Noname manuscript No.
(will be inserted by the editor)

A Survey on Intrinsic Images

Delving Deep Into Lambert and Beyond

Elena Garces · Carlos Rodriguez-Pardo · Dan Casas · Jorge Lopez-Moreno

Received: date / Accepted: date

Abstract Intrinsic imaging or intrinsic image decomposition has traditionally been described as the problem of decomposing an image into two layers: a reflectance, the albedo invariant color of the material; and a shading, produced by the interaction between light and geometry. Deep learning techniques have been broadly applied in recent years to increase the accuracy of those separations. In this survey, we overview those results in context of well-known intrinsic image data sets and relevant metrics used in the literature, discussing their suitability to predict a desirable intrinsic image decomposition.

Although the Lambertian assumption is still a foundational basis for many methods, we show that there is increasing awareness on the potential of more sophisticated physically-principled components of the image formation process, that is, optically accurate material models and geometry, and more complete inverse light transport estimations. We classify these methods in terms of the type of decomposition, considering the priors and models used, as well as the learning architecture and methodology driving the decomposition process. We also provide insights about future directions for research, given the recent advances in neural, inverse and differentiable rendering techniques.

Elena Garces
SEDDI & Universidad Rey Juan Carlos, Madrid, Spain
E-mail: elena.garces@seddi.com

Carlos Rodriguez-Pardo
SEDDI & Universidad Carlos III, Madrid, Spain
E-mail: carlos.rodriguezpardo.jimenez@gmail.com

Dan Casas
Universidad Rey Juan Carlos, Madrid, Spain
E-mail: dan.casas@urjc.es

Jorge Lopez-Moreno
Universidad Rey Juan Carlos & SEDDI, Madrid, Spain
E-mail: jorge@jorg3.com

1 Introduction

Images, as two-dimensional projections that depict the world around us, can be described as a harmonious combination of colors, shades, and shadows. Understanding how an image is generated by the complex interaction between light and matter has been a subject of research for decades: the light rays that carry all the information about a given scene are integrated into the RGB values of the camera sensor, turning the process of recovering back the original scene into an ill-posed problem. In the computer graphics and vision literature, two main approaches target the problem of recovering the underlying properties of the scene elements (such as lights, geometry, and materials): inverse rendering, and intrinsic decomposition methods. Although they share a common root, these two problems have been traditionally tackled from two perspectives resulting in different outcomes.

Inverse rendering methods have the goal of representing the scene digitally in a way that it allows to photo-realistically *re-render* novel views of it. This means estimating the parameters required by the render equation such as geometry, lights, materials, or the camera model. This is extremely challenging given a single image as input, even forward render engines sometimes merely approximate the complex light phenomena. Traditional approaches relied on manual intervention to aid modeling arbitrary geometries [101], or established priors about the shape narrowing the scope to *e.g.* faces [148, 13], humans [61], flat materials [33, 32], or single objects [48]. Nowadays, inverse rendering has undergone a major disruption in the way the problem is being tackled: First, deep neural networks, as powerful universal approximators, have reduced the need to define the scene elements explicitly, as shown by neural rendering methods [126]. Second, differentiable renders [75, 100, 144, 85], by allowing to compute direct derivatives of the

images with respect to arbitrary scene parameters, have enabled physically-based end-to-end inverse parameter estimation. Neural rendering techniques are reaching high degrees of realism for reproducing any kind of scene [93,87], although this comes at the cost of limiting the manipulation of the scene parameters. Differentiable rendering, although it cannot yet cope with arbitrary scene setups, is showing promising results towards this end.

Intrinsic decomposition, which can be seen as a simplification of inverse rendering for general scenes, aims to provide interpretable intermediate representations that prove useful for intelligent vision systems or to allow local material edits that do not require changes in lighting or viewpoint. The intrinsic scene model [7] described the world as the combination of three *intrinsic* components –surface reflectance, distance or surface orientation, and incident illumination. A fundamental observation for defining these layers is that the human visual system understands them independently of viewing and lighting conditions, even if it is not familiarized with the scene or the objects. In practice, most of the methods have referred to intrinsic image decomposition as the problem of separating the reflectance and shading layers, assuming a Lambertian world. The Retinex theory [71,51] was fundamental for many of the algorithms developed during the last two decades [38,40,11,15,10] providing some basic priors about how shading and reflectance typically behave in our retina: a change in reflectance cause sharp gradients, while a change in shading cause smooth gradient variations. Until recently, most methods relied only on such kinds of cues (or heuristics) derived from low-level understanding of the physics of the light or empirical observations. With the advent of deep learning, current solutions have posed the problem as end-to-end network architectures which learn to predict the reflectance and shading layers given huge amounts of data as training. A key difference with respect to traditional solutions is that learned models also take into account the global semantics of the scene, while previous methods performed mostly at the local level (gradients and edges).

Deep learning-based solutions for the intrinsic decomposition problem have facilitated more complex scene models beyond the Lambertian one [88,130,50,70], and also to estimate some of the scene elements, such as illumination and geometry [56,139,114,146,77]. While this will ultimately enable photo-realistic arbitrary scene manipulations (goal shared with image-based inverse rendering), some critical aspects make evaluating the contributions of each new method a difficult process: there is a variety of datasets that contain diverse types of objects, scenes, and labels; non-standardized and low-quality quantitative metrics to compare the methods with; and a lack of a unified methodology to universally evaluate the progress.

With this survey, we would like to review the current status of learning-based solutions that may serve to inspire and guide future research in the fields of computer graphics and vision. In particular: First, we review the connection of the intrinsic decomposition problem from a forward and inverse physically-based rendering perspective, hoping that this will clarify doubts and inspire more complex approaches outside the Lambertian assumption. Second, we show a taxonomy of current datasets, learning strategies, and architectures, putting them in context of traditionally non-learning-based solutions, we also discuss their main advantages and limitations. Third, we gather quantitative evaluations of these methods according to commonly used metrics and show qualitative results for difficult cases. Finally, we conclude with open research opportunities.

In addition to the review presented in the paper, we provide a web project which will contain the compendium of datasets, metrics, papers and their performance, which we will keep updated with the latest research.

Related Surveys Intrinsic image decomposition problem has been reviewed before in the STAR report of Bonneel et al. [14], where several –non-deep learning-based– algorithms were evaluated in the context of image editing tasks: logo removal, shadow removal, texture replacement, and wrinkles attenuation. Since then, dozens of new papers have tackled the problem from a purely data-driven perspective. Our work is complementary to theirs, as we review the approaches not covered there, which propose solutions based on deep learning frameworks. Neural rendering has been reviewed in a recent survey [126]. Similarly, inverse rendering and image-based rendering has been widely studied in several surveys: for generic scenes [103], for particular applications like faces [148] or materials [32], and for image-based 3D object reconstruction [48]. Despite the vast amount of papers tackling both the problem of intrinsic decomposition and inverse rendering, the explicit connection between both fields has not been addressed before. Moreover, as we will discuss in this paper, the connection of intrinsic imaging with the actual optical properties of materials is very relevant, and thus, a survey on its representation and acquisition [46] could be a good reading for practitioners in the field.

Scope In this survey, we cover several recent papers that propose a deep learning-based solution to estimate the intrinsic components of the scene given a single image as input. We thoroughly review those which include quantitative metrics of performance, either using the IIW dataset [10], which contains indoor real scenes and scores given by human raters, or using the MIT Intrinsic dataset [45], which contains isolated painted figurines. Although most of the papers discussed use the Lambertian material model, some of

the most recent ones include more complex ones or model other scene elements such as illumination or geometry. This latter approach resembles image-based inverse rendering methods, so we further discuss the ones which explicitly model part of the scene elements (illumination, material, geometry). Among these methods, we include a brief overview of the ones which target particular domains (faces, humans, flat materials). Finally, we do not review methods that target specific applications such as relighting, colorization or texture editing, or intrinsic decomposition methods which do not provide quantitative performance comparisons, unless they prove useful to facilitate the discussion.

2 Theoretical Background

If we look at any simple scene surrounding us, such as the photograph in Figure 1, we can find a plethora of optical interactions: indirect lighting (color bleeding), internal scattering in translucent objects, caustics, anisotropic and glossy reflections, etc. which are far from the traditional assumptions in intrinsic imaging of diffuse (lambertian) shading, direct lighting and diffuse albedo materials.

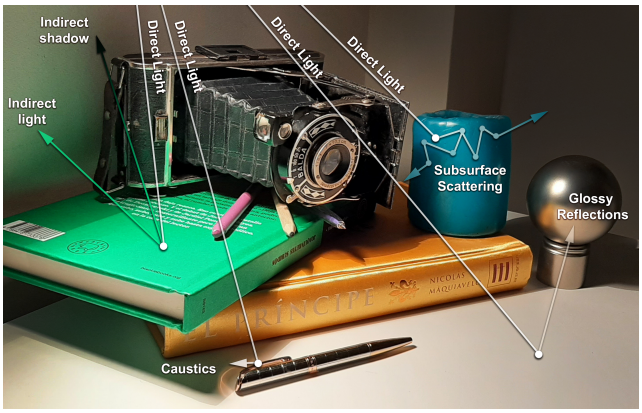


Fig. 1 Example of light transport and material interactions. Secondary bounces of light produce reflection caustics (chrome pen) and color bleeding from the green book. The wax candle exhibits multiple internal (subsurface) scattering of photons. The yellow silk fabric of the book cover shows specular anisotropic reflections due to yarn orientation. Figure inspired by [110].

In this section we provide an overview of the theoretical background behind the image formation model, its derivation for non-diffuse materials, and the link with physically-based and inverse rendering. As we will show in following sections, only a few of these physical aspects are considered in the reviewed methods, but we think that this theory is relevant for the discussion of future lines of research.

For a deeper dive into any of the concepts quite briefly described in the following subsection, we recommend reading the book on physically based rendering by Pharr et al.[104].

2.1 Physically-based rendering

The color and luminosity at any point of an image, the incoming irradiance, is proportional to the sum of the outgoing radiance from all the visible points of the scene towards the camera sensor at the corresponding pixel, I, resulting from multiple interactions between light and matter in the scene. Naturally, this is a simplification: even if we consider the camera lenses and color filters as part of the scene, the interaction of irradiance and the sensor point affects the result, and both electronic and film cameras have specific additional image formation steps which can be simulated. In physically-based rendering, the general approach to compute this value is to use Monte Carlo estimators of the pixel and shading integral [60], which has the general form of:

$$I = \int_{\chi} f(x, \Theta)\, dx \tag{1}$$

where f is a function that defines the radiance towards pixel I and is defined on a domain χ, generally a unit sphere, or the set of all the surfaces (A) in the scene, and depends on the scene parameters Θ, which might include the definition of the geometry (normals, z-depth, vertices), material (albedo, BRDF), or illumination sources – far-field environment lighting, or 3D light emitters (point, area, objects).

The value of I is defined for a given λ, which is the spectral band of the camera sensor. We could define it for as many bands as desired (including non-visible ones) but in practice, the majority of camera sensors mimics the human visual system and are commonly three-band: $\lambda \in \{R,G,B\}$. In subsequent rendering equations, we will simplify the notation by assuming a single-band, and omitting the term λ.

Equation 1 is an integral of integrals (see Figure 2): to account for all the light arriving at a surface point p_1 to a sensor pixel at p_0, we have to estimate all the contributions of light from all the surfaces of the scene, recursively tracing paths bouncing in surfaces (p_2, p_3...p_n) until we reach the light emitted by a source $L_e(p_n \rightarrow p_{n-1})$. This is referred as the Light Transport Equation (LTE) in rendering and it is another way of seeing the equation 1. To compute the radiance reaching the pixel I, that is from p_1 to p_0 in Figure 2, we would need to solve:

$$L(p_1 \rightarrow p_0) = \sum_{k=1}^{\infty} P(\bar{p}_k) \tag{2}$$

with $P(\bar{p}_k)$ being the radiance scattered over a path $\bar{p}_k$ with $n+1$ vertices (p_0,p_1,p_2, p_3...p_n) and computed as:

$$P(\bar{p}_k) = \underbrace{\int_A \int_A \dots \int_A}_{n-1} L_e(p_n \rightarrow p_{n-1}) T(\bar{p}_k) dA(p_2) \dots dA(p_n)$$

(3)

Remember that we can integrate over solid angles in the unit sphere, or surfaces (A) of the scene, being $dA(p_i)$ the differential area at point p_i. Note the term $T(\bar{p}_k)$, named *throughput* of the path: the fraction of radiance from the light source that arrives at the camera after all of the scattering at vertices between them. The total transmitted energy will be reduced at each interaction event, as some wavelengths (colors) are absorbed or scattered away from the observer. This is a common trade-off decision in many Monte Carlo rendering engines; longer paths per sample are costly to compute, while contributing with less and less energy with each additional vertex, but in some scenes, they might be very relevant to reduce variance (noise) and converge with less samples to an accurate image.

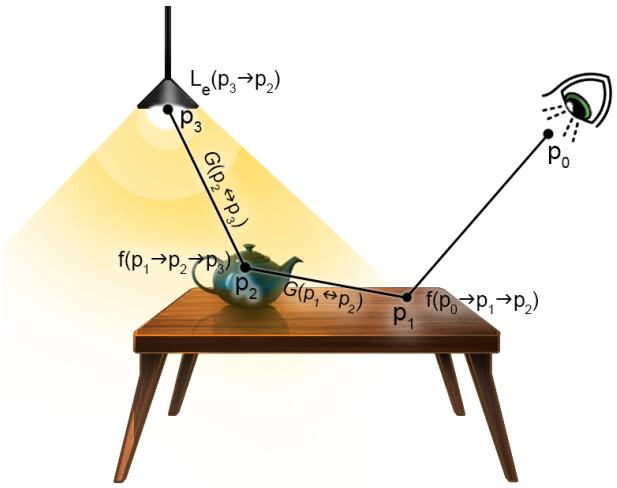


Fig. 2 Example of light transport, connecting a light source p_3 to a pixel I at p_0. Multiple paths like this one will need to be explored to provide a good statistical estimate of the radiance from p_1 to p_0. Figure inspired by [104].

Moreover, if there are interactions with non-opaque materials (human skin, cloth, marble) or participating media, such as liquids or smoke, we have to integrate volumetric scattering interactions of photons along the path between the light sources and the camera sensor pixel, requiring a more complex mathematical model such as the Radiative Transport Equation (RTE).

2.2 Geometry and Space

If we take a look at Figure 2, we can observe that materials are distributed on discrete 3D objects: the table and the cup, and even the light source if we consider it as an emissive material. We can assume that one of the most important Θ parameters is geometry, usually in the form of **3D vertices and edges**, **normals** (N) or **depth** maps (D). Please note that, in contrast to a a full 3D mesh, a single camera-view depth image (*a.k.a.* Z-buffer) is an incomplete definition because the non-visible surfaces are undefined and the reflected light paths cannot be traced behind the visible objects. For example, Nimier et al.[100] require multiple views of a smoke volume in order to reconstruct its 3D density distribution by inverse rendering.

There are additional ways of defining this geometry, such as implicit surfaces, but the material distribution is more complex when there are no clear surface boundaries. That is the case of heterogeneous participating materials: human skin, airlight, mixed liquids, smoke, etc., where light transport between two points has to be in turn evaluated along the path to account for all the possible scattering effects. For instance, imagine a small cloud of vapor between points p_1 and p_2 in Figure 2, at each infinitesimal step along the path between those points, there will be a possibility of absorption (collision with a water particle) that will reduce the energy, but there is also a possibility of receiving incoming energy (not emitted by p_2), as the cloud itself is receiving direct illumination from the light source and multiple scattering events distribute the light across its volume.

In order to compute the rendering equation in volumetric media, a distribution of parameters is required at any point of the space, not only on the 3D surfaces. The usual representations include solid 3D implicit functions (for instance 3D Perlin noise, used in cloud procedural generation), meshes and distance fields, or discrete volumetric grids which store the scattering probability function at any point of the space (also known as *phase function*) by means of voxels.

2.3 Materials

Each time the light interacts with a material, there is a loss of energy and a transformation of the original wavelength reflected towards the observed direction. In rendering, the result depends on the intrinsic material response for those two angles: incident light and viewing direction (*e.g.*: the camera, or another element of the scene). For *surfaces*, this response is modeled with a Bidirectional Reflectance Distribution Function (BRDF) $f_r(\mathrm{x}, \omega_i, \omega_o)$ which yields at each particular 3D surface point x, and for each incident direction ω_i, the fraction of reflected radiance observed from a direction ω_o.

The total reflected radiance L at any point x can be obtained by integrating with Equation 4 over the positive hemisphere Ω^+, to sample the whole incident light attenuated by the cosine term (dot product between the incident light direction L_i and the normal of the surface N) [60]:

$$L(\mathrm{x}, \omega_o) = \int_{\Omega^+} f_r(\mathrm{x}, \omega_i, \omega_o) L_i(\mathrm{x}, \omega_i)(\omega_i \cdot N)\, d\omega_i \tag{4}$$

Note that by integrating the computed radiance $L(\mathrm{x}, \omega_o)$ of the points sampled from p_0 at the camera sensor ($x = p_1, \omega_o = p_1 - p_0$, see Figure 2), we are obtaining the irradiance at the sensor and the corresponding image pixel values (I in equation 1). Naturally, even the pixels themselves can be sampled several times and integrated over the camera sensor with another Monte Carlo estimator to minimize aliasing effects.

The BRDF can be extended with a Bidirectional Transmittance Distribution Function (BTDF) to conform a full Bidirectional Scattering Function (BSDF), defined in the full sphere Ω. The model can be further extended to account for Surface Scattering phenomena (BSSDF).

These functions have a minimum of four dimensions (input-output pair directions in polar coordinates) and usually three RGB values as output. It is thus technically possible to choose a discrete set of (ω_i, ω_o) orientations an create a lookup table to interpolate the response of the material, which is captured with multiple light and camera positions (*e.g.*with a gonioreflectometer). It is evident that the storage becomes a major drawback for tabulated data, which can only be reduced through a significant reduction of quality. Moreover, we are considering only homogeneous surface materials, which is not often the case in actual scenes (*e.g.*, a printed paper). Those spatially-varying values (svBRDF) can be stored in a stack of textures; a Bidirectional Distribution Texture Function (BTF), increasing the dimensions and size of the table.

Beyond direct compression techniques, the most successful approach in graphics has been the use of analytic N-dimensional functions to approximate the reflectance and scattering distributions. The simplest of them, Lambertian diffuse, and Phong specular shading are also well known in the computer vision community. These functions leverage symmetry (isotropy) to model the material with a few parameters. For instance, a Lambertian material only requires to know the intrinsic albedo, a sort of base color, while the Phong model requires three additional parameters for the specular component (*e.g.*, shininess). In the following subsections, we will review the most common and simple material assumptions used in recent papers, finalizing with the most sophisticated inverse material models which are starting to be studied in our field.

2.3.1 Lambertian Assumption

The Lambertian assumption is the most common material reflectance simplification used to tackle the problem of intrinsic image decomposition. It consists on assuming that the BRDF of a surface is constant in all directions (diffuse) and, consequently, the observed light radiance does not depend on the viewpoint. Therefore, we can omit ω_o in the surface reflectance model f_r used in Equation 4. If the surface is diffuse, then $f_r(\omega_i) = \frac{\rho_d}{2\pi}$, with ρ_d denoting the diffuse albedo: the constant ratio of incident light which is reflected in any direction, independently of the view point ω_o. The image pixel value I is then given by:

$$I = \underbrace{\frac{\rho_d}{\pi}}_{\mathrm{A}} \underbrace{\int_{\Omega^+} L_i(\omega_i)(\omega_i \cdot N)\, \mathrm{d}\omega_i}_{S} \tag{5}$$

The intrinsic model then can be defined as,

$$I = A \cdot S \tag{6}$$

where S contains all the shading variations due to the geometry of the local surface w.r.t. light direction. In some cases, the shading image should contain contributions of all the lights in the scene ($S_1 + S_2 + ... + S_K$), which for discrete directional lights, can be deterministically estimated with a linear summation:

$$S = \sum_{1}^{K} L_i(\omega_i)(\omega_i \cdot N) \tag{7}$$

In the case of a more realistic illumination representation, such as environment lighting, or indirect light, the shading computation requires to sample the whole hemisphere Ω^+, often recursively sampling other surfaces to approximate the integral of the incoming light. The shading component S within Equation 5 in the integral form, is difficult to compute and not so easily invertible and differentiable, so until recently, most intrinsic decomposition methods assumed the simpler formula described in Equations 6 and 7. Note that the illumination visibility is not considered in most cases (E.g.: cast shadows).

2.3.2 Non-Lambertian Assumption

There are two possible sources producing a Lambertian shading, either a surface which has an extremely rough micro-geometry, and thus reflects light equally in multiple random directions at any differential patch of the surface, or a very diffuse light source, coming from any direction with equal intensity (*e.g.*, a foggy day). Both scenarios can be combined (Equation 4): the shiniest object in a foggy day will look quite diffuse, while even the most diffuse materials tend to project specular reflections under focused lighting from certain view angles. However, the majority of materials in the world are not Lambertian: even the most diffuse surface will exhibit Fresnel reflections when observed at grazing angles. Therefore, most surfaces will show the view-dependent effects that are classified as specular reflections. This separation between specular and Lambertian is rather pragmatic, but arbitrary, as even a simple microfacet model (shown in Figure 3) requires multiple analytic 3D lobes to approximate the 3D reflectance response for an infinitesimal incoming light ray (ω_i). The term *specular* is usually

applied to narrow lobes with high probability of scattering radiance, producing high luminance values at pixels (highlights). This family of materials are of the general form:

$$L(\omega_o) = \int_{\Omega^+} \underbrace{f_r(\omega_i,\omega_o)(\omega_i \cdot N)}_{f_{NL}} L_i(\omega_i)\, \mathrm{d}\omega_i \tag{8}$$

$$f_{NL}(\omega_i,\omega_o,N) = f_d(\omega_i,N) + f_s(\omega_i,\omega_o,N) \tag{9}$$

where f_{NL} is a non-lambertian BRDF composed by two components: f_d, a diffuse isotropic lobe, and f_s, a specular lobe which depends on the camera viewpoint (ω_o).

*Dichromatic Reflection Model.*This particular Non-Lambertian model [88,130] separates the object in two reflection components (S_d, S_s), but considers that the specular component S_s might have a color α_s which could differ from the color of the reflected light:

$$L(\omega_o) = \alpha_d \underbrace{\int_{\Omega^+} f_d(\omega_i,N) L_i(\omega_i) \mathrm{d}\omega_i}_{S_d} \tag{10}$$

$$+\alpha_s \underbrace{\int_{\Omega^+} f_s(\omega_i,\omega_o,N) L_i(\omega_i) \mathrm{d}\omega_i}_{S_s} \tag{11}$$

$$I = \alpha_d S_d + \alpha_s S_s \tag{12}$$

This is the case of metallic materials which, unlike dielectric ones, will show specular reflections with a change in wavelength. Some additional effects such as colored inter-reflections might be also captured in all layers.

*Phong and Blinn-Phong.*The dichromatic model can be extended with one of the most adopted analytic approximations, either Phong (f^P_{NL}) or Blinn-Phong (f^P_{NL}), which could be estimated with Monte Carlo integration and arbitrary lighting, or analytically computed with directional light sources:

$$f^P_{NL}(\omega_i,\omega_o,N) = \alpha_d(\omega_i \cdot N) + \alpha_s(h \cdot N)^k \tag{13}$$

$$f^{BP}_{NL}(\omega_i,\omega_o,N) = \alpha_s(\omega_i \cdot N) + \alpha_s(\mathrm{r} \cdot \mathrm{v})^k \tag{14}$$

where α_d and α_s are the colors of the diffuse and the specular reflections, the halfway vector $h = \frac{\omega_i+\omega_o}{||\omega_i+\omega_o||}$ depends on the light direction ω_i and the view direction ω_o. The size of the specular lobe is determined by the scalar term $k \in \mathbb{R}$.

2.3.3 Beyond Dichromatic Models: Physically-based Materials

Naturally, the breath of materials which can be synthesized with the previous models is very limited and not quite realistic in most cases. The advent of physically-based materials has introduced many variations [50] of the original micro-facets models [131], which assume that a surface is composed of many very tiny facets that reflect light perfectly. By controlling the statistical distribution of their orientations, the roughness of the surface varies from mirror-like to almost diffuse. Additional optical properties are introduced in these models: Fresnel view-dependent reflectivity, multiple specular lobes, metalness (conductive materials such as gold, change the color of the highlights), multiple reflection and refraction lobes, etc.

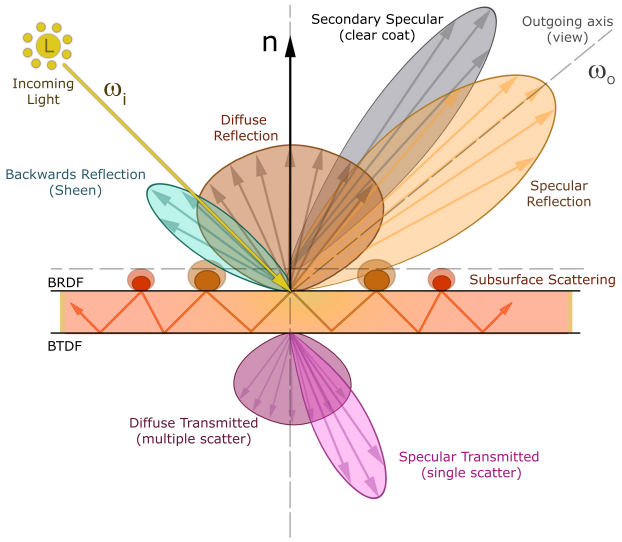


Fig. 3 2D depiction of a physically-principled BSDF theorical model. In most standard representations, the continuous reflectance 4D function is discretized into a combination of analytic lobes (Cosine, GGX) which can be easily computed and used for sampling purposes. Note that subsurface scattering (photons traveling through the medium) is not composed by multiple lobes. They are depicted for descriptive purposes, but it is rather approximated by a constant value or a diffusion profile, if not explicitly computed by simulating multiple scattering events with path tracing or photon mapping.

The separation of the lobes described in the multi-lobed physically-based model shown in Figure 3 is not arbitrary, grouping reflections and refractions which share the same orientation and energy level. For instance, the main diffuse lobe is grouping multiple different orientations which are not view-dependent and share the same intensity and color. If it covers the full hemisphere with a cosine-like ratio, it is often referred as *Lambertian*. Likewise, the specular transmitted lobe is grouping a view-dependent peak that the observer would only see when the translucent surface is be-

tween the light source and the camera. Even if it is often called a *single scatter* lobe, likely multiple internal scattering bounces of light are also included in this group.

2.4 Illumination

The illumination is a significant contributor to the shading term (S) in most decompositions. From a rendering perspective, as shown in Equation 15, the computation of the pixel radiance, L, requires to consider both the emitters L_e and the irradiance: the integral of all the incoming lighting at the observed point. This incoming illumination is often neglected, considering only point light or directional analytic emitters, which simplify the shading computation by removing L_i from the integral. If f_r is also assumed to be Lambertian, only the form factor given by the cosine of the surface normal and the light direction remains.

$$L(\mathrm{x},\omega_o) = L_e(\mathrm{x},\omega_o) + \int_{\Omega^+} f_r(\mathrm{x},\omega_i,\omega_o) L_i(\mathrm{x},\omega_i) \max(\omega_i \cdot N, 0)\, d\omega_i \tag{15}$$

In actual scenes, the incoming lighting is a combination of emitted or reflected illumination from distant objects (far field) and local surfaces close to the observed area (near field). The former is usually approximated in computer graphics with environment lighting, often based in High-Dynamic-Range (HDR) images mapped into an infinite sphere or cube surrounding the scene, while the latter can be derived from the far field illumination, by simulating the local secondary light bounces. If the geometry does not change, an environment map can be stored at multiple scene locations and distances, to include near field effects more accurately (Spatially Varying Environment Maps), although at a great memory cost, and only for static scenes.

To reduce the sampling and size of environment maps and simplify the computation, Ramamoorthi and Hanrahan [106] proposed their compression with Spherical Harmonics (SH), a set of orthonormal basis functions defined on the spherical domain (elevation θ and azimuth ϕ angles). Thus the equation 16 describes the irradiance E as the sum of bases weighted by the cosine decay term A_l and the illumination coefficient $L_{l,m}$. By changing the representation of the bases $\hat{Y}$ to polynomial coordinates of a unit normal $n = (x,y,z)^T$, this becomes an efficient vector dot product operation (Equation 17). With the required modifications, this strategy is feasible with other orthogonal basis functions on the sphere.

$$E(\theta,\phi) = \sum_{l,m} \hat{A}_l L_{l,m} Y_{l,m}(\theta,\phi) \tag{16}$$

$$E = \hat{Y}^T L \tag{17}$$

$$E = T^T L \tag{18}$$

If we want to account for near field occlusion and interreflection, it is possible to precompute those local interactions, because they depend on the object geometry and materials, and not on far field illumination. This family of techniques is known as *precomputed radiance transfer* (PRT) [120]: they precompute multiple events of light transport (see Figure 2) into the T term in Equation 18 with Monte Carlo path-tracing. In this fashion, each pixel will have secondary light bounces stored in a light transport map. If only the visibility term $V(\omega_i)$ is considered for T, the method will be storing the ambient occlusion shadows, but not colored interreflections.

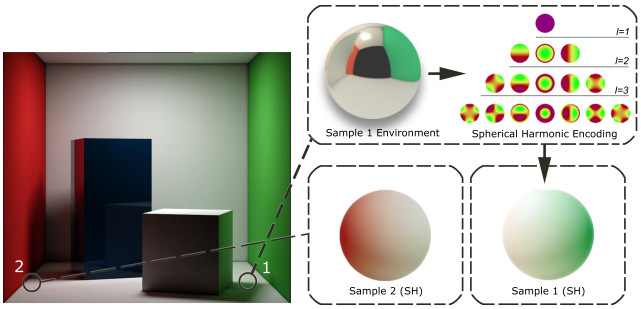


Fig. 4 Example of spatially varying illumination encoding with Spherical Harmonics (SH). The incoming lighting can be computed globally for the whole scene (far field), or locally, at multiple points (near field) as we show for the two samples near each colored wall. If we project the irradiance (top row) into an SH basis we obtain a diffuse low-frequency representation (examples in bottom row).

In Figure 4, we can see a pyramid of spherical harmonics bases $Y_{l,m}$ with different coefficients. It is important to know that, although usually nine bases are considered enough to account for 99% of the far-field irradiance at diffuse surfaces, this percentage is significantly smaller in glossy surfaces (requiring many more coefficients). Moreover, a small number of coefficients will never account for high frequency effects, such as cast shadows from high-frequency light sources (E.g: a point light representing the sun), even producing ringing artifacts if we try to increase the accuracy by adding more bases.

There are other popular basis in rendering such as Haar Wavelets or Spherical Gaussians (SG) [133], which also have very interesting properties. Most of the methods analyzed in this survey use the spherical harmonics encoding (more details in Section 3).

3 Single Image Inverse Appearance Reconstruction

Given a single image, the general goal of inverse appearance reconstruction is to obtain a set of parameters, that, for a function (known or not), produce the same original image.

It can be argued that even a simple image auto-encoder performs appearance reconstruction, just by learning projections (functions) to and from deep latent space variables (parameters). However, one of the most desirable properties for those parameters in computer graphics is *editability* (*e.g.*, change the color of a wall, dim the illumination, remove a specular highlight), and such neural parameters would, in most cases, lack any meaningful semantics or intuitive control over specific components of the image formation. This is a well-known limitation in the field of *neural rendering* [126], where the generality and editability required to create novel images with a neural network is very constrained because all the light, geometry, and material interaction events described in Section 2 are learned and embedded in an implicit unknown function. These models are often trained with specific parameter variations (*e.g.*, face relighting with environment illumination images, view synthesis with multiple lightfield views) and thus, any novel output image is limited to the parameter space sampling considered in the original training set.

The traditional approach to decompose an image into editable components is to mimic the optical process of image formation. However, as can be inferred from Equation 3, this problem is generally ill-posed because the number of physical parameters to infer is substantially bigger than the number of known values in the system, resulting in several ambiguities. For example, the color of a pixel might be caused by the color of the light source or the color –reflectance– of the material. Most of the previous work assumes gray-scale lighting, while only a minority assume colored lighting ([15, 95, 5]). Other ambiguities derive from the non-orthogonal parameter space, with multiple combinations yielding the same pixel value. This ambiguity, sometimes referred as the scale ambiguity, has been addressed by previous work [95] in the loss function, using a Scale Invariant L2 loss instead of the regular L2 (further details in Section 7), or imposing priors on the albedo layer by means of bilateral filtering or L1 losses.

The complexity of the inverse reconstruction problem will be thus determined by the rendering function used to model the scene. As such, given a single RGB image as input, the tendency has been to simplify the rendering model, reducing it to the components with the most significant contribution to the final appearance (*e.g.*, choosing direct over indirect illumination, diffuse Lambertian over complex BSDF models, or surface-level geometry instead of micro-geometry). This has ultimately lead to the *intrinsic* image formation models as described in Section 3.1. Any optical phenomena at the target image which is not reproducible by those models are thus dismissed, and either accumulated as errors in the wrong intrinsic parameter layer or stored as a residual image, trading simplicity in the parameter estimation for precision in the reconstruction. For example, the *intrinsic diffuse model*, with albedo and shading as the only unknown parameters, is unable to capture effects like specular highlights or inter-reflections. Multiple nuanced errors are accumulated by such a simple assumption: if applied to the decomposition of a translucent object, it will generate smoother normals than the actual surface, due to the natural blurring of gradients produced by subsurface photon scattering [31]. To capture those effects, along with the possibility to modify the geometry or lighting of the scene in post-processing, a complex image formulation with lights, materials, normals, and scene depth as controllable parameters is required.

3.1 Intrinsic Image Formation Models

From a practical standpoint, most of the methods that tackle the problem of inverse appearance reconstruction for generic scenes or objects can be classified into several categories, according to the image formation model they assume (see Figure 5). All the methods reviewed in this survey are classified according to these models in Table 2.

Intrinsic Diffuse, $\mathrm{I}_d(I|A,S)$. The input image is parameterized by the albedo (A) and the shading (S) images according to Equation 6. This model assumes the materials of the objects in the scene are purely Lambertian. The majority of the methods shown in Table 2 chose this model due to several reasons: reduced number of unknown parameters, existing priors derived from classical approaches, existing labeled datasets, and the assumption that the isotropic The Lambertian model is a good approximation to the dominant reflectance in everyday scenes [95, 147, 149, 64, 97, 56, 8, 25, 37, 78, 79, 86, 12, 74, 139].

Intrinsic Residual, $\mathrm{I}_r(I|A,S,R)$. An image can be decomposed into an additive combination of the multiple bounces of light. The first bounce of light is directly reflected from the surface before reaching the camera sensor (direct illumination), the subsequent bounces of light hit or penetrate other surfaces a variable number of times producing global illumination effects like color bleeding, subsurface scattering, or caustics (see Figure 1). In the literature of intrinsic image decomposition there is a body of work focused on isolating global illumination effects, generally simplified to diffuse interreflections on lambertian surfaces. Early methods altered the illumination of the scene in order to estimate the light transport between pixels. For instance, Seitz et al. [113] build upon shape-from-interreflection methods, computing

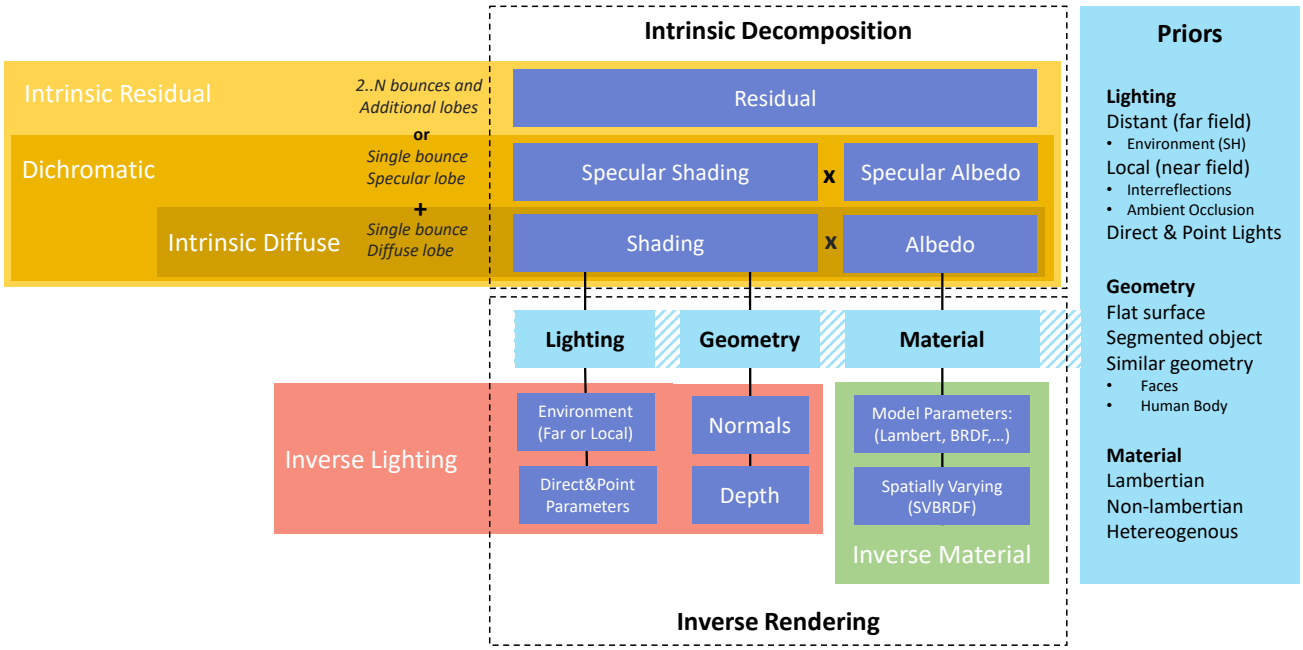


Fig. 5 Taxonomy of single image inverse appearance reconstruction methods: intrinsic decomposition and inverse rendering. The intrinsic diffuse model assumes Lambertian materials and jointly couples light and geometry interaction through a shading image. The intrinsic residual captures a residual image of the sum of the other non-diffuse effects. In between, several methods in the literature have used the dichromatic reflection model, where only a colored specular reflection is taken into account. Inverse rendering methods aim at recovering the full parameterized scene (camera, lighting, geometry, and material) to synthesize novel views or to enable predictably edits. To simplify the problem, many methods impose priors over the scene elements, *e.g.*: distant lighting through spherical harmonics illumination, or known proxy geometries (faces, human bodies, flat surfaces, etc.). Note that the boundary between intrinsic decomposition and inverse rendering is really fuzzy, as recent intrinsic methods are aided by implicit modeling of some scene elements (lighting through environment maps, or geometry through normals). A key difference between both approaches is that pure inverse rendering methods have the goal of modifying all the scene parameters, while intrinsic decomposition targets a more physically correct estimation of the albedo layer.

the forward propagation of light (N bounces) and finding the cancellation operator which removes the effects of each interreflection, from a set of images in which an individual scene point is illuminated by a narrow beam of light. Similarly, Bo et al. [30] used a projector to illuminate areas of the scene and capture the multiple interactions of radiance from the target area with other surfaces, requiring a number of captures linearly dependent on the amount of recursive light bounces being estimated. The smooth nature of interreflections and user-based constraints are leveraged by Carroll et al. [19] with compelling results, albeit quite sensitive to the correct placement of the strokes that guide their iterative weighted least-squares optimization process. If we consider video sequences, initial clustering can be propagated to additional frames leveraging temporal information, in such fashion, Ye et al. [138] used a Bayesian Maximum a Posteriori formulation. Similarly, Meka et al. [91] relied on iterative reweighted least squares to achieve real time decomposition. However, only direct illumination is assumed, assigning interreflections to shading variations. Finally, many methods rely on multiple views of the same scene, with different viewpoints or varying illumination conditions. In this line, Duchene et al. [35] obtained separated illumination layers from outdoor photographs accounting for secondary light bounces from close surfaces, sun illumination with cast shadows, and indirect light from the sky, by propagating values in image space supported by an approximate 3D reconstruction from multiple camera views of the same scene. All these methods require information of the same scene from multiple sources, whether it is additional illumination, a novel point of view, or pixel annotation by user intervention. For instance, in the case of mirror-like reflections, the traditional approaches have estimated the shape of the reflective surface either by introducing coded lighting with projectors [4], or capturing from multiple viewpoints, as shown by Godard et al. [44], even accounting for self-interreflections. It is only recently, with the development of differentiable rendering algorithms capable of physically-based inverse global illumination, such as Mitsuba [100], that interreflections have started to become tractable in single image decomposition.

In addition to albedo and shading parameters, the intrinsic residual model, as defined by Equation 12, introduces an extra term (R) to account for all the remaining optical effects, both due to multiple interactions of light *e.g.*, ambient occlusion, color bleeding from inter-reflections, caus-

tics, and to account for additional material reflectance components *e.g.*, specular reflections, translucency, scattering effects, etc.[117,90,114,146,77]. While none of the deep learning-based methods use it, the dichromatic reflection model (Section 2.3.2) was also used to account for metallic objects and colored speculars in traditional approaches [130, 88,9].

Inverse Lighting, $\mathrm{I}_I(S|N,L,D)$. A more expressive set of methods further parameterize the shading as the result of the interaction between the normal map (N), the illumination (L), or a depth map (D). This implies an extra level of complexity, as the amount of unknowns in the system increases. Some methods assume distant lighting (far-field) and thus use an approximate reconstruction of the environment map (E) [114], while others use spherical harmonics (SH) with a fixed number of coefficients [139] or assume directional lighting (dirL) [56] parameterized by its 3D position and intensity. Unlike other methods, which evaluate the render equation during training, Janner et al. [56] also learn the *render shader* within the deep network architecture, akin to *neural rendering*. Most of these methods reconstruct the surface normal N, as it is a required component to compute the shading. Few approaches have started to consider near-field illumination effects, such as ambient occlusion [61] or local incoming illumination [146,77], as per-pixel environment maps encoded with Spherical Harmonics or Spherical Gaussians (svSH, svSG).

Most approaches combine multiple formation models, often with coupled estimation of geometry and light sources. For instance, Sengupta *et al.* [114] train one network to predict an intermediate representation of normals, environment maps, and albedo map, using a closed-form Lambertian shader for far-field lighting. This method also relies on the intrinsic residual model: in a second self-supervised step, they train a residual network that takes as input both the normals and environment map along with the predicted albedo, and estimates the remaining residual illumination effects which were not captured by the first pass (e.g. inter-reflections, cast shadows, near-field illumination).

Inverse Material. Beyond the Lambertian material model, there are a few methods which introduce more complex materials. Meka *et al.* [90] use Blinn-Phong (Section 2.3.2) and regress the shininess coefficient (s-BP). Sengupta *et al.* [114] train the network with a dataset that contains glossy objects rendered using Phong. In this case, the use of non-Lambertian materials only serves to reinforce the estimation of the intrinsic residual term, as the network is not designed to explicitly estimate Phong material parameters. The more complete method so far is the work of Li *et al.* [77], which assume a physically-based microfacet material model [62, 50] and predicts the albedo and roughness (r_{SVBRDF}) parameters, besides illumination, depth, and normals. Also related to this problem are the methods that estimate a coupled representation of material and illumination using reflectance maps [52,108].

3.2 Semantic Priors

Some works have focused on specific scenarios to enforce semantic priors and, therefore, simplify the challenging problem of inverse appearance reconstruction. This simplification allows to leverage known parametric models to represent surfaces, deformation, and appearances, which significantly reduce the complexity of the problem. Notice that such scenario-specific methods are built on top of the image formation models described in Section 3.1. In this section we discuss several domains where learning-based methods have been proposed (*e.g.*, flat materials and single objects, faces, and humans). We explicitly link each of the methods discussed below with its corresponding underlying image formation model from Section 3.1, and provide insights about the benefits of using semantic priors.

3.2.1 Flat Materials and Single Objects

Estimating complex material parameters can be done much more easily by targeting single planar materials [32,76,81, 132] or isolated objects [82]. Our survey is mainly focused on arbitrary scenes for which it is not possible to make such geometric assumptions. Nevertheless, for consistency with the literature, we overview the methods which evaluate their performance on the MIT Intrinsic dataset [45], or use such dataset for training (Section 4.1). There are two exceptions: the work of Meka *et al.* [90], which we review to connect the diffuse intrinsic decomposition model with more complex svBRDF material models; and the work of Janner *et al.* [56] as serves to link the problem with neural rendering methods.

Several existing methods estimate a microfacet svBRDF model from one or several images of the material captured with a mobile device. Deschaintre *et al.* [27,28,29] present a framework based on deep neural networks (UNets) trained using self-supervision and render losses. Gao *et al.* [39] uses a similar framework augmenting the training data using rendered views of the material. Recently, Guo *et al.* [47] demonstrate that GANs, StyleGAN2 [63]) in particular, can be powerful frameworks for estimating the reflectance properties enabling material editions using a learned latent space.

3.2.2 Faces

Many methods leverage the seminal work of Blanz and Vetter [13] on modeling 3D faces with a low-dimensional 3D morphable model (3DMM) to incorporate geometry priors

to solve the intrinsic decomposition problem. Importantly, assuming a tight cropped image of the face, most of existing works in the area of intrinsic faces are able to train their model directly from unlabeled images in-the-wild by computing the pixel-wise difference of input and predicted image.

In the context of the different simplifications of the image formation model described at Section 3, initial methods on faces focus on the *intrinsic diffuse* simplification (*i.e.*, estimate albedo A and shading S) [118,128], while more sophisticated approaches predict the components of the *intrinsic residual* model (*e.g.*, also predict specular or noise) [137]. In order to estimate the shading layer, most methods also formulate an *inverse lighting* problem and estimate normals and lighting [128].

Shu *et al.* [118] learn a subspace capable of representing face image with explicit disentanglement of normal, shading, and albedo components. This enables seamless edits to face images, including manipulation of expression, and adding glasses or beard. Even if it is a self-supervised method, Shu *et al.* require intermediate constraints to prevent the network to converge to naive solutions such as shading to be constant and albedo capturing all the appearance. To this end, they introduce a weak supervision strategy based on enforcing the estimated normals to be closed to those extracted from a 3DMM. Tewari *et al.* [128] also use 3DMM to reconstruct faces from monocular images by a model fitting approach. They propose a carefully designed subspace with a latent parameters that match to a semantic encoding of facial expression, shape, illumination and albedo. Illumination is modeled using Spherical Harmonics [94] with nine coefficients which tends to produce oversmooth results.

SfSNet [115] proposes an architecture to learn to separate albedo and normal layers. Their key observation is that in previous networks [117] all high-frequency details are passed through the skip-connections. Therefore, the latent representation is unable to figure out whether fine details such as wrinkles or beards are due to shading or albedo. Consequently, they propose a new architecture that learns to separate both low and high frequency details into normal and albedo to obtain a meaningful subspace. This is used along with the original image to predict lighting represented with spherical harmonics. This model reconstructs more detailed shape and reflectance than MoFA [128] because it is not limited by the 3DMM prior. Similarly, in their subsequent work, Tewari *et al.* [127] also propose a method that is not bounded by the underlying 3DMM prior. They propose and end-to-end trainable system that uses 3DMM just as a regularizer and learns corrective space for out-of-space generalization. Despite using the same illumination model as [128], the corrective spaces enables the estimation of geometry, reflectance and lighting of higher quality, but it still assumes a Lambertian reflectance. Follow-up research [125] further improve upon the use of priors and learn from scratch an appearance and geometry model to estimate the surface, albedo, and illumination of unconstrained images with unprecedented level of detail. Key to their success is a new graph-based multi-level face representation. They use both a coarse shape deformation graph and a high-resolution surface mesh, where each vertex has a color value that encodes the facial appearance. Despite the impressive results of these works, they are all based on the assumption of a distant and smooth illumination and purely Lambertian surface properties, which prevents the modeling of any residual component (*e.g.*, specular effect), which is a fundamental part intrinsic imaging.

Yamaguchi *et al.* [137] go one step forward and focus on learning to infer *high-resolution* facial reflectance, including albedo and specular layers, and fine-scale geometry from an unconstrained image. In contrast to other approaches [118, 128,115,127], their model goes beyond the Lambertian assumption, and accounts for non-trivial lighting effects such as ambient occlusion and subsurface scattering. To this end, they use two identical architectures to extract specular and albedo textures, arguing that these components capture different optical features of the skin and therefore a single network easily fails in modeling conflicting features. Additionally, they incorporate an image completion step that generates complete texture maps.

A different approach to learn an *intrinsic residual* image formation model for faces is to circumvent the use of explicit image parameters (*e.g.*, S, A, or N) altogether, and attempt to learn a specific input-output model for a particular task. This allows the learned model to account for non-Lambertian reflectance or go beyond Spherical Harmonics (SH) illumination. Sun *et al.* [122] train a network that takes as a input a single image of a faces and a target illumination, and directly predict the relit image. Similarly, Zhou *et al.* [145] propose a deep architecture to relight single images conditioned to a target lighting expressed in SH. Meka *et al.* [89] also propose a deep learning-based approach to learn a mapping between spherical gradient images and the one-light-at-a-time (OLAT) image from a particular direction. Even if no explicit reflectance model is imposed, the residual component is captured with a task-specific perceptual loss trained to pick up specularities and high frequency details. Nestmeyer *et al.* [98] propose a hybrid approach where the diffuse component is represented with an explicit model, and the residual is unconstrained and modeled with a neural network. This allows for effects that are not predictable by the BRDF, such as subsurface scattering and indirect light.

3.2.3 Full Human Body

Most of the above-discussed methods for intrinsic decomposition of faces share a common limitation: they ignore light occlusion. While this is an acceptable assumption for non-articulated and rather convex surfaces such as faces, full human bodies often present self-shadows and self-occlusions which requires more complex illumination models. This is specially evident in exceptions such as the work of Nestmeyer et al. [98] which consider a binary mask encoding the visibility from a point light source. The shadows cast by the nose or the head on the neck increase the realism of the relighting results. Few method exist that tackle such challenge with a learning-based strategy. Kanamori *et al.* [61] propose a method that is able to learn and encode light occlusion from masked full body images. Despite being a deep-learning based approach, their loss functions explicitly minimize the intrinsic components of the images (*e.g.*, albedo, light, and transport map).

4 Datasets

One of the critical pieces of any learning-based approach is the data available for training. A dataset of sufficient variety, significance, and size is required regardless of the formulation of the learning problem, e.g., supervised, semi-supervised or unsupervised. The complexity of the intrinsic decomposition problem makes creating labeled datasets of all the individual components a highly challenging task, as this type of data cannot be freely obtained from the natural world, neither is easy to gather from human annotations.

The very first dataset with *explicit* labels for reflectance and shading was created in a laboratory setup where a few small painted figurines were coated with neutral gray to create shading images [45]. For a long time, it was the only ground truth dataset available for quantitative evaluations, and not until recently, the boost in performance and quality of physically-based rendering engines along with the flourishing of 3D datasets has facilitated the creation of larger, more complex, and heterogeneous data for training. The lack of labeled data motivated the use of alternative learning-based solutions. In this regard, the main approach has been to leverage semantic knowledge of the scene content and learn from relative measurements instead of regressing absolute radiance values. Relative measurements of the scene appearance can be found relatively easily, for example, humans are quite skilled at judging whether two surfaces are made of the same material despite illumination variations [10]. This property (the *Albedo Invariance*), which is key to disambiguate the contribution of the intrinsic components, can also be exploited given existing and freely available datasets such as time-lapse sequences.

In this section, we review the most common datasets which are being used for training and evaluation purposes. The core datasets discussed here correspond to those explicitly published and made available to the public. Nevertheless, many methods build on these datasets to create their own without publishing it. In the following, the description of each dataset includes the methods that use them along with existing derivations. We organize the datasets according to the following properties:

- **Size (# 3D Models, # Sequences, # Imgs)**: Total number of images (Imgs), sequences of the same scene with varying illumination or viewpoint (Sequences), or number of 3D renderable scenes. Note that ShapeNet [21] and SUNCG [121] are datasets of renderable objects/scenes so the amount of images generated to train the models depend on the particular method.
- **Scene Content**: Images included in the dataset might be of individual objects (*obj*) or complex scenes. In the latter case, some datasets might contain indoor scenes (*ind*), outdoor scenes (*outd*), or a combination (*any*).
- **Scene Syn/Real**: The image can be *synthetic* or *real*. In the former case, some of the datasets use physically-based rendering engines, while other have been generated with non-photorealistic ones.
- **Source**: The constraints used for training might be automatically generated from the data (*auto*), or come from *human* annotations.
- **Labeling**: Some datasets provide *sparse* annotations of just a few pixels of the images, while others provide *dense* per-pixel annotations.
- **Constraints**: the data can be labeled with *explicit* (or *absolute*) values for each of the unknown parameters, or can be used as a way to extract *relative* relationships for the intrinsic components within a single image, or across several images of the same scene. If the latter, the dataset might be additionally organized in *sequences*.

The reminder of this section is organized according to the *scene content*, namely objects or general scenes. The other properties will be mentioned within the description of each dataset. Please refer to Table 1 for a comprehensive summary of datasets and their properties, and Figure 6 for a selection of representative images.

4.1 Objects

As a way to constrain the problem, some methods have limited the domain to isolated objects. This enables the use of additional priors about the underlying geometry and shape reducing the number of possible solutions and enabling more complex materials, and illumination models. Such is the case for methods targeted at flat surfaces, faces or humans. Although we briefly review them in Section 3.2, in

Fig. 6 Representative images for the most relevant and publicly available datasets

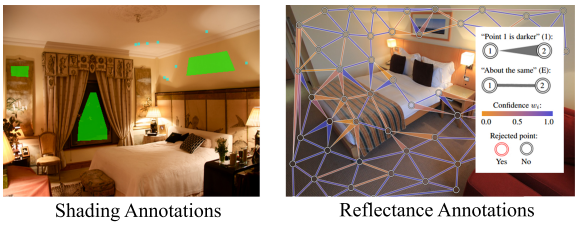
Table 1 Summary of datasets used in the literature of deep learning-based intrinsic decompositions which are avaible to the public. [1] The render engine used to generated the images of this dataset is a non-photorealistic one. More details about the definition of each category in the accompanying text.

Dataset	# 3D Models	# Seqs	# Imgs	Max Image Size (approx)	Scene	Syn/Real	Source	Labeling	Constraints
ShapeNet [21]	4000	*	*	*	obj	syn	auto	dense	abs/rel
MIT Intrinsics [45]	-	10	220	600	obj	real	auto	dense	abs/rel
SUNCG [121]	40k	*	*	*	ind	syn	auto	dense	abs
MPI Sintel [18]	-	-	890	1024×436	any	syn[2]	auto	dense	abs
CGIntrinsics [78]	-	-	20k	640×480	ind	syn	auto	dense	abs
IIW [10]	-	-	5230	512	ind	real	human	sparse	rel
SAW [64]	-	-	6677	512	ind	real	human	sparse	rel
BigTime [79]	-	155	6500	1080	any	real	auto	dense	rel
MegaDepth [80]	*	200	150k	1080	outd	real	auto	dense	rel

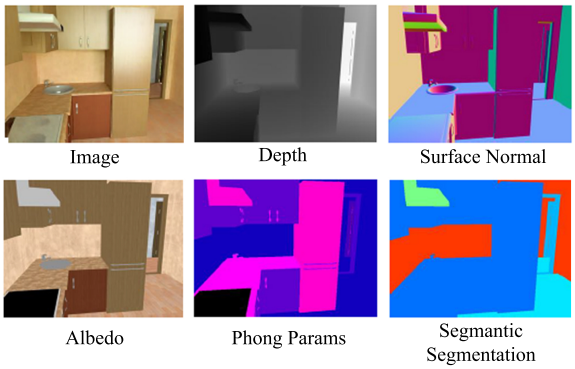
this survey we focus on the methods that deal with arbitrary object shapes or provide quantitative errors on common metrics. Another advantage of dealing with objects instead of generic scenes is that it becomes less complex to capture/render variations (or sequences) of images of the object as seen under different perspectives or viewpoints. Such is the case for the two datasets reviewed below.

MIT Intrinsic [45]. Contains real images of 20 objects with ground truth albedo and shading captured under 11 different directional light sources, resulting in 220 images (20 sequences). Shading images were obtained by painting the object with gray spray. The objects were photographed within a controlled setup which minimized indirect illumination and allowed easy alignment between different shots. Even though it is a small dataset, the majority of methods have used it for training or fine-tuning their models. There are two divisions of this dataset that people have consistently used to compare performance: the Barron split [5], which divides each sequence by image, and the Direct Intrinsics split [95], which divides the sequences by objects.

ShapeNet [21]. The ShapeNet dataset is a richly annotated dataset of 3D objects with albedo maps of over 4000 object categories. Different methods have rendered the available 3D models into datasets of different sizes, objects variety, and illuminations. They all leverage a physically-based rendering engine (e.g. Mitsuba[55] or Blender Cycles) as well as provide dense labels. Additionally, the ShapeNet 3D dataset has been used to generate sequences of images of the same object under different illumination conditions, providing relative constraints for the learning formulation. Meka *et al.* [90] render 100k images of 55 objects using Blinn-Phong materials, and 45 indoor environment maps. Shi *et al.* [117] render more than 2M images of 30k objects, using Phong materials, and 98 environment maps. They perform category-specific training using four objects (car, chair, airplane, and sofa), and evaluate cross-category generaliza-



(a) Sparse labels



(b) Dense labels

Fig. 7 Difference between sparse and dense annotations. (a) Sparse labeling provided by human annotators. On the left, in Green: regions of near-constant shading but with possibly varying reflectance. In Red: edges due to discontinuities in shape (surface normal or depth). In Cyan: edges due to discontinuities in illumination (cast shadows). Original images by *herry* and *uggboy* @ Flicker. (b) Dense per pixel labels provided by a synthetic dataset [114].

tion. Janner *et al.* [56] also study cross-category generalization, and use Blender Cycles to render Lambertian materials *on demand* in the unsupervised setup. Baslamisli *et al.* [8] render 20k images of different 3D models assigning random colors to the albedo materials to introduce more variety. Ma *et al.* [86] further leverage this dataset to generate multi-illuminant training sequences by randomly picking 10 different light positions per each of the 10 selected object categories (each one containing 100 objects). The amount of training samples obtained thanks to ShapeNet is huge, however, none of the methods published the splits so they are not available for comparing performance across different methods.

4.2 Scenes

Dealing with arbitrary types of scenes is the ultimate goal of intrinsic decomposition methods. However, the appearance of a material under different types of illumination can change dramatically (e.g. clear-sky vs cloudy day, or natural vs artificial lighting). Most of the existing datasets in this category, in particular, the ones that are synthetically generated, describe indoor scenes. In contrast to the ShapeNet dataset presented in the previous section, generating on-the-fly samples for arbitrary scenes is much more expensive, consequently, existing synthetic datasets in this category are static and provide absolute learning constraints (MPI Sintel, SUNCG, CGIntrinsics). Relative comparisons have been gathered from crowd-sourcing experiments (IIW, SAW) or from the physics of the image formation (BigTime, MegaDepth).

MPI Sintel. This synthetic dataset [18] of animation scenes was originally designed for optical flow evaluation, however, thanks to providing the albedo layer, it proved useful for the evaluation of intrinsic decomposition methods. As the original renders contained complex lighting effects (specular highlights, inter-reflections, etc.), it was re-synthesized to appear purely *lambertian* and coherent [23, 95]. It contains a total of 890 images from 18 scenes with around 50 frames each. Like MIT Intrinsic, there are two known splits: the *scene split*, placing the whole scene (all frames) either completely in training or completely in testing; and the *image split*, where the same scene will appear both in train and test, placing different frames in each split. Methods have consistently used the same splits of this dataset for training and evaluating their methods.

SUNCG. The SUNCG dataset [121] contains 40k manually created indoor environments with dense volumetric semantic annotations. Likewise ShapeNet, the 3D models of this dataset have served as a baseline to several methods which have rendered the scenes under different illumination conditions and materials. Zhou *et al.* [146] have used it to render 58949 images of Lambertian surfaces using Mitsuba [55] for which they also have the albedo, normal map, depth, and shading generated by setting all the materials to diffuse and reflectance to 1. Sengupta *et al.* [114] also departs from the SUNCG dataset, enriching the existing material models using Phong [70]. It contains 230k images of indoor scenes physically-based rendered under multiple outdoor environment maps. It further provides the same scene rendered under both diffuse and specular settings, as well as labels about normal maps, depth, Phong model parameters, semantic and glossiness segmentation. In order to reduce ray-tracing timings, the method has used deep denoising [20].

CGIntrinsics [78]. Taking 3D models and textures of indoor scenes from the SUNCG dataset [121], this dataset contains 20k images (and albedo maps) of physically-based renderings using path tracing with global illumination. The dataset also provides: the set of 50 synthetic scenes provided by Bonneel *et al.* [14], code to compute the shading image from the render and the albedo, the segmentation of each image into superpixels [2], the training split, and the precomputed bilateral embedding used in the paper to guarantee shading smoothness.

Intrinsic Images in the Wild (IIW). IIW dataset [10] is a sparse large-scale set of relative reflectance judgments of indoor scenes collected via crowdsourcing, containing

over 900k comparisons across 5000 photos. Along with the dataset, the authors provide a metric for evaluating the algorithms performance, the Weighted Human Disagreement Rate (WHDR), which measures the percent of human judgments that the method predicts incorrectly, weighted by the confidence of each judgment. Although it has the limitation of the sparsity of the annotations, this dataset is being used consistently for comparing performance.

Shading Annotations in the Wild (SAW). Following the methodology of IIW, SAW dataset [64] contains 15k sparse annotations of shading gradients over 6000 images of indoor environments: smooth shading, normal/depth discontinuity, and shadow boundary. Using Precision-Recall metrics, this dataset can also be used across methods to compare performance.

BigTime [79]. Large dataset of real image sequences of both indoor and outdoor scenes under varying illumination. Sky, and dynamic objects such as pets, people, and cars were masked out. It contains a total of 145 sequences from indoor scenes and 50 from outdoor scenes, yielding a total of over 6500 images.

MegaDepth [80]. Contains 150k images of 200 different landmarks which have been reconstructed using state-of-the-art Shape-from-Motion and Multi-View-Stereo methods. Each image is accompanied by its depth map and camera parameters, so that it is possible to reconstruct the scene from a vast range of view points.

5 Learning Formulation

The problem of inferring the intrinsic components of a scene given a single image has been traditionally (*i.e.,* before deep learning algorithms flourished) addressed with optimization-based methods that make assumptions about the contents of the scene or the physics of the imaging process. For example, some methods assume clear-sky illumination [102,38,5], monochromatic lighting [15,5,22], piecewise smooth reflectance [10,11,112], or force areas with similar texture or chromaticity to have the same reflectance [40,143]. These assumptions, formulated as statistical priors, were combined within optimization frameworks such as Conditional Random Fields (CRFs) [66,10], multi-scale gradient-based solvers [5], or closed-form systems of equations [143]. However, as in many computer vision problems, finding the optimal solution is computationally very complex, and the use of ad-hoc priors and heuristic parameters narrows the scope and the generalization capabilities of the solution. For example, the common assumption that an edge might be produced by either a change of reflectance or a change in shading [51,124,45], overlooks the fact that both changes might occur simultaneously as it happens at occlusion boundaries [40].

In recent years, Convolutional Neural Networks (CNNs) [72,73] have become the state-of-the-art models for solving many different computer vision tasks, such as object segmentation, image classification or image-to-image translation [140,119,54]. A key factor to their success is the fact that they internally learn hierarchical patterns that represent image features at multiple scales, in a way that loosely mimics the behavior of the visual cortex in mammals. Additionally, the recent boost in the size and variety of datasets, as well as the expansion of available computing power, have made those models ubiquitous in computer vision problems. Consequently, in the intrinsic image decomposition literature, CNNs have gradually substituted or complemented traditional hand-crafted priors and assumptions, to features that are directly learned from data. We refer to the problem of learning an intrinsic image decomposition from images using CNNs as Deep Intrinsic Images. There are several ways in which the solution to this problem can be approximated by CNNs.

In this section, we describe the methods according to the formulation of the learning problem, or similarly, in the way data and prior knowledge can be leveraged to train a model capable of providing a solution. Despite the variety of existing inverse image formation models described in Section 2, all the methods share similar learning strategies, which are frequently compatible and combined in the objective function to provide the best performance. We have identified four main complimentary strategies which have been used in the literature so far:

- Weak-Supervision: use human judgments about the perception of materials and illumination in images (Section 5.1).
- Full-Supervision: leverage labeled datasets in full regression frameworks in order to learn statistical priors over the parameters domain (Section 5.2).
- Self-Supervision: include an image formation loss in order to guarantee that the target parameters will effectively reconstruct the original input image (Section 5.3).
- Priors: explicitly model prior knowledge about the nature of each individual intrinsic component (Section 5.4).

Figure 8 shows an overview of these strategies and how they relate to the existing datasets. Although the majority of the datasets serve a single purpose, some of them can be used to feed in more than one learning strategy, particularly the *ShapeNet* dataset [21] as it enables the generation of samples *on-the-fly* during training. Table 2 shows the methods covered in this survey according to the previous categorization. Note that only a few methods follow a single strategy, while the majority of them combine three or four, thus, most of them will be discussed in more than one section.

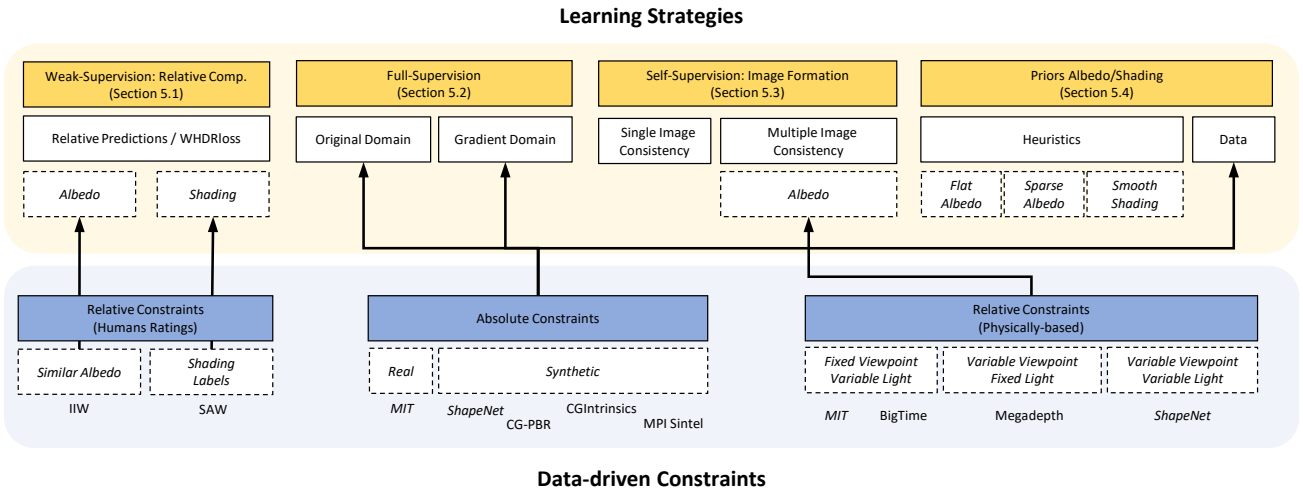


Fig. 8 Learning strategies and relationship with data-driven constraints.

Table 2 Summary of methods presented in this survey. The explanation of the models is presented in Section 3.1. A full description of the learning strategy and neural architectures is presented in Sections 5 and 6.

Method	Model				Learning Strategy										Network Architecture	
					Weak-S		Full-S		Self-S		Priors					
	Diffuse I_d	Residual I_r	Lighting L	Material	RP	$WHDR_{loss}$	O	∇	SIC	MIC	A_f	A_{sp}	S_{sm}	Data	C	I2IT
Narihira *et al.* [96]	-	-	-	-	*S*										P-W	
Narihira *et al.* [95]	✓						*AS*	*A*								B
Zhou *et al.* [147]	✓				*A*										P-W	
Zoran *et al.* [149]	✓				*A*										P-W	
Kovacs *et al.* [64]	✓				*S*											B
Nestmeyer *et al.* [97]	✓					*A*					✓					B
Shi *et al.* [117]		✓					I_r *AS*	*A*								E-D
Janner *et al.* [56]	✓		dirL				*ANL*		✓							E-D
Meka *et al.* [90]		✓	E	BP			I_d R s-BP		✓							E-D
Baslamisli *et al.* [8]	✓						*AS*	*AS*	✓							E-D
Cheng *et al.* [25]	✓						*AS*		✓		✓	✓	✓			Res
Fan *et al.* [37]	✓					*A*	*AS*	$ASP_{\nabla A}$			✓					Res
Li *et al.* [79]	✓								✓	✓	✓		✓			E-D
Yu *et al.* [139]	✓		SH				*AN*		✓	✓		✓				E-D
Li *et al.* [78]	✓					*AS*	*AS*	*AS*	✓		✓		✓			E-D
Ma *et al.* [86]	✓						*AS*	$P_{\nabla A}$	✓	✓						E-D
Bi *et al.* [12]	✓						*AS*		✓	✓	✓					E-D
Lettry *et al.* [74]	✓						*AS*	*AS*	✓							Res
Sengupta *et al.* [114]		✓	E	Phong		*A*	*ANL*		✓							E-D
Zhou *et al.* [146]	✓		svSH			*AS*	*ASNL*	*ASN*	✓							E-D
Liu *et al.* [84]	✓								✓		✓			✓		Res
Li *et al.* [77]	✓	✓	svSG	svBRDF		*A*	*ANDL* r_{svBRDF}		✓		✓					E-D

5.1 Weak-Supervision: Learn from Relative Human Judgments

As opposed to computers that only understand absolute color values, humans are very skilled in judging whether two surfaces are made of the same material despite variations in illumination [71]. This ability, known as color constancy, has been exploited in two recent datasets, the IIW [10] and the SAW [64] datasets, that contain human judgments about regions of images sharing *similar albedo*, or with assigned *shading labels* (smooth shading, normal/depth discontinuity, and shadow boundary), respectively. These judgments are provided as sparse sets of pairwise comparisons and accompanied by a confidence score, the *Weighted Human Disagreement Rate* (*WHDR*) [96], that takes into account disagreements between the raters. Several methods have used these relative constraints to train, fine-tune or evaluate their models. Below we detail the two main strategies to leverage such weak supervision to train deep networks.

5.1.1 Relative Predictions [RP]

Early approaches use this kind of data to train models that predict relative scores between two image regions. Then, they combine the output of these sparse predictions within

existing inference frameworks to provide smooth estimations. Narihira *et al.* [96] (trained on IIW) predict local lightness relationships by means of a CNN used as feature descriptor combined with ridge ranking regression. They provide relative predictions, but unlike other methods, do not attempt to reconstruct the intrinsic components. Using the SAW dataset, Kovacs *et al.* [64] predict the source of a shading gradient –smooth shading, normal/depth discontinuity, and shadow boundary– using a CNN along with a linear classifier. Then, they use the local prediction within a classical Retinex formulation [143] to estimate the intrinsic components. Zhou *et al.* [147] builds on a dense CRF framework [66, 10] initialized by the output of a siamese network used as binary classifier which behaves as a prior for reflectance. Concurrently, Zoran *et al.* [149] propose a framework to reason about ordinal relationships between local patches of the images; besides depth estimation, they prove to be successful on the IIW dataset for intrinsic decomposition. In this case, they use quadratic programming to propagate and smooth the estimations using superpixels [2].

5.1.2 Relative Comparisons as Weak Supervision [WHDRloss]

The second group of methods that use data from relative human judgments leverages such weak annotations for an extra supervision in the learning loss to constrain the reflectance, the shading, or both, depending on the dataset used. Instead of combining the prediction with external frameworks, these methods estimate the intrinsic components using only the predictions of their neural networks. To this end, they use the (*WHDR*) which, as mentioned, is a distance metric developed to evaluate the quality of the automatic estimations with respect to the human ratings. Nestmeyer *et al.* [97] proposed the only method to use this loss without further supervision to provide dense estimations within an end-to-end deep framework. They were followed by other methods [37, 114, 78, 146], which included this strategy as a form of weak-supervision to fine-tune the models trained with other sources of data.

5.2 Full-Supervision: Learn from Labeled Data

The use of labeled datasets for training machine learning models is a common training strategy, often necessary for obtaining successful models. The majority of methods mentioned in this section use absolute label values in regression losses. That is, for each labeled image, the error function is penalized if the estimated intrinsic components do not conform with the ground truth labeled ones. Only using this kind of supervision [95, 117], however, is not a guarantee that the estimated components will faithfully reconstruct the input image. Consequently, the majority of methods combine full regression with other forms of supervision, as shown in Table 2. An alternative way of using labeled data without explicitly connecting the labels in the loss function by means of regression is proposed by Liu *et al.* [84]. In such solution, large sets of images of each unknown intrinsic component were used to train individual latent spaces representations of each domain. We discuss such case in Section 5.4.

Jointly training with real and synthetically generated data is becoming a popular and successful trend (see Table 3). Synthetic data may be useful to teach the model the global shapes and common geometries, while real data is necessary to fill the domain gap necessary for the methods to work well with images taken under real-world illuminations and materials.

As common with any image processing algorithms, the regression of the intrinsic components can be done in the original domain of the image pixels or in the gradient domain. This may help the neural networks learn a better representation of the problem, but also hinder their learning capabilities by restricting the space of solutions they can find. We have classified the methods in this section according to whether the regression problem is formulated in the original domain of the intrinsic parameters, or whether they are transformed to the gradient domain before applying ℓ^1, ℓ^2, or perceptual regression losses. Note that other transformations applied to the intrinsic components before performing regression are also possible, for example, applying a bilateral filter to the reflectance layer. We discuss such cases in Section 5.4.

5.2.1 Original Domain [O]

Most of the methods regress the intrinsic components using either ℓ^1 or ℓ^2 losses. This strategy is sometimes referred as Direct Intrinsic [95, 117, 8, 37, 79, 86, 12, 74, 146], and can be used to initialize just part of the network [56, 114, 90].

5.2.2 Gradient Domain [∇]

Inspired by classical Retinex approaches [51, 45], which obtain the shading image by first predicting the gradients and then integrating them with Poisson-like reconstruction, a set of methods explicitly regress the gradients of the components. This loss has been applied only to the albedo [95, 117], or both the albedo and shading layers [8, 37, 78, 74, 146]. The probabilistic version of working in the gradient domain is presented by the methods that predict the probability of an albedo gradient ($P_{\nabla A}$). Fan *et al.* [37] complement a Direct Intrinsic network with a Guidance Network that is trained to predict binary albedo edges emulating the response of applying a L1 flattening [11], method used as ground truth. Similarly, Ma *et al.* [86] explicitly predict a

soft assignment mask with the probability of an albedo edge which, unlike before [37], is trained using the ground truth albedo and not a flattened version of it.

5.3 Self-Supervision: Learn from the Image Formation Model

There are several problems that make a fully supervised approach with explicit labels insufficient: First, it is not determined how perceptual differences will be distributed between the different intrinsic components. Second, there is no guarantee that the reconstructed image from the inferred components will exactly match the input. Finally, full regression methods require a huge amount of labeled data in order to generalize reasonably. In this context, self-supervised strategies formulated as the *Render Loss* or the *Image Formation Loss* have evolved to address these limitations. The key of these strategies is to introduce, at training time, per-pixel reconstruction of the input image as an additional signal to guide the learning process. This operation can be done for each single image of the training dataset (*Single-Image Consistency*), or across images for multi-image datasets with relative constraints (*Multi-Image Consistency*). Thus, it is critical to perform this reconstruction efficiently, as it has to be performed thousands of times during training. We discuss the trade-offs of choosing an image formation model in Section 3 and recent trends that allow more expressive models in Section 9. Here we assume that reconstructing the image from a set of intrinsic parameters can be done in a negligible amount of time.

5.3.1 Single-Image Consistency [SVC]

One of the most critical decisions in the problem of single image inverse reconstruction is to choose the model that should reconstruct the scene. As discussed in Section 3, there is a trade-off between the desired complexity of the target scene –geometry, material, illumination–, and the model complexity. For this reason, most of the methods dealing with arbitrary scenes choose simple formulations, either the intrinsic diffuse or intrinsic residual, with limited number of parameters that can be evaluated in real time. The *intrinsic diffuse* model receives as input the albedo and shading images, which, when multiplied together, reconstruct the input image [25,78,74,79,139,86,12,74,8]. The *intrinsic residual* has an additional parameter to capture the specular reflections and other light effects that do not belong to the diffuse behavior of light [90,114]. Regardless of the inverse model and the internal architecture used to capture the intermediate steps of the physics of the image formation process (more details in Section 3 and Section 6), reconstructions losses guarantee that the estimated components will be able to reconstruct the input image.

5.3.2 Multi-Image Consistency [MIC]

One important characteristic of the diffuse albedo of the materials is that its value remains constant despite variations of other scene properties such as the illumination or the view angle. This property has been exploited as an extra form of supervision by leveraging existing datasets which might not be specifically collected for the purpose of intrinsic decomposition and lack of explicit labels.

Time-lapse sequences, *i.e.* sequences of images of a static scene with varying illumination, have been widely used for years as input to estimate the intrinsic components. Weiss *et al.* [136] learn from the data that shading images of outdoor scenes convolved with a derivate filter are sparse, and apply it as a prior to estimate the intrinsic components. Sunkavalli *et al.* [123] additionally decomposed the scene into shadows, shading and reflectance under the assumption of clear-sky illumination. Later on, Laffont *et al.* [68] used the albedo invariance to constrain the decomposition within a classical optimization framework. *Multi-view sequences* contain the same scene under different perspectives or viewpoints. This kind of datasets have been mostly gathered with the purpose of 3D reconstruction, although occasionally used within the context of intrinsic decomposition. The method of Laffont *et al.* [69] used online photo collections as input to guide the decomposition, leveraging cues from partial 3D reconstruction. Duchene *et al.* [35] further provide a full 3D model of the scene enabling relighting applications of outdoor scenes. As opposed to the methods reviewed in this survey, which leverage these datasets for training only, these approaches require as input the whole sequence of images to decompose a single view of the scene.

The key idea of Multi-Image Consistency is to combine different estimations for reflectance and shading images taken from the same *sequence* but from different *images* of the sequence. The most frequent approach is to combine these cues with fully supervised training [12,86]. Li *et al.* [79] propose the only method that not requiring any explicitly labeled data for training but heavily relying on heuristic priors (as described in the next section). Finally, Yu *et al.* [139] are the first to use a multi-view stereo dataset which contains rich variations in illumination to train a single image inverse method, capable of recovering albedo, normals and two spherical harmonic lighting coefficients. The network is mainly trained in a self-supervised manner by cross-projecting the views using depth maps and camera projection matrices, imposing coherency in the reconstruction and in the inferred albedo, as previous methods do [79, 86,12].

5.4 Priors

Priors are the existing beliefs about a problem. In classical non-learning based approaches, the priors about each intrinsic component were, first, observed from the data and then modeled as hand-crafted image heuristics, taking into account each phenomena independently. For example, it was observed that under natural daylight and for narrow-band camera sensors, pixels with the same reflectance form a single line in logRGB space. Such observation was used, for example, to identify shadows boundaries [38]. An extensive overview of these priors and assumptions from a classical perspective is presented in Bonneel *et al.* [14]. Using deep learning architectures has deemed the use of such priors unnecessary in the majority of situations, as they are now implicitly learned by the deep model during training. However, due the complexity of the inverse reconstruction problem, some of these priors have proven to be still useful nowadays. In the following we focus on these heuristic priors, as well as present an different approach to learn them from the data.

5.4.1 Albedo is Piece-wise Flat [A_f]

Observing that the human visual system perceives colors locally and constantly with independence of the illumination conditions, the Retinex theory [71] was fundamental for the development of the most popular computational prior on the albedo, which assumes that it is piece-wise flat, of high frequency, and sparse [15, 11]. In a deep learning formulation, this prior can be applied in two ways: First, as an additional loss term that applies to the albedo only, being the L1 loss the most natural way to impose such constrain [79, 78, 86, 25]. Second, as an explicit filtering operation applied over such component to guide the learning process in a more aggressive way. In the latter case, the l1 flattening algorithm [11] or different variations of the bilateral filter [41, 105, 129] have been the most popular and successful strategies [12, 25, 37, 97].

5.4.2 Albedo is Sparse [A_{sp}]

In order to reduce the complexity of the problem, previous work relied on two strategies also related to the appearance of the albedo component. First, assuming that similar chromaticities values of the input image are likely to have the same albedo values, and second, that the amount of different albedos within a natural image is sparse [40, 10, 116]. Yu *et al.* [139] explicity take into account the former by apply a pixel-wise weighted penalty according to chromaticity values of the input image. The latter was implicitly considered by Cheng *et al.* [25], who use a deep perceptual loss in order to preserve textured details [59].

5.4.3 Shading is Smooth [S_{sm}]

Also derived from Retinex, and from the assumption of convex and smooth 3D geometries, this prior assumes that smooth image variation are mostly due to the interaction of light with a continuous smooth surface [51]. Existing work has modeled it with the ℓ^1 or ℓ^2 norms [25, 78], or minimizing second-order shading gradients [86].

5.4.4 Data-driven Priors [Data]

In a learning-based approach, the priors are directly learned from the available data as probability distributions. In a fully-supervised setup, the most common way to leverage labeled data is by feeding the network with one image along with its corresponding intrinsic components. After enough data and iterations of the training process, the deep neural network will have learned an internal and *coupled* implicit representation of each of the intrinsic components. An alternative way to use labeled data for training was proposed by Liu *et al.* [84]. They present a deep network architecture that do no require aligned relationships between the intrinsic components and the input image in order to learn the prior distribution of each of them. Instead, it feeds in the network with independent sets of images of each intrinsic component. In this regard, the prior information is given by the actual data, and the work of understanding the unique features of each component is a learning task.

6 Deep Neural Network Architectures

In the previous section, we discussed the problem of deep intrinsic decomposition from a learning formulation perspective. The design of the deep neural network architecture that is used to learn to decompose shading and reflectance is also important, as different network architectures are biased towards finding solutions of different characteristics. The neural network architecture categorization proposed herein is summarized in Table 2. In this section, we will describe the methods according to their network architecture design choices. We have identified two main groups that have been used in the literature so far:

- Networks designed to learn **pairwise comparisons** between patches of an image (Section 6.1).
- Networks designed to perform an **image-to-image translation**, from the input image to either an intermediate representation of the decomposition, or to the full image decomposition (Section 6.2).

6.1 Pairwise Comparison

(P-W C) As discussed in Section 5.1, early deep intrinsic decomposition methods relied on sparse local judgments of either reflectance or shading. Some methods proposed deep architectures that learned to solve that exact problem: given two patches of an image, find their relative magnitude of lightness. Narihira *et al.* [96] fine-tune an Alexnet network [67] pre-trained on Imagenet [26], and use its last fully-connected layer as a feature descriptor vector of the input image. To compare the relative lightness of two patches of images, they perform ridge-ranking regression using their two feature vectors as input. Other methods argue that using only local patches does not provide the model with enough information, as the context of the whole image is lost. Zhou *et al.* [147] propose a three-stream deep convolutional architecture that performs the relative lightness prediction by combining –inside a shared vector– the features extracted of three images: the two patches and the global input image. This vector is enriched with the spatial coordinates of the two patches. A set of fully connected layers uses this shared vector to predict the relative lightness. Similarly, Zoran *et al.* [149] propose a multi-stream convolutional architecture that receives both patches, the global image, as well as masks for both patches, a bounding box and a *region of interest* image. The predictions of the convolutional networks within the architecture are then aggregated in a feature vector, used by a block of fully-connected layers to perform the relative lightness prediction.

6.2 Image-to-Image Translation

Instead of relying on pairwise comparisons between patches of images, many methods perform a direct prediction of either the shading, the reflectance, or both maps, in an end-to-end fashion. This type of framework reduces the amount of post-processing needed to complete the intrinsic decomposition, and are more easily integrated with other differentiable modules (e.g. a differentiable render, differentiable filtering layers, etc.), which may help increase their learning capabilities. We have divided those image-to-image translation networks into three groups, depending on how they learn the mapping between input image and intrinsic decomposition, from a network architecture perspective. As we will see in Section 6.2.1, earlier methods rely on a simple set of connected convolutional layers to solve this problem, without the use of any skip or mirror connections. Previous work [54, 111, 99] show that the use of skip connections between layers that represent features at different scales is helpful for preserving rich details on the output maps. Those skip connections can be included in encoder-decoder architectures (Section 6.2.2) or using residual connections (Section 6.2.3).

6.2.1 Baseline Methods

(**B I2IT**) Some methods propose a deep architecture similar to AlexNet [67] or VGG-16 [119], which are comprised of a set of convolutional layers, followed by a block of fully-connected layers. Such is the case of the architecture in Kovacs *et al.* [64], where they use a VGG network (trained on Image-Net) and add to it 3 fully-connected layers, which help complete their shading prediction. The use of fully-connected layers is limiting, as it forces the input image to have some specific dimensions. Consequently, multiple methods use only convolutional layers for their predictions. Nestmeyer *et al.* [97] propose a fully-convolutional architecture, which outputs a reflectance intensity prediction, then transformed to reflectance and shading maps using differentiable operations. The use of a fully-convolutional architecture is also proposed in Narihira *et al.* [95]. Their method performs directly the decomposition estimation by combining two networks that process the input image at different scales.

6.2.2 Encoder-Decoder Networks

(**E-D I2IT**) A common approach for deep intrinsic image decomposition is to use an encoder-decoder architecture [111]. Those architectures are composed of an *encoder*, which translates the input image to a dense, information-rich, latent space, which is then transformed into another image by a *decoder*. In the deep intrinsic images literature, all the methods that use encoder-decoder architectures transform the input image using a shared encoder, followed by different decoders for every map they want to predict. To preserve rich spatial details, they enhance their models with skip connections between mirrored layers in the encoder and decoder networks. Learning different decoders for the shading and reflectance layers is a common approach [78, 8, 86]. Different enhancements can be performed to this architecture, such as predicting a residual layer [117], an illumination color [79], an environment map [114], or sharing activation values between decoders [12]. This type of architecture can also be used to predict normal [56, 139, 77, 146] or specular maps [90, 77], which can then be processed to estimate shading.

6.2.3 Residual Networks

(**Res I2IT**) A different way of preserving rich multi-scale spatial information is by using residual skip connections [49]. In a residual block within a convolutional neural network, the input is processed by a set of convolutional and non-linear activation layers, the result of which is added to the original input. This helps the flow of gradients during back-propagation, and adds spatial details to image-to-image translation networks. Instead of only using these skip

connections between mirrored layers, as in the methods in Section 6.2.2, those skip connections are used in every layer of the network. Residual networks for intrinsic image decomposition has been proposed in different fashions. The model of Fan *et al.* [37] contains two residual networks: one which predicts albedo intensities, and other that estimates the probability of each pixel corresponding to an edge on the albedo. A fully-convolutional residual network is proposed in Lettry *et al.* [74], which predicts the shading of the input image. The residual network proposed in Cheng *et al.* [25] predicts a laplacian pyramid, which is then collapsed into the decomposition prediction. The architecture proposed in [84] performs the intrinsic decomposition by leveraging residual deep latent spaces for unsupervised image domain translations. It is worth mentioning that the Intrinsic Residual Network (IRN) proposed in Sengupta *et al.* [114] uses residual layers as part of their encoder-decoder architecture.

Table 3 Methods presented in the paper and datasets used for Training (T) and Evaluation (E) purposes. *These datasets contain only isolated objects. [1] Uses BigTime [79]; [2] Uses MegaDepth [80]; [3] Uses a custom multi-illumination dataset.

Method	Synthetic			Real		
	*ShapeNet-D	MPI Sintel	CG	*MIT Intrinsic	IIW/SAW	Other
Narihira *et al.* [96]					TE	
Narihira *et al.* [95]		TE		TE		
Zhou *et al.* [147]					TE	
Zoran *et al.* [149]					TE	
Kovacs *et al.* [64]					TE	
Nestmeyer *et al.* [97]					T	
Shi *et al.* [117]	T			TE		
Janner *et al.* [56]	T					
Meka *et al.* [90]	T					
Baslamisli *et al.* [8]	T			E		
Cheng *et al.* [25]		TE		TE		
Fan *et al.* [37]		TE		TE	TE	
Li *et al.* [79]				TE	E/E	T[1]
Yu *et al.* [139]					E	T[2]
Li *et al.* [78]			TE		TE/TE	
Ma *et al.* [86]	T			TE		
Bi *et al.* [12]		TE			E	T[3]
Lettry *et al.* [74]		TE		TE		
Sengupta *et al.* [114]			TE		TE	
Zhou *et al.* [146]			TE		TE/TE	
Liu *et al.* [84]		TE	TE	TE	E	
Li *et al.* [77]			TE		TE	

7 Evaluation

The quality of an intrinsic decomposition method is challenging to evaluate. As we argue in Sections 2 and 3, one of the reasons is that the common assumption about Lambertian materials does not hold in real life. Consequently, the methods wrongly assign the additional light effects either to the shading or to the albedo layers. The existence of explicitly labeled datasets has facilitated the task, which was done for a long time by means of visual side-by-side comparisons, and most of the methods nowadays quantify similarity to the ground truth using pixel-wise errors metrics. As opposed to using explicitly labeled datasets, several methods have evaluated the algorithmic performance by comparing it with humans doing the same task. These forms of evaluation have some trade-offs that we discuss in Section 8. In the following, we describe the error metrics along with the reported performance of each of the reviewed works. We will report the error metrics for the two most commonly used datasets to compare different approaches for intrinsic decomposition.

7.1 Pixel-wise Error Metrics

A common way of evaluating image regression models is through pixel-wise error metrics. In them, the regressed output and the target image are compared pixel-by-pixel, using a given distance metric, such as the ℓ^1 or ℓ^2 norms. Pixel-wise error metrics provide an estimation of the quality of the output of the model that fails to take into account spatial information, as individual pixels are considered to be independent from their neighboring pixels. Consequently, such error metrics are not able to provide an estimation of the perceptual quality of the decomposition, in ways that deep perceptual-aware metrics are capable of (see [42,53,141]).

Nevertheless, pixel-wise distances have been widely used in the intrinsic decomposition literature, particularly for evaluating the quality of the results obtained on the MIT dataset [45]. Pixel-wise approaches can be used for evaluating the quality of intrinsic decomposition methods for this dataset, because it contains densely-annotated data, as discussed in Section 4. In particular, the most common way of reporting errors for this dataset has been through the Scale Invariant L2 Loss, which is a modified version of the ℓ^2 norm that ignores the scale of the (log) shading and albedo maps. When averaging this loss over overlapping windows, the Local Mean Squared Error (LSME) error can be computed. A more detailed description of these error metrics is found in [95].

In Table 4, a comprehensive comparison of LMSE metrics on the MIT dataset can be found. In particular, we detail the self-reported average LMSE (averaged between shading and albedo LSME), for multiple train and test splits. Many of the methods do not provide information about the test dataset they are evaluating their method on, so we simply include the self-reported evaluation metrics, as well as their own comparisons with other methods in the intrinsic decomposition literature. As previously discussed, the MIT dataset

does not contain sufficient data for a deep learning model to learn from. Consequently, a common approach is to train the deep learning model using other datasets, and evaluate on a test split of the MIT dataset. Nevertheless, many methods further fine-tune their model using a portion of the MIT dataset, so as to improve the results on its test split. We include both the baseline and the fine-tuned models' results for the methods that perform this fine-tuning on the MIT dataset. As it can be seen, despite the fact that comparing methods on this dataset is difficult (because in many cases the train/test split is unreported), there is a trend towards smaller LMSE errors on this dataset. This may indicate that there has been progress on the quality of the deep learning models and learning frameworks used to train them.

7.2 Human Disagreement Metrics

A different way of evaluating the quality of the intrinsic decomposition performed by machine learning models is by comparing the results of their outputs to judgments of humans that were asked to assess the relative lightness of two patches of an image. Such approach is helpful for datasets in which no ground-truth values are available, but for which sparse humans annotations can be found.

A dataset with these characteristics is the Intrinsic Images in the Wild (IIW) dataset [10], which is annotated by humans using relative reflectance values, as described in Section 4. In the images in this dataset, there are sparse human annotations of the relative lightness (lighter, darker, same lightness) of pairs of patches of the image. This dataset has been widely used on the intrinsic decomposition literature, and it provides a specific train/test split, which helps make comparisons fair.

To evaluate the quality of the results on the images in this dataset, most methods have used the *Weighted Human Disagreement Rate* (WHDR), which compares the proportion of human judgments that the models disagrees with, weighted by the confidence of each pairwise judgments. This error metric may be more perceptually accurate than using simple pixel-wise metrics, but it fails to account for the intensity of the relative lightness values (e.g. how much lighter one patch of the image is compared to another), and it only measures the quality of the predictions on a subset of the images, as annotations are sparse.

A summary of the self-reported WHDR values for the methods that use the IIW dataset for evaluation can be found on Table 5. For each method, we provide their self-reported WHDR value and the values they report for other methods, as the latter do not necessarily agree with their own self-reported values. As it can be seen, there is a downward trend of the error metrics, which indicates that the most recent approaches, that include self-supervision, better priors, or more sophisticated neural architectures, may be more suitable for performing the intrinsic decomposition task. However, smaller values on this error metric does not necessarily mean that the decomposition performed by the model is physically accurate, as we will discuss in Section 8.

8 Discussion

As shown in Sections 5 and 7, recent methods are showing considerable advances in their ability to learn with non explicitly labeled datasets, purely computer graphics datasets, estimate extra scene elements (such as normals or depth), and take into account more complex light-surface interactions beyond Lambertian material models. Nevertheless, this progress is not clearly obvious by looking at quantitative metrics or qualitative results, often showing very different outcomes for similar error values. In this section, our aim is to discuss this problem as well as other factors which are making it challenging to objectively track advances in this field.

8.1 Outperforming Learning Strategies

As discussed in detail in Section 5, there are several ways to use datasets and priors to train a neural network to address the intrinsic decomposition problem. In the following, we will discuss how the different strategies have been combined by the most successful methods according to Section 7. In particular, to show generalization capabilities of the different training strategies we will center the discussion around the methods that show compelling results on the IIW dataset without using such data for training (see Table 5 †).

According to the WHDR error reported in Table 5, the method of Bi *et al.* [12] is the most compelling one. It combines a twofold strategy using full and self-supervision (single-image and multi-image), as well as synthetic data from the MPI Sintel and real data from a custom built multi-illuminant dataset. It further includes prior information for the albedo layer by means of a learned bilateral filter [6]. The method of Liu *et al.* [84] is the second best with a score under twenty. Their approach is fully unsupervised and their key idea is to train latent feature spaces for a domain-invariant image-to-image translation architecture. Their models are trained using mostly synthetic images of ShapeNet, MPI Sintel, and CG dataset, and real from MIT Intrinsics.

The methods of Yu *et al.* [139] and Li *et al.* [77] which perform similarly also follow a similar approach to encode and infer explicit illumination using spherical harmonics. The former by means of a single environment map, while the latter encodes it in a spatially-varying basis. The idea of encoding a spatially-varying material map is further followed

Table 4 Average LMSE (between shading and reflectance maps) reported for the MIT dataset [45], for multiple train/test splits. We only show the results for each paper's best-performing method. Many authors provide the results of their deep learning model with and without fine-tuning their weights on a training set of the MIT dataset. We include both results both results for fairness when comparing methods. $*$ [95] also splits the test set by objects, instead of by images, as done in [5]. † These methods are all evaluated using the same train/test split, which is unknown.

Method	[5] split		Other / unknown split		Error reported for other methods			
	Baseline	Finetune on [45]	Baseline	Finetune on [45]	[5]	[95]	[117]	[37]
Narihira *et al.* [96]		0.0218	0.0234*	0.0224*	0.0125	0.0218*		
Shi *et al.* [117]			0.0535†	0.0375†	0.0292	0.044	0.0375	
Baslamisli *et al.* [8] (IntrinsicNet)			0.0226†			0.044	0.0535	
Cheng *et al.* [25]	0.0133	0.0121			0.0125	0.0239	0.0271	0.02
Fan *et al.* [37]	0.0203				0.0125		0.0271	0.0203
Li *et al.* [79]	0.0297				0.0292	0.044	0.0372	
Ma *et al.* [86]			0.0105†	0.0063†	0.005		0.0098	
Lettry *et al.* [74]				0.00055		0.00091		

Table 5 Reported WHDR error for the IIW dataset [10]. †These methods were not trained on IIW human ratings. $*$ indicate the values obtained using the test set proposed in [149].

Method	WHDR	Error reported for other methods								
		[95]	[96]	[147]	[149]	[97]	[117]	[37]	[79]	[12]
Narihira *et al.* [96]	18.1		18.1							
Zhou *et al.* [147]	15.7		18.1	15.7						
Zoran *et al.* [149]*	17.86				17.86					
Nestmeyer *et al.* [97]	17.69			19.95	17.85	17.69				
Fan *et al.* [37]	14.45			19.95	17.85	17.69		14.45		
Li *et al.* [78]	14.8	37.3		19.9		17.7*/19.5	59.4			17.7
Sengupta *et al.* [114]	16.7			19.9		19.5				
Li *et al.* [77]	15.93/†21.99									
Li *et al.* [79]	†20.3	37.3	18.1	15.7*/19.9			59.4		20.3	
Bi *et al.* [12]	†17.18	40.9		19.9	17.85	17.69	54.44			17.18
Yu *et al.* [139]	†21.4	37.3		19.9		19.5	59.4	14.5		
Zhou *et al.* [146]	15.2			19.9		19.5		15.4		
Liu *et al.* [84]	†18.69		18.1	15.7				14.45	20.3	20.94

by Li's which leverages a large synthetic dataset for training. Yu's method is however trained to obtain a pure diffuse model using real scenes from MegaDepth dataset for which normals and albedo maps have been estimated using multi-view stereo. Finally, the method of Li *et al.* [79], while performing similarly to the above one, is the most different in terms of learning strategy. It relies on a multi-image consistency loss trained using timelapse sequences and smoothness priors for the shading and reflectance layers.

In light of these results, it is worth noting that the most successful strategy seems to be the combination of self-supervised learning along with fine-tuning on a specific dataset or domain. Using priors on the albedo layer by means of some form of bilateral filter appears to be also an interesting approach to reduce the amount of training data required. The main disadvantage of self-supervision for this problem is that the reconstruction loss needs to be computed efficiently during the progress of training, difficulting the use of global-illumination based reconstruction losses. Taking into account global-illumination effects requires the computation of multiple light bounces that is prohibitive in this context, which is the reason why the majority of the methods rely on direct illumination and lambertian materials models. As discussed in Section 9.3, a promising research direction in this area involves the use of differentiable rendering techniques.

8.2 Training and Testing Data

It is well known that one of the reasons for the superior performance of CNNs is their capability of capturing local and global –semantic, object level– features. CNNs are empowered by two main factors: translation invariance and hierarchical definition of image-level patterns. Although important for several computer vision applications, such as image classification or semantic segmentation, those properties might be problematic for the intrinsic decomposition problem.

Semantic dependence during training impacts the generalization of the method to arbitrary scenes. First, the diversity of objects and materials of the scene is given by its

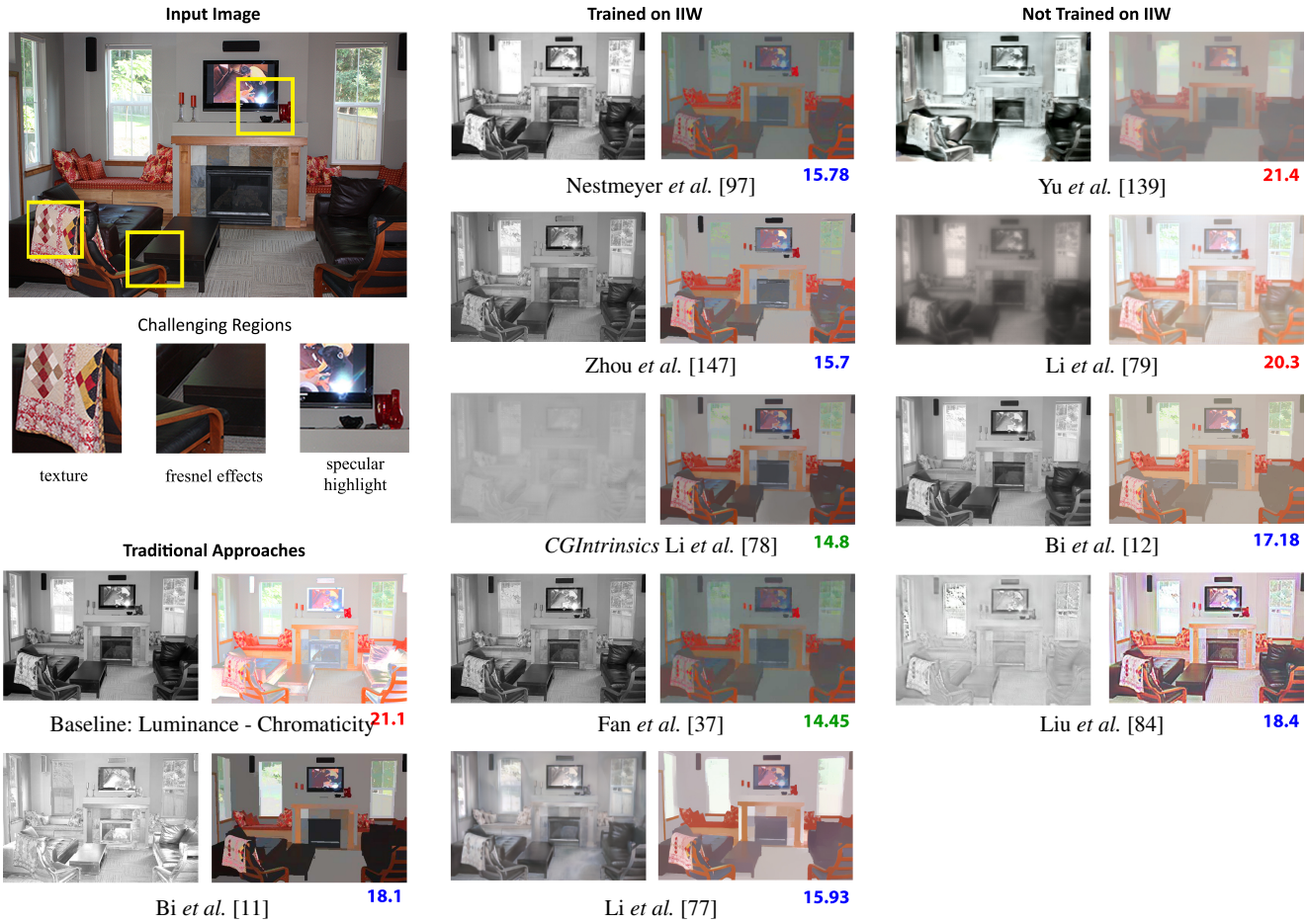


Fig. 9 Qualitative comparison with an input image from the IIW dataset. Here we compare results between traditional non-learning based solutions, methods trained or fine-tuned on the IIW dataset, and methods not trained on such dataset. The quantitative error is shown below (errors below 15 on green, above 20 on red). Note that a similar score does not necessarily mean the same qualitative result, as shown by the two methods with green score. Three interesting effects are highlighted in the input image. We recommend the reader to zoom-in and analyze the results of the method at these specific areas. Image ID: *quiltsalad_3711222369*

semantics, so a network trained with a dataset only containing indoor scenes will not likely generalize to outdoor scenes. In other words, the prior distribution of scene components learned by the network will not cover the diversity of scenes that are present in the real world. A similar problem is faced with illumination, as the same object and material might have a vastly different appearance under different lighting conditions, so if the training data is not targeted at disentangling this relationship, the network will perform inadequately. Such a limitation is common in most deep intrinsic decomposition models. Currently, many of the existing datasets, either synthetic (SUNCG, CGIntrinsics) or real (IIW, SAW), used for training and testing contain a majority of indoor scenes. The methods of Sengupta *et al.* [114] and Li *et al.* [77] account for this bias and limit the scope of the contributions to indoor scenes, as well as capturing an inverse model of indoor lighting through environment maps.

Testing on data outside the training set (without fine-tuning) is done by only a few methods: on the MIT dataset [8], and IIW dataset [12,79,139,84]. As shown in Table 3, the majority of methods are both trained and tested using the same dataset. In Figure 9, we can appreciate better quantitative performance for methods both trained and tested on IIW. Nevertheless, the qualitative improvement is less clearly observable, suggesting that the quantitative evaluation done to those algorithms may not account for perceptual factors. We discuss this issue next.

8.3 Quantitative Errors as Indicator of Performance

Quantitative error metrics are the common way of measuring the performance of machine learning algorithms. However, as explained in Section 7, the intrinsic decomposition problem lacks properly established benchmarks that facilitate these comparisons. This is evidenced in Tables 4 and 5 where we can observe a diverse range of reported errors for the same method. There are two key factors that might be the

cause of this discrepancy: first, choosing different splits for the train/test sets, and second, the selection of the reported error metric when the methods report several errors for different training conditions (e.g. with and without fine-tuning with a specific dataset).

In the following discussion, we will focus on the IIW dataset and reported WHDR errors, shown in Table 5, as it contains the most consistent train/test splits and is the most common. As it can be seen, there is a consistent trend towards small errors in most recent approaches that include complex inverse models or network architectures. However, a smaller error value does not necessarily correlate with a more physically accurate decomposition. Furthermore, similar error values do not necessarily imply similar decompositions, as we discuss next.

In Figure 9, we show the intrinsic components of one of the scenes of the IIW dataset. We compare the performance among learning-based methods: 1) trained or fine-tuned with IIW human annotations, 2) methods only tested on that dataset, and 3) two baseline decompositions that do not require training data, first, using the luminance channel as shading component and, second, the decomposition provided by Bi *et al.*[11] which relied on a clustering-based strategy. In terms of quantitative metrics, it is shown that methods trained on such dataset perform slightly better. However, the differences between methods with similar error values are noticeable, for example, for methods trained on IIW, between Li's [78] and Fan's[37], and between Yu's [139] and Li's[79]. At the same time, it seems quite difficult to qualitatively judge which decomposition is more accurate. It could be argued that Li's[78], Bi's [11] and Liu's [84] produce the best results as the shading images have less albedo remnants. Bonneel *et al.* [14] already exposed this problem and presented a thorough evaluation of intrinsic decomposition methods in the context of image editing tasks. In their study, none of the existing methods was robust enough to be used for such kind of operations.

8.4 The Influence of the Material Model

For a long time, intrinsic decomposition methods assumed Lambertian material models, a design decision which impacted also the variability of the datasets used to train and evaluate the methods. As we have shown in Section 2, this significantly simplifies the image formation model, making the inverse problem tractable from a practical standpoint. However, the use of these models and datasets have had two main problems: First, as *the model is not rich enough*, it will not be able to inversely reproduce the scene correctly. For example, looking at the world around us, it is easy to discover a majority of non-Lambertian surfaces: glasses, plastics, or cloth, among others. Second, if *the data used for training does not contain enough variability and realism*, it will be hard to predict how the methods will behave on real scenes containing a broad range of material types.

Only a few of the latest methods on intrinsic decomposition have incorporated more complex material models. Meka *et al.* [90] estimate the parameters of a Blinn-Phong material model, however, as it just deals with single isolated objects, a fair comparison is unfeasible. The two methods that deal with general scenes and complex materials are Sengupta *et al.* [114] that incorporate the specular lobe of Phong shader, and the method of Li *et al.* [77] that goes a step further and estimates the parameters of a spatially-varying microfacet BRDF. It is worth mentioning that the latter performs reasonably well on the IIW dataset after fine-tuning, and even provides the error estimation for a model which wasn't fine-tuned on such a dataset, suggesting a positive trend in attempting to generalize to new scenes. In Figure 9, we observe that while the reflectance layer looks coherent, the shading layer tends to have artifacts due to the increased complexity of having to estimate an underlying geometry of the scene in the form of normals and lighting. These extra layers (normals and lighting) are nevertheless beneficial to improve the editability of the scene, for example, to change illumination or material properties.

When a specular image is tested with a method that assumes a Lambertian model, the specular highlight will be wrongly placed either completely or partially in either component, because the intrinsic diffuse model does not consider such a property of the materials. We can see an example of such effect in the television highlight of Figure 9. Note that all the methods shown in that figure, except Li *et al.* [77], follow the intrinsic diffuse model. At the same time, the Lambertian assumption and the *classical* intrinsic decomposition framework has yielded a set of explicitly labeled datasets which describe a limited set of Lambertian materials [45, 78].

9 Research Opportunities and Future Directions

Deep learning has changed the way the intrinsic decomposition problem is addressed, transitioning from sometimes *manual* or ad-hoc heuristics to implicit rules defined by the data. Among the data-driven methods reviewed, we have also observed a trend, from simple regression-based approaches to recent methods that take into account the physics of light, for example, modeling complex scene elements: geometry through normals, lighting using spherical harmonics, or non-lambertian material models using microfacets. Nevertheless, the world is highly complex and the problem of fully reconstructing arbitrary scenes from single images is still in an open research area. In the following, we discuss, among other topics, how the field can benefit from novel machine learning techniques learn more efficiently from data, initiatives that promote replicability and

facilitate comparative evaluation frameworks, as well as the potential of differentiable and neural rendering.

9.1 Enhancing Generalization

Addressing the problem of intrinsic image decomposition with a purely data-driven approach is particularly risky: the content of the dataset guides the variability and amount of object-material and light-material phenomena that a model learns to disambiguate. Some approaches have focused on indoor scenes [10,78,114] dominated by furniture, decorative objects, and certain types of lighting; and others have limited the scope to a subset of objects within the ShapeNet dataset [21]. Recently, domain specific datasets have been created for outdoor road scenes [65], or planar materials [32]. In contrast to the brute force approach of training a massive network able to generalize blindly for each domain (akin to language models [16]), a more environment-friendly and scalable approach would include combining the variety of domain-specific network in a single framework without needing to re-train the whole system, *e.g.*, pre-classifying the scenes to decide the best networks from an ensemble, defining accessible APIs, or using networks designed for cross-domain generalization, as proposed by Rebuffi *et al.* [107].

Most intrinsic decomposition datasets are smaller than many popular datasets used in the computer vision literature, such as ImageNet [26] or COCO [83]. The patterns found by early layers in CNNs trained on ImageNet have been found to generalize to multiple computer vision problems, which could benefit deep learning models applied to the intrinsic decomposition problems. Pre-training a CNN on ImageNet using recent unsupervised [43] or self-supervised learning [24] approaches, then fine-tune such network for the intrinsic decomposition problem, could provide a solution for the lack of sufficient training data in the datasets we discussed.

Besides, generative adversarial networks have proven successful at generating photo-realistic images in many domains [63]. Recent work on neural rendering and inverse graphics [142] suggests that those generative models learn representations in which geometry, light and texture can be disentangled. Such approaches could be used to generate new synthetic data samples that allow for larger intrinsic decomposition datasets. Furthermore, some data points are more informative than others. Active learning methods [17, 134] could help efficiently generate new (either synthetic or real) samples that help intrinsic decomposition methods improve their generalization capabilities.

Additionally, most neural network architectures used for intrinsic decomposition are CNNs. Recent work on attention mechanisms indicate that traditional convolutional layers, while powerful, may be limiting the potential of deep learning for computer vision problems [34,135,58]. Moving beyond traditional deep CNN architectures may provide intrisic decomposition algorithms with more sophisticated inductive biases.

9.2 Evaluation Frameworks

As we have seen in the previous section, existing evaluation metrics [10,45] are not necessarily representative of the complexity of the real world, neither capture consistently the actual accuracy and consistency of the decomposition results (see Figure 9). A particularly interesting opportunity in this field would include creating a common evaluation framework (a benchmark, or a challenge), following existing initiatives in the computer vision and graphics communities [36,109,1,92].

Creating an evaluation dataset and metrics for the intrinsic decomposition problem is challenging, as the intrinsic components do not freely exist in the nature. We propose several ideas to facilitate this process. First, leverage hyper-realistic computer generated scenes, like existing works [78, 14,114] but shaping the data in the form of a benchmark so that it is easily accessible, and comparable. The main limitations in this case are the render engine not being able to reproduce physically complex light phenomena, the cost of manually creating a variety of scenes of several semantics, as well the required render time. Second, exploit scene characteristics such as the albedo invariance to illumination changes. This can be done by leveraging time-lapse or multi-view scenes ([80,79]) which are more easily accessible, so that the final evaluation should take into account the quality of the reconstruction as well as the invariance of the albedo layer (mimicking the multi-view learning strategy presented in Section 5.3). Finally, an ideal dataset for a benchmark would also include: a variety of scene semantics (indoor, outdoors, single objects, humans, etc.), the same scene with multiple illuminations, and complex light phenomena useful to understand the path of light. Those characteristics might also serve as intermediate low-level step towards a higher-level task, *e.g.*, specular layers, scattering effects, a layer for caustics, or a layer for shadows or occlusion boundaries. Having these layers would be useful for many other tasks, such as material editing, object segmentation, or scene compositing.

9.3 Beyond Lambert

We have observed in recent works an increasing use of complex illumination models [139,114,146,77], learning from synthetic datasets that contain both global illumination and inter-reflections, and some relying on the *intrinsic residual* model to enable spatially varying light effects. In these

methods, light transport simulation has been restricted to real time techniques like directional lights and spherical harmonics environments, with limitations in realism at high frequency shadows, occlusions and inter-reflections.

The advent of differentiable ray tracing methods [75, 3, 100], have opened a new opportunity to include accurate light transport path simulations, not only as a costly part of the error function in the learning process, but also as a neural network component. Also, the success of the neural networks to learn parts of the render equation [56, 114, 77, 146] have a great potential to encapsulate complex lighting interactions, as shown by neural rendering techniques [126].

The definition of reflectance has also reached increasing levels of sophistication, the material being no longer just a base albedo color, but also including specular or Phong coefficients in more recent methods[90, 77]. However, there is great potential for improvement. The success in estimating physically-based material models for flat surfaces [27] or objects [82] suggest an interesting avenue for future work in such direction in order to generalize such findings to arbitrary scenes. Without a full BSDF model [132] and its multiple parameters linked to the reflectance layer, some intrinsic decomposition applications such as relighting will never produce realistic results. For instance, the human relighting technique by Kanamori et al. [61] relies on advanced illumination (spherical harmonics and local occlussion) but fails to produce realistic relighting because their material model lacks transmittance and subsurface parameters which are paramount for the final appearance of human skin and cloth.

We have barely mentioned several optical phenomena in this survey: participating media, caustics, subsurface scattering (highly relevant in many materials such as skin, cloth or liquids) [57], energy transfer between wavelengths (reradiance, fluorescence), polarization, interference (iridescence), etc. Although we have seen impressive results in the recent years for problems such as inferring the scattering parameters an density of volumetric media [100], many of these effects will require further advances of the inverse rendering field to be applied to intrinsic imaging.

The challenges ahead include finding an adequate balance between inverse and neural rendering techniques, and the right distribution of parameters into multiple intrinsic layers for each desired application. For instance, editing applications will required more physically-principled decompositions through inverse rendering, so more parameters are modifiable. On the other side, constrained parts of the problem, like restricted geometries (faces, bodies), partial light transport paths, or material shaders, will surely benefit from neural rendering strategies.

10 Conclusions

Deep learning has changed the way the intrinsic decomposition problem is addressed, transitioning from sometimes *manual* or ad-hoc heuristics to implicit rules defined by the data. In this survey, we have reviewed this transition discussing learning frameworks, architectures, and datasets used, putting them in the context of traditional non-learning-based solutions. Through this revision, we have identified several problems that might prevent the field to develop further: the semantic dependence on the training data, the uncorrelation between current evaluation metrics and qualitative results, and the limitations of the widely used Lambertian material model to capture complex materials and light phenomena.

In light of the recent advances in neural rendering and differentiable rendering, we also believe that the intrinsic decomposition would greatly benefit from a physically-based –inverse rendering– perspective. For this reason, in this survey, we also provide a thorough explanation of the physics of light from a rendering and inverse rendering perspective, as well as make explicit connections with other inverse methods that deal with specific domains such as faces, humans, flat materials, or objects.

Acknowledgments Elena Garces was partially supported by a Torres Quevedo Fellowship (PTQ2018-009868). The work was also funded in part by the Spanish Ministry of Science (RTI2018-098694-B-I00 VizLearning).

References

1. Robust Vision Challenge 2020. Accessed: 2020-10-31
2. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11), 2274–2282 (2012)
3. Azinovic, D., Li, T.M., Kaplanyan, A., Niessner, M.: Inverse path tracing for joint material and lighting estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
4. Balzer, J., Höfer, S., Beyerer, J.: Multiview specular stereo reconstruction of large mirror surfaces. In: CVPR 2011, pp. 2537–2544 (2011). DOI 10.1109/CVPR.2011.5995346
5. Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(8), 1670–1687 (2014)
6. Barron, J.T., Poole, B.: The fast bilateral solver. In: European Conference on Computer Vision, pp. 617–632. Springer (2016)
7. Barrow, H., Tenenbaum, J., Hanson, A., Riseman, E.: Recovering intrinsic scene characteristics. Computer Vision Systems **2**(3-26), 2 (1978)
8. Baslamisli, A.S., Le, H.A., Gevers, T.: Cnn based learning using reflection and retinex models for intrinsic image decomposition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6674–6683 (2018)

9. Beigpour, S., Van De Weijer, J.: Object recoloring based on intrinsic image estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 327–334. IEEE (2011)
10. Bell, S., Bala, K., Snavely, N.: Intrinsic images in the wild. ACM Transactions on Graphics (TOG) **33**(4), 1–12 (2014)
11. Bi, S., Han, X., Yu, Y.: An l 1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. ACM Transactions on Graphics (TOG) **34**(4), 1–12 (2015)
12. Bi, S., Kalantari, N.K., Ramamoorthi, R.: Deep hybrid real and synthetic training for intrinsic decomposition. Computer Graphics Forum (Proc. Eurographics Symposium on Rendering) (2018)
13. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, pp. 187–194 (1999)
14. Bonneel, N., Kovacs, B., Paris, S., Bala, K.: Intrinsic decompositions for image editing. In: Computer Graphics Forum (Proc. Eurographics STAR), vol. 36, pp. 593–609 (2017)
15. Bousseau, A., Paris, S., Durand, F.: User-assisted intrinsic images. In: Proceedings of the 2009 SIGGRAPH Asia Conference, pp. 1–10 (2009)
16. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)
17. Brust, C.A., Käding, C., Denzler, J.: Active learning for deep object detection. arXiv preprint arXiv:1809.09875 (2018)
18. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 611–625. Springer (2012)
19. Carroll, R., Ramamoorthi, R., Agrawala, M.: Illumination decomposition for material recoloring with consistent interreflections. In: ACM SIGGRAPH 2011 papers, pp. 1–10 (2011)
20. Chaitanya, C.R.A., Kaplanyan, A.S., Schied, C., Salvi, M., Lefohn, A., Nowrouzezahrai, D., Aila, T.: Interactive reconstruction of monte carlo image sequences using a recurrent denoising autoencoder. ACM Transactions on Graphics (TOG) **36**(4), 1–12 (2017)
21. 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)
22. Chang, J., Cabezas, R., Fisher, J.W.: Bayesian nonparametric intrinsic image decomposition. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 704–719. Springer (2014)
23. Chen, Q., Koltun, V.: A simple model for intrinsic image decomposition with depth cues. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 241–248 (2013)
24. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020)
25. Cheng, L., Zhang, C., Liao, Z.: Intrinsic image transformation via scale space decomposition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 656–665 (2018)
26. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. Ieee (2009)
27. Deschaintre, V., Aittala, M., Durand, F., Drettakis, G., Bousseau, A.: Single-image svbrdf capture with a rendering-aware deep network. ACM Transactions on Graphics (ToG) **37**(4), 1–15 (2018)
28. Deschaintre, V., Aittala, M., Durand, F., Drettakis, G., Bousseau, A.: Flexible svbrdf capture with a multi-image deep network. In: Computer Graphics Forum, vol. 38, pp. 1–13. Wiley Online Library (2019)
29. Deschaintre, V., Drettakis, G., Bousseau, A.: Guided fine-tuning for large-scale material transfer. Computer Graphics Forum (Proceedings of the Eurographics Symposium on Rendering) **39**(4) (2020). URL `http://www-sop.inria.fr/reves/Basilic/2020/DDB20`
30. Dong, B., Dong, Y., Tong, X., Peers, P.: Measurement-based editing of diffuse albedo with consistent interreflections. ACM Trans. Graph. **34**(4) (2015)
31. Dong, B., Moore, K.D., Zhang, W., Peers, P.: Scattering parameters and surface normals from homogeneous translucent materials using photometric stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2291–2298 (2014)
32. Dong, Y.: Deep appearance modeling: A survey. Visual Informatics **3**(2), 59–68 (2019)
33. Dong, Y., Tong, X., Pellacini, F., Guo, B.: Appgen: interactive material modeling from a single image. In: Proceedings of the 2011 SIGGRAPH Asia Conference, pp. 1–10 (2011)
34. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
35. Duchêne, S., Riant, C., Chaurasia, G., Moreno, J.L., Laffont, P.Y., Popov, S., Bousseau, A., Drettakis, G.: Multiview intrinsic images of outdoors scenes with an application to relighting. ACM Transactions on Graphics (TOG) **34**(5) (2015)
36. Erofeev, M., Gitman, Y., Vatolin, D., Fedorov, A., Wang, J.: Perceptually motivated benchmark for video matting. In: Proceedings of the British Machine Vision Conference (BMVC), pp. 99.1–99.12. BMVA Press (2015)
37. Fan, Q., Yang, J., Hua, G., Chen, B., Wipf, D.: Revisiting deep intrinsic image decompositions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8944–8952 (2018)
38. Finlayson, G.D., Drew, M.S., Lu, C.: Intrinsic images by entropy minimization. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 582–595. Springer (2004)
39. 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 (TOG) **38**(4), 1–15 (2019)
40. Garces, E., Munoz, A., Lopez-Moreno, J., Gutierrez, D.: Intrinsic images by clustering. In: Computer Graphics Forum, vol. 31, pp. 1415–1424 (2012)
41. Gastal, E.S., Oliveira, M.M.: Adaptive manifolds for real-time high-dimensional filtering. ACM Transactions on Graphics (TOG) **31**(4), 1–13 (2012)
42. Gatys, L.A., Ecker, A.S., Bethge, M., Hertzmann, A., Shechtman, E.: Controlling perceptual factors in neural style transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3985–3993 (2017)
43. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 (2018)
44. Godard, C., Hedman, P., Li, W., Gabriel, J.: Multi-view reconstruction of highly specular surfaces in uncontrolled environments. In: 3DV (2015). DOI 10.1109/3DV.2015.10

45. Grosse, R., Johnson, M.K., Adelson, E.H., Freeman, W.T.: Ground truth dataset and baseline evaluations for intrinsic image algorithms. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2335–2342. IEEE (2009)
46. Guarnera, D., Guarnera, G., Ghosh, A., Denk, C., Glencross, M.: Brdf representation and acquisition. Computer Graphics Forum **35**(2), 625–650 (2016)
47. Guo, Y., Smith, C., Hašan, M., Sunkavalli, K., Zhao, S.: Materialgan: Reflectance capture using a generative svbrdf model. ACM Trans. Graph. **39**(6), 254:1–254:13 (2020)
48. Han, X., Laga, H., Bennamoun, M.: Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
49. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
50. Hill, S., McAuley, S., Burley, B., Chan, D., Fascione, L., Iwanicki, M., Hoffman, N., Jakob, W., Neubelt, D., Pesce, A., Pettineo, M.: Physically based shading in theory and practice. In: ACM SIGGRAPH 2015 Courses, SIGGRAPH '15. Association for Computing Machinery, New York, NY, USA (2015)
51. Horn, B.K.: Determining lightness from an image. Computer graphics and image processing **3**(4), 277–299 (1974)
52. Horn, B.K., Sjoberg, R.W.: Calculating the reflectance map. Applied optics **18**(11), 1770–1779 (1979)
53. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1501–1510 (2017)
54. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
55. Jakob, W.: Mitsuba renderer (2010)
56. Janner, M., Wu, J., Kulkarni, T.D., Yildirim, I., Tenenbaum, J.: Self-supervised intrinsic image decomposition. In: Advances in Neural Information Processing Systems, pp. 5936–5946 (2017)
57. Jensen, H.W., Marschner, S.R., Levoy, M., Hanrahan, P.: A practical model for subsurface light transport. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques, pp. 511–518 (2001)
58. Jetley, S., Lord, N.A., Lee, N., Torr, P.H.: Learn to pay attention. arXiv preprint arXiv:1804.02391 (2018)
59. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 694–711. Springer (2016)
60. Kajiya, J.T.: The rendering equation. In: Proceedings of the 13th annual conference on Computer graphics and interactive techniques, pp. 143–150 (1986)
61. Kanamori, Y., Endo, Y.: Relighting humans: occlusion-aware inverse rendering for full-body human images. ACM Transactions on Graphics (TOG) **37**(6), 1–11 (2018)
62. Karis, B., Games, E.: Real shading in unreal engine 4. Proc. Physically Based Shading Theory Practice **4**, 3 (2013)
63. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8110–8119 (2020)
64. Kovacs, B., Bell, S., Snavely, N., Bala, K.: Shading annotations in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6998–7007 (2017)
65. Krähenbühl, P.: Free supervision from video games. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2955–2964 (2018)
66. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in Neural Information Processing Systems, pp. 109–117 (2011)
67. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
68. Laffont, P.Y., Bazin, J.C.: Intrinsic decomposition of image sequences from local temporal variations. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 433–441 (2015)
69. Laffont, P.Y., Bousseau, A., Paris, S., Durand, F., Drettakis, G.: Coherent intrinsic images from photo collections. ACM Transactions on Graphics (TOG) **31**(6), 1–11 (2012)
70. Lafortune, E.P., Willems, Y.D.: Using the modified phong reflectance model for physically based rendering (1994)
71. Land, E.H., McCann, J.J.: Lightness and retinex theory. Josa **61**(1), 1–11 (1971)
72. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 541–551 (1989)
73. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
74. Lettry, L., Vanhoey, K., Van Gool, L.: Darn: a deep adversarial residual network for intrinsic image decomposition. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1359–1367. IEEE (2018)
75. Li, T.M., Aittala, M., Durand, F., Lehtinen, J.: Differentiable monte carlo ray tracing through edge sampling. ACM Trans. Graph. (Proc. SIGGRAPH Asia) **37**(6), 222:1–222:11 (2018)
76. Li, X., Dong, Y., Peers, P., Tong, X.: Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. ACM Transactions on Graphics (ToG) **36**(4), 1–11 (2017)
77. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2475–2484 (2020)
78. Li, Z., Snavely, N.: Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 371–387 (2018)
79. Li, Z., Snavely, N.: Learning intrinsic image decomposition from watching the world. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9039–9048 (2018)
80. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2041–2050 (2018)
81. Li, Z., Sunkavalli, K., Chandraker, M.: Materials for masses: Svbrdf acquisition with a single mobile phone image. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 72–87 (2018)
82. 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 (TOG) **37**(6), 1–11 (2018)
83. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 740–755. Springer (2014)
84. Liu, Y., Li, Y., You, S., Lu, F.: Unsupervised learning for intrinsic image decomposition from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)

85. Loubet, G., Holzschuch, N., Jakob, W.: Reparameterizing discontinuous integrands for differentiable rendering. ACM Transactions on Graphics (TOG) **38**(6), 1–14 (2019)
86. Ma, W.C., Chu, H., Zhou, B., Urtasun, R., Torralba, A.: Single image intrinsic decomposition without a single intrinsic image. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 201–217 (2018)
87. Martin-Brualla, R., Radwan, N., Sajjadi, M.S., Barron, J.T., Dosovitskiy, A., Duckworth, D.: Nerf in the wild: Neural radiance fields for unconstrained photo collections. arXiv preprint arXiv:2008.02268 (2020)
88. Maxwell, B.A., Friedhoff, R.M., Smith, C.A.: A bi-illuminant dichromatic reflection model for understanding images. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2008)
89. 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 (TOG) **38**(4) (2019)
90. Meka, A., Maximov, M., Zollhoefer, M., Chatterjee, A., Seidel, H.P., Richardt, C., Theobalt, C.: Lime: Live intrinsic material estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6315–6324 (2018)
91. Meka, A., Zollhöfer, M., Richardt, C., Theobalt, C.: Live intrinsic video. ACM Trans. Graph. **35**(4) (2016)
92. Merzbach, S., Klein, R.: Bonn appearance benchmark. The Eurographics Association (2020)
93. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)
94. Müller, C.: Spherical harmonics, vol. 17. Springer (2006)
95. Narihira, T., Maire, M., Yu, S.X.: Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In: Proceedings of the IEEE international conference on computer vision, pp. 2992–2992 (2015)
96. Narihira, T., Maire, M., Yu, S.X.: Learning lightness from human judgement on relative reflectance. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2965–2973 (2015)
97. Nestmeyer, T., Gehler, P.V.: Reflectance adaptive filtering improves intrinsic image estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6789–6798 (2017)
98. Nestmeyer, T., Lalonde, J.F., Matthews, I., Lehrmann, A.: Learning physics-guided face relighting under directional light. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)
99. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 483–499. Springer (2016)
100. Nimier-David, M., Vicini, D., Zeltner, T., Jakob, W.: Mitsuba 2: A retargetable forward and inverse renderer. ACM Transactions on Graphics (TOG) **38**(6), 1–17 (2019)
101. Oh, B.M., Chen, M., Dorsey, J., Durand, F.: Image-based modeling and photo editing. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques, pp. 433–442 (2001)
102. Omer, I., Werman, M.: Color lines: Image specific color representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. II–II. IEEE (2004)
103. Patow, G., Pueyo, X.: A survey of inverse rendering problems. In: Computer Graphics Forum, vol. 22, pp. 663–687 (2003)
104. Pharr, M., Jakob, W., Humphreys, G.: Physically Based Rendering: From Theory to Implementation, 3rd edn. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2016)
105. Poole, B., Barron, J.T.: The fast bilateral solver. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 617–632 (2016)
106. Ramamoorthi, R., Hanrahan, P.: An efficient representation for irradiance environment maps. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '01, p. 497–500. Association for Computing Machinery, New York, NY, USA (2001)
107. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Learning multiple visual domains with residual adapters. In: Advances in Neural Information Processing Systems, pp. 506–516 (2017)
108. Rematas, K., Ritschel, T., Fritz, M., Gavves, E., Tuytelaars, T.: Deep reflectance maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4508–4516 (2016)
109. Rhemann, C., Rother, C., Wang, J., Gelautz, M., Kohli, P., Rott, P.: A perceptually motivated online benchmark for image matting. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1826–1833. IEEE (2009)
110. Ritschel T Dachsbacher C, G.T.K.J.: The state of the art in interactive global illumination. Computer Graphics Forum **31**(1) (2012)
111. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, pp. 234–241. Springer (2015)
112. Rother, C., Kiefel, M., Zhang, L., Schölkopf, B., Gehler, P.V.: Recovering intrinsic images with a global sparsity prior on reflectance. In: Advances in Neural Information Processing Systems, pp. 765–773 (2011)
113. Seitz, S., Matsushita, Y., Kutulakos, K.: A theory of inverse light transport. In: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, vol. 2, pp. 1440–1447 Vol. 2 (2005)
114. Sengupta, S., Gu, J., Kim, K., Liu, G., Jacobs, D.W., Kautz, J.: Neural inverse rendering of an indoor scene from a single image. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 8598–8607 (2019)
115. Sengupta, S., Kanazawa, A., Castillo, C.D., Jacobs, D.W.: Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6296–6305 (2018)
116. Shen, L., Yeo, C.: Intrinsic images decomposition using a local and global sparse representation of reflectance. In: CVPR 2011, pp. 697–704. IEEE (2011)
117. Shi, J., Dong, Y., Su, H., Yu, S.X.: Learning non-lambertian object intrinsics across shapenet categories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1685–1694 (2017)
118. Shu, Z., Yumer, E., Hadap, S., Sunkavalli, K., Shechtman, E., Samaras, D.: Neural face editing with intrinsic image disentangling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5541–5550 (2017)
119. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
120. Sloan, P.P., Kautz, J., Snyder, J.: Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. ACM Transactions on Graphics (TOG) **21**(3), 527–536 (2002)
121. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1746–1754 (2017)

122. 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 (TOG) **38**(4) (2019)
123. Sunkavalli, K., Matusik, W., Pfister, H., Rusinkiewicz, S.: Factored time-lapse video. In: ACM SIGGRAPH 2007 papers, pp. 101–es (2007)
124. Tappen, M.F., Freeman, W.T., Adelson, E.H.: Recovering intrinsic images from a single image. In: Advances in Neural Information Processing Systems, pp. 1367–1374 (2003)
125. Tewari, A., Bernard, F., Garrido, P., Bharaj, G., Elgharib, M., Seidel, H.P., Pérez, P., Zöllhofer, M., Theobalt, C.: FML: Face Model Learning from Videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10812–10822 (2019)
126. Tewari, A., Fried, O., Thies, J., Sitzmann, V., Lombardi, S., Sunkavalli, K., Martin-Brualla, R., Simon, T., Saragih, J., Nießner, M., Pandey, R., Fanello, S., Wetzstein, G., Zhu, J.Y., Theobalt, C., Agrawala, M., Shechtman, E., Goldman, D.B., Zollhöfer, M.: State of the Art on Neural Rendering. Computer Graphics Forum **39**(2), 701–727 (2020)
127. Tewari, A., Zollhöfer, M., Garrido, P., Bernard, F., Kim, H., Pérez, P., Theobalt, C.: Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2549–2559 (2018)
128. Tewari, A., Zollöfer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., Christian, T.: MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
129. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 839–846. IEEE (1998)
130. Tominaga, S.: Dichromatic reflection models for a variety of materials. Color Research & Application **19**(4), 277–285 (1994)
131. Torrance, K.E., Sparrow, E.M.: Theory for off-specular reflection from roughened surfaces. Josa **57**(9), 1105–1114 (1967)
132. Vidaurre, R., Casas, D., Garces, E., Lopez-Moreno, J.: Brdf estimation of complex materials with nested learning. In: IEEE Winter Conference on Applications of Computer Vision (WACV) (2019)
133. Wang, J., Ren, P., Gong, M., Snyder, J., Guo, B.: All-frequency rendering of dynamic, spatially-varying reflectance **28**(5), 1–10 (2009)
134. Wang, K., Zhang, D., Li, Y., Zhang, R., Lin, L.: Cost-effective active learning for deep image classification. IEEE Transactions on Circuits and Systems for Video Technology **27**(12), 2591–2600 (2016)
135. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
136. Weiss, Y.: Deriving intrinsic images from image sequences. In: Proceedings of the IEEE International Conference on Computer Vision, vol. 2, pp. 68–75. IEEE (2001)
137. Yamaguchi, S., Saito, S., Nagano, K., Zhao, Y., Chen, W., Olszewski, K., Morishima, S., Li, H.: High-fidelity facial reflectance and geometry inference from an unconstrained image. ACM Transactions on Graphics (TOG) **37**(4), 1–14 (2018)
138. Ye, G., Garces, E., Liu, Y., Dai, Q., Gutierrez, D.: Intrinsic video and applications. ACM Trans. Graph. **33**(4) (2014)
139. Yu, Y., Smith, W.A.: Inverserendernet: Learning single image inverse rendering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3155–3164 (2019)
140. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Zhang, Z., Lin, H., Sun, Y., He, T., Mueller, J., Manmatha, R., et al.: Resnest: Split-attention networks. arXiv preprint arXiv:2004.08955 (2020)
141. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
142. Zhang, Y., Chen, W., Ling, H., Gao, J., Zhang, Y., Torralba, A., Fidler, S.: Image gans meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. arXiv preprint arXiv:2010.09125 (2020)
143. Zhao, Q., Tan, P., Dai, Q., Shen, L., Wu, E., Lin, S.: A closed-form solution to retinex with nonlocal texture constraints. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(7), 1437–1444 (2012)
144. Zhao, S., Jakob, W., Li, T.M.: Physics-based differentiable rendering: from theory to implementation. In: ACM SIGGRAPH 2020 Courses, pp. 1–30 (2020)
145. Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D.W.: Deep Single-Image Portrait Relighting. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 7194–7202 (2019)
146. Zhou, H., Yu, X., Jacobs, D.W.: Glosh: Global-local spherical harmonics for intrinsic image decomposition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 7820–7829 (2019)
147. Zhou, T., Krahenbuhl, P., Efros, A.A.: Learning data-driven reflectance priors for intrinsic image decomposition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3469–3477 (2015)
148. Zollhöfer, M., Thies, J., Garrido, P., Bradley, D., Beeler, T., Pérez, P., Stamminger, M., Nießner, M., Theobalt, C.: State of the art on monocular 3d face reconstruction, tracking, and applications. In: Computer Graphics Forum, vol. 37, pp. 523–550 (2018)
149. Zoran, D., Isola, P., Krishnan, D., Freeman, W.T.: Learning ordinal relationships for mid-level vision. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 388–396 (2015)