Single Image Intrinsic Decomposition Without a Single Intrinsic Image

Wei-Chiu Ma[1,2](✉), Hang Chu[3], Bolei Zhou[1], Raquel Urtasun[2,3], and Antonio Torralba[1]

[1] Massachusetts Institute of Technology, Cambridge, USA
weichium@mit.edu
[2] Uber Advanced Technologies Group, Pittsburgh, USA
[3] University of Toronto, Toronto, Canada

Abstract. Intrinsic image decomposition—decomposing a natural image into a set of images corresponding to different physical causes—is one of the key and fundamental problems of computer vision. Previous intrinsic decomposition approaches either address the problem in a fully supervised manner, or require multiple images of the same scene as input. These approaches are less desirable in practice, as ground truth intrinsic images are extremely difficult to acquire, and requirement of multiple images pose severe limitation on applicable scenarios. In this paper, we propose to bring the best of both worlds. We present a two stream convolutional neural network framework that is capable of learning the decomposition effectively in the absence of any ground truth intrinsic images, and can be easily extended to a (semi-)supervised setup. At inference time, our model can be easily reduced to a single stream module that performs intrinsic decomposition on a single input image. We demonstrate the effectiveness of our framework through extensive experimental study on both synthetic and real-world datasets, showing superior performance over previous approaches in both single-image and multi-image settings. Notably, our approach outperforms previous state-of-the-art single image methods while using only 50% of ground truth supervision.

Keywords: Intrinsic decomposition · Unsupervised learning
Self-supervised learning

1 Introduction

In a scorching afternoon, you walk all the way through the sunshine and finally enter the shading. You notice that there is a sharp edge on the ground and the appearance of the sidewalk changes drastically. Without a second thought, you realize that the bricks are in fact identical and the color difference is due to the variation of scene illumination. Despite merely a quick glance, humans have the remarkable ability to decompose the intricate mess of confounds, which our visual

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11218, pp. 211–229, 2018.
https://doi.org/10.1007/978-3-030-01264-9_13

world is, into simple underlying factors. Even though most people have never seen a single intrinsic image in their lifetime, they can still estimate the intrinsic properties of the materials and reason about their relative albedo effectively [6]. This is because human visual systems have accumulated thousands hours of implicit observations which can serve as their priors during judgment. Such an ability not only plays a fundamental role in interpreting real-world imaging, but is also a key to truly understand the complex visual world. The goal of this work is to equip computational visual machines with similar capabilities by emulating humans' learning procedure. We believe by enabling perception systems to disentangle *intrinsic* properties (*e.g.* albedo) from *extrinsic* factors (*e.g.* shading), they will better understand the physical interactions of the world. In computer vision, such task of decomposing an image into a set of images each of which corresponds to a different physical cause is commonly referred to as *intrinsic decomposition* [4].

Despite the inverse problem being ill-posed [1], it has drawn extensive attention due to its potential utilities for algorithms and applications in computer vision. For instance, many low-level vision tasks such as shadow removal [14] and optical flow estimation [27] benefit substantially from reliable estimation of albedo images. Advanced image manipulation applications such as appearance editing [48], object insertions [24], and image relighting [49] also become much easier if an image is correctly decomposed into material properties and shading effects. Motivated by such great potentials, a variety of approaches have been proposed for intrinsic decomposition [6,17,28,62]. Most of them focus on monocular case, as it often arises in practice [13]. They either exploit manually designed priors [2,3,31,41], or capitalize on data-driven statistics [39,48,61] to address the ambiguities. The models are powerful, yet with a critical drawback—requiring ground truth for learning. The ground truth for intrinsic images, however, are extremely difficult and expensive to collect [16]. Current publicly available datasets are either small [16], synthetic [9,48], or sparsely annotated [6], which significantly restricts the scalability and generalizability of this task. To overcome the limitations, multi-image based approaches have been introduced [17,18,28,29,55]. They remove the need of ground truth and employ multiple observations to disambiguate the problem. While the unsupervised intrinsic decomposition paradigm is appealing, they require multi-image as input both during training and at inference, which largely limits their applications in real world.

In this work, we propose a novel approach to learning intrinsic decomposition that requires neither ground truth nor priors about scene geometry or lighting models. We draw connections between single image based methods and multi-image based approaches and explicitly show how one can benefit from the other. Following the derived formulation, we design an unified model whose *training* stage can be viewed as an approach to *multi-image intrinsic decomposition*. While at *test* time it is capable of decomposing arbitrary *single* image. To be more specific, we design a two stream deep architecture that observes a pair of images and aims to explain the variations of the scene by predicting the correct intrinsic decompositions. No ground truth is required for learning. The model reduces to a single stream network during inference and performs single image intrinsic decomposition. As the problem is under-constrained, we derive multiple objective functions based on image formation model to constrain the solution space and aid the learning process. We show that by regularizing the model carefully, the intrinsic images emerge automatically. The learned representations are not only comparable to those learned under full supervision, but can also serve as a better initialization for (semi-)supervised training. As a byproduct, our model also learns to predict whether a gradient belongs to albedo or shading without any labels. This provides an intuitive explanation for the model's behavior, and can be used for further diagnoses and improvements (Fig. 1).

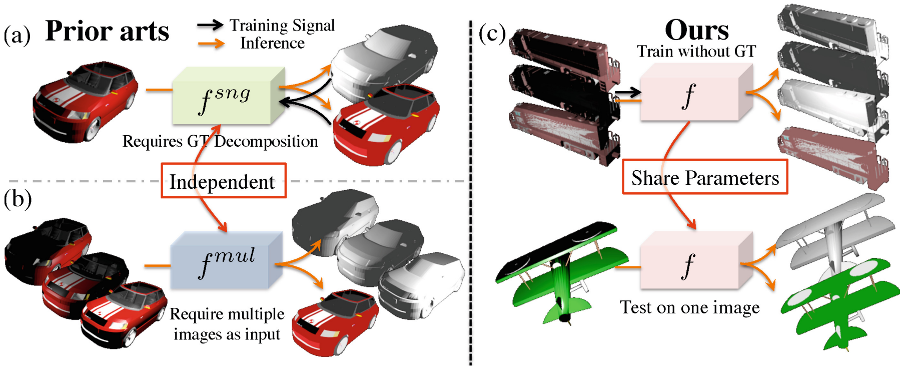

Fig. 1. Novelties and advantages of our approach: Previous works on intrinsic image decomposition can be classified into two categories, (a) single imaged based and (b) multi-image based. While single imaged based models are useful in practice, they require ground truth (GT) for training. Multi-image based approaches remove the need of GT, yet at the cost of flexibility (*i.e.*, always requires multiple images as input). (c) Our model takes the best of both world. We do not need GT during training (*i.e.*, training signal comes from input images), yet can be applied to arbitrary single image at test time.

We demonstrate the effectiveness of our model on one large-scale synthetic dataset and one real-world dataset. Our method achieves state-of-the-art performance on multi-image intrinsic decomposition, and significantly outperforms previous deep learning based single image intrinsic decomposition models using only 50% of ground truth data. To the best of our knowledge, we are the first attempt to bridge the gap between the two tasks and learn an intrinsic network without any ground truth intrinsic image.

2 Related Work

Intrinsic Decomposition. The work in intrinsic decomposition can be roughly classified into two groups: approaches that take as input only a single image [3,31,37,39,48,50,61,62], and algorithms that require addition sources of input [7,11,23,30,38,55]. For single image based methods, since the task is completely under constrained, they often rely on a variety of priors to help disambiguate the problem. [5,14,31,50] proposed to classify images edges into either albedo or shading and use [19] to reconstruct the intrinsic images. [34,41] exploited texture statistics to deal with the smoothly varying textures. While [3] explicitly modeled lighting conditions to better disentangle the shading effect, [42,46] assumed sparsity in albedo images. Despite many efforts have been put into designing priors, none of them has succeeded in including all intrinsic phenomenon. To avoid painstakingly constructing priors, [21,39,48,61,62] propose to capitalize on the feature learning capability of deep neural networks to learn the statistical priors directly from data. Their method, however, requires massive amount of labeled data, which is expensive to collect. In contrast, our deep learning based method requires no supervision. Another line of research in intrinsic decomposition leverages additional sources of input to resolve the problem, such as using image sequences [20,28–30,55], multi-modal input [2,11], or user annotations [7,8,47]. Similar to our work, [29,55] exploit a sequence of images taken from a fixed viewpoint, where the only variation is the illumination, to learn the decomposition. The critical difference is that these frameworks require multiple images for both training and testing, while our method rely on multiple images only during training. At test time, our network can perform intrinsic decomposition for an arbitrary single image.

Unsupervised/Self-supervised Learning from Image Sequences/Videos. Leveraging videos or image sequences, together with physical constraints, to train a neural network has recently become an emerging topic of research [15,32,44,51,52,56–59]. Zhou *et al.* [60] proposed a self-supervised approach to learning monocular depth estimation from image sequences. Vijayanarasimhan *et al.* [53] extended the idea and introduced a more flexible structure from motion framework that can incorporate supervision. Our work is conceptually similar to [53,60], yet focusing on completely different tasks. Recently, Janner *et al.* [21] introduced a self-supervised framework for transferring intrinsics. They first trained their network with ground truth and then fine-tune with reconstruction loss. In this work, we take a step further and attempt to learn intrinsic decomposition in a fully unsupervised manner. Concurrently and independently, Li and Snavely [33] also developed an approach to learning intrinsic decomposition without any supervision. More generally speaking, our work is in spirit similar to visual representation learning whose goal is to learn generic features by solving certain pretext tasks [22,43,54].

3 Background and Problem Formulation

In this section, we first briefly review current works on single image and multi-image intrinsic decomposition. Then we show the connections between the two tasks and demonstrate that they can be solved with a single, unified model under certain parameterizations.

3.1 Single Image Intrinsic Decomposition

The single image intrinsic decomposition problem is generally formulated as:

$$\hat{\mathcal{A}}, \hat{\mathcal{S}} = f^{sng}(\mathcal{I}; \boldsymbol{\Theta}^{sng}), \tag{1}$$

where the goal is to learn a function f that takes as input a natural image $\mathcal{I}$, and outputs an albedo image $\hat{\mathcal{A}}$ and a shading image $\hat{\mathcal{S}}$. The hat sign $\hat{\cdot}$ indicates that it is the output of the function rather than the ground truth. Ideally, the Hadamard product of the output images should be identical to the input image, *i.e.* $\mathcal{I} = \hat{\mathcal{A}} \odot \hat{\mathcal{S}}$. The parameter $\boldsymbol{\Theta}$ and the function f can take different forms. For instance, in traditional Retinex algorithm [31], $\boldsymbol{\Theta}$ is simply a threshold used to classify the gradients of the original image $\mathcal{I}$ and f^{sng} is the solver for Poisson equation. In recent deep learning based approaches [39,48], f^{sng} refers to a neural network and $\boldsymbol{\Theta}$ represents the weights. Since these models require only a single image as input, they potentially can be applied to various scenarios and have a number of use cases [13]. The problem, however, is inherently ambiguous and technically ill-posed under monocular setting. Ground truths are required to train either the weights for manual designed priors [6] or the data-driven statistics [21]. They learn by minimizing the difference between the GT intrinsic images and the predictions.

3.2 Multi-image Intrinsic Decomposition

Another way to address the ambiguities in intrinsic decomposition is to exploit multiple images as input. The task is defined as:

$$\hat{\mathbf{A}}, \hat{\mathbf{S}} = f^{mul}(\mathbf{I}; \boldsymbol{\Theta}^{mul}), \tag{2}$$

where $\mathbf{I} = \{\mathcal{I}_i\}_{i=1}^{N}$ is the set of input images of the same scene, and $\hat{\mathbf{A}} = \{\hat{\mathcal{A}}_i\}_{i=1}^{N}$, $\hat{\mathbf{S}} = \{\hat{\mathcal{S}}_i\}_{i=1}^{N}$ are the corresponding set of intrinsic predictions. The input images $\mathbf{I}$ can be collected with a moving camera [27], yet for simplicity they are often assumed being captured with a static camera pose under varying lighting conditions [29,36]. The extra constraint not only gives birth to some useful priors [55], but also open the door to solving the problem in an unsupervised manner [18]. For example, based on the observation that shadows tend to move and a pixel in a static scene is unlikely to contain shadow edges in multiple images, Weiss [55] assumed that the median gradients across all images belong to albedo and solve the Poisson equation. The simple algorithm works well on shadow removal, and was further extend by [36] to combine with Retinex algorithm (W+Ret) to produce better results. More recently, Laffont and Bazin [29] derived several energy functions based on image formation model and formulate the task as an optimization problem. The goal simply becomes finding the intrinsic images that minimize the pre-defined energy. Ground truth data is not required under many circumstances [18,29,55]. This addresses one of the major difficulties in learning intrinsic decomposition. Unfortunately, as a trade off, these models rely on multi-image as input all the time, which largely limits their applicability in practice.

3.3 Connecting Single and Multi-image Based Approaches

The key insight is to use a same set of parameters $\mathbf{\Theta}$ for both single image and multi-image intrinsic decomposition. Multi-image approaches have already achieved impressive results without the need of ground truth. If we can transfer the learned parameters from multi-image model to single image one, then we will be able to decompose arbitrary single image without any supervision. Unfortunately, previous works are incapable of doing this. The multi-image parameters $\mathbf{\Theta}^{mul}$ or energy functions are often dependent on all input images $\mathbf{I}$, which makes them impossible to be reused under single image setting. With such motivation in mind, we design our model to have the following form:

$$f^{mul}(\mathbf{I};\mathbf{\Theta}) = g(f^{sng}(\mathcal{I}_1;\mathbf{\Theta}), f^{sng}(\mathcal{I}_2;\mathbf{\Theta}), ..., f^{sng}(\mathcal{I}_N;\mathbf{\Theta})), \tag{3}$$

where g denotes some parameter-free, pre-defined constraints applied to the outputs of single image models. By formulating the multi-image model f^{mul} as a composition function of multiple single image model f^{sng}, we are able to share the same parameters $\mathbf{\Theta}$ and further learn the single image model through multi-image training without any ground truth. The high-level idea of sharing parameters has been introduced in W+Ret [36]; however, our work exists three critical differences: first and foremost, their approach requires ground truth for learning, while ours does not. Second, they encode the information across several observations at the input level via some heuristics. In contrast, our aggregation function g is based on image formation model, and operates directly on the intrinsic predictions. Finally, rather than employing the relatively simple Retinex model, we parameterize f^{sng} as a neural network, with $\mathbf{\Theta}$ being its weight, and g being a series of carefully designed, parameter-free, and differentiable operations. The details of our model are discussed in Sect. 4 and the differences between our method and several previous approaches are summarized in Table 1.

Table 1. Summary of different intrinsic decomposition approaches.

Methods	Supervision	Training input	Inference input	Learnable parameter $\boldsymbol{\Theta}$
Retinex [31]	✓	Single image	Single image	Gradient threshold
CNN [21,39,48]	✓	Single image	Single image	Network weights
CRF [6,61]	✓	Single image	Single image	Energy weights
Weiss [55]	✕	Multi-image	Multi-image	None
W+RET [36]	✓	Multi-image	Multi-image	Gradient threshold
Hauagge *et al.* [18]	✕	Multi-image	Multi-image	None
Laffont *et al.* [29]	✕	Multi-image	Multi-image	None
Our method	✕	Multi-image	Single image	Network weights

4 Unsupervised Intrinsic Learning

Our model consists of two main components: the *intrinsic network* f^{sng}, and the *aggregation function* g. The intrinsic network f^{sng} produces a set of intrinsic representations given an input image. The differentiable, parameter-free aggregation function g constrains the outputs of f^{sng}, so that they are plausible and comply to the image formation model. As all operations are differentiable, the errors can be backpropagated all the way through f^{sng} during training. Our model can be trained even no ground truth exists. The training stage is hence equivalent to performing multi-image intrinsic decomposition. At test time, the trained intrinsic network f^{sng} serves as an independent module, which enables decomposing an arbitrary single image. In this work, we assume the input images come in pairs during training. This works well in practice and an extension to more images is trivial. We explore three different setups of the aggregation function. An overview of our model is shown in Fig. 2.

4.1 Intrinsic Network f^{sng}

The goal of the intrinsic network is to produce a set of reliable intrinsic representations from the input image and then pass them to the aggregation function for further composition and evaluation. To be more formal, given a single image $\mathcal{I}_1$, we seek to learn a neural network f^{sng} such that $(\hat{\mathcal{A}}_1, \hat{\mathcal{S}}_1, \hat{\mathcal{M}}_1) = f^{sng}(\mathcal{I}_1; \boldsymbol{\Theta})$, where $\mathcal{A}$ denotes albedo, $\mathcal{S}$ refers to shading, and $\mathcal{M}$ represents a soft assignment mask (details in Sect. 4.2).

Following [12,45,48], we employ an encoder-decoder architecture with skip links for f^{sng}. The bottom-up top-down structure enables the network to effectively process and consolidate features across various scales [35], while the skip links from encoder to decoder help preserve spatial information at each resolution [40]. Since the intrinsic components (*e.g.* albedo, shading) are mutual dependent, they share the same encoder. In general, our network architecture is similar to the Mirror-link network [47]. We, however, note that this is not the only feasible choice. Other designs that disperse and aggregate information in different manners may also work well for our task. One can replace the current structure with arbitrary network as long as the output has the same resolution as the input. We refer the readers to supp. material for detailed architecture.

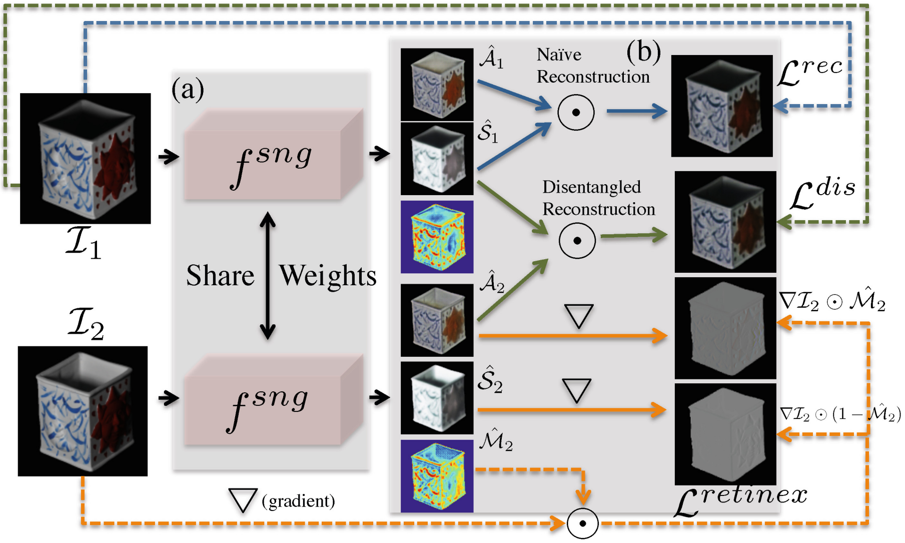

Fig. 2. Network architecture for training: Our model consists of *intrinsic networks* and *aggregation functions*. (a) The siamese intrinsic network takes as input a pair of images with varying illumination and generate a set of intrinsic estimations. (b) The aggregation functions compose the predictions into images whose ground truths are available via pre-defined operations (*i.e.* the orange, green, and blue lines). The objectives are then applied to the final outputs, and the errors are backpropagated all the way to the intrinsic network to refine the estimations. With this design, our model is able to learn intrinsic decomposition without a single ground truth image. Note that the model is symmetric and for clarity we omit similar lines. The full model is only employed during training. At test time, our model reduces to a single stream network f^{sng} (pink) and performs single image intrinsic decomposition. (Color figure online)

4.2 Aggregation Functions g and Objectives

Suppose now we have the intrinsic representations predicted by the intrinsic network. In order to evaluate the performance of these estimations, whose ground truths are unavailable, and learn accordingly, we exploit several differentiable aggregation functions. Through a series of fixed, pre-defined operations, the aggregation functions re-compose the estimated intrinsic images into images which we have ground truth for. We can then compute the objectives and use it to guide the network learning. Keeping such motivation in mind, we design the following three aggregation functions as well as the corresponding objectives.

Naive Reconstruction. The first aggregation function simply follows the definition of intrinsic decomposition: given the estimated intrinsic tensors $\hat{\mathcal{A}}_1$ and $\hat{\mathcal{S}}_1$, the Hadamard product $\hat{\mathcal{I}}_1^{rec} = \hat{\mathcal{A}}_1 \odot \hat{\mathcal{S}}_1$ should flawlessly reconstruct the original input image $\mathcal{I}_1$. Building upon this idea, we employ a pixel-wise regression loss $\mathcal{L}_1^{rec} = \|\hat{\mathcal{I}}_1^{rec} - \mathcal{I}_1\|_2$ on the reconstructed output, and constrain the network to learn only the representations that satisfy this rule. Despite such objective greatly reduce the solution space of intrinsic representations, the problem is still highly under-constrained—there exists infinite images that meet $\mathcal{I}_1 = \hat{\mathcal{A}}_1 \odot \hat{\mathcal{S}}_1$. We thus employ another aggregation operation to reconstruct the input images and further constrain the solution manifold.

Disentangled Reconstruction. According to the definition of intrinsic images, the albedo component should be invariant to illumination changes. Hence given a pair of images $\mathcal{I}_1, \mathcal{I}_2$ of the same scene, ideally we should be able to perfectly reconstruct $\mathcal{I}_1$ even with $\hat{\mathcal{A}}_2$ and $\hat{\mathcal{S}}_1$. Based on this idea, we define our second aggregation function to be $\hat{\mathcal{I}}_1^{dis} = \hat{\mathcal{A}}_2 \odot \hat{\mathcal{S}}_1$. By taking the albedo estimation from the other image yet still hoping for perfect reconstruction, we force the network to extract the illumination invariant component automatically. Since we aim to disentangle the illumination component through this reconstruction process, we name the output as disentangled reconstruction. Similar to naive reconstruction, we employ a pixel-wise regression loss $\mathcal{L}_1^{dis}$ for $\hat{\mathcal{I}}_1^{dis}$.

One obvious shortcut that the network might pick up is to collapse all information from input image into $\hat{\mathcal{S}}_1$, and have the albedo decoder always output a white image regardless of input. In this case, the albedo is still invariant to illumination, yet the network fails. In order to avoid such degenerate cases, we follow Jayaraman and Grauman [22] and incorporate an additional embedding loss $\mathcal{L}_1^{ebd}$ for regularization. Specifically, we force the two albedo predictions $\hat{\mathcal{A}}_1$ and $\hat{\mathcal{A}}_2$ to be as similar as possible, while being different from the randomly sampled albedo predictions $\hat{\mathcal{A}}_{neg}$.

Gradient. As natural images and intrinsic images exhibit stronger correlations in gradient domain [25], the third operation is to convert the intrinsic estimations to gradient domain, *i.e.* $\nabla\hat{\mathcal{A}}_1$ and $\nabla\hat{\mathcal{S}}_1$. However, unlike the outputs of the previous two aggregation function, we do not have ground truth to directly supervise the gradient images. We hence propose a self-supervised approach to address this issue.

Our method is inspired by the traditional Retinex algorithm [31] where each derivative in the image is assumed to be caused by either change in albedo or that of shading. Intuitively, if we can accurately classify all derivatives, we can then obtain ground truths for $\nabla\hat{\mathcal{A}}_1$ and $\nabla\hat{\mathcal{S}}_1$. We thus exploit deep neural network for edge classification. To be more specific, we let the intrinsic network predict a *soft* assignment mask $\mathcal{M}_1$ to determine to which intrinsic component each edge belongs. Unlike [31] where a image derivative can only belong to either albedo or shading, the assignment mask outputs the *probability* that a image derivative is caused by changes in albedo. One can think of it as a soft version of Retinex algorithm, yet completely data-driven without manual tuning. With the help of the soft assignment mask, we can then generate the "pseudo" ground truth

$\nabla\mathcal{I} \odot \hat{\mathcal{M}}_1$ and $\nabla\mathcal{I} \odot (1 - \hat{\mathcal{M}}_1)$ to supervise the gradient intrinsic estimations. The Retinex loss[1] is defined as follows:

$$\mathcal{L}_1^{retinex} = \|\nabla\hat{\mathcal{A}}_1 - \nabla\mathcal{I} \odot \hat{\mathcal{M}}_1\|_2 + \|\nabla\hat{\mathcal{S}}_1 - \nabla\mathcal{I} \odot (1 - \hat{\mathcal{M}}_1)\|_2 \qquad (4)$$

The final objective thus becomes:

$$\mathcal{L}_1^{final} = \mathcal{L}_1^{rec} + \lambda_d\mathcal{L}_1^{dis} + \lambda_r\mathcal{L}_1^{retinex} + \lambda_e\mathcal{L}_1^{ebd}, \qquad (5)$$

where λ's are the weightings. In practice, we set $\lambda_d = 1$, $\lambda_r = 0.1$, and $\lambda_e = 0.01$. We select them based on the stability of the training loss. $\mathcal{L}_2^{final}$ is completely identical as we use a siamese network structure.

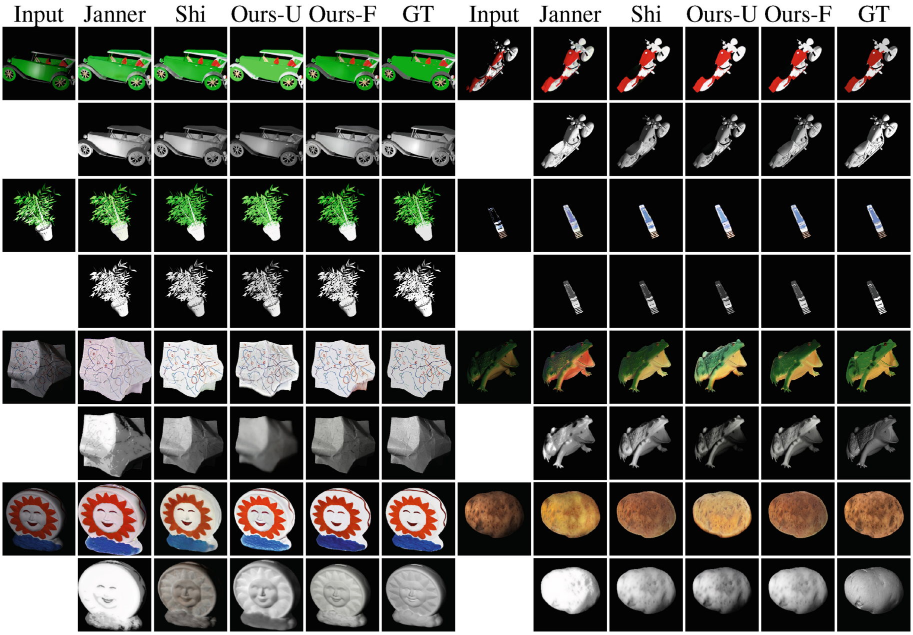

Fig. 3. Single image intrinsic decomposition: Our model (`Ours-U`) learns the intrinsic representations without any supervision and produces best results after fine-tuning (`Ours-F`).

4.3 Training and Testing

Since we only supervise the output of the aggregation functions, we do not enforce that each decoder in the intrinsic network solves its respective subproblem (*i.e.* albedo, shading, and mask). Rather, we expect that the proposed network structure encourages these roles to emerge automatically. Training the

[1] In practice, we need to transform all images into *logarithm* domain before computing the gradient and applying Retinex loss. We omit the log operator here for simplicity.

network from scratch without direction supervision, however, is a challenging problem. It often results in semantically meaningless intermediate representations [49]. We thus introduce additional constraints to carefully regularize the intrinsic estimations during training. Specifically, we penalize the L1 norm of the gradients for the albedo and minimize the L1 norm of the second-order gradients for the shading. While $\|\nabla\hat{\mathcal{A}}\|$ encourages the albedo to be piece-wise constant, $\|\nabla^2\hat{\mathcal{S}}\|$ favors smoothly changing illumination. To further encourage the emergence of the soft assignment mask, we compute the gradient of the input image and use it to supervise the mask for the first four epochs. The early supervision pushes the mask decoder towards learning a gradient-aware representation. The mask representations are later freed and fine-tuned during the joint self-supervised training process. We train our network with ADAM [26] and set the learning rate to 10^{-5}. We augment our training data with horizontal flips and random crops.

Extending to (Semi-)supervised Learning. Our model can be easily extended to (semi-)supervised settings whenever a ground truth is available. In the original model, the objectives are only applied to the final output of the aggregation functions and the output of the intrinsic network is left without explicit guidance. Hence, a straightforward way to incorporate supervision is to directly supervise the intermediate representation and guide the learning process. Specifically, we can employ a pixel-wise regression loss on both albedo and shading, *i.e.* $\mathcal{L}^A = \|\hat{\mathcal{A}} - \mathcal{A}\|_2$ and $\mathcal{L}^S = \|\hat{\mathcal{S}} - \mathcal{S}\|_2$.

5 Experiments

5.1 Setup

Data. To effectively evaluate our model, we consider two datasets: one larger-scale synthetic dataset [21,48], and one real world dataset [16]. For synthetic dataset, we use the 3D objects from ShapeNet [10] and perform rendering in Blender[2]. Specifically, we randomly sample 100 objects from each of the following 10 categories: airplane, boat, bottle, car, flowerpot, guitar, motorbike, piano, tower, and train. For each object, we randomly select 10 poses, and for each pose we use 10 different lightings. This leads to in total of $100 \times 10 \times 10 \times C_2^{10} = 450K$ pairs of images. We split the data by objects, in which 90% belong to training and validation and 10% belong to test split.

The MIT Intrinsics dataset [16] is a real-world image dataset with ground truths. The dataset consists of 20 objects. Each object was captured under 11 different illumination conditions, resulting in 220 images in total. We use the same data split as in [39,48], where the images are split into two folds by objects (10 for each split).

[2] We follow the same rendering process as [21]. Please refer to their paper for more details.

Metrics. We employ two standard error measures to quantitatively evaluate the performance of our model: the standard mean-squared error (MSE) and the local mean-squared error (LMSE) [16]. Comparing to MSE, LMSE provides a more fine-grained measure. It allows each local region to have a different scaling factor. We set the size of the sliding window in LSME to 12.5% of the image in each dimension.

5.2 Multi-image Intrinsic Decomposition

Since no ground truth data has been used during training, our training process can be viewed as an approach to multi-image intrinsic decomposition.

Baselines. For fair analysis, we compare with methods that also take as input a sequence of photographs of the same scene with varying illumination conditions. In particular, we consider three publicly available multi-image based approaches: Weiss [55], W+Ret [36], and Hauagge *et al.* [17].

Results. Following [16,29], we use LMSE as the main metric to evaluate our multi-image based model. The results are shown in Table 2. As our model is able to effectively harness the optimization power of deep neural network, we outperform all previous methods that rely on hand-crafted priors or explicit lighting modelings.

Table 2. Comparison against multi-image based methods.

Methods	Average LMSE	
	MIT	ShapeNet
Weiss [55]	0.0215	0.0632
W+Ret [36]	0.0170	0.0525
Hauagge *et al.* [18]	0.0155	-
Hauagge *et al.* [17]	0.0115	0.0240
Laffont *et al.* [29]	0.0138	-
Our method	**0.0097**	**0.0049**

5.3 Single Image Intrinsic Decomposition

Baselines. We compare our approach against three state-of-the-art methods: Barron *et al.* [3], Shi *et al.* [48], and Janner *et al.* [21]. While Barron *et al.* handcraft priors for shape, shading, albedo and pose the task as an optimization problem. Shi *et al.* [48], and Janner *et al.* [21] exploit deep neural network to

Table 3. Comparison against single image-based methods on ShapeNet: Our unsupervised intrinsic model is comparable to [3]. After fine-tuning, it achieves state-of-the-art performances.

Methods	Supervision	MSE			LMSE		
	Amount	Albedo	Shading	Average	Albedo	Shading	Average
Barron *et al.* [3]	100%	0.0203	0.0232	0.0217	0.0066	0.0043	0.0055
Janner *et al.* [21]	100%	0.0119	0.0145	0.0132	0.0028	0.0037	0.0032
Shi *et al.* [48]	100%	0.0076	0.0122	0.0099	0.0018	0.0032	0.0024
Our method (U)	0%	0.0174	0.0310	0.0242	0.0050	0.0070	0.0060
Our method (F)	100%	**0.0064**	**0.0100**	**0.0082**	**0.0016**	**0.0025**	**0.0020**

learn natural image statistics from data and predict the decomposition. All three methods require ground truth for learning.

Results. As shown in Tables 3 and 4, our unsupervised intrinsic network f^{sng}, denoted as `Ours-U`, achieves comparable performance to other deep learning based approaches on MIT Dataset, and is on par with Barron *et al.* on ShapeNet. To further evaluate the learned unsupervised representation, we use it as initialization and fine-tune the network with ground truth data. The fine-tuned representation, denoted as `Ours-F`, significantly outperforms all baselines on ShapeNet and is comparable with Barron *et al.* on MIT Dataset. We note that MIT Dataset is extremely hard for deep learning based approaches due to its scale. Furthermore, Barron *et al.* employ several priors specifically designed for the dataset. Yet with our unsupervised training scheme, we are able to overcome the data issue and close the gap from Barron *et al.* Some qualitative results are shown in Fig. 3. Our unsupervised intrinsic network, in general, