoor scenes[5]. We also select 18 best-aligned pairs from the LAVAL FACE&LIGHTING Database [15], resulting in the LIGHTTEST dataset containing 84 outdoor and 84 indoor pairs in total. Examples from the LIGHTTEST are shown in Fig. 4.

4.3 In-the-wild Face Images

Additionally, we use face images from the FFHQ [14] dataset for qualitative evaluations only. We call this dataset INTHEWILD. Note that besides the availability of GT HDR environment lighting, another difference between INTHEWILD and LIGHTTEST is that portraits in INTHEWILD are processed by completely unknown ISPs, including tone reproduction (or camera response function), making linear images inaccessible. In this case, we assume a $\gamma = 2.2$ during tone reproduction and apply inverse gamma corrections accordingly to obtain approximately linear images as input to the methods.

5 EXPERIMENTAL RESULTS

In this section, we conduct ablation studies and comparisons among methods on datasets introduced in Sec. 4.

5.1 Evaluation Protocol and Comparing Methods

Quantitative evaluation on intrinsic decomposition. We perform an ablation study of our tetrad intrinsic decomposition module and a quantitative comparison with HyFRIS [22] using our synthetic validation dataset SYNVALID in Sec. 5.2 (Table 2). HyFRIS [22] is the state-of-the-art method for simultaneously estimating shape, reflectance, and illumination from a single portrait. We use their code to obtain their results. We employ the root mean squared error (RMSE), structural similarity (SSIM [80]), learned perceptual image patch similarity (LPIPS [71]), and mean angular error in degrees (MAngE) as error metrics. LPIPS and SSIM are computed after gamma correction, while RMSE is computed in linear space. Due to the inherent scale ambiguity between diffuse albedo $\mathbf{A}$ and diffuse shading $\mathbf{D}$ [58], we also use scale-invariant (SI-) versions of RMSE by adjusting the mean value of estimated components to match the label.

Quantitative evaluation on lighting estimation. We conduct an ablation study of our spherical lighting estimation module using SYNVALID in Sec. 5.2 (Table 3), and a quantitative comparison using the real dataset LIGHTTEST in Sec. 5.3 (Table 4). We compare our SPLiT with 4 state-of-the-art methods on single portrait lighting estimation, which include FaceProbe [23], LIDP [34], the method proposed by Sztrajman *et al.* [37] abbreviated as HDRLE, and HyFRIS [22]. For FaceProbe [23], we run their released code to obtain their results. The authors of LIDP [34] kindly provided their results on LIGHTTEST[6]. The authors of HDRLE [37] kindly provided their results on the LAVAL FACE&LIGHTING Database [15], which is part of our test dataset LIGHTTEST. Due to the lighting representation of HyFRIS [22], we use the rendered diffuse sphere under their SH lighting and the specular sphere under their 22 pre-defined distant light basis. Following prior work [34], [52], we evaluate lighting estimation by rendering 3 spheres (diffuse $\mathbf{L}^{d}$, glossy $\mathbf{L}^{s}$, and mirror $\mathbf{L}$) under the predicted lighting and computing errors on them. As in [34], the mirror sphere is clipped to $[0, 1]$ to focus on the perceivable LDR appearance of the environment map. Note that this approach allows for the evaluation of both the ambient texture of the environment map through the mirror sphere and the light source quality by comparing the other 2 spheres. We use RMSE, SI-RMSE, Fréchet Inception Distance (FID) [81], and mean angular error of the dominant light direction in degrees[7] (MAngE) as error metrics. The FID score,

[5] We have obtained the written consent for scientific usage of subjects' portraits.

[6] The results of LIDP do not contain those on the LAVAL FACE&LIGHTING [15] dataset and photos of some of our captured subjects, because we are not allowed to redistribute them.

[7] We exclude cloudy panoramas when computing the dominant light direction error of outdoor scenes.

TABLE 2
Ablation study on the design of our tetrad intrinsic decomposition module (Sec. 3.2) and comparison with HyFRIS [22] on intrinsic image decomposition performance. The metrics are computed on SYNVALID. Each "→" represents an individual network, with input on the left side and output on the right side. The best results are marked in **bold**, separately among rows with GT (only compare rows 4-6) and without GT in the input.

Pipeline/Method	Diffuse albedo $\mathbf{A}$			Normal $\mathbf{N}$	Diffuse shading $\mathbf{D}$			Specular $\mathbf{S}$
	LPIPS↓	RMSE↓	SI-RMSE↓	MAngE↓	SSIM↑	RMSE↓	SI-RMSE↓	RMSE↓
(1) $\mathbf{I} \to \mathbf{A}, \mathbf{N}$	0.0821	0.0392	0.0288	7.0325	–	–	–	–
(2) $\mathbf{I} \to \mathbf{A}$	0.0887	0.0430	0.0320	–	–	–	–	–
(3) $\mathbf{I} \to \mathbf{N}$	–	–	–	7.7027	–	–	–	–
(4) $\mathbf{I}, \mathbf{A}_{\mathrm{gt}}, \mathbf{N}_{\mathrm{gt}} \to \mathbf{D}, \mathbf{S}$	–	–	–	–	**0.9792**	0.0547	**0.0425**	0.0103
(5) $\mathbf{I}, \mathbf{A}_{\mathrm{gt}} \to \mathbf{D}, \mathbf{S}$	–	–	–	–	0.9747	**0.0532**	0.0470	**0.0095**
(6) $\mathbf{I}, \mathbf{N}_{\mathrm{gt}} \to \mathbf{D}, \mathbf{S}$	–	–	–	–	0.9707	0.1067	0.0713	0.0136
(7) $\mathbf{I} \to \mathbf{D}, \mathbf{S}$	–	–	–	–	0.9583	0.1138	0.0784	0.0148
(8) $\{\mathbf{I} \to \mathbf{A}, \mathbf{N}\} \to \mathbf{D}, \mathbf{S}$ †	0.0821	0.0392	0.0288	7.0325	0.9697	0.0882	0.0572	0.0131
(9) $\{\mathbf{I} \to \mathbf{A}, \mathbf{N}\} \to \mathbf{D}, \mathbf{S}$	**0.0800**	**0.0371**	**0.0278**	**6.9958**	**0.9715**	**0.0799**	**0.0536**	**0.0110**
(10) $\mathbf{I} \to \mathbf{A}, \mathbf{N}, \mathbf{D}, \mathbf{S}$	0.0847	0.0396	0.0292	7.5034	0.9699	0.0844	0.0580	0.0112
(11) HyFRIS [22]	0.1764	0.0957	0.0524	20.476	0.9177	0.3297	0.1249	0.0340

† Without combined fine-tuning described in Sec. 3.5.

TABLE 3
Ablation study on the inputs and design of our spherical lighting estimation module (Sec. 3.4). The first column shows the input of the light source network f_{src}. The ambient texture network f_{amb} is used only in row 6. The best results are marked in **bold**. Each entry shows measurements of indoor (left) and outdoor (right) data, separated by a "/".

Input	Environment map $\mathbf{L}$			Specular sphere $\mathbf{L}^{\mathrm{s}}$		Diffuse sphere $\mathbf{L}^{\mathrm{d}}$	
	FID↓	RMSE↓	MAngE↓	RMSE↓	SI-RMSE↓	RMSE↓	SI-RMSE↓
(1) $\mathbf{I}$	172.2/166.4	0.220/0.160	26.53/17.90	0.036/0.042	0.036/0.024	0.128/0.099	0.092/0.058
(2) $\mathbf{N}, \mathbf{D}, \mathbf{S}$	154.8/170.6	0.208/0.144	25.69/6.935	0.032/0.025	0.031/0.023	0.106/0.090	0.060/0.048
(3) $\mathbf{N}^{\circ}, \mathbf{D}^{\circ}$	157.8/158.8	0.211/0.137	25.97/5.854	0.033/0.021	0.033/0.020	0.090/0.060	0.057/0.037
(4) $\mathbf{N}^{\circ}, \mathbf{S}^{\circ}$	151.8/160.9	0.199/0.132	24.73/5.838	0.031/0.022	0.031/0.021	0.094/0.069	0.058/0.041
(5) $\mathbf{N}^{\circ}, \mathbf{D}^{\circ}, \mathbf{S}^{\circ}$	154.6/159.5	**0.197**/**0.129**	**24.26**/5.024	**0.030**/**0.019**	**0.030**/**0.018**	**0.079**/**0.055**	**0.053**/**0.034**
(6) $\mathbf{N}^{\circ}, \mathbf{D}^{\circ}, \mathbf{S}^{\circ}$ $(+f_{\mathrm{amb}})$	**98.70**/**106.4**	0.247/0.192	**24.26**/**4.946**	0.031/0.021	0.034/0.021	0.108/0.066	0.065/0.042

like LPIPS, is computed on gamma-corrected images. The LAVAL SKY HDR database [3] is not used in training f_{amb} or computing FID scores because the lower hemisphere of environment maps in this dataset is not recorded. Comparing methods do not necessarily predict the same absolute level of lighting due to different training datasets or rendering specifications, as the pixel values of lighting are not defined according to an absolute unit (*e.g.*, $\mathrm{W} \cdot \mathrm{sr}^{-1} \cdot \mathrm{m}^{-2}$) or a shared reference. For example, the mean value of unadjusted results of LIDP [34] on LIGHTTEST is $17.5\times$ on average compared to the captured lighting. Thus, for a more informative and fair comparison, we adjust the mean value of all lighting to a fixed level before qualitative or quantitative comparison on LIGHTTEST and INTHEWILD.

Qualitative evaluation. Qualitative evaluations on both intrinsic decomposition (Fig. 5) and lighting estimation (Fig. 6) are performed on datasets LIGHTTEST and INTHEWILD, which contain real portrait images.

5.2 Analysis Using Synthetic Images

We conduct a comprehensive validation of both intrinsic decomposition and lighting estimation performances on our SYNVALID dataset. Additionally, we quantitatively compare our intrinsic module against HyFRIS [22].

