DOI: 10.1111/cgf.14283

COMPUTER GRAPHICS *forum*
Volume 40 (2021), number 6 pp. 315–331

Deep Neural Models for Illumination Estimation and Relighting: A Survey

Farshad Einabadi, Jean-Yves Guillemaut and Adrian Hilton

Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, England
{f.einabadi, j.guillemaut, a.hilton}@surrey.ac.uk

Abstract

Scene relighting and estimating illumination of a real scene for insertion of virtual objects in a mixed-reality scenario are well-studied challenges in the computer vision and graphics fields. Classical inverse rendering approaches aim to decompose a scene into its orthogonal constituting elements, namely scene geometry, illumination and surface materials, which can later be used for augmented reality or to render new images under novel lighting or viewpoints. Recently, the application of deep neural computing to illumination estimation, relighting and inverse rendering has shown promising results. This contribution aims to bring together in a coherent manner current advances in this conjunction. We examine in detail the attributes of the proposed approaches, presented in three categories: scene illumination estimation, relighting with reflectance-aware scene-specific representations and finally relighting as image-to-image transformations. Each category is concluded with a discussion on the main characteristics of the current methods and possible future trends. We also provide an overview of current publicly available datasets for neural lighting applications.

Keywords: neural relighting, neural illumination estimation

ACM CSS: • Computer graphics → Neural networks

1. Introduction

With the recent surge in augmented and mixed reality applications in games and digital effects, the demand for more plausible lighting has become the centre of attention. Traditionally, trained artists carefully tune the final lighting effects in post-production in an artistic and plausible, but manual and time-consuming manner. [Deb08] suggests recording the scene lighting as high dynamic range (HDR) *light probes* so that it can later be used for harmonizing the appearance of inserted virtual objects to the real scene. This concept has been improved over the years and is still being deployed in various applications. Both manual fine tuning of lights and probing are offline and costly in terms of time and the required equipment. Inverse rendering approaches aim to address these shortcomings by decomposing a scene into its building components, potentially including lighting parameters, from one or a few images. However, this has proved to be a challenge considering the ill-posed nature and the dimensionality of the problem formulation. Normally, simplifying priors are therefore imposed in terms of specific lighting or material models.

Deep neural models have recently shown promising results in various computer vision and graphics tasks, among others neural rendering [TFT*20], neural inverse rendering [YS19], scene representation [SMT*20], relighting [LMF*19], etc. Despite being in its infancy, deep learning is believed by the community to be naturally suitable for the inherent high-dimensionality of visual computing problems. For example, Murmann et al. [MGAD19] show a deep autoencoder without any domain knowledge in graphics can potentially relight an indoor scene to a desired target illumination. However, the main challenge here remains to obtain training data at scale and diversity with appropriate metadata – as regards to data labelling, self-supervision has been explored in the literature [SGK*19]. Nevertheless, many recent efforts aim to resolve this matter by generating realistic looking, large, synthetic datasets [PGZ*19], HZBS20]. In this case, the intrinsic issue is the generalization ability of the trained models to unseen real scenes.

This article specifically focuses on lighting estimation, harmonization and relighting applications in conjunction with recent advances in deep neural computing. We have categorized and precisely compared the attributes of the recent contributions. The rest of this section will provide more details on the scope and respectively the structure of this review and the comparison terminology. This report concludes with a discussion on current and future research trends in this field.

Figure 1: *Examples of relighting and illumination estimation for compositing virtual objects using deep neural networks. In order from left to right images from [GHS*19, GSH*19, PGZ*19, ZFT*20, HZBS20, PPYW20]*

1.1. Scope of the review

The focus of this review is on neural lighting applications and therefore it does not survey neural rendering approaches per se. Refer to Tewari et al. [TFT*20] for a comprehensive state-of-the-art on neural rendering and its applications. Also, neural geometry reconstruction [PHC*19, ZYW*19] and similarly neural material reflectance modelling [GLD*19a, KCW*18] are out of scope of this review.

1.2. Structure of the review

This review is structured in four main sections. Section 2 addresses the problem of scene illumination estimation. More specifically, the lighting parameters of a scene are estimated from one or more input images and then potentially used to render augmented virtual objects in a mixed reality scenario. Neural inverse rendering techniques are also included in this section only when they explicitly estimate the parameters of a lighting model. Here we further categorize the contributions based on their use case, that is the scene content for general scenes (indoor and/or outdoor) or scenes with specific contents, for example human faces, certain object(s), etc.

Section 3 and Section 4 both address the problem of scene relighting. Section 3 covers different approaches that model and store *scene-specific* reflectance fields which are used for rendering a scene under novel unseen lightings via *image-based* methods, or novel unseen lightings and viewpoints via *geometry-aware* methods.

The relighting approaches presented in Section 4 are not scene-specific and unlike the inverse rendering methods, do not decompose a scene into its constituting elements to facilitate relighting. On the other hand, they rely on direct *image-to-image* transformations to relight a scene under a desired target lighting condition. As a result, these approaches do not necessarily explicitly estimate the lighting parameters or model the scene reflectance field. A further partitioning is considered in this section based on the scene content, and whether the approaches address either or both of the shading and shadow harmonization sub-problems.

Furthermore, we provide an overview of publicly available datasets for the purpose of scene illumination estimation and relighting in Section 5.

1.3. Comparison terminology

For comparing the attributes of the examined contributions, we adopt the terminology of the neural rendering review by Tewari et al. [TFT*20] with some modifications and extensions. Below is a summary of the attributes and their possible values.

- Required Data, Network Input and Network Output: LDR image (I), HDR image (H), LDR panorama (P), HDR panorama (N), stereo images (S), colour gradient images (T), binary mask (B), video (V), lighting (L), depth (D), 3D information (G), for example mesh, normals, face models, volumetric scene representations, etc., material (M), camera (C), locale (E)
- Scene Content: general indoor (I), general outdoor (O), human face (F), human hands (H), human body or garment (B), single or multiple objects (J)
- Lighting Model: dominant direction (D), distance (T), colour (C), HDR illumination map (M), spherical harmonics (H) [RH01], spherical Gaussian (G) [LSR*20], Hosek and Wilkie sky model (S) [HW12], Lalonde-Matthews outdoor illumination model (O) [LM14], reflection maps (R) [MH84], reflectance fields (F), volumetric RGBα representation (V)
- Spatially Varying Lighting Effects
- Explicit Controllable Lighting
- Temporal Coherency
- Computer Graphics Module: differentiable (D), non-differentiable (N)
- Training Dataset: real or based on real data (R), synthetic (S)
- Generality

Possible attribute values are extended compared to Tewari et al. [TFT*20] to include more details. More specifically, as this article is concerned with lighting, we extend the original terminology by including the used or inferred lighting model, and whether or not it is modelling spatially varying effects. The lighting model is not necessarily an immediate output of the neural model. Furthermore, we report if a training dataset is real or synthetic. If images are warped or cropped from a real dataset, they are still considered real. Synthetic datasets, on the other hand, benefit from real or synthetic scene representations to render new content.

Modifications to the comparison terminology of Tewari et al. [TFT*20] are: (1) *Required Data* is only for the training phase of the system. Note that *All* required data can still be derived by considering both the *Required Data* and the *Network Input*; (2) *Temporal Coherency* is not explicitly enforced in the network design; (3) Explicit Controllable Lighting cannot be defined for some contributions as they do not directly address neural rendering, but illumination estimation.

Table 1: *Surveyed methods compared based on the terminology of Section 1.3.*

	Required data	Network input	Network output	Scene content	Lighting model	Spatially varying	Controllable lighting	Temporal coherency	CG module	Training dataset	Generality
Bi et al. [BXS*20a]	I	E	LG	J	F	√	√	✗	D	R	✗
Boss et al. [BJK*20]	IBLDGM	IB	LGM	J	G	✗	–	✗	D	S	√
Calian et al. [CLG*18]	ING	I	L	F	M	✗	–	✗	✗	RS	√
Chalmers et al. [CZMR20]	PN	I	L	IO	MR	✗	–	✗	✗	R	√
Cheng et al. [CSC*18]	N	I	L	IO	H	✗	–	✗	✗	R	√
Gardner et al. [GSY*17]	PN	I	L	I	M	✗	–	✗	✗	R	√
Gardner et al. [GHS*19]	NLD	I	L	I	DTC	√	–	✗	✗	R	√
Garon et al. [GSH*19]	ND	IE	LD	I	H	√	–	✗	✗	R	√
Hold-Geoffroy et al. [HSH*17]	P	I	L	O	S	✗	–	✗	✗	R	√
Hold-Geoffroy et al. [HGAL19]	PN	I	L	O	M	✗	–	✗	✗	R	√
Jin et al. [JDL*20]	PL	I	L	O	O	✗	–	✗	✗	R	√
Kán and Kafumann [KK19]	ID	I	L	IJ	D	✗	–	√	✗	S	√
Kanamori and Endo [KE18]	IBLM	IB	LM	B	H	√	–	✗	N	S	√
LeGendre et al. [LMF*19]	VL	I	L	IO	M	✗	–	√	D	R	√
Li et al. [LGC*19]	P	I	L	I	G	✗	–	✗	✗	S	√
Li et al. [LSR*20]	ILDGM	I	LDGM	I	G	√	–	✗	D	S	√
Liu et al. [LLQX20]	NIGM	I	LGM	J	H	✗	–	✗	D	R	√
Liu et al. [LLZ*20]	IB	IB	IB	J	–	–	–	✗	N	S	√
Marques et al. [MCV18]	IL	I	L	H	H	✗	–	✗	✗	S	√
Meka et al. [MHP*19]	ITL	TL	I	F	D	✗	√	✗	✗	R	√
Murmann et al. [MGAD19]	H	H	H	I	–	–	✗	✗	✗	R	√
Park et al. [PPYW20]	IPNL	I	L	J	M	✗	–	✗	N	RS	√
Philip et al. [PGZ*19]	IL	IL	I	O	D	✗	√	✗	N	S	√
Reddy et al. [RTO*20]	ILGC	ILGC	I	F	D	–	√	✗	D	R	√
Sengupta et al. [SGK*19]	ILGM	I	LGM	I	M	✗	–	✗	D	RS	√
Sheng et al. [SZB20]	LG	LB	I	J	–	–	✗	✗	✗	S	√
Sial et al. [SBVS20]	IL	I	L	IJ	DC	✗	–	✗	✗	S	√
Song and Funkhouser [SF19]	PND	IE	L	I	M	√	–	✗	D	R	√
Srinivasan et al. [SMT*20]	IPSC	S	LG	I	V	√	✗	✗	D	S	✗
Sun et al. [SBT*19]	IL	IL	IL	F	M	✗	√	✗	✗	R	√
Sun et al. [SLL*20]	P	I	L	IO	H	✗	–	✗	✗	R	√
Wang et al. [WWL19]	I	I	I	I	–	–	–	✗	✗	S	√
Wang et al. [WSL*20]	IL	I	I	I	–	–	–	✗	N	S	√
Weber et al. [WPL18]	ING	IG	L	J	M	✗	–	✗	✗	RS	√
Wei et al. [WCD*20]	IND	IG	LM	J	M	✗	–	✗	N	S	√
Xu et al. [XSHR18]	IL	IL	I	J	D	✗	√	✗	✗	S	√
Xu et al. [XLZ20]	N	I	L	IO	H	✗	–	✗	✗	R	√
Yi et al. [YZTL18]	ING	I	L	F	M	✗	–	✗	✗	RS	√
Yu and Smith [YS19]	IN	I	LGM	O	H	✗	–	✗	D	R	√
Yu et al. [YME*20]	I	I	LGM	O	M	✗	√	✗	D	R	√
Zhan et al. [ZLZ*20]	IB	IB	IL	J	H	✗	–	✗	N	R	√
Zhan et al. [ZYW*21]	ND	I	L	I	DTC	√	–	✗	✗	R	√
Zhang et al. [ZSHG*19]	PN	I	L	O	O	✗	–	✗	✗	R	√
Zhang et al. [ZFT*20]	ILGC	ILGC	I	B	F	√	√	✗	D	R	✗
Zhou et al. [ZHSJ19]	IL	IL	IL	F	H	✗	√	✗	✗	S	√
Zhu et al. [ZHZ*20]	IBLGM	IB	LGM	B	H	✗	–	✗	N	S	√

Table 1 summarizes the key differences between the methods covered in this review based on the above criteria. Regarding the attribute values, note that *Not Applicable* is denoted with –, and *Yes* and *No* respectively with √ and ✗.

2. Illumination Estimation for Mixed Reality

The methods in this section are divided based on their assumptions on the scene content into two groups: *general* scenes with no explicit assumption on the scene content, except the indoor or outdoor constraint; and those assuming a specific content present in the scene such as human face(s), hand(s) or other certain objects of known geometry and/or material. Note that neural inverse rendering is also addressed in this section, as it too estimates illumination.

2.1. General scenes

Approaches presented in this section make no explicit assumptions on the scene content, geometry or appearance, aside from the indoor or outdoor constraint.

2.1.1. *Indoor scenes*

Gardner et al. [GSY*17] were the first to estimate an HDR illumination map from a single limited field of view (FOV) low dynamic range (LDR) image, “in the wild,” for indoor environments. There are no explicit assumptions on the scene geometry, materials or the illumination. The proposed approach consists of a two-stage trainable deep encoder-decoder network with two decoder heads – one predicting the light direction/intensity and the other one predicting an RGB panorama. In the first stage of training, light directions are learned from LDR panoramic images of the SUN360 dataset [XEOT12]. The second stage refines the network to predict the HDR intensities using a newly recorded HDR illumination map dataset consisting of 2100 images: the Laval Indoor HDR panorama dataset.

Sial et al. [SBVS20] estimate the colour and direction (pan/tilt) of a single dominant light source from an input LDR image of an indoor scene using a simple DNN. For training, a synthetic dataset is generated based on the synthetic dataset of Sial et al. [SBV20]. Experiments show the practicality of the approach, both quantitatively and qualitatively, on synthetic and real images (dataset of Murmann et al. [MGAD19]).

Li et al. [LGC*19] estimate the parameters of a spherical Gaussian lighting model (128 parameters) using a simple CNN – given as input an LDR image of an indoor scene. The synthetic training data is generated based on the SUN360 dataset [XEOT12]. The network is trained both with a lighting parameters loss, and a new *glossy* loss in order to maintain the high frequency features of lighting during training. Experiments show improvements to [GSY*17] in terms of shading effects; shadows are not compared.

Kán and Kafumann [KK19] also estimate a dominant light direction with a DNN but relative to the camera viewpoint and in a live augmented reality (AR) system. The proposed network is based on ResNet [HZRS16] and takes an RGB-D image as input. A filtering method motivated by Dante and Brookes [DB03] is applied on the predictions of each frame to remove outliers and smooth the light direction. This leads to a temporally coherent lighting estimation in AR applications.

2.1.2. *Indoor scenes: modelling spatially varying illumination*

Modelling indoor illumination with only one illumination map is inherently limited and cannot cover spatially varying localized lighting effects due to near light sources or nearby geometries. Gardner et al. [GSY*17] seek to address this problem by warping the LDR/HDR illumination maps to be suitable for the exact position of the scene where the virtual object is going to be composited. Yet, Gardner et al. [GHS*19] show with a user study that their proposed approach, which regresses geometric and photometric properties of discrete light sources in an indoor scene, leads to more plausible composites of virtual objects.

Gardner et al. [GHS*19] train a DNN to predict the direction, distance, size, and colour of a pre-defined number of light sources from a single LDR input image. To do so, the Laval Indoor HDR panorama dataset [GSY*17] is further manually labelled for four types of predefined discrete light sources and pixel-wise depth information. This labelled dataset is then employed for the training of the proposed neural model.

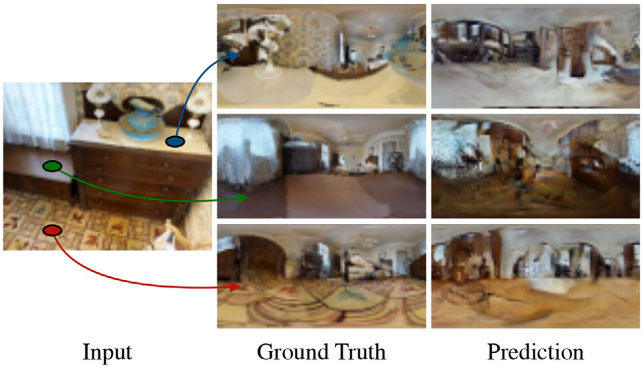

Figure 2: *Song and Funkhouser [SF19] predict separate HDR illumination maps for each locale. Images taken from Song and Funkhouser [SF19].*

The inherent problem with a discrete light source model is its inability to represent realistic global illumination effects. It is however generating spatially consistent lighting effects when compositing virtual objects into the scene.

Song and Funkhouser [SF19] also address the issue of lack of modelling spatially varying illumination in indoor scenes by Gardner et al. [GSY*17]. However, in contrast to Gardner et al. [GHS*19], they seek to regress a separate illumination map for any selected point in the visible scene (see Figure 2). The task is challenging in nature but Song and Funkhouser [SF19] propose to tackle it in multiple steps. In the first step, partial scene geometry is estimated from an input LDR image which is then used to warp the image to a partial LDR illumination map – Gardner et al. [GSY*17] assume a sphere for warping due to lack of geometry. The partial LDR illumination map is completed by another network, and yet transformed to HDR by another one. Inferred HDR illumination maps by Song and Funkhouser [SF19] have much more details compared to [GSY*17] and are *locale-aware*. To train the proposed model, a dataset is generated based on the high quality RGB-D and HDR panoramas of Matterport3D dataset [CDF*17].

In the same direction, Garon et al. [GSH*19] infer instead the parameters of a fifth order spherical harmonics (SH) illumination model given an LDR image and a selected point in the scene, therefore also locale-aware. The proposed neural model has two paths: local and global. Extracted features from each path are concatenated to regress the SH model parameters (similar to Cheng et al. [CSC*18]), among others. The training dataset is generated synthetically based on the SUNCG depth dataset [SYZ*17]. In addition, 20 more scenes are recorded with 79 spatially varying HDR light probes for testing. Relighting/compositing results show – as expected – substantial improvements to Gardner et al. [GSY*17] regarding modelling local changes in illumination.

GMLight by Zhan et al. [ZYW*21] also regresses a locale-aware, parametric light model of an indoor scene. A proposed encoder estimates the distribution, depth and the intensity of individual light sources as well as an ambient term, given as input a single LDR image. These parameters are then used in a generative scheme to render illumination maps for the specific locations of the scene. An innovative *geometric mover's* loss, "inspired by earth mover's distance," is introduced to train the proposed system (the encoder and the generator) in an end-to-end manner. The Laval Indoor HDR panorama dataset [GSY*17] is used for the training. Compared to the work of Gardner et al. [GHS*19] and Garon et al. [GSH*19], experiments show that the generated illumination maps have more details for different locations in the scenes, and are therefore more suitable for the realistic insertion of virtual objects.

Very similar to GMLight [ZYW*21], EMLight by Zhan et al. [ZZY*20] also predicts a parametric lighting model of an indoor scene but it fails to generate locally-correct illumination maps due to the lack of depth information in the proposed model.

Lighthouse [SMT*20] shows, by a compositing and relighting experiment, improvements over previous methods [GHS*19, SF19, GSH*19]. The key idea is to generate a volumetric scene representation from a narrow-baseline stereo camera such as those available on the rear side of a mobile phone. For more details, see Section 3.2.

2.1.3. *Outdoor scenes*

Hold-Geoffroy et al. [HSH*17] propose a DNN to estimate parametric sun-sky model of Hosek and Wilkie [HW12] – specifically the sun position, skylight exposure factor and turbidity (amount of atmospheric aerosols) – given a single input LDR image. Virtual objects can thereby be composited into the real scene with an illumination map generated based on these parameters. Training is done based on the estimated parameters of the model of Hosek and Wilkie [HW12] on the outdoor panoramas of the LDR SUN360 dataset [XEOT12]. The proposed approach fails in predicting outdoor lighting in cases of lack of illumination cues in the input image, or co-existence of specular surfaces or complex geometries [HSH*17].

In a very similar approach, Zhang et al. [ZSHG*19] base their work on the Lalonde-Matthew outdoor illumination model [LM14] which has separate sun and sky parameters. They propose a deep network, dubbed CropNet, to regress the illumination parameters for crops of the LDR SUN360 dataset [XEOT12]. For training of CropNet, another DNN (PanoNet) is employed to generate ground truth data (here parameters of Lalonde-Matthew model) for the input LDR panoramas of the SUN360 dataset. Experimental compositing results show improvements to Hold-Geoffroy et al. [HSH*17] for the claimed "all-weather" conditions. However, soft shadows are not handled gracefully with this illumination model. The approach proposed by Jin et al. [JDL*20] is very similar to that of Zhang et al. [ZSHG*19].

As opposed to the above parametric outdoor illumination models, SkyNet [HGAL19] aims to estimate a *non-parametric* HDR illumination map from a single input LDR outdoor image using a three-stage learning method. The key difference to the low-dimensional parametric models [HSH*17], ZSHG*19, JDL*20] is that an illumination map can potentially model more sky lighting conditions such as partially cloudy or fully overcast [HGAL19]. The proposed neural system is trained on the Laval HDR sky database [LM14] and SUN360 LDR outdoor panoramas [XEOT12]. Hold-Geoffroy et al. [HGAL19] also record a new HDR outdoor panorama dataset for evaluating their approach. The evaluation results show significant improvements compared to [HSH*17] regarding the shading of the composited objects into the outdoor scenes; however, the proposed approach fails to regenerate the hard shadows of the parametric model of [HSH*17]. Hold-Geoffroy et al. [HGAL19] also report lack of texture in the estimated sky illumination maps.