Regarding the intrinsic module. We aim to demonstrate the effectiveness of the cascaded design of the tetrad intrinsic decomposition module, which enhances our method's performance on both intrinsic decomposition and subsequent lighting estimation. We maintain the loss functions for each intrinsic component, the training scheme, and individual network architectures, only modifying the combination of networks and the input and output of each individual network. Our proposed design (row 9) is compared with other variants in Table 2. Row 8 is the cascade of row 1 and row 4, replacing GT in input with estimations, and row 9 is the fine-tuned version of row 8. The parameter amount of row 10 is enlarged by approximately $2\times$. We make the following observations:

1) A conjunctive estimation of $\{\mathbf{A}, \mathbf{N}\}$ outperforms separated estimations (rows 1-3). This seemingly counter-intuitive phenomenon may arise from the correlation between $\mathbf{A}$ and $\mathbf{N}$, which could guide the network to correctly attribute local discontinuities to facial features or light transfer and expand the scope of loss terms applied to a single component (*e.g.*, we observe that VGG loss applied on $\mathbf{A}$ also sharpens $\mathbf{N}$).

2) Compared to directly estimating $\{\mathbf{D}, \mathbf{S}\}$ from $\mathbf{I}$ (row 7), providing the estimated lighting-independent intrinsics $\{\mathbf{A}, \mathbf{N}\}$ (rows 9) significantly improves the performance on estimating $\{\mathbf{D}, \mathbf{S}\}$, which is more crucial for our lighting estimation task. This improvement may stem from the inherently constrained space of $\{\mathbf{A}, \mathbf{N}\}$, which is easier to learn in a data-driven manner from our SYNTRAIN dataset. This constrained space alleviates the color ambiguity between $\mathbf{A}$ and $\mathbf{D}$. Furthermore, comparing row 4 and row 9, we observe that when GT of $\mathbf{A}$ is provided, which completely resolves the ambiguity between $\mathbf{A}$ and $\mathbf{D}$, the decomposition performance can be further enhanced.

3) Comparison using different intrinsics as inputs of f_{DS} in rows 4-7 shows that directly incorporating $\mathbf{N}$ into the input may not be as beneficial as $\mathbf{A}$. This is reasonable because instead of directly appearing in Eq. (1), $\mathbf{N}$ affects the

shading in a more obscure and implicit way for the network.

As a consequence, the proposed cascade (row 9) surpasses all other practical pipelines, including the previous state-of-the-art method HyFRIS [22] (row 11), quantitatively on intrinsic decomposition.

Regarding the lighting estimation module. Our light source network f_{src} takes the constant $\mathbf{N}^{\circ}$ and spherically distributed $\{\mathbf{D}^{\circ}, \mathbf{S}^{\circ}\}$ as input. To validate the advantages of individual intrinsics and the normal-indexed distribute operation in lighting estimation, we compare the proposed pipeline with f_{src} taking different inputs in Table 3. The loss functions and architectures are identical, and all models use the fixed decoder trained on GT lighting. The difference lies in the input of f_{src}. To accommodate varying amounts and sizes of inputs, we perform average pooling on codes of different sizes before feeding them into the decoder. We make the following observations:

1) The model in row 1, which takes only the face image $\mathbf{I}$ as input (therefore is an end-to-end lighting estimator), has an increment of the same parameter amount as our intrinsic module. However, it performs the worst against other input settings, highlighting that excluding the lighting-independent $\mathbf{A}$ from $\mathbf{I}$ and decomposing it into $\{\mathbf{D}, \mathbf{S}\}$ (as in the model in row 2) could significantly facilitate lighting estimation.

2) The effect of distributing intrinsics for improving lighting estimation accuracy is also demonstrated. Compared to non-distributed inputs (row 2), the distributed components (row 5) yield better results using 64 times smaller input. This indicates that the spherical $\{\mathbf{D}^{\circ}, \mathbf{S}^{\circ}\}$ are condensed versions that are easier to learn lighting from and are more computationally efficient compared to $\{\mathbf{D}, \mathbf{S}\}$.

3) The comparison using different spherical intrinsics (rows 3-5) further demonstrates that both $\{\mathbf{D}, \mathbf{S}\}$ reveal useful clues for lighting estimation, and the model with all the $\{\mathbf{D}, \mathbf{S}\}$ as inputs (row 5) outperforms the other two models.

4) Comparing row 5 (without f_{amb}) and row 6 (with f_{amb}), we observe that although the inclusion of f_{amb} results in a slight decrease in the accuracy of $\mathbf{L}^{\mathrm{s}}$ and $\mathbf{L}^{\mathrm{d}}$, there is a significant improvement in the FID score. The worsening of RMSE on $\mathbf{L}$ is reasonable because this metric prefers an average image when estimating details is difficult. Moreover, as shown by the visual results in Fig. 6, the f_{amb} contributes to a substantial enhancement in visual quality compared to the model without f_{amb}, thus revealing its usefulness. For more qualitative evaluations regarding our ambient texture network, including comparisons of rendered mirror-like objects and using randomly picked real panoramas as textures, please check the supplementary material.

5.3 Analysis Using Real Images

Qualitative evaluation on intrinsic decomposition. On LIGHTTEST and INTHEWILD we qualitatively compare the intrinsic decomposition between our method and HyFRIS [22]. Fig. 5 shows that our model produces more reasonable and detailed predictions on all four intrinsics compared to HyFRIS [22]. Our $\mathbf{A}$ is free of highlights and the detailed $\mathbf{N}$ faithfully follows the face shape in $\mathbf{I}$. Additionally, our model can produce sharp estimations on both $\mathbf{D}$ and $\mathbf{S}$, while those from HyFRIS [22] tend to be smooth due to their low-frequency lighting representation. Nevertheless, on images from INTHEWILD (the upper two groups in Fig. 5), the diffuse reflection color sometimes bleeds into our $\mathbf{S}$. We believe this is caused by the unknown ISP of these in-the-wild images, making linear images inaccessible and breaking our assumption of linear inputs (Eq. (1)).

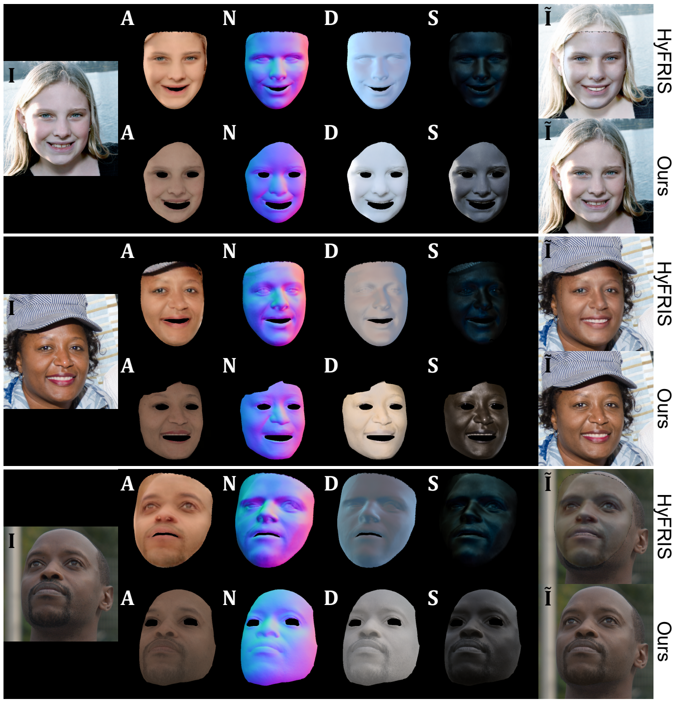


Fig. 5. Intrinsic decomposition results from our model and HyFRIS [22] on INTHEWILD and LIGHTTEST.

TABLE 4
Comparison among methods on lighting estimation using LIGHTTEST. The best results are marked in **bold**. Each entry shows measurements on indoor (left) and outdoor (right) data, separated by a “/”.

Method	Env. map $\mathbf{L}$	Specular $\mathbf{L}^{\mathrm{s}}$	Diffuse $\mathbf{L}^{\mathrm{d}}$
	SI-RMSE↓	SI-RMSE↓	SI-RMSE↓
HyFRIS [22]	– / –	0.060/0.055	0.119/0.139
HDRLE [37]	– /0.318	– /0.034	– /0.145
FaceProbe [23]	0.336/0.338	0.077/0.046	0.231/0.195
LIDP [34]	0.307/0.258	0.043/0.042	0.121/0.135
Ours	**0.268/0.180**	**0.029/0.018**	**0.073/0.069**

Comparison on lighting estimation. We compare our model with LIDP [34], FaceProbe [23], HDRLE [37], and HyFRIS [22] concerning lighting estimation both qualitatively on LIGHTTEST and INTHEWILD (see Fig. 6) and quantitatively on LIGHTTEST (see Table 4). Quantitative results in Table 4 show that our method outperforms all these methods on lighting estimation. Qualitative results in Fig. 6 also demonstrate that our method can recover accurate light sources with correct amounts and positions and realistic ambient textures in the estimated environment map. For example, our SPLiT succeeds in recovering two lights in row 1 col 1 (where LIDP [34] produces a blurry halo) and more accurate sun positions in row 2, probably because SPLiT extracts lighting-dependent intrinsics and condenses them into more friendly representations to lighting estimation.

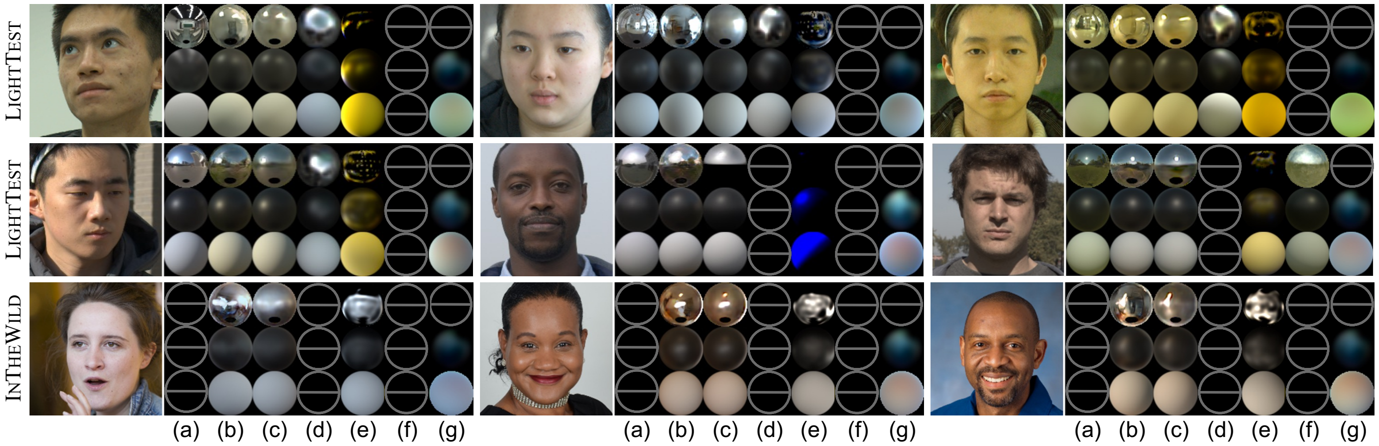

LightTest
LightTest
InTheWild
(a) (b) (c) (d) (e) (f) (g)
(a) (b) (c) (d) (e) (f) (g)
(a) (b) (c) (d) (e) (f) (g)

Fig. 6. Comparison among methods on lighting estimation using LightTest and InTheWild. For each set of results, we show from top to bottom $\mathbf{L}$, $\mathbf{L}^{\mathrm{s}}$, and $\mathbf{L}^{\mathrm{d}}$. Lighting estimates from left to right: (a) ground truth, (b) Ours, (c) Ours w/o f_{amb}, (d) LIDP [34], (e) FaceProbe [23], (f) HDRLE [37], and (g) HyFRIS [22]. Some entries (indicated as "⊖") are unavailable (see Sec. 5.1). Please zoom-in in the electronic version for details.

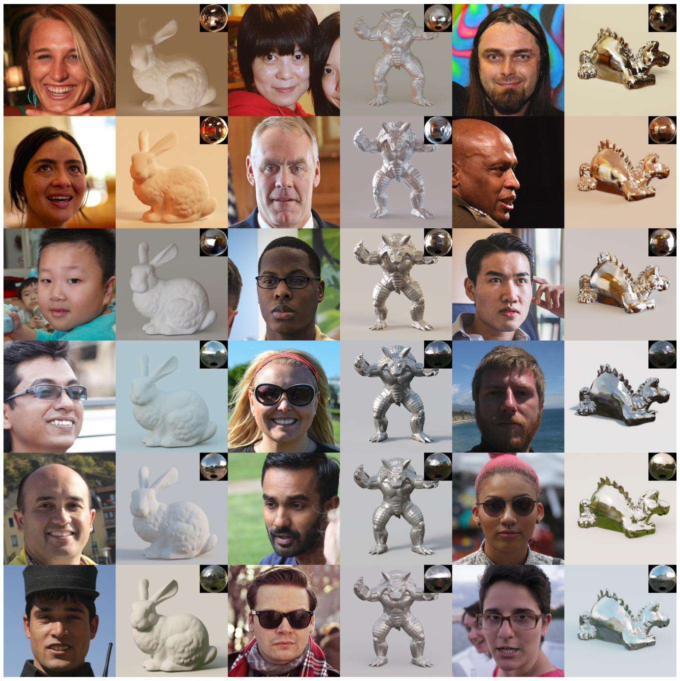


Fig. 7. We use the estimated lighting (shown in the top-right corner alongside each rendered object) to relight virtual objects. The first 3 rows are indoor scenes, while the last 3 rows are outdoor scenes. Please zoom-in in the electronic version for details.

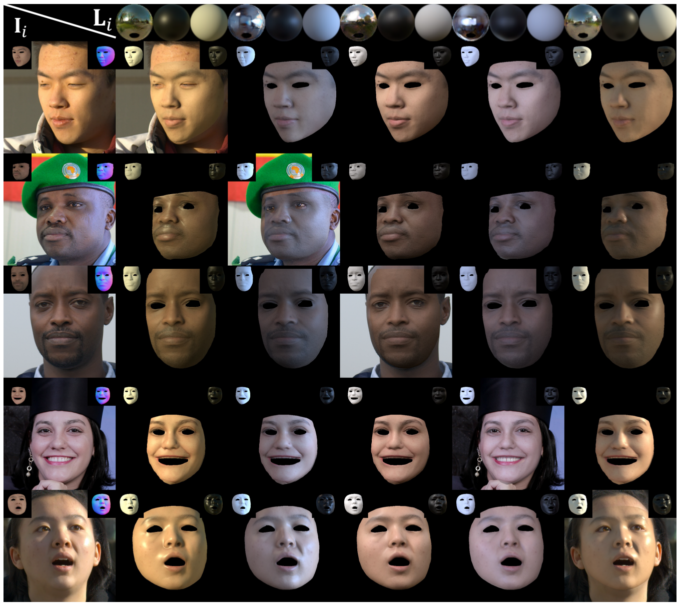


Fig. 8. We transfer the lighting $\mathbf{L}_i$ estimated from one input face image $\mathbf{I}_i$ to other face images. The leftmost column shows the input face images, while the uppermost row shows the estimated lighting in the same order. For each lighting, we show from left to right $\mathbf{L}$, $\mathbf{L}^{\mathrm{s}}$, and $\mathbf{L}^{\mathrm{d}}$. The remaining space shows a grid of relit or reproduced (on the diagonal) face images. For the input face image, we show the estimated albedo $\mathbf{A}$ and surface normal $\mathbf{N}$ at the top-left and top-right corner, respectively. For the relit face image, we show the rendered diffuse shading $\widetilde{\mathbf{D}}$ and specular reflection $\widetilde{\mathbf{S}}$ at the top-left and top-right corner, respectively. Please zoom-in in the electronic version for details.

6 Applications

We show potential applications enabled by our lighting estimations in this section, including virtual object insertion and lighting transfer.

6.1 Virtual Object Relighting

As a prevalent application for lighting estimation, we present virtual object relighting results on InTheWild in Fig. 7. We employ the lighting estimations derived from our model to render three objects with varying shapes and reflective properties: a matte Bunny, a plastic-like Armadillo, and a glossy Dragon. The rendered objects can then be realistically blended into the original image by alpha compositing. As illustrated in Fig. 7, our SPLiT model is applicable to a diverse array of portraits, encompassing various poses, skin tones, and lighting conditions. Both the Bunny and the Armadillo reflect the accurately estimated light source characteristics, including direction, intensity, and color. For instance, near-frontal lights with distinct colors are observed in row 1, column 1, and row 3, column 1, while a potent white light emanates from the rear left in row 2, column 2. Sun directions are also correctly estimated in row 5, column 1, and row 5, column 2, as evidenced by the cast shadows on the ground. Moreover, the glossy Dragon exhibits rich ambient textures, such as the trees in row 5, column 3. Notably, our model is capable of handling facial accessories, like hats in row 6, column 1, and sunglasses in row 6, column 2, despite their absence in the training set.

6.2 Lighting Transfer

To illustrate another potential application, we present the results of lighting transfer using our estimated $\mathbf{A}, \mathbf{N}$, and $\mathbf{L}$. It is crucial to emphasize that the primary objective of displaying the lighting transfer results is to provide an intuitive demonstration of the accuracy and detail of our estimated intrinsic components and lighting. Our method does not include a neural rendering layer to photo-realistically render the same individual under novel lighting conditions. Furthermore, it does not recover complete 3D shapes and facial BRDF parameters, which are necessary in a non-neural rendering process, as they are not employed in lighting estimation. We use normal integration [82] to acquire the approximate depth (we do not guarantee the correctness of its scale) to show shadowing effects due to lighting changes and use empirical BRDF parameters during the relighting process. Consequently, our method is unsuitable for comparison with other automatic portrait relighting techniques, such as [38], [41], [60], which incorporate dedicated neural rendering layers to generate realistic relit portraits.

For two input images $\mathbf{I}_1, \mathbf{I}_2$, we initially apply our method to obtain their surface normal $\mathbf{N}_1, \mathbf{N}_2$, diffuse albedo $\mathbf{A}_1, \mathbf{A}_2$, and lighting $\mathbf{L}_1, \mathbf{L}_2$. Subsequently, we relight face $\mathbf{I}_1$ under lighting $\mathbf{L}_2$ (and vice versa) using the following non-neural rendering process. We maintain the estimated $\mathbf{A}_1$ and $\mathbf{N}_1$ constant while modifying $\mathbf{D}$ and $\mathbf{S}$ under the non-Lambertian reflectance model $\mathbf{I} = \mathbf{A} \odot \mathbf{D} + \mathbf{S}$. The relit shading $\widetilde{\mathbf{D}}_2$ and $\widetilde{\mathbf{S}}_2$ are rendered by Blender Cycles [77] using the new lighting $\mathbf{L}_2$ and the depth obtained via normal integration [82] of $\mathbf{N}_1$. We employ the same BRDF model used when generating our SynTrain dataset. The relit image $\widetilde{\mathbf{I}}_2$ is achieved by computing $(\mathbf{A}_1 \odot \widetilde{\mathbf{D}}_2 + \widetilde{\mathbf{S}}_2)$. This procedure can also be employed to reproduce the input image by using $\mathbf{L}_1$ instead of $\mathbf{L}_2$. The results are displayed in Fig. 8, illustrating that our estimated intrinsic components and lighting can be utilized to transfer lighting to another portrait with a completely different illumination condition.

7 Conclusion

We present a method to estimate lighting from an input human face image, by estimating a tetrad of face intrinsics and predicting an environment map from two spherically distributed lighting-dependent intrinsics. Our method is physically interpretable and gives accurate and detailed results on lighting estimation and intrinsic decomposition.