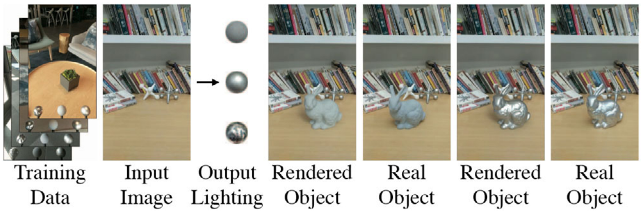

Figure 3: *LeGendre et al. [LMF*19] augment virtual objects into a real scene. Images taken from LeGendre et al. [LMF*19].*

2.1.4. *Indoor or outdoor scenes*

A common application area for light estimation in general indoor or outdoor scenarios is mobile AR. We cover a number of these contributions below.

Cheng et al. [CSC*18] propose a novel mobile AR solution that uses both front and rear cameras of a mobile device to regress the parameters of a forth order SH illumination model. The proposed model consists of two feature-extraction convolutional neural networks (CNNs), respectively for the two input images, and an illumination estimation CNN that takes as input the concatenated feature maps of both images. A dataset of 200 HDR illumination maps are recorded for training and testing. Experiments show improvements to both [GSY*17] and [HSH*17] for indoor and outdoor scenes, respectively. Also experimental results show the benefit of using a rendering-based loss in addition to the loss computed on SH parameters.

Similar to [CSC*18], Sun et al. [SLL*20] and Xu et al. [XLZ20] also estimate the parameters of an SH lighting model for indoor/outdoor scenarios using a CNN trained with a rendering-based loss. The approach of Xu et al. [XLZ20] is tested for real-time relighting on mobile devices. Sun et al. [SLL*20] introduce a new warping method similar to [GSY*17] to compensate for local changes in illumination in indoor scenes.

LeGendre et al. [LMF*19] propose a mobile AR solution for relighting in both indoor and outdoor scenarios. The proposed approach entails a DNN which is trained with a rendering-based loss to predict an HDR illumination map from a single input LDR image (see Figure 3). The capture setup consists of a mobile phone and three light probe spheres with varying reflectance properties (visible in the FOV of camera) assembled on a rig – which is then used to record the training dataset. The results show improvements in compositing virtual objects compared to [GSY*17] and [HSH*17] in the case of indoor and outdoor scenes, respectively. The known

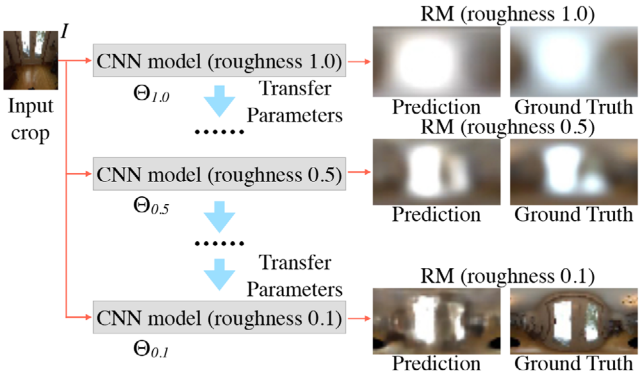

Figure 4: *Chalmers et al. [CZMR20] utilise stacked CNNs to predict reflection maps. Diagram content and images taken from Chalmers et al. [CZMR20].*

limitations are however (1) inability to model the spatially varying illumination in indoor scenes and (2) lack of generalization to the input images captured with other cameras with regard to white balancing and FOV.

For improving the performance of compositing in indoor/outdoor AR, Chalmers et al. [CZMR20] propose to model environment lighting as Reflection Maps (RMs) [MH84] – pre-lit lookup tables for objects with different material roughness. A novel multi-level *stacked* CNN is designed and trained to predict RMs with specific roughness at each level (see Figure 4). The outdoor training dataset is generated synthetically based on LDR panoramas from the SUN360 database [XEOT12] – after conversion to HDR. Extensive experiments show comparable augmentation results to LeGendre et al. [LMF*19]. The main disadvantage of such an approach is however the dependence on a material model (here Phong [Pho75]) with a roughness parameter.

2.2. Scenes including known object(s)

To estimate scene lighting, methods presented in this section assume certain contents with known geometry and/or reflectance properties are directly visible in the captured images, for example human face(s), hand(s), or other objects.

2.2.1. *Human face*

Calian et al. [CLG*18] propose a neural model to estimate a parametric model or a high frequency HDR spherical map of outdoor illumination from a single LDR image of a human face. They assume a high quality 3D reconstruction of the face – for detecting self-shadowing. A neural model is trained to regress incident illumination and face albedo parameters – with strong priors imposed on both to achieve realistic results. They evaluate their approach on a novel synthetic and a recorded real dataset. The experiments show the superiority of the regressed HDR illumination maps in comparison to parametric lighting models such as SH or Lalonde-Matthew [LM14] for augmentation of virtual objects into the scene.

Yi et al. [YZTL18] also estimate a non-parametric illumination map from a single image of a human face (similar to [CLG*18]). However, they benefit from a deep neural model to regress highlight areas of the face which can be traced back to the scene to form an illumination map. These maps are then further refined to remove the blurring effect caused by the diffuse properties of human skin. Experimental results show improvement for compositing compared to [GSY*17] in the case of indoor scenes and [HSH*17] for outdoor environments. Yi et al. [YZTL18] show that the extra capability of this approach compared to the previous work [CLG*18, LMF*19, GSY*17, HSH*17, GZA*20] is that if there are multiple faces in a scene, one can potentially triangulate for the light sources from multiple inferred illumination maps.

2.2.2. *Human hands*

Marques et al. [MCV18] address the problem of realistic virtual scene and hand relighting in mixed reality games. The proposed approach assumes that the user's hands are visible in the scene and convey enough information from the surrounding lighting. The segmented user's hands are fed to a trained deep CNN to regress the parameters of a second order SH lighting model. The CNN is trained on a synthetically generated hands dataset. Experiments show improvements over the previous work of Marques et al. [MDVC18] which is limited to estimating only a single point light source from the same input images.

2.2.3. *Other objects*

Weber et al. [WPL18] present a neural model to estimate an HDR illumination map of a scene from a single visible object with known geometry and reflectance. This inherently imposes limitations in data preparation and training phases. The neural model of [WPL18] consists of a deep autoencoder and a deep CNN. The autoencoder learns the latent space of an HDR illumination map, while the CNN transforms an LDR input image to the autoencoder latent space – similar to the “T-Network” architecture of Girdhar et al. [GFRG16]. To overcome the lack of training data for the autoencoder network, [WPL18] generates 21,000 novel HDR illumination maps. A subset of HDR illumination maps from the Laval Indoor HDR panorama dataset [GSY*17] are manually labelled for illumination surfaces and their depth using EnvyDepth [BCD*13]. The CNN is trained with pairs of synthetic renderings of five selected 3D models with three different materials under random illumination maps. Experiments show sharper illumination maps compared to the parametric SH light model.

Park et al. [PPYW20] propose a step-by-step neural approach to estimate an HDR illumination map from a single LDR image of a scene, as long as it contains an object with homogeneous (uniform) material with no dominant concavity. The presented method first infers an average incident illumination map given the cropped image of an object using an importance sampling technique. It therefore factors out the effects of the shape and material for the next stages of the method. The second and third stages respectively regress LDR and HDR illumination maps of the scene using two additional neural models. The proposed approach shows improvements in rendering the appearance of augmented virtual objects over Meka et al.

[MMZ*18] and Georgoulis et al. [GRR*18] for both indoor and outdoor scenarios. The neural models are trained with a proposed publicly available synthetic dataset.

2.3. Neural inverse rendering

Methods presented in this section, while estimating scene attributes such as reflectance or shape, reproduce lighting parameters as a by-product.

2.3.1. *Indoor scenes*

The neural model of Sengupta et al. [SGK*19] estimates lighting, albedo and normals of an indoor scene from a single input image. Their main contribution is a *differentiable residual appearance* renderer which reconstructs high-frequency scene details such as shadows and inter-reflections from the aforementioned scene components – to facilitate a self-supervised training mechanism. Another contribution is a photorealistic, synthetic dataset based on SUNCG [SYZ*17] which improves the dataset of Zhang et al. [ZSY*17] in terms of natural lighting, materials and denoised renderings. With regard to lighting specifically, experiments show quantitative improvement over the work of Gardner et al. [GSY*17] for the predicted low-frequency HDR illumination maps over synthetic and real data.

Li et al. [LSR*20] employ a cascade neural model consisting of two autoencoder networks, one for material and geometry and the other one for lighting, plus a *differentiable* renderer and a refinement module. For indoor scenes, the proposed approach can regress parameters of both spatially varying lighting and BRDF models. In contrast, Sengupta et al. [SGK*19] use the Phong [Pho75] model. In particular, lighting is estimated as parameters of spatially varying spherical Gaussian lobes which, as shown in the experiments, can handle high-frequency lighting effects on objects much better than the fourth order SH parameters. In terms of compositing virtual objects into real scenes, comparisons show more plausible shading and shadows compared to [GSH*19] and [GSY*17]. Note that [GSY*17] does not model spatially varying lighting. The main contribution of [LSR*20] is a synthetically, *GPU* generated dataset of images with global illumination and various complex spatially varying BRDFs based on scene geometries from the SUNCG dataset [SYZ*17].

2.3.2. *Outdoor scenes*

InverseRenderNet by Yu and Smith [YS19] addresses the same problem formulation but in arbitrary outdoor scenes. The key idea is to use multiview stereo (MVS) based on an available dataset [LS18b] to supervise neural inverse rendering training in a Siamese, self-supervised fashion. The selected lighting model is a second order SH which is acceptable for modelling low frequency properties of lighting. Figure 5 shows the extracted scene elements from a single input image and then used for re-rendering the shading.

Yu et al. [YME*20] improve InverseRenderNet [YS19] with an HDR illumination map as illumination model instead of SH. In addition, foreground monuments are segmented using a novel sky detection DNN. Also, a new outdoor relighting dataset is recorded consisting of 10 views and six lighting directions for each scene. Experiments show plausible relighting of outdoor scenes in general and specifically more reconstruction details compared to InverseRenderNet [YS19].

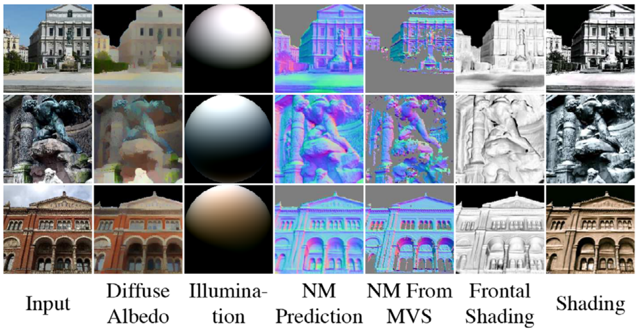

Figure 5: *InverseRenderNet [YS19] infers scene elements from a single image. Images taken from Yu and Smith [YS19].*

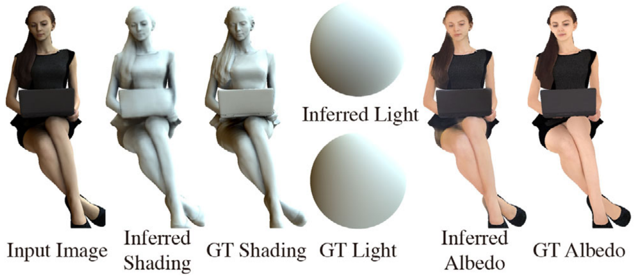

Figure 6: *Kanamori and Endo [KE18] perform inverse rendering even for unseen sitting poses. Images taken from Kanamori and Endo [KE18]*

2.3.3. *Clothing or fashion*

Kanamori and Endo [KE18] estimate illumination, shape and albedo from a single image of a posed human (see Figure 6). Potential applications are relighting or lighting transfer. The main idea is based on the precomputed radiance transfer work of Sloan et al. [SKS02] and second order SH based lighting to infer light occlusions on body parts or apparel. The network consists of an encoder with two separate decoders, one for albedo estimation and the other one for light transport estimation. The training is achieved end-to-end with synthetic data. Compared to the state-of-the-art, this work produces plausible (darker) shading for concave parts of human body or clothing wrinkles which otherwise would have been rendered much brighter. This is due to the consideration of occlusion in the estimated light transfer maps.

Zhu et al. [ZHZ*20] address the inverse rendering problem for complex-textured fashion images. A generative adversarial concept [GPAM*14] is employed on a *local* and a *global* level – to regress complex local shading variations while avoiding artifacts caused by neglecting global illumination conditions. The proposed network is

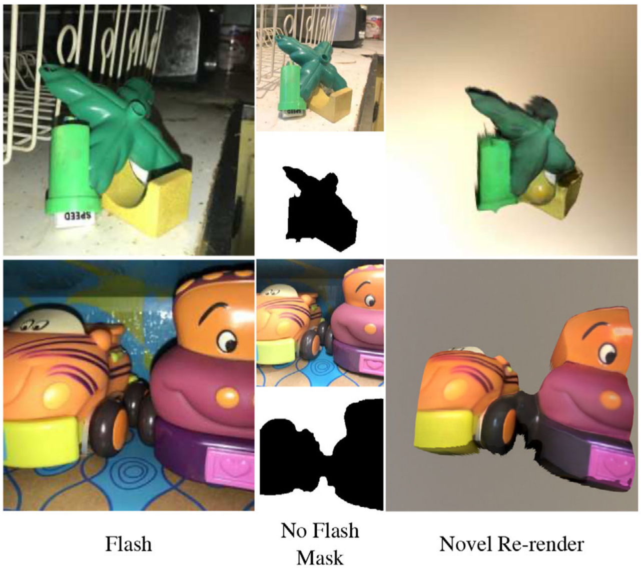

Figure 7: *Boss et al. [BJK*20] re-render objects from two input images (with and without flash light) and their respective masks. Images taken from Boss et al. [BJK*20]*

trained on a synthetically generated dataset. Experiments show the suitability of the approach for texture replacement on fashion images.

2.3.4. *Single or multiple object(s)*

Boss et al. [BJK*20] focus only on single objects and propose a *two-shot* capture method (with and without flash) to regress illumination, shape and spatially varying BRDF. The environment illumination is modelled with magnitudes of a spherical Gaussian model. The network architecture follows a cascade concept (similar to Li et al. [LSR*20]), that is regressed shape is used for illumination estimation and thereafter both shape and illumination are used for the reflectance model. The network is trained on a large number of procedurally generated geometries rendered with complex materials and HDR illumination maps. Figure 7 shows sample results.

Liu et al. [LLQX20] also address inverse rendering from single or multiple objects on planar surfaces. The lighting model is a fifth order SH and a novel "differentiable screen-space rendering" method is proposed for a self-supervised training phase. Object normals, albedo, roughness and the respective lighting parameters are regressed in a conventional encoder-decoder scheme. The model is trained with synthetically generated images based on the Matterport3D dataset [CDF*17].

Wei et al. [WCD*20] propose a neural model to regress an illumination map lighting model from the existing objects' appearance and depth information. A key idea is to mirror and project the highlight regions of an intermediate regressed specular shading to the illumination map through a network with *angular* convolution layers. The mirroring is facilitated by an estimated normal map given the local depth information. The network is trained on synthetically rendered scenes of 600 *artist* modelled objects lit under real and synthetic HDR illumination maps. The experiments show the generalization ability of the network in regressing the illumination map for synthetic and real scenes with various lighting conditions such as a dominant light, multiple lights, area lights, etc. Also, a comparison to LeGendre et al. [LMF*19] and Gardner et al. [GSY*17] shows a more precise reconstruction of the specularities for an inserted virtual object.

Meka et al. [MMZ*18] do not explicitly estimate illumination. They present a neural architecture consisting of seven DNNs which learn separate diffuse and specular albedo of an object from a single input image in indoor environments (without manual segmentation). This method can potentially estimate an incident illumination map (diffuse and specular) only if depth information is provided. Depth information facilitates the mapping of inferred (diffuse and mirror-like) illumination components onto a spherical map. The proposed method can be used live in mixed-reality applications.

2.3.5. *Notable considerations*

There are also a few other neural inverse rendering methods that do not estimate lighting parameters, for example Gao et al. [GLD*19a] and Kang et al. [KCW*18] only for spatially varying BRDFs and Li et al. [LXR*18] and Kang et al. [KXH*19] for spatially varying BRDFs and shape. Also, intrinsic image decomposition approaches such as Zhou et al. [ZKE15] only estimate shading and reflectance and therefore no explicit lighting parameters. Similarly, the estimated *reflectance* maps of Rematas et al. [RRF*16] bind lighting and surface reflectance in an inseparable representation. The same is valid for the neural *appearance* maps of Maximov et al. [MRLF19]. The appearance maps [MRLF19] are however view-independent – in contrast to the reflectance maps [RRF*16].

It is also worth mentioning neural rendering approaches that are physically motivated, for example Chen et al. [CCZ*20] and Granskog et al. [GRPN20]. They have therefore lighting parametrization orthogonal to the representations of other scene components. However, they should not be confused to be performing inverse rendering. For the state-of-the-art on neural rendering, the reader is referred to Tewari et al. [TFT*20].

2.4. Discussion

Estimating illumination from a single or a number of input images has been extensively investigated in an inverse rendering formulation. Deep learning not only provides solutions to this classic problem, but also addresses the illumination estimation from input image(s) with little or no explicit domain knowledge in computer vision or graphics. With regard to neural inverse rendering, approaches such as those of Sengupta et al. [SGK*19] and Yu et al. [YME*20] can already estimate all scene components from a single input image. All of the methods proposed in this category have generality, that is they are not instance specific. There is a recent trend to infer spatially varying lighting effects, especially in indoor scenes, where can be significant changes in lighting even in nearby locations [GHS*19].

Some recent work exploits differentiable CG modules, for example differentiable renderers, in their proposed architecture [YS19, SGK*19]. This is beneficial for self-supervised learning with "in the wild" collected data, as generating supervision metadata still remains a challenge.

As neural processing architectures become more affordable on mobile devices, neural illumination estimation methods for more realistic AR rendering is on the rise. However, so far, not enough attention has been given to temporally coherent estimations which would avoid noise and outliers in lighting predictions. Recently, Kán and Kafumann [KK19] apply a filtering mechanism on the output of the neural network to temporally *smooth* the inferred lighting. However, this is not achieved by the network itself. LeGendre et al. [LMF*19] show samples of temporal coherency as an output of their neural model, despite the fact it is not explicitly enforced in the learning optimization. The majority of approaches consider estimation from each input image independently and do not address temporal coherence for image sequences.

3. Relighting with Reflectance-Aware Scene-Specific Representations

Reflectance-aware scene representations model the reflectance properties of a specific scene, either purely based on images or with the consideration of scene geometry. This enables the two different use cases of relighting and novel view synthesis which will be discussed in the following.

3.1. Image-based relighting

Debevec et al. [DHT*00] in their seminal work propose using a light stage setup to sample the reflectance field of a human face for relighting under novel lighting conditions from one of the existing sparse camera viewpoints. Their work is based on the additive behaviour of light transport. However, in addition to the costly setup of a light stage, a dense sampling of a subject with a one-light-at-a-time (OLAT) technique requires a substantial amount of time while the subject is required to remain still. Nevertheless, the proposed approach is able to generate photorealistic renderings of human faces. In addition, OLAT images are often used as ground truth for comparison to other learning-based systems [MHP*19, ZFT*20].

Ren et al. [RDL*15], on the other hand, propose to employ only a small number of images taken by a commodity single-lens reflex camera and therefore alleviate the need for a light stage setup. The main idea is to regress the different parts of a 4D light transport function by different ensembles of neural networks of a proposed design. The system is trained separately on three real, indoor datasets containing scenes with various materials – captured under a fixed camera viewpoint and a moving light. A relighting experiment on these scenes – using the respective trained models – show the ability of the proposed method to regress images of the same viewpoint under new light source positions.

3.2. Geometry-aware relighting and novel view synthesis

Purely image-based rendering approaches [DHT*00, RDL*15] fail to synthesize the scene under novel unseen viewpoints without having a proper notion of geometry. Recent advances [GLD*19b, ZFT*20] leverage explicit pre-acquired geometry, while Bi et al. [BXS*20a] and Srinivasan et al. [SMT*20] introduce a geometry and reflectance-aware scene representation.

The Relightables by Guo et al. [GLD*19b] combine common approaches for volumetric performance capture of humans, for example [PKC*17] with an image-based rendering from a light stage setup to overcome the issue of relightability and photorealism of free-viewpoint video. Their proposed light stage is additionally equipped with high-speed, high-resolution depth sensors to reconstruct precise meshes of the performer from multiple views. The mesh is tracked over time and relit in arbitrary new environments using the two captured images under Colour Gradient illumination similar to [MHP*19]. A learning-based approach is employed for background subtraction of the subjects in the light stage to avoid the colour spill of a green screen. The Relightables suffer from an inability to reconstruct thin geometrical structures, or highly specular or transparent materials.

Zhang et al. [ZFT*20] train a DNN to estimate the 6D light transport function of a full human body for not only relighting but also novel view synthesis of a subject in a light stage setup. Their approach reconstructs a high quality base mesh of each subject similar to Guo et al. [GLD*19b]. In addition, a base diffuse image is created by combining all available OLAT images of the subject. The proposed network then infers only the non-diffuse residuals to the base diffuse reconstruction as cosine maps [ZFT*20], given a target lighting and viewpoint. Compared to Guo et al. [GLD*19b], the benefit of such an approach is its tolerance to imperfectly reconstructed geometry for, for example hair, and having no parametric prior (cosine lobes in [GLD*19b]) for the reflectance model of the scene.

Neural Reflectance Fields [BXS*20a] provide a general scene representation that tackles the problems of free-viewpoint synthesis and relighting. The work formulates a reflectance-aware ray marching framework similar to Bi et al. [BXS*20b]. It is equipped with a Deep Multilayer Perceptron (MLP) to regress the volume density, normal and parameters of a *differentiable* reflectance model at each shading point on a camera ray. In contrast to [BXS*20b] which has a fixed step size ray marching, volume density allows adaptive sampling of shading points on a camera ray. In contrast to light stage setups [DHT*00, MHP*19, GLD*19b, ZFT*20] which are costly, the MLP is trained with a few images captured from a scene with a collocated camera-light setup – here a mobile phone with flash photography. The experimental results show photorealistic renderings of different scenes such as toys, human faces and furry objects under arbitrary unseen lighting and viewpoints.

Similar to [BXS*20a], Lighthouse by Srinivasan et al. [SMT*20] also generates a "multiscale volumetric lighting representation" of a scene, however from only a pair of narrow-baseline stereo images. This representation is later used to infer local spherical illumination maps for relighting purposes. The proposed neural model consists of two networks to predict visible scene geometry, as well as unobserved scene contents. The training dataset however does not include depth data. A key benefit of having such a scene representation is that the rendered local illumination maps are readily coherent for all scene positions.

Gao et al. [GCD*20] also leverage hand-held *multiview* stereo with flash photography to reconstruct a proxy geometry of an object with complex material on which a learned *neural texture* [TZN19] is projected. In contrast to the *deferred* neural rendering of Thies et al. [TZN19], Gao et al. [GCD*20] introduce an additional *lighting pass* in which the aforementioned textured geometry is multiplied by the *radiance cues* – “global illumination renderings of a rough proxy geometry of the scene for a small set of basis materials and lit by the target lighting” – and eventually passed to the neural renderer. The proposed method is trained separately on nine recorded real scenes and three synthetic ones, all containing objects of complex materials. Experiments show plausible renderings of the scenes under novel views and/or novel directional lights or illumination maps.

More recently but limited to human faces skin area, Reddy et al. [RTO*20] propose a neural face *full* reflectance field model that can be trained on monocular images from a light stage dataset [WMP*06]. The learned neural model is parametrized over a desired light direction, viewpoint and instance face geometry to reconstruct the respective OLAT images – which can later be used for relighting. Experiments show the suitability of the reflectance field representation for modelling specularities and self-shadowing. This work is not scene-specific per se – as the face geometry is required as an input to the proposed neural reflectance model. It is however surveyed here due to the strong conceptual similarity to the other approaches in this section.

3.3. Discussion

Reflectance-aware scene-specific representations are well studied in computer graphics, for example for measuring reflectance fields. The neural approach to learn these representations is, as expected, instance specific, that is a new model must be learned if the content of the scene changes. This can be time consuming with regard to data capture and/or learning time. On the positive side, these representations guarantee *spatial coherency*, which means, once a scene is learned, generated novel views [BXS*20a], or illumination maps [SMT*20] are coherent for all desired viewpoints or locations.

4. Relighting as Image-to-Image Transformations

This section covers relighting with image-to-image transforms that do not rely on neural scene-specific representation, nor image decomposition and inverse rendering techniques. The first two parts address the problem of relighting a scene given one or a few input images and potentially a controllable target lighting condition. The first part specifically contains general indoor/outdoor scenarios, while the second focuses on human faces. The third part addresses the problem of illumination harmonization between overlaid foreground objects into background plates.

4.1. General scenes

Xu et al. [XSHR18] propose a DNN model to generate a scene under a novel single directional light – given as input, sparse base lighting images taken under a few pre-defined single light directions. The proposed model consists of two networks: Sample-Net and Relight-Net. During training, Sample-Net learns five optimal lighting directions needed for generating the scene under a target novel illumination – in an end-to-end scheme with both networks present. During inference, only Relight-Network is employed with a target lighting direction. The training dataset is rendered synthetically with densely sampled light directions in the upper hemisphere, procedurally generated scenes (of nine objects) and arbitrary complex spatially varying BRDF material models. Experiments show successful relighting of real objects, as well.

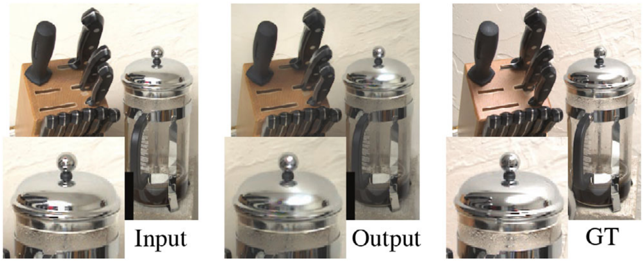

Figure 8: *Murmann et al. [MGAD19] relight an input image from another light direction. Images taken from Murmann et al. [MGAD19]*

The work of Xu et al. [XSHR18] is in nature very similar to image-based relighting methods, for example Debevec et al. [DHT*00]; it however requires a much smaller number of samples. Also, there is no need for a complex light stage capture setup and the trained model is not scene-specific. One downside of [XSHR18] is, on the other hand, its inability to reproduce cast shadows for “highly non-convex geometry.”

Murmann et al. [MGAD19] propose to train a U-Net [RFB15] like encoder-decoder DNN to directly regress HDR images which are lit from another direction, given an input HDR image (see Figure 8). The network architecture is simple and does not require a target lighting direction as input; however, a separate network must therefore be trained for each desired target lighting direction. The network is trained and tested on the proposed multi-illumination dataset for indoor scenarios. For more details on the proposed dataset, see Section 5.

4.1.1. *General scenes: modelling shadows*

To achieve more realistic results, complex neural architectures such as Philip et al. [PGZ*19] and Wang et al. [WSL*20] explicitly remove shadows from the source image and regenerate them under a target illumination condition in the inferred image.

Philip et al. [PGZ*19] address the problem of outdoor scenes relighting by first reconstructing a *proxy* geometry of a scene from public internet images. The approximate geometry is then used to create RGB shadow images of the scene for both the source and target (distant) illumination. These shadow images are refined by two separate but similar ResNets [HZRS16]. The relighting happens through a third ResNet which takes as input refined shadow images, a reference image and its illumination buffers (e.g. normal map, reflection features, etc.) and infers the relit image. The neural model is trained on a synthetically generated dataset consisting of 10 photorealistic scenes. Experiments show the benefit of using shadow masks/networks compared to a bare ResNet for direct

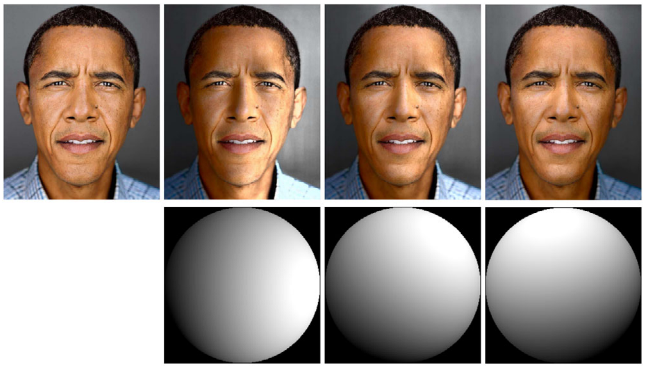

Figure 9: *Zhou et al. [ZHSJ19] generate a new portrait image under novel lighting. Images taken from Zhou et al. [ZHSJ19]*

image-to-image lighting transfer. They also show that first removing shadows, and then casting new ones from the approximate proxy geometry does not produce visually comparable results. Also, the network generalises to real footage such as those taken from a drone. Note that this work does not perform inverse rendering, rather it benefits from a reconstructed geometry as input to the neural model.

Similarly, the neural model put forward by Wang et al. [WSL*20] for indoor scenes has two paths: one for removing the effects of illumination and the other one for estimating shadow priors. The output of these two paths are concatenated and re-rendered by a third network. The model is trained and tested on the virtual image dataset for illumination transfer (VIDIT) [HZBS20] and has achieved the best PSNR in the Advances in Image Manipulation (AIM) 2020 challenge [HZS*20]. For more information on VIDIT, refer to Section 5.

4.2. Human face

Sun et al. [SBT*19] focus on portrait relighting in unconstrained environments using a single input image fed into a U-Net [RFB15] like encoder-decoder neural network. The encoder sub-network finds the illumination map under which the portrait is taken. The decoder sub-network can therefore either use the user-edited original illumination map, or a novel one to generate the relit portrait. The training of the network is end-to-end and achieved using a dataset of one-light-at-a-time (OLAT) [DHT*00] images captured from subjects in a light stage with seven cameras. Key results are visually plausible relighted front-facing portraits and generalization to unseen real-world selfie photos.

Very similar to [SBT*19, Zhou et al. ZHSJ19] also use an Hourglass network [NYD16] for the same problem formulation. They however train the network with a realistic, synthetically generated dataset, dubbed portrait relighting (DPR). DPR is generated using the photorealistic Ratio Image-based rendering algorithm [SR01] on the images of CelebA-HQ dataset [KALL18]. Experiments show generalization to real portraits (see Figure 9).

Meka et al. [MHP*19] also address the image-to-image face relighting but with a reflectance-aware approach. To reduce the sampling time of a subject in a light stage and therefore enable fast capture of dynamic performances, Meka et al. [MHP*19] propose to train a deep neural network (DNN) that takes as input only two images captured under Colour Gradient illumination [Fyf09], accompanied by a desired lighting direction, and can infer the corresponding OLAT image. The OLAT images captured for the training phase are partially compensated for the subject's movements by calculating dense optical flow-fields on the so-called *tracking* frames. Tracking frames are captured in between OLATs where all light stage lamps are on. The results for both the estimated OLATs and the novel view synthesis are comparable to the ground truth light stage images.

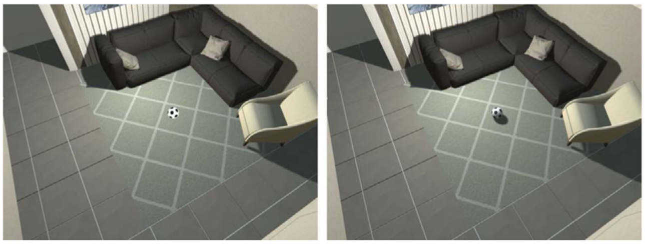

Figure 10: *Wang et al. [WWL19] harmonize foreground objects to a background. Images taken from Wang et al. [WWL19]*

4.3. Harmonizing foreground to background

Wang et al. [WWL19, Zhan et al. [ZLZ*20] as well as Nicolet et al. [NPD20] address the illumination consistency/harmonization problem in mixed reality scenes: given as input a foreground object naively pasted in an image, the solution seeks to harmonize the shading and shadows of the object with the surrounding scene (see Figure 10). Partial solutions that only consider shadow generation are proposed by Liu et al. [LLZ*20] for hard shadows and by Sheng et al. [SZB20] for soft ones. These methods are discussed below.

4.3.1. *Shading and shadows*

The model of Wang et al. [WWL19] is trained and tested on a (non-photorealistic) synthetic dataset generated in Unity3D; it covers various single point light source positions, camera viewpoints and scene content (from 3D Warehouse [Tri20]). The architecture is based on a generator-discriminator concept [GPAM*14] and trained with three different losses: adversarial, features matching and perceptual. Experiments show the capability of the network for illumination/style transfer from backgrounds to inserted objects – without any explicit inverse rendering (see Figure 10).

Zhan et al. [ZLZ*20] also address the general problem of illumination harmonization between composited, foreground objects and the background – therefore, handling both shadows and shading. However, they estimate illumination parameters of an SH model as an intermediate output; They benefit from such estimation in two separate shadow and texture decoders on a local level. Locally harmonized foreground objects are then transferred to the global level in the final image. Similar to Wang et al. [WWL19], a generator-discriminator-based loss is used for training the DNNs. Experiments are performed using various real datasets and show plausible compositing results.

Work of Nicolet et al. [NPD20] is based on multiple usage of the pre-trained relighting network of Philip et al. [PGZ*19]. In the first pass, the relighting network is employed for shading the inserted object in accordance to the illumination of the reference scene; The second pass, however, generates coherent shadows for the inserted objects and harmonizes the composited scene.

4.3.2. *Shadows only*

On the other hand, ARShadowGAN by Liu et al. [LLZ*20] seeks to generate only shadows from an input image and the respective mask of an inserted object. The main contribution is annotating a real dataset containing inserted virtual objects from ShapeNet [CFG*15]. The annotation includes the ground truth shadows and occluders of the real objects, in addition to the scene camera and lighting calibration. Similar to [WWL19, ZLZ*20], a generator-discriminator scheme is used for modelling the shadow generation process in real images without inverse rendering. The main downsides of the proposed model are: not handling the shading of the inserted object, and failing to generate correct shadows where there exist large dark areas in the input image [LLZ*20].

Sheng et al. [SZB20] consider realistic soft shadow generation without compositing. The solution is based on an encoder-decoder network and is trained on synthetically generated data from common classes of objects like humans, chairs, etc. The model does not generalize to unseen object classes and similar to ARShadowGAN [LLZ*20] can only generate plausible shadows for planar receivers.

4.4. Discussion

In this section, the problem formulation is inherently more challenging because these methods do not intend to decompose a scene into its comprising elements, and rather on the contrary, seek to learn a model that performs an image-to-image transform. Nevertheless, face relighting with deep neural models has proven to be successful [ZHSJ19]. For more general scenes, Philip et al. [PGZ*19] and Murmann et al. [MGAD19] show satisfactory results respectively for outdoor and indoor scenes.

Approaches in this category have generality but do not consider temporal coherency. A more recent problem formulation is how to harmonize illumination effects such as shading and/or shadows between a naively inserted foreground object and the surrounding background plate. In addition to illumination, the geometrical aspects of such insertions, for example camera perspective and scale, remain to be further studied.

5. Datasets

Deep neural models are heavily dependent on training data of large scale and diversity to be able to generalize to unseen scenarios. Methods reviewed in this survey address this concern differently. Common practices are to generate a synthetic dataset [HZBS20, PGZ*19, ZSY*17, LYS*20, LLZ*20, ZHSJ19, GSH*19, GZA*20] or to record a real one [MGAD19, GSY*17, LM14, XEOT12, WMP*06]. In the following a few substantial contributions are discussed in more detail. Table 2 summarizes the key attributes of these contributions.

5.1. Real datasets

The dataset of multi-illumination images in the wild [MGAD19] consists of 1000 indoor (e.g. home, or office) scenes which are illuminated under 25 different lighting directions. Various lighting directions are generated by servoing a mounted flash on a mirror-less digital camera. Dataset images are captured with almost 20 steps of dynamic range and include the appearance of a chrome and a diffuse probe in the scene as illumination ground truth. The proposed dataset is tested successfully in applications such as single-image illumination estimation, direct image relighting and mixed-illuminant white balancing [MGAD19].

The Laval indoor HDR panorama dataset [GSY*17] consists of 2100 HDR 360° panoramas from various indoor scenes, for example houses, apartments, factories, etc. The HDR content covers 22 steps of scene dynamic range which is deemed appropriate for indoor environments.

The Laval HDR sky database [LM14] covers 1,850 images of 22 outdoor landmark scenes under 350 different illumination conditions. “Each image has high dynamic range, is radiometrically and geometrically calibrated, and is aligned with its corresponding light probe.”

The scene understanding 360° panorama (SUN360) database [XEOT12] contains 67,583 LDR panoramas in 80 indoor or outdoor categories e.g. indoor theatre, outdoor field, etc.

5.2. Synthetically generated datasets

The virtual image dataset for illumination transfer (VIDIT) [HZBS20] is synthetically generated to create a baseline for comparison of ‘illumination manipulation methods.’ It contains 390 scenes, each captured with five different colour temperatures, at eight different incident light directions, resulting in a total of 15,600 images. The dataset covers both indoor and outdoor scenes and a variety of materials. Each rendered image is accompanied by ground truth metadata such as depth, lighting parameters, etc.

5.3. Dataset generation

More universal methods for dataset generation are recently getting closer attention. For example, Gkitsas et al. [GZA*20] leverage the synergy of *uncoupled* datasets of two categories: (a) HDR illumination maps, for example the Laval Indoor HDR panorama dataset [GSY*17] and (b) LDR scenes including corresponding normals, for example [KZS*19]. Normally, datasets from (a) are limited in their size and scene variety and datasets from (b) have implicit lighting information already in the captured images. Gkitsas et al. [GZA*20] estimate the SH lighting parameters through image-based relighting of scenes from dataset (b) (assuming a global Lambertian albedo) with blended illumination maps from dataset (a). In

Table 2: *Selected dataset contributions compared using the terminology of Section 1.3.*

	# Scenes	# Images	Data	HDR	Scene content	Real/Synthetic	Publicly available
ARShadowGAN [LLZ*20]	–	3000	IB	✗	IOJ	S	partial
Laval HDR Sky Database [LM14]	22	1850	N	√	O	R	√
Laval Indoor HDR Panorama Dataset [GSY*17]	2100	–	N	√	I	R	√
Multi-Illumination Dataset [MGAD19]	1000	25000	H	√	I	R	√
OpenRooms [LYS*20]	1506	118343	HLDGMC	√	I	S	planned
Philip et al. [PGZ*19]	10	73500	IL	✗	O	S	✗
Portrait Relighting Dataset (DPR) [ZHSJ19]	27627	138135	ILG	✗	F	S	√
SUN360 (Core/Extended) [XEOT12]	80/369	67583/–	P	✗	IO	R	√
VIDIT [HZBS20]	390	15600	IL	✗	IO	S	√
Zhang et al. [ZSY*17]	45622	568793	ILDGMC	✗	I	S	√

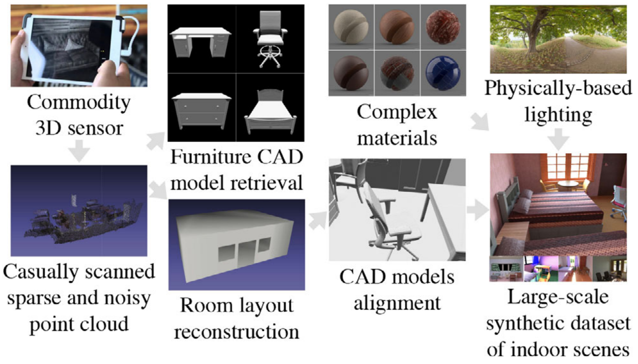
Commodity 3D sensor
Furniture CAD model retrieval
Complex materials
Physically-based lighting
Casually scanned sparse and noisy point cloud
Room layout reconstruction
CAD models alignment
Large-scale synthetic dataset of indoor scenes

Figure 11: *OpenRooms [LYS*20] generates synthetic dataset of indoor scenes. Diagram content and images taken from Li et al. [LYS*20]*

this manner, they increase the amount of supervision to their proposed end-to-end trainable neural model.

For indoor scenes, in a novel work, Li et al. [LYS*20] suggest generating datasets of HDR images under variety of lighting conditions and materials using available casual 3D scans of the scenes. The key idea is to reconstruct the room layout from 3D scans and to retrieve and align furniture CAD models into it. Complex materials can be assigned to these scene geometries and then lit under various lighting parameters. Figure 11 shows an overview of this process. The authors show the generalization of their proposed approach to real scenes by training the state-of-the-art neural model of Li et al. [LSR*20] on a synthetically generated dataset based on ScanNet [DCS*17]. Experiments show comparable results to [LSR*20, SGK*19, LS18a] for inverse rendering and to [LSR*20, GHS*19, GSH*19] for mixed reality compositing.

MLIC-Synthetizer [DDG19] is a plugin for Blender to generate physically-based synthetic multi-light image collections which can be used in different application scenarios, amongst them relighting. Each generated image comes with its own ground truth data such as depth or specularity maps. The plugin also provides the possibility of changing materials programmatically.

6. Conclusion

In this paper we have reviewed, compared and discussed the recently proposed deep neural methods for scene illumination estimation and relighting in three major categories based on approach and use case. Also, we provided an overview of publicly available datasets in the respective field and how they are potentially generated, if synthetic.

The main advantage of deep neural approaches for the aforementioned tasks is the fact that once a model is trained, relighting images or estimating a scene's lighting are significantly faster respectively compared to classical global-illumination-based rendering or optimization-based inverse rendering. In addition, the other equally important advantage, specifically for relighting and lighting harmonization applications, is that the accurate models of the scene components are not necessarily needed – in fact, deep neural models can learn these representations given the required mathematical capacity and diverse training data at scale.

Considering the ultimate aim of the fast relighting of dynamic scenes of arbitrary content, we are still in the early stages of research. Notable limitations of the current state-of-the-art using deep neural models, and potential future research directions, among others, are: considering only a single image versus sequences, that is lack of temporal coherency in predictions; limitations on realism of the generated images; being limited to certain scene content, for example human faces, indoor or outdoor scenes, etc.; limitations on the complexity of the illumination models, for example only considering directional lights; and limited existing training datasets in the senses of scale, realism, and diversity.

More specifically, the related mobile AR lighting applications are diverse: TV/film post-production or on-set relighting/compositing, seamless high-quality integration of real/synthetic imagery, real-time interactive applications such as AR games, etc. The main challenges include the ones mentioned above but with the emphasis on the photorealism and complex lighting for TV/film productions, and performance for interactive media.

Acknowledgement

This research was supported by UKRI EPSRC Platform Grant EP/P022529/1.

References

[BCD*13] Banterle F., Callieri M., Dellepiane M., Corsini M., Pellacini F., Scopigno R.: EnvyDepth: An interface for recovering local natural illumination from environment maps. *Computer Graphics Forum 32*, *7* (2013), 411–420.

[BJK*20] Boss M., Jampani V., Kim K., Lensch H. P., Kautz J.: Two-shot spatially-varying BRDF and shape estimation. In *CVPR* (2020).

[BXS*20a] Bi S., Xu Z., Srinivasan P., Mildenhall B., Sunkavalli K., Hašan M., Hold-Geoffroy Y., Kriegman D., Ramamoorthi R.: Neural reflectance fields for appearance acquisition. *arXiv preprint arXiv:2008.03824* (2020).

[BXS*20b] Bi S., Xu Z., Sunkavalli K., Hašan M., Hold-Geoffroy Y., Kriegman D., Ramamoorthi R.: Deep reflectance volumes: Relightable reconstructions from multi-view photometric images. In *ECCV* (2020).

[CCZ*20] Chen Z., Chen A., Zhang G., Wang C., Ji Y., Kutulakos K. N., Yu J.: A neural rendering framework for free-viewpoint relighting. In *CVPR* (2020).

[CDF*17] Chang A., Dai A., Funkhouser T., Halber M., Niebner M., Savva M., Song S., Zeng A., Zhang Y.: Matterport3D: Learning from rgb-d data in indoor environments. In *3DV* (2017), pp. 667–676.

[CFG*15] Chang A. X., Funkhouser T., Guibas L., Hanrahan P., Huang Q., Li Z., Savarese S., Savva M., Song S., Su H., et al.: ShapeNet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015).

[CLG*18] Calian D. A., Lalonde J.-F., Gotardo P., Simon T., Matthews I., Mitchell K.: From faces to outdoor light probes. *Computer Graphics Forum 37*, *2* (2018), 51–61.

[CSC*18] Cheng D., Shi J., Chen Y., Deng X., Zhang X.: Learning scene illumination by pairwise photos from rear and front mobile cameras. *Computer Graphics Forum 37*, *7* (2018), 213–221.

[CZMR20] Chalmers A., Zhao J., Medeiros D., Rhee T.: Reconstructing reflection maps using a Stacked-CNN for mixed reality rendering. *IEEE Transactions on Visualization and Computer Graphics* (2020).

[DB03] Dante A., Brookes M.: Precise real-time outlier removal from motion vector fields for 3d reconstruction. In *ICIP* (2003), vol. 1, pp. I–393.

[DCS*17] Dai A., Chang A. X., Savva M., Halber M., Funkhouser T., Niessner M.: ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR* (2017).

[DDG19] Dulecha T. G., Dall'Alba A., Giachetti A.: MLIC-Synthetizer: a synthetic multi-light image collection generator. In *STAG* (2019), pp. 105–106.

[Deb08] Debevec P.: Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *SIGGRAPH Classes* (2008).

[DHT*00] Debevec P., Hawkins T., Tchou C., Duiker H.-P., Sarokin W., Sagar M.: Acquiring the reflectance field of a human face. In *SIGGRAPH* (2000), p. 145–156.

[Fyf09] Fyffe G.: Cosine lobe based relighting from gradient illumination photographs. In *SIGGRAPH Posters* (2009).

[GCD*20] Gao D., Chen G., Dong Y., Peers P., Xu K., Tong X.: Deferred neural lighting: Free-viewpoint relighting from unstructured photographs. *ACM Transactions on Graphics 39*, *6* (2020).

[GFRG16] Girdhar R., Fouhey D. F., Rodriguez M., Gupta A.: Learning a predictable and generative vector representation for objects. In *ECCV* (2016), pp. 484–499.

[GHS*19] Gardner M., Hold-Geoffroy Y., Sunkavalli K., Gagné C., Lalonde J.: Deep parametric indoor lighting estimation. In *ICCV* (2019), pp. 7174–7182.

[GLD*19a] GAO D., Li X., Dong Y., Peers P., Xu K., Tong X.: Deep inverse rendering for high-resolution SVBRDF estimation from an arbitrary number of images. *ACM Transactions on Graphics 38*, *4* (2019).

[GLD*19b] Guo K., Lincoln P., Davidson P., Busch J., Yu X., Whalen M., Harvey G., Orts-Escolano S., Pandey R., Dourgarian J., Tang D., Tkach A., Kowdle A., Cooper E., Dou M., Fanello S., Fyffe G., Rhemann C., Taylor J., Debevec P., Izadi S.: The Relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics 38*, *6* (2019).

[GPAM*14] Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y.: Generative adversarial nets. In *NIPS*. 2014, pp. 2672–2680.

[GRPN20] Granskog J., Rousselle F., Papas M., Novák J.: Compositional neural scene representations for shading inference. *ACM Transactions on Graphics 39*, *4* (2020).

[GRR*18] Georgoulis S., Rematas K., Ritschel T., Gavves E., Fritz M., Van Gool L., Tuytelaars T.: Reflectance and natural illumination from single-material specular objects using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence 40*, *8* (2018), 1932–1947.

[GSH*19] Garon M., Sunkavalli K., Hadap S., Carr N., Lalonde J.: Fast spatially-varying indoor lighting estimation. In *CVPR* (2019), pp. 6901–6910.

[GSY*17] Gardner M.-A., Sunkavalli K., Yumer E., Shen X., Gambaretto E., Gagné C., Lalonde J.-F.: Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics 36*, *6* (2017).

[GZA*20] Gkitsas V., Zioulis N., Alvarez F., Zarpalas D., Daras P.: Deep lighting environment map estimation from spherical panoramas. In *CVPR Workshops* (2020).

[HGAL19] Hold-Geoffroy Y., Athawale A., Lalonde J.-F.: Deep sky modeling for single image outdoor lighting estimation. In *CVPR* (2019).

[HSH*17] Hold-Geoffroy Y., Sunkavalli K., Hadap S., Gambaretto E., Lalonde J.: Deep outdoor illumination estimation. In *CVPR* (2017), pp. 2373–2382.

[HW12] Hosek L., Wilkie A.: An analytic model for full spectral sky-dome radiance. *ACM Transactions on Graphics 31*, *4* (2012).

[HZBS20] Helou M. E., Zhou R., Barthas J., Süsstrunk S.: VIDIT: Virtual image dataset for illumination transfer. *arXiv preprint arXiv:2005.05460* (2020).

[HZRS16] He K., Zhang X., Ren S., Sun J.: Deep residual learning for image recognition. In *CVPR* (2016).

[HZS*20] Helou M. E., Zhou R., Süsstrunk S., Timofte R., Afifi M., Brown M. S., Xu K., Cai H., Liu Y., Wang L.-W., et al.: AIM 2020: Scene relighting and illumination estimation challenge. *ECCV Workshops* (2020).

[JDL*20] Jin X., Deng P., Li X., Zhang K., Li X., Zhou Q., Xie S., Fang X.: Sun-sky model estimation from outdoor images. *Journal of Ambient Intelligence and Humanized Computing* (2020), 1–12.

[KALL18] Karras T., Aila T., Laine S., Lehtinen J.: Progressive growing of gans for improved quality, stability, and variation. In *ICLR* (2018).

[KCW*18] Kang K., Chen Z., Wang J., Zhou K., Wu H.: Efficient reflectance capture using an autoencoder. *ACM Transactions on Graphics 37*, *4* (2018).

[KE18] Kanamori Y., Endo Y.: Relighting humans: Occlusion-aware inverse rendering for full-body human images. *ACM Transactions on Graphics 37*, *6* (2018).

[KK19] Kán P., Kafumann H.: Deeplight: light source estimation for augmented reality using deep learning. *The Visual Computer 35*, *6-8* (2019), 873–883.

[KXH*19] Kang K., Xie C., He C., Yi M., Gu M., Chen Z., Zhou K., Wu H.: Learning efficient illumination multiplexing for joint capture of reflectance and shape. *ACM Transactions on Graphics 38*, *6* (2019).

[KZS*19] Karakottas A., Zioulis N., Samaras S., Ataloglou D., Gkitsas V., Zarpalas D., Daras P.: 360 surface regression with a hyper-sphere loss. In *3DV* (2019), pp. 258–268.

[LGC*19] Li M., Guo J., Cui X., Pan R., Guo Y., Wang C., Yu P., Pan F.: Deep spherical gaussian illumination estimation for indoor scene. In *MMAsia* (2019).

[LLQX20] Liu C., Li Z., Quan S., Xu Y.: Lighting estimation via differentiable screen-space rendering. In *VR Workshops* (2020), pp. 575–576.

[LLZ*20] Liu D., Long C., Zhang H., Yu H., Dong X., Xiao C.: ARShadowGAN: Shadow generative adversarial network for augmented reality in single light scenes. In *CVPR* (2020).

[LM14] Lalonde J., Matthews I.: Lighting estimation in outdoor image collections. In *3DV* (2014), vol. 1, pp. 131–138.

[LMF*19] LeGendre C., Ma W.-C., Fyffe G., Flynn J., Charbonnel L., Busch J., Debevec P.: DeepLight: Learning illumination for unconstrained mobile mixed reality. In *SIGGRAPH Talks* (2019).

[LS18a] Li Z., Snavely N.: CGIntrinsics: Better intrinsic image decomposition through physically-based rendering. In *ECCV* (2018).

[LS18b] Li Z., Snavely N.: Megadepth: Learning single-view depth prediction from internet photos. In *CVPR* (2018).

[LSR*20] Li Z., Shafiei M., Ramamoorthi R., Sunkavalli K., Chandraker M.: Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and SVBRDF from a single image. In *CVPR* (2020), pp. 2472–2481.

[LXR*18] Li Z., Xu Z., Ramamoorthi R., Sunkavalli K., Chandraker M.: Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics 37*, *6* (2018).

[LYS*20] Li Z., Yu T.-W., Sang S., Wang S., Bi S., Xu Z., Yu H.-X., Sunkavalli K., Hašan M., Ramamoorthi R., Chandraker M.: OpenRooms: An end-to-end open framework for photorealistic indoor scene datasets. *arXiv preprint arXiv:2007.12868* (2020).

[MCV18] Marques B. A. D., Clua E. W. G., Vasconcelos C. N.: Deep spherical harmonics light probe estimator for mixed reality games. *Computers and Graphics 76* (2018), 96–106.

[MDVC18] Marques B. A. D., Drumond R. R., Vasconcelos C. N., Clua E.: Deep light source estimation for mixed reality. In *VISIGRAPP* (2018), vol. 1: GRAPP, pp. 303–311.

[MGAD19] Murmann L., Gharbi M., Aittala M., Durand F.: A dataset of multi-illumination images in the wild. In *ICCV* (2019), pp. 4079–4088.

[MH84] Miller G. S., Hoffman C. R.: Illumination and reflection maps. In *SIGGRAPH Course Notes* (1984).

[MHP*19] Meka A., Häne C., Pandey R., Zollhöfer M., Fanello S., Fyffe G., Kowdle A., Yu X., Busch J., Dourgarian J., Denny P., Bouaziz S., Lincoln P., Whalen M., Harvey G., Taylor J., Izadi S., Tagliasacchi A., Debevec P., Theobalt C., Valentin J., Rhemann C.: Deep reflectance fields: High-quality facial reflectance field inference from color gradient illumination. *ACM Transactions on Graphics 38*, *4* (2019).

[MMZ*18] Meka A., Maximov M., Zollhöfer M., Chatterjee A., Seidel H., Richardt C., Theobalt C.: LIME: Live intrinsic material estimation. In *CVPR* (2018), pp. 6315–6324.

[MRLF19] Maximov M., Ritschel T., Leal-Taixé L., Fritz M.: Deep appearance maps. In *ICCV* (2019), pp. 8728–8737.

[NPD20] Nicolet B., Philip J., Drettakis G.: Repurposing a relighting network for realistic compositions of captured scenes. In *I3D* (2020).

[NYD16] Newell A., Yang K., Deng J.: Stacked hourglass networks for human pose estimation. In *ECCV* (2016), pp. 483–499.

[PGZ*19] Philip J., Gharbi M., Zhou T., Efros A. A., Drettakis G.: Multi-view relighting using a geometry-aware network. *ACM Transactions on Graphics 38*, *4* (2019).

[PHC*19] Pan J., Han X., Chen W., Tang J., Jia K.: Deep mesh reconstruction from single rgb images via topology modification networks. In *ICCV* (2019).

[Pho75] Phong B. T.: Illumination for computer generated pictures. *Communications of the ACM 18*, *6* (1975), 311–317.

[PKC*17] Prada F., Kazhdan M., Chuang M., Collet A., Hoppe H.: Spatiotemporal atlas parameterization for evolving meshes. *ACM Transactions on Graphics 36*, *4* (2017).

[PPYW20] Park J., Park H., Yoon S., Woo W.: Physically-inspired deep light estimation from a homogeneous-material object for mixed reality lighting. *IEEE Transactions on Visualization and Computer Graphics 26*, *5* (2020), pp. 2002–2011.

[RDL*15] Ren P., Dong Y., Lin S., Tong X., Guo B.: Image based relighting using neural networks. *ACM Transactions on Graphics 34*, *4* (2015).

[RFB15] Ronneberger O., Fischer P., Brox T.: U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI* (2015), pp. 234–241.

[RH01] Ramamoorthi R., Hanrahan P.: An efficient representation for irradiance environment maps. In *SIGGRAPH* (2001), pp. 497–500.

[RRF*16] Rematas K., Ritschel T., Fritz M., Gavves E., Tuytelaars T.: Deep reflectance maps. In *CVPR* (2016), pp. 4508–4516.

[RTO*20] Reddy M. B. R., Tewari A., Oh T.-H., Weyrich T., Bickel B., Seidel H.-P., Pfister H., Matusik W., Elgharib M., Theobalt C.: Monocular reconstruction of neural face reflectance fields. *arXiv preprint arXiv:2008.10247* (2020).

[SBT*19] Sun T., Barron J. T., Tsai Y.-T., Xu Z., Yu X., Fyffe G., Rhemann C., Busch J., Debevec P., Ramamoorthi R.: Single image portrait relighting. *ACM Transactions on Graphics 38*, *4* (2019).

[SBV20] Sial H. A., Baldrich R., Vanrell M.: Deep intrinsic decomposition trained on surreal scenes yet with realistic light effects. *Journal of the Optical Society of America A 37*, *1* (2020), pp. 1–15.

[SBVS20] Sial H. A., Baldrich R., Vanrell M., Samaras D.: Light direction and color estimation from single image with deep regression. *London Imaging Meeting* (2020).

[SF19] Song S., Funkhouser T.: Neural illumination: Lighting prediction for indoor environments. In *CVPR* (2019), pp. 6911–6919.

[SGK*19] Sengupta S., Gu J., Kim K., Liu G., Jacobs D., Kautz J.: Neural inverse rendering of an indoor scene from a single image. In *ICCV* (2019), pp. 8597–8606.

[SKS02] Sloan P.-P., Kautz J., Snyder J.: Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. In *SIGGRAPH* (2002), pp. 527–536.

[SLL*20] Sun Y. k., Li D., Liu S., Cao T. C., Hu Y. S.: Learning illumination from a limited field-of-view image. In *ICME Workshops* (2020), pp. 1–6.

[SMT*20] Srinivasan P. P., Mildenhall B., Tancik M., Barron J. T., Tucker R., Snavely N.: Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In *CVPR* (2020).

[SR01] Shashua A., Riklin-Raviv T.: The quotient image: class-based re-rendering and recognition with varying illuminations. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*, *2* (2001), pp. 129–139.

[SYZ*17] Song S., Yu F., Zeng A., Chang A. X., Savva M., Funkhouser T.: Semantic scene completion from a single depth image. In *CVPR* (2017).

[SZB20] Sheng Y., Zhang J., Benes B.: SSN: Soft shadow network for image compositing. *arXiv preprint arXiv:2007.08211* (2020).

[TFT*20] Tewari A., Fried O., Thies J., Sitzmann V., Lombardi S., Sunkavalli K., Martin-Brualla R., Simon T., Saragih J. M., Nießner M., Pandey R., Fanello S. R., Wetzstein G., Zhu J., Theobalt C., Agrawala M., Shechtman E., Goldman D. B., Zollhöfer M.: State of the art on neural rendering. *Computer Graphics Forum 39*, *2* (2020), pp. 701–727.

[Tri20] Trimble Inc: 3D Warehouse, 2020. 3dwarehouse.sketchup.com.

[TZN19] Thies J., Zollhöfer M., Nießner M.: Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics 38*, *4* (2019).

[WCD*20] Wei X., Chen G., Dong Y., Lin S., Tong X.: Object-based illumination estimation with rendering-aware neural networks. *ECCV* (2020).

[WMP*06] Weyrich T., Matusik W., Pfister H., Bickel B., Donner C., Tu C., McAndless J., Lee J., Ngan A., Jensen H. W., Gross M.: Analysis of human faces using a measurement-based skin reflectance model. *ACM Transactions on Graphics 25*, *3* (2006), pp. 1013–1024.

[WPL18] Weber H., Prévost D., Lalonde J.: Learning to estimate indoor lighting from 3D objects. In *3DV* (2018), pp. 199–207.

[WSL*20] Wang L.-W., Siu W.-C., Liu Z.-S., Li C.-T., Lun D. P.: Deep relighting networks for image light source manipulation. *arXiv preprint arXiv:2008.08298* (2020).

[WWL19] Wang X., Wang K., Lian S.: Deep consistent illumination in augmented reality. In *ISMAR Adjunct* (2019), pp. 189–194.

[XEOT12] Xiao J., Ehinger K. A., Oliva A., Torralba A.: Recognizing scene viewpoint using panoramic place representation. In *CVPR* (2012), pp. 2695–2702.

[XLZ20] Xu D., Li Z., Zhang Y.: Real-time illumination estimation for mixed reality on mobile devices. In *VR Workshops* (2020), pp. 702–703.

[XSHR18] Xu Z., Sunkavalli K., Hadap S., Ramamoorthi R.: Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics 37*, *4* (2018).

[YME*20] Yu Y., Meka A., Elgharib M., Seidel H.-P., Theobalt C., Smith W. A. P.: Self-supervised outdoor scene relighting. In *ECCV* (2020).

[YS19] Yu Y., Smith W. A. P.: InverseRenderNet: Learning single image inverse rendering. In *CVPR* (2019).

[YZTL18] Yi R., Zhu C., Tan P., Lin S.: Faces as lighting probes via unsupervised deep highlight extraction. In *ECCV* (2018).

[ZFT*20] Zhang X., Fanello S., Tsai Y.-T., Sun T., Xue T., Pandey R., Orts-Escolano S., Davidson P., Rhemann C., Debevec P., et al.: Neural light transport for relighting and view synthesis. *ACM Transactions on Graphics* (2020).

[ZHSJ19] Zhou H., Hadap S., Sunkavalli K., Jacobs D.: Deep single-image portrait relighting. In *ICCV* (2019), pp. 7193–7201.

[ZHZ*20] Zhu X., Han X., Zhang W., Zhao J., Liu L.: Learning intrinsic decomposition of complex-textured fashion images. In *ICME* (2020), pp. 1–6.

[ZKE15] Zhou T., Krähenbühl P., Efros A. A.: Learning data-driven reflectance priors for intrinsic image decomposition. In *ICCV* (2015), pp. 3469–3477.

[ZLZ*20] Zhan F., Lu S., Zhang C., Ma F., Xie X.: Adversarial image composition with auxiliary illumination. In *ACCV* (2020).

[ZSHG*19] Zhang J., Sunkavalli K., Hold-Geoffroy Y., Hadap S., Eisenman J., Lalonde J.-F.: All-weather deep outdoor lighting estimation. In *CVPR* (2019).

[ZSY*17] Zhang Y., Song S., Yumer E., Savva M., Lee J.-Y., Jin H., Funkhouser T.: Physically-based rendering for indoor scene understanding using convolutional neural networks. In *CVPR* (2017).

[ZYW*19] Zheng Z., Yu T., Wei Y., Dai Q., Liu Y.: DeepHuman: 3d human reconstruction from a single image. In *ICCV* (2019).

[ZYW*21] Zhan F., Yu Y., Wu R., Zhang C., Lu S., Shao L., Ma F., Xie X.: GMLight: Lighting estimation via geometric distribution approximation. *arXiv preprint arXiv:2102.10244* (2021).

[ZZY*20] Zhan F., Zhang C., Yu Y., Chang Y., Lu S., Ma F., Xie X.: EMLight: Lighting estimation via spherical distribution approximation. *arXiv preprint arXiv:2012.11116* (2020).