Limitation. As a lighting estimation method, SPLiT assumes distant and uniform illumination on the whole face, so that near field illumination, and occlusions or interreflections from non-face objects (*e.g.*, hats, clothes with strong hue) are not modeled. Also, since our focus is lighting estimation, we do not either estimate a 3D face model for novel view synthesis, or spatially-varying BRDF parameters such as roughness for re-rendering. Our method depends on a face mask and will produce false light sources if the mask is oversized and bright background pixels are recognized as strong specularities. But this can be automatically eliminated by aggressively eroding the mask since we experimentally find our method robust to undersized masks (see supplementary material).

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 62136001, 62088102 and Program for Youth Innovative Research Team of BUPT No. 2023QNTD02.

References

[1] M. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambaretto, C. Gagné, and J. Lalonde, "Learning to predict indoor illumination from a single image," in *ACM Transactions on Graphics*, 2017.

[2] M. Gardner, Y. Hold-Geoffroy, K. Sunkavalli, C. Gagné, and J. Lalonde, "Deep parametric indoor lighting estimation," in *Proc. of International Conference on Computer Vision*, 2019.

[3] Y. Hold-Geoffroy, A. Athawale, and J. Lalonde, "Deep sky modeling for single image outdoor lighting estimation," in *Proc. of Computer Vision and Pattern Recognition*, 2019.

[4] Z. Li, M. Shafiei, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker, "Inverse rendering for complex indoor scenes: shape, spatially-varying lighting and SVBRDF from a single image," in *Proc. of Computer Vision and Pattern Recognition*, 2020.

[5] J. Zhu, F. Luan, Y. Huo, Z. Lin, Z. Zhong, D. Xi, R. Wang, H. Bao, J. Zheng, and R. Tang, "Learning-based inverse rendering of complex indoor scenes with differentiable Monte Carlo raytracing," in *Proc. of ACM SIGGRAPH Asia*, 2022.

[6] H. Weber, M. Garon, and J. Lalonde, "Editable indoor lighting estimation," in *Proc. of European Conference on Computer Vision*, 2022.

[7] Z. Li, J. Shi, S. Bi, R. Zhu, K. Sunkavalli, M. Hasan, Z. Xu, R. Ramamoorthi, and M. Chandraker, "Physically-based editing of indoor scene lighting from a single image," in *Proc. of European Conference on Computer Vision*, 2022.

[8] G. Wang, Y. Yang, C. C. Loy, and Z. Liu, "StyleLight: HDR panorama generation for lighting estimation and editing," in *Proc. of European Conference on Computer Vision*, 2022.

[9] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. of ACM SIGGRAPH*, 1999.

[10] S. Alotaibi and W. A. P. Smith, "BioFaceNet: deep biophysical face image interpretation," in *Proc. of British Machine Vision Conference*, 2019.

[11] N. Tsumura, N. Ojima, K. Sato, M. Shiraishi, H. Shimizu, H. Nabeshima, S. Akazaki, K. Hori, and Y. Miyake, "Image-based skin color and texture analysis/synthesis by extracting hemoglobin and melanin information in the skin," in *ACM Transactions on Graphics*, 2003.

[12] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003.

[13] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "FaceScape: a large-scale high quality 3D face dataset and detailed riggable 3D face prediction," in *Proc. of Computer Vision and Pattern Recognition*, 2020.

[14] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. of Computer Vision and Pattern Recognition*, 2019.

[15] D. A. Calian, J. Lalonde, P. F. U. Gotardo, T. Simon, I. A. Matthews, and K. Mitchell, "From faces to outdoor light probes," in *Computer Graphics Forum*, 2018.

[16] B. Egger, S. Schönborn, A. Schneider, A. Kortylewski, A. Morel-Forster, C. Blumer, and T. Vetter, "Occlusion-aware 3D morphable models and an illumination prior for face image analysis," in *International Journal of Computer Vision*, 2018.

[17] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt, "Self-supervised multi-level face model learning for monocular reconstruction at over 250 Hz," in *Proc. of Computer Vision and Pattern Recognition*, 2018.

[18] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Pérez, and C. Theobalt, "MoFA: model-based deep convolutional face autoencoder for unsupervised monocular reconstruction," in *Proc. of International Conference on Computer Vision*, 2017.

[19] A. Tewari, F. Bernard, P. Garrido, G. Bharaj, M. Elgharib, H. Seidel, P. Pérez, M. Zollhöfer, and C. Theobalt, "FML: face model learning from videos," in *Proc. of Computer Vision and Pattern Recognition*, 2019.

[20] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs, "SfSNet: learning shape, reflectance and illuminance of faces "in the wild"," in *Proc. of Computer Vision and Pattern Recognition*, 2018.

[21] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras, "Neural face editing with intrinsic image disentangling," in *Proc. of Computer Vision and Pattern Recognition*, 2017.

[22] Y. Zhu, C. Li, S. Li, B. Shi, and Y. W. Tai, "Hybrid face reflectance, illumination, and shape from a single image," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[23] R. Yi, C. Zhu, P. Tan, and S. Lin, "Faces as lighting probes via unsupervised deep highlight extraction," in *Proc. of European Conference on Computer Vision*, 2018.

[24] H. Barrow, J. Tenenbaum, A. Hanson, and E. Riseman, "Recovering intrinsic scene characteristics from images," 1978.

[25] J. Shen, X. Yang, Y. Jia, and X. Li, "Intrinsic images using optimization," in *Proc. of Computer Vision and Pattern Recognition*, 2011.

[26] B. K. P. Horn, "Determining lightness from an image," in *Computer Graphics and Image Processing*, 1974.

[27] Y. Weiss, "Deriving intrinsic images from image sequences," in *Proc. of International Conference on Computer Vision*, 2001.

[28] J. T. Barron and J. Malik, "Shape, illumination, and reflectance from shading," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[29] S. Sengupta, J. Gu, K. Kim, G. Liu, D. W. Jacobs, and J. Kautz, "Neural inverse rendering of an indoor scene from a single image," in *Proc. of International Conference on Computer Vision*, 2019.

[30] Y. Yu and W. A. P. Smith, "InverseRenderNet: learning single image inverse rendering," in *Proc. of Computer Vision and Pattern Recognition*, 2019.

[31] P. Tan, S. Lin, L. Quan, and H. Shum, "Highlight removal by illumination-constrained inpainting," in *Proc. of International Conference on Computer Vision*, 2003.

[32] C. Li, S. Lin, K. Zhou, and K. Ikeuchi, "Specular highlight removal in facial images," in *Proc. of Computer Vision and Pattern Recognition*, 2017.

[33] Y. Zhu, Y. Zhang, S. Li, and B. Shi, "Spatially-varying outdoor lighting estimation from intrinsics," in *Proc. of Computer Vision and Pattern Recognition*, 2021.

[34] C. LeGendre, W. Ma, R. Pandey, S. R. Fanello, C. Rhemann, J. Dourgarian, J. Busch, and P. E. Debevec, "Learning illumination from diverse portraits," in *Proc. of ACM SIGGRAPH Asia Technical Communications*, 2020.

[35] T. Li, M. Aittala, F. Durand, and J. Lehtinen, "Differentiable Monte Carlo ray tracing through edge sampling," in *ACM Transactions on Graphics*, 2018.

[36] D. Vicini, S. Speierer, and W. Jakob, "Path replay backpropagation: differentiating light paths using constant memory and linear time," in *ACM Transactions on Graphics*, 2021.

[37] A. Sztrajman, A. Neophytou, T. Weyrich, and E. Sommerlade, "High-dynamic-range lighting estimation from face portraits," in *Proc. of International Conference on 3D Vision*, 2020.

[38] Z. Wang, X. Yu, M. Lu, Q. Wang, C. Qian, and F. Xu, "Single image portrait relighting via explicit multiple reflectance channel modeling," in *ACM Transactions on Graphics*, 2020.

[39] P. E. Debevec, T. Hawkins, C. Tchou, H. Duiker, W. Sarokin, and M. Sagar, "Acquiring the reflectance field of a human face," in *Proc. of ACM SIGGRAPH*, 2000.

[40] F. Zhan, C. Zhang, Y. Yu, Y. Chang, S. Lu, F. Ma, and X. Xie, "EMLight: Lighting estimation via spherical distribution approximation," in *Proc. of Association for the Advancement of Artificial Intelligence*, 2021.

[41] T. Sun, J. T. Barron, Y. Tsai, Z. Xu, X. Yu, G. Fyffe, C. Rhemann, J. Busch, P. E. Debevec, and R. Ramamoorthi, "Single image portrait relighting," in *ACM Transactions on Graphics*, 2019.

[42] C. Li, T. Ngo, and H. Nagahara, "Inverse rendering of translucent objects using physical and neural renderers," 2023.

[43] V. Rudnev, M. Elgharib, W. A. P. Smith, L. Liu, V. Golyanik, and C. Theobalt, "NeRF for outdoor scene relighting," in *Proc. of European Conference on Computer Vision*, 2022.

[44] K. Zhang, F. Luan, Q. Wang, K. Bala, and N. Snavely, "PhySG: Inverse rendering with spherical gaussians for physics-based material editing and relighting," in *Proc. of Computer Vision and Pattern Recognition*, 2021.

[45] J. Lalonde, A. A. Efros, and S. G. Narasimhan, "Estimating the natural illumination conditions from a single outdoor image," in *International Journal of Computer Vision*, 2012.

[46] Y. Hold-Geoffroy, K. Sunkavalli, S. Hadap, E. Gambaretto, and J. Lalonde, "Deep outdoor illumination estimation," in *Proc. of Computer Vision and Pattern Recognition*, 2017.

[47] B. Yu, S. Yang, X. Cui, S. Dong, B. Chen, and B. Shi, "MILO: Multi-bounce inverse rendering for indoor scene with light-emitting objects," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[48] Z. Wang, T. Shen, J. Gao, S. Huang, J. Munkberg, J. Hasselgren, Z. Gojcic, W. Chen, and S. Fidler, "Neural fields meet explicit geometric representations for inverse rendering of urban scenes," in *Proc. of Computer Vision and Pattern Recognition*, 2023.

[49] J. Zhang, K. Sunkavalli, Y. Hold-Geoffroy, S. Hadap, J. Eisenmann, and J. Lalonde, "All-weather deep outdoor lighting estimation," in *Proc. of Computer Vision and Pattern Recognition*, 2019.

[50] P. Yu, J. Guo, F. Huang, C. Zhou, H. Che, X. Ling, and Y. Guo, "Hierarchical disentangled representation learning for outdoor illumination estimation and editing," in *Proc. of International Conference on Computer Vision*, 2021.

[51] J. Tang, Y. Zhu, H. Wang, J.-H. Chan, S. Li, and B. Shi, "Estimating spatially-varying lighting in urban scenes with disentangled representation," in *Proc. of European Conference on Computer Vision*, 2022.

[52] C. LeGendre, W. Ma, G. Fyffe, J. Flynn, L. Charbonnel, J. Busch, and P. E. Debevec, "DeepLight: learning illumination for unconstrained mobile mixed reality," in *Proc. of Computer Vision and Pattern Recognition*, 2019.

[53] M. Garon, K. Sunkavalli, S. Hadap, N. Carr, and J. Lalonde, "Fast spatially-varying indoor lighting estimation," in *Proc. of Computer Vision and Pattern Recognition*, 2019.

[54] S. Song and T. A. Funkhouser, "Neural illumination: Lighting prediction for indoor environments," in *Proc. of Computer Vision and Pattern Recognition*, 2019.

[55] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Proc. of European Conference on Computer Vision*, 2020.

[56] M. Boss, R. Braun, V. Jampani, J. T. Barron, C. Liu, and H. P. A. Lensch, "NeRD: Neural reflectance decomposition from image collections," in *Proc. of International Conference on Computer Vision*, 2021.

[57] Y. Zhang, J. Sun, X. He, H. Fu, R. Jia, and X. Zhou, "Modeling indirect illumination for inverse rendering," in *Proc. of Computer Vision and Pattern Recognition*, 2022.

[58] E. Garces, C. Rodríguez-Pardo, D. Casas, and J. Lopez-Moreno, "A survey on intrinsic images: Delving deep into lambert and beyond," in *International Journal of Computer Vision*, 2022.

[59] S. Yamaguchi, S. Saito, K. Nagano, Y. Zhao, W. Chen, K. Olszewski, S. Morishima, and H. Li, "High-fidelity facial reflectance and geometry inference from an unconstrained image," in *ACM Transactions on Graphics*, 2018.

[60] R. Pandey, S. Orts-Escolano, C. LeGendre, C. Häne, S. Bouaziz, C. Rhemann, P. E. Debevec, and S. R. Fanello, "Total relighting: learning to relight portraits for background replacement," in *ACM Transactions on Graphics*, 2021.

[61] T. Nestmeyer, J. Lalonde, I. A. Matthews, and A. M. Lehrmann, "Learning physics-guided face relighting under directional light," in *Proc. of Computer Vision and Pattern Recognition*, 2020.

[62] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A framework for perceiving and processing reality," in *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition*, 2019.

[63] C. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," in *Proc. of Computer Vision and Pattern Recognition*, 2020.

[64] P. E. Debevec, P. Graham, J. Busch, and M. T. Bolas, "A single-shot light probe," in *International Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2012, Los Angeles, California, USA, August 5-9, 2012, Talks Proceedings*, 2012.

[65] M. R. K. Dastjerdi, Y. Hold-Geoffroy, J. Eisenmann, and J. Lalonde, "Everlight: Indoor-outdoor editable HDR lighting estimation," in *Proc. of International Conference on Computer Vision*, 2023.

[66] H. Yu, S. Agarwala, C. Herrmann, R. Szeliski, N. Snavely, J. Wu, and D. Sun, "Accidental light probes," in *Proc. of Computer Vision and Pattern Recognition*, 2023.

[67] G. Somanath and D. Kurz, "HDR environment map estimation for real-time augmented reality," in *Proc. of Computer Vision and Pattern Recognition*, 2021.

[68] Z. Li, Z. Xu, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker, "Learning to reconstruct shape and spatially-varying reflectance from a single image," in *ACM Transactions on Graphics*, 2018.

[69] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *Proc. of International Conference on Medical Image Computing and Computer Assisted Intervention*, 2015.

[70] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. of International Conference on Learning Representations*, 2015.

[71] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. of Computer Vision and Pattern Recognition*, 2018.

[72] K. Rematas, T. Ritschel, M. Fritz, E. Gavves, and T. Tuytelaars, "Deep reflectance maps," in *Proc. of Computer Vision and Pattern Recognition*, 2016.

[73] M. R. K. Dastjerdi, Y. Hold-Geoffroy, J. Eisenmann, S. Khodadadeh, and J. Lalonde, "Guided co-modulated GAN for 360° field of view extrapolation," in *Proc. of International Conference on 3D Vision*, 2022.

[74] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. of Computer Vision and Pattern Recognition*, 2018.

[75] H. Weber, D. Prévost, and J. Lalonde, "Learning to estimate indoor lighting from 3D objects," in *Proc. of International Conference on 3D Vision*, 2018.

[76] L. Zhang, Q. Zhang, M. Wu, J. Yu, and L. Xu, "Neural video portrait relighting in real-time via consistency modeling," in *Proc. of International Conference on Computer Vision*, 2021.

[77] Blender Foundation, "Cycles open source production rendering," https://www.cycles-renderer.org. Accessed: 2023-04-23.

[78] B. Burley, "Physically-based shading at Disney," in *Proc. of ACM SIGGRAPH*, 2012.

[79] J. Stumpfel, A. Jones, A. Wenger, C. Tchou, T. Hawkins, and P. E. Debevec, "Direct HDR capture of the sun and sky," in *Proc. of ACM SIGGRAPH*, 2004.

[80] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, 2004.

[81] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. of Neural Information Processing Systems*, 2017.

[82] X. Cao, B. Shi, F. Okura, and Y. Matsushita, "Normal integration via inverse plane fitting with minimum point-to-plane distance," in *Proc. of Computer Vision and Pattern Recognition*, 2021.

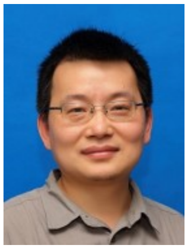
Fan Fei received the BS degree from Peking University in 2022. He is currently pursuing PhD degree at Peking University. His research interests include inverse rendering and scene reconstruction.

Yean Cheng received the BE degree from Tsinghua University in 2021. He is a MSc student at Peking University. His research interests include computational photography, scene lighting estimation, and implicit neural rendering.

Yongjie Zhu received the BE and ME degrees from Beijing University of Posts and Telecommunications in 2019 and 2022. He is currently an algorithm engineer at Alibaba Group (Beijing). His research interest is physics based computer vision.

Qian Zheng is a tenure track professor of the College of Computer Science and Technology, Zhejiang University, China. He received the BE and PhD degrees from Zhejiang University in 2011 and 2017, respectively. From 2018 to 2022, he was a research fellow with the ROSE lab, Nanyang Technological University. His research interests include neuromorphic computing and computer vision. He has co-authored more than 40 papers. He serves as an associate editor of neurocomputing and a reviewer of many journals and conferences, such as T-PAMI, IJCV, CVPR, ICCV, SIGGRAPH, NeurIPS, ICLR. He was a guest editor of Frontiers in Neuroscience Neuroprosthetics. He is a member of IEEE.

Si Li received the PhD degree from Beijing University of Posts and Telecommunications in 2012. Now she is an Associate Professor in School of Artificial Intelligence, Beijing University of Posts and Telecommunications. Her current research interests include computer vision, natural language processing, and multimodal artificial intelligence.



Gang Pan is a Professor of the College of Computer Science and Technology, Zhejiang University, the Executive Deputy Director of the State Key Laboratory of Brain-Machine Intelligence. He received the BE and PhD degrees from Zhejiang University in 1998 and 2004 respectively. His interests include artificial intelligence, brain-inspired computing, brain-machine interfaces, and pervasive computing. He has co-authored more than 150 refereed papers, and has more than 60 patents granted. Dr. Pan is a recipient of NSFC for Distinguished Young Scholars, IEEE TCSC Award for Excellence (Middle Career Researcher), CCF-IEEE CS Young Scientist Award, TOP-10 Achievements in Science and Technology in Chinese Universities, National Science and Technology Progress Award, and several best paper awards. He serves as an associate editor of IEEE Trans. Neural Networks and Learning Systems, Cognitive Neurodynamics, etc. He is a senior member of IEEE.

Boxin Shi received the BE degree from the Beijing University of Posts and Telecommunications, the ME degree from Peking University, and the PhD degree from the University of Tokyo, in 2007, 2010, and 2013. He is currently a Boya Young Fellow Assistant Professor and Research Professor at Peking University, where he leads the Camera Intelligence Lab. Before joining PKU, he did research with MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University, National Institute of Advanced Industrial Science and Technology, from 2013 to 2017. His papers were awarded as Best Paper Runner-Up at ICCP 2015 and selected as Best Papers from ICCV 2015 for IJCV Special Issue. He is an associate editor of TPAMI/IJCV and an area chair of CVPR/ICCV. He is a senior member of IEEE.

SPLiT: Single Portrait Lighting Estimation via a Tetrad of Face Intrinsics Supplementary Material

Fan Fei#, Yean Cheng#, Yongjie Zhu, Qian Zheng, Si Li, Gang Pan, *Senior Member, IEEE*, and Boxin Shi*, *Senior Member, IEEE*

A Projection types of environment maps

Our method adopts the spherical projection rather than the equirectangular projection as the type of environment map. Here we discuss different types of projection for environment maps that can be used in lighting-related computer vision problems.

The three most popular types of environment mapping may have been cube mapping, equirectangular mapping, and spherical mapping. While cube mapping using the six faces of a cube as the map shape is a frequently used type in computer graphics, we find that it has not been used much as lighting representations in computer vision. This could be a consequence of its enormous seams, *i.e.*, the paired edges separated in the flattened 2D image space but are instead connected when projected back onto 3D space. Such seams present a significant challenge to the convolutional neural networks, as their receptive field is forced to terminate on these seams, causing undesirable discontinuities on different sides of the seams in the 3D space when projected back [5]. The seams are still present in the equirectangular mapping and spherical mapping, but are much less – these two mappings have only one seam each (the horizontal edge of the equirectangular mapping corresponding to a line in 3D space, and the circular edge of the spherical mapping corresponding to a point), while cube mappings have seven seams corresponding to seven lines in 3D space. It may be the above fact that has led to the preference for equirectangular environment maps [2], [6], [7] and sphere maps [8], [9], [10] over cube maps in lighting estimation works, despite that cube maps suffer less from image distortion. Furthermore, it is worth noting that most lighting estimation methods that use equirectangular mapping choose to place the LFoV input image (which contains the region in front of the camera) at the center of the panoramic environment map [1], [2], [3], [4], thus the vertical edge of the environment map is behind the camera when projected back to 3D. This leads to the manifestation of the strong vertical edge when rendering mirror-like objects. In contrast, the distortion caused by the seams in the spherical mapping is in front of the camera, so the distortion will not be reflected on the rendered mirror-like object. As examples of the discontinuities caused by seams, in Fig. A we show the equirectangular environment maps estimated by four state-of-the-art lighting estimation methods [1], [2], [3], [4] and the rendered highly glossy objects. Undesirable strong vertical edges can be easily seen in the rendered teapot with a mirror-like surface.

With cube mapping no longer in our consideration, we believe that the remaining two types of mapping are both acceptable. We finally choose the sphere mapping rather than the equirectangular mapping, taking into account the characteristics of our pipeline. After the normal-indexed distribute operation (**Sec 3.3**), the input of our lighting estimation module (**Sec 3.4**) should be the lighting-dependent intrinsic components projected onto a predefined geometry, either a standard sphere or the equirectangular coordinate system. We find that the spherical form is more friendly for human perception, and probably for neural networks as well, than the equirectangular form (similar to prefiltered irradiance maps and reflection maps [11]), as the spherically distributed components are identical to images of diffuse or glossy balls illuminated by the scene lighting (ignoring noises and missing pixels). The $\{\mathbf{D}^\circ, \mathbf{S}^\circ\}$ components in Fig. 2 in the manuscript give a good example. This perceptional friendliness of the sphere mapping may have inspired some recent illumination estimation methods [4], [8], [9], [12] to provide visual comparisons on their estimations using three different balls (mirror, glossy, diffuse) rendered under lighting to be compared, which we also perform in **Sec. 5.3.**

#Fan Fei and Yean Cheng contribute equally to this work.
**Boxin Shi is the corresponding author.*

- *Fan Fei, Yean Cheng, and Boxin Shi are with National Key Laboratory for Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing 100871, China. E-mails: {feifan_eecs, shiboxin}@pku.edu.cn, cya17@stu.pku.edu.cn.*
- *Yean Cheng and Boxin Shi are also with AI Innovation Center, School of Computer Science, Peking University, Beijing 100871, China.*
- *Yongjie Zhu and Si Li are with School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China. E-mails: {zhuyongjie, lisi}@bupt.edu.cn.*
- *Qian Zheng and Gang Pan are with The State Key Lab of Brain-Machine Intelligence and College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China. Emails: {qianzheng, gpan}@zju.edu.cn.*

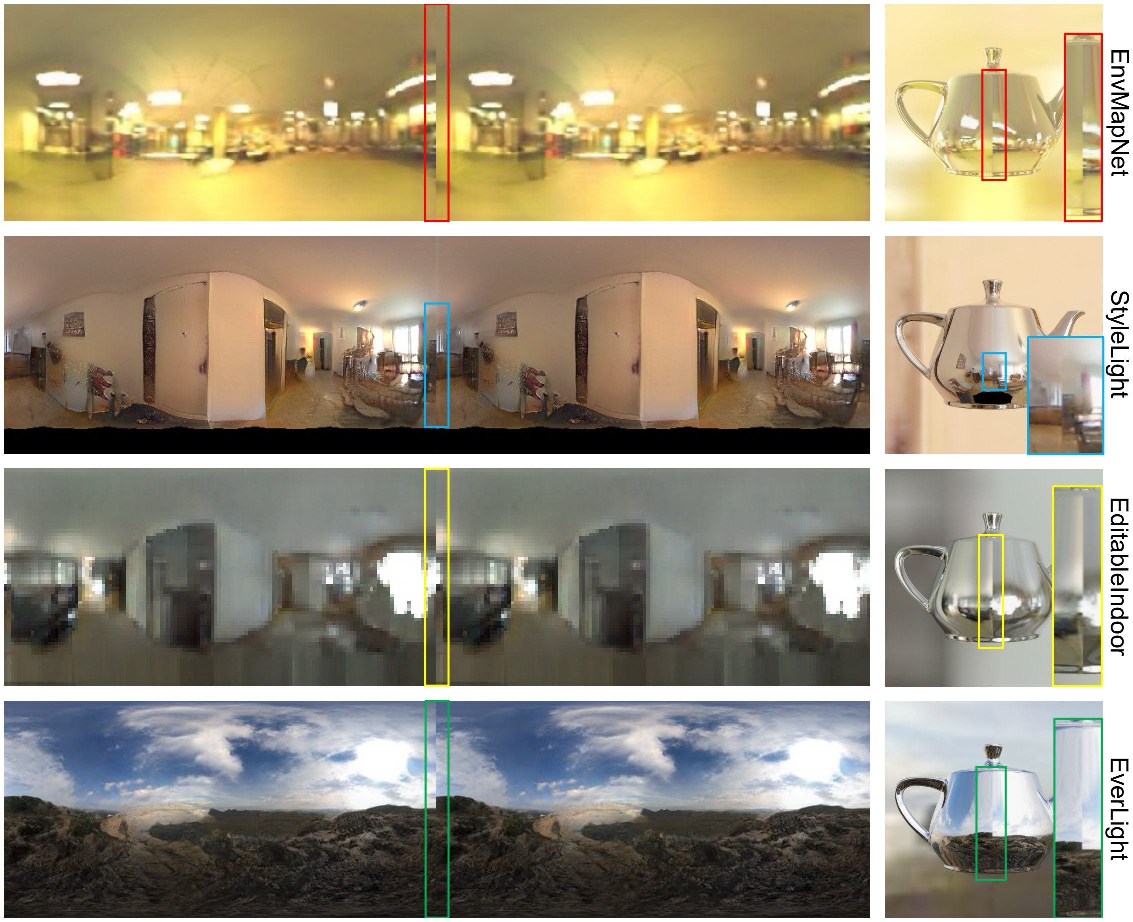


Fig. A. Seams in the estimated equirectangular environment map cause strong vertical edges in the rendered mirror-like objects. Left: the estimated environment map (LDR versions copied from each paper), repeat twice to show the discontinuity at the vertical edge. Right: teapot with a mirrored surface rendered under the environment map (note that LDR environment maps basically suffice to render LDR images of a mirror-like object). We show the results of four recent lighting estimation methods: EnvMapNet [1], StyleLight [2], EditableIndoor [3], and EverLight [4]. We show no cube environment map results as we do not find a lighting estimation method using the cube map representation.

B Implementation Details

In this section, we elaborate on the implementation details, including loss weights, network architectures, and training schemes, of our intrinsic decomposition module and lighting estimation module. The network architectures are also shown in Fig. B. We also recommend that readers check our released code if they have any questions about the architecture.

B.1 Tetrad Intrinsic Decomposition Module

Loss function weights are set as $\lambda_{\mathrm{A}} = \lambda_{\mathrm{N}} = \lambda_{\mathrm{D}} = \lambda_{\mathrm{rec}_1} = 1, \lambda_{\mathrm{A}_{\mathrm{vgg}}} = \lambda_{\mathrm{D}_{\mathrm{vgg}}} = 0.1, \lambda_{\mathrm{S}} = 10$ to balance different loss terms to be of similar scales. Both f_{AN} and f_{DS} share a modified U-Net [13] architecture, comprising a 6-layer encoder, a bottleneck layer, and a 6-layer decoder with base filter number $= 48$. The encoder and decoder layers consist of down-sampling by blur-pooling [14] or bilinear up-sampling, followed by a 3×3 convolution of stride 1. Every convolution operation, except the final one, is succeeded by a batch normalization [15], a SiLU non-linearity [16], and a channel-wise dropout [17] of rate 0.1. To transform the unbounded output values of networks into feasible intrinsic components, sigmoid, vector normalization, and softplus functions are applied to $\mathbf{A}$, $\mathbf{N}$, and $\mathbf{D}/\mathbf{S}$ components, respectively.

f_{AN} and f_{DS} are initially trained separately, followed by combined training without freezing layers. This training process, using the Adam optimizer [18], involves a linear learning-rate warm-up stage, reaching a base learning rate of 10^{-3} after the first epoch. The learning rate is then multiplied by $0.1\times$ after each subsequent epoch. The training lasts for a total of 10 epochs, with the first 5 epochs for separate training and the last 5 for combined training. The networks are trained with a batch size of 8, and the entire training process takes approximately 30 hours on a single NVIDIA GeForce RTX 3090 GPU.

B.2 Spherical Lighting Estimation Module

Loss function weights are set as $\lambda_{\mathrm{m}} = 0.6, \lambda_{\mathrm{d}} = \lambda_{\mathrm{s}} = 0.2, \lambda_{\mathrm{feat}} = 10, \lambda_{\mathrm{adv}} = \lambda_{\mathrm{vgg}} = 1$ to balance different loss terms. All spherical lighting representations and distributed components are sized 64×64. We train this module separately on indoor and outdoor data using the same architecture and training scheme. The light source network f_{src}, implemented as a T-network [19], [20], consists of two parts: an autoencoder and an estimator. The autoencoder takes GT HDR environment maps as input and is trained using

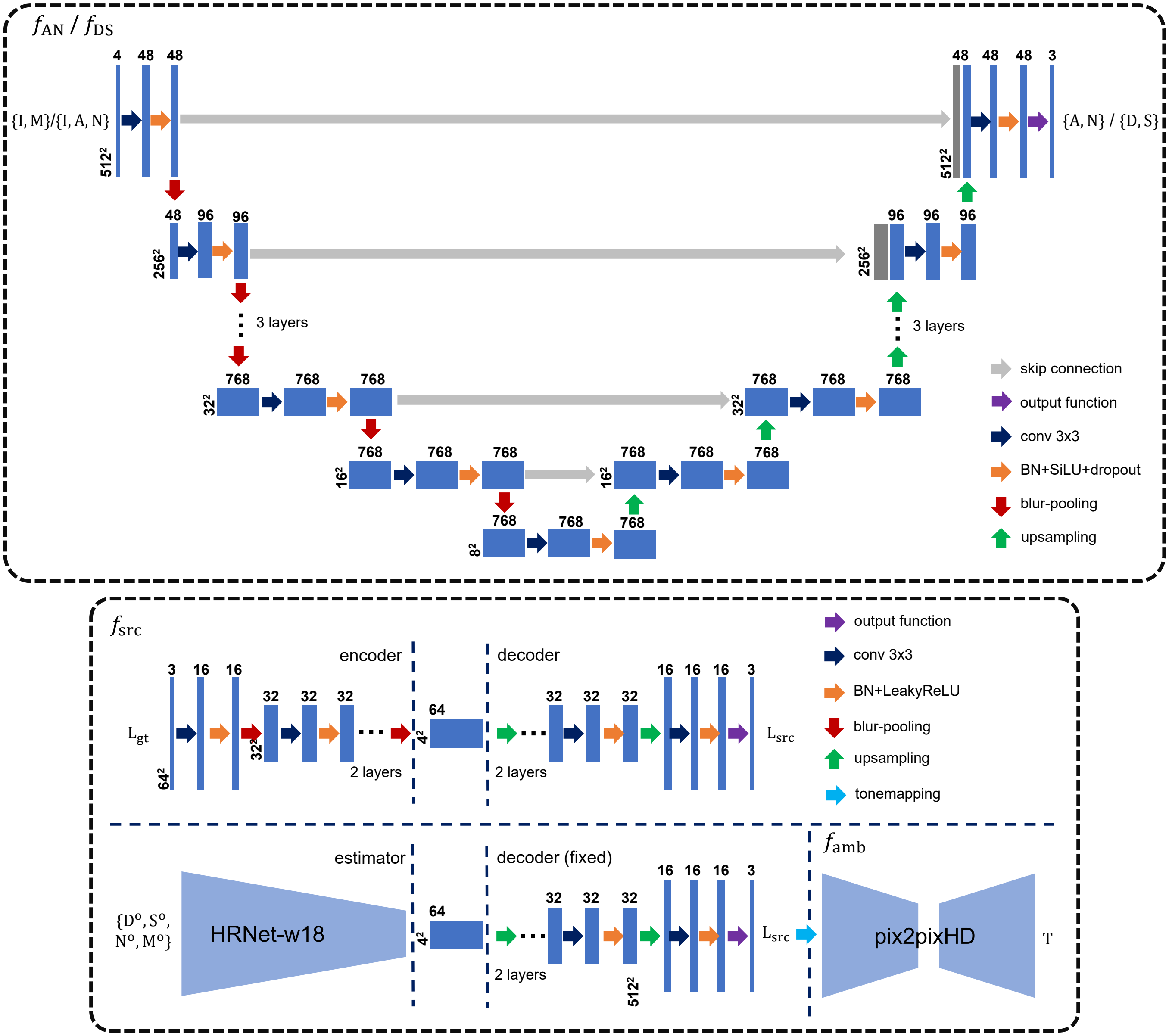


Fig. B. The architecture diagrams of our intrinsic decomposition module (upper) and our lighting estimation module (lower).

$\mathcal{L}_{\mathrm{src}}$ in Eq. (8), while the estimator takes $\mathbf{D}^{\circ}, \mathbf{S}^{\circ}, \mathbf{N}^{\circ}, \mathbf{M}^{\circ}\}$ as input and generates a latent code close to the code generated by the encoder. After training, the encoder is discarded, and the estimator is combined with the decoder to form f_{src}. The autoencoder comprises a 5-layer encoder and a 4-layer decoder with base filter number $= 16$. Blur pooling and bilinear up-sampling are used in the encoder and decoder, respectively, and every convolution is followed by a batch normalization and a leakyReLU [21]. The encoder generates a latent code with the size $4^2 \times 64 = 1024$. The estimator, employing the HRNet-w18 [22] architecture with a 3×3 convolutional head, produces a latent code of the same size. For the ambient texture network, we use the local enhancer architecture in pix2pixHD [23] with base filter number $= 16$ as the backbone of the generator f_{amb}, while the multi-scale discriminator with three scales and base filter number $= 32$ serves as discriminator f_{disc}.

We use Adam [18] optimizer for training. The autoencoder is trained for 200 epochs, while the estimator is first trained for 200 epochs taking distributed GT components as input, then fine-tuned for another 200 epochs taking distributed estimations from the trained intrinsic module as input, with an initial learning rate of 10^{-3} and step learning rate decay. The ambient texture network is trained for 70 epochs on the output of the autoencoder with an initial learning rate of 2×10^{-4} and a linear learning rate decay. The networks are trained with a batch size of 256, and the entire training process takes approximately 30 hours on a single NVIDIA GeForce RTX 3090 GPU.

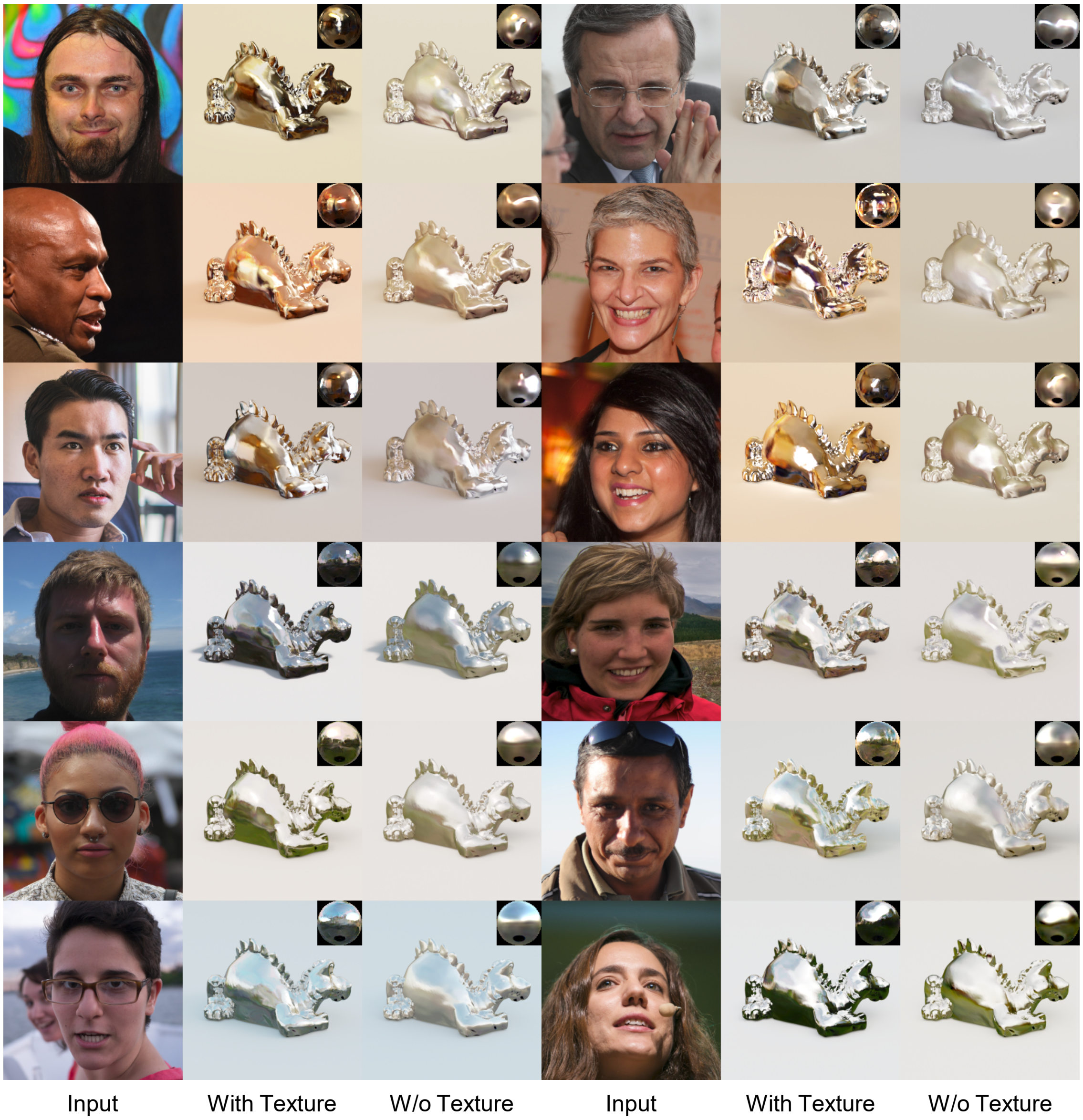


Fig. C. We use the estimated lighting (shown in the top-right corner alongside each rendered object) with texture ($\mathbf{L}$ in Eq. (10)) and without texture ($\mathbf{L}_{\text{src}}$ in Eq. (10)) to relight mirror-like virtual objects. The first 3 rows are indoor scenes, while the last 3 rows are outdoor scenes.

C More evaluation of the texture network

In **Sec. 3.4** in the main paper, we introduce the ambient texture network to fill in high-frequency and realistic textures in case that the estimated lighting will be used to render mirror-like objects in applications. In this section, we provide more qualitative evaluations of this network to show its effectiveness.

C.1 Comparison on Mirror-like Objects

In Fig. C, we show side-by-side mirror-like objects rendered using environment maps either with the texture ($\mathbf{L}$) hallucinated by the texture network or without the texture ($\mathbf{L}_{\text{src}}$). We can observe that the rendered objects using the textured environment maps manifest more clear and realistic appearances, while those rendered using the environment maps without texture are more blurry.

C.2 Comparison with Random Real Textures

A seemingly plausible substitute for our ambient texture network is randomly picking a real panorama in the training lighting dataset, and using it as $\mathbf{T}$ in Eq. (10). We experiment with this strategy, and as expected, using randomly picked real panoramas yields a huge improvement in FID score (from 98.7/106.4 to 18.27/29.55 for indoor/outdoor), since the details are now as fine as real panoramas, but with the unfair advantage of using an additional database instead of only the trained neural network in the test. Moreover, in the qualitative comparison of the generated environment maps using different strategies in Fig. D, we find that although using random real texture leads to a more detailed texture, the texture does not fit the light source since the randomly picked texture is light source-agnostic. For example, in outdoor scenes (the right half of Fig. D), the bright part in the random texture is usually not collocated with the sun position in the light source image, resulting in two visible suns in

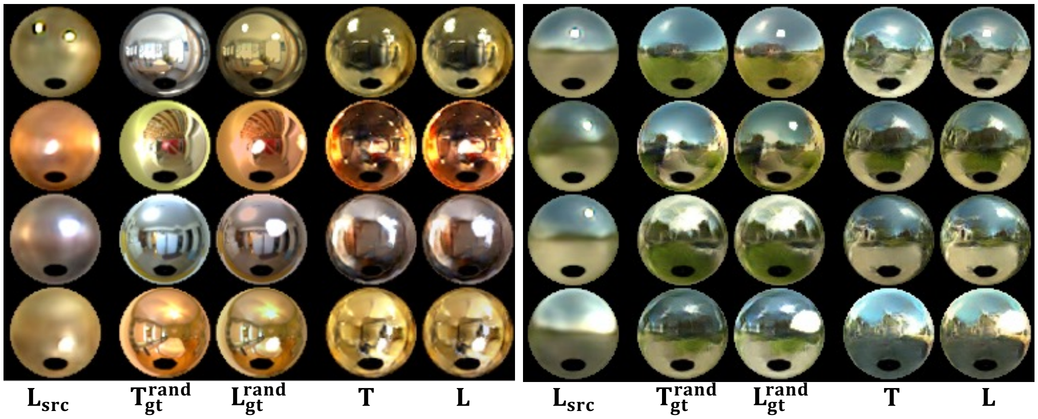

$\mathbf{L}_{\mathbf{src}}$
$\mathbf{T}_{\mathbf{gt}}^{\mathbf{rand}}$
$\mathbf{L}_{\mathbf{gt}}^{\mathbf{rand}}$
$\mathbf{T}$
$\mathbf{L}$
$\mathbf{L}_{\mathbf{src}}$
$\mathbf{T}_{\mathbf{gt}}^{\mathbf{rand}}$
$\mathbf{L}_{\mathbf{gt}}^{\mathbf{rand}}$
$\mathbf{T}$
$\mathbf{L}$

Fig. D. Comparison between filling light source $\mathbf{L}_{\text{src}}$ with texture $\mathbf{T}$ generated by our ambient texture network f_{amb} and with random real texture $\mathbf{T}_{\text{gt}}^{\text{rand}}$ from randomly picked real panorama. The left part shows indoor scenes, while the right part shows outdoor scenes.

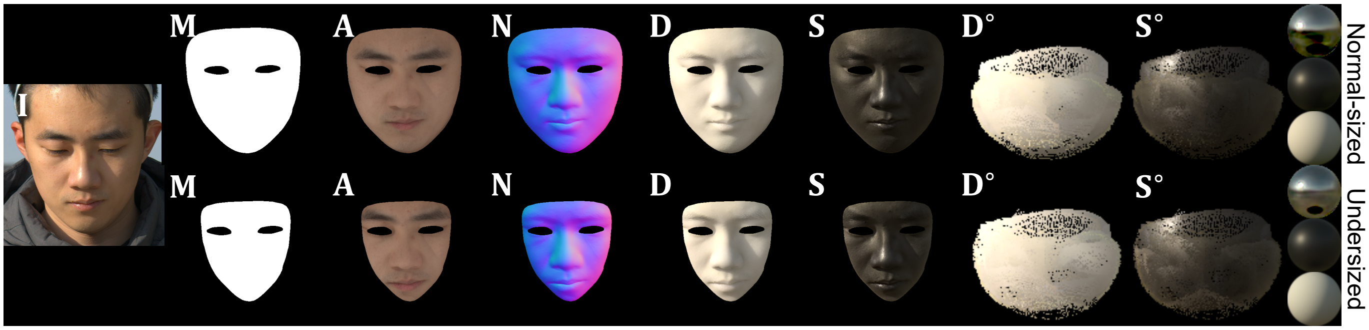


Fig. E. We show the robustness of SPLiT to undersized input masks. Compared to the results estimated using normal-sized masks as an input (upper row), in the results using the undersized mask (lower row), although the estimated face intrinsics $\{\mathbf{A}, \mathbf{N}, \mathbf{D}, \mathbf{S}\}$ are shrunk accordingly, the distributed intrinsics $\{\mathbf{D}^{\circ}, \mathbf{S}^{\circ}\}$ are barely affected. Therefore, the estimated lighting (w/o f_{amb}) is also unaffected.

the sky (rows 1, 2); light sources and textures for sunny and cloudy days are sometimes mistakenly interleaved (rows 3, 4). These phenomena are not that obvious in indoor scenes (the left half of Fig. D), but still render the occurrence of the light sources more abrupt and unnatural.

D Robustness to undersized masks

One limitation of our method is that it depends on a face mask and will produce false light sources if the mask is oversized and bright background pixels are recognized as strong specularities (*i.e.*, background leaking into the estimated light sources). However, we show in Fig. E that our method is robust to undersized masks, thus the above leaking problem can be automatically eliminated by aggressively eroding the mask.

References

[1] G. Somanath and D. Kurz, "HDR environment map estimation for real-time augmented reality," in *Proc. of Computer Vision and Pattern Recognition*, 2021.

[2] G. Wang, Y. Yang, C. C. Loy, and Z. Liu, "StyleLight: HDR panorama generation for lighting estimation and editing," in *Proc. of European Conference on Computer Vision*, 2022.

[3] H. Weber, M. Garon, and J. Lalonde, "Editable indoor lighting estimation," in *Proc. of European Conference on Computer Vision*, 2022.

[4] M. R. K. Dastjerdi, Y. Hold-Geoffroy, J. Eisenmann, and J. Lalonde, "Everlight: Indoor-outdoor editable HDR lighting estimation," in *Proc. of International Conference on Computer Vision*, 2023.

[5] K. Tateno, N. Navab, and F. Tombari, "Distortion-aware convolutional filters for dense prediction in panoramic images," in *Proc. of European Conference on Computer Vision*, 2018.

[6] Y. Hold-Geoffroy, A. Athawale, and J. Lalonde, "Deep sky modeling for single image outdoor lighting estimation," in *Proc. of Computer Vision and Pattern Recognition*, 2019.

[7] M. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambaretto, C. Gagné, and J. Lalonde, "Learning to predict indoor illumination from a single image," in *ACM Transactions on Graphics*, 2017.

[8] C. LeGendre, W. Ma, G. Fyffe, J. Flynn, L. Charbonnel, J. Busch, and P. E. Debevec, "DeepLight: learning illumination for unconstrained mobile mixed reality," in *Proc. of Computer Vision and Pattern Recognition*, 2019.

[9] C. LeGendre, W. Ma, R. Pandey, S. R. Fanello, C. Rhemann, J. Dourgarian, J. Busch, and P. E. Debevec, "Learning illumination from diverse portraits," in *Proc. of ACM SIGGRAPH Asia Technical Communications*, 2020.

[10] P. E. Debevec, P. Graham, J. Busch, and M. T. Bolas, "A single-shot light probe," in *International Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2012, Los Angeles, California, USA, August 5-9, 2012, Talks Proceedings*, 2012.

[11] R. Pandey, S. Orts-Escolano, C. LeGendre, C. Häne, S. Bouaziz, C. Rhemann, P. E. Debevec, and S. R. Fanello, "Total relighting: learning to relight portraits for background replacement," in *ACM Transactions on Graphics*, 2021.

[12] H. Yu, S. Agarwala, C. Herrmann, R. Szeliski, N. Snavely, J. Wu, and D. Sun, "Accidental light probes," in *Proc. of Computer Vision and Pattern Recognition*, 2023.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *Proc. of Interna-*

tional Conference on Medical Image Computing and Computer Assisted Intervention, 2015.

[14] R. Zhang, "Making convolutional networks shift-invariant again," in *Proc. of International Conference on Machine Learning*, 2019.

[15] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proc. of International Conference on Machine Learning*, 2015.

[16] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," in *Neural Networks*, 2018.

[17] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," in *Journal of Machine Learning Research*, 2014.

[18] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proc. of International Conference on Learning Representations*, 2015.

[19] H. Weber, D. Prévost, and J. Lalonde, "Learning to estimate indoor lighting from 3D objects," in *Proc. of International Conference on 3D Vision*, 2018.

[20] A. Sztrajman, A. Neophytou, T. Weyrich, and E. Sommerlade, "High-dynamic-range lighting estimation from face portraits," in *Proc. of International Conference on 3D Vision*, 2020.

[21] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. of International Conference on Machine Learning*, 2010.

[22] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[23] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. of Computer Vision and Pattern Recognition*, 2018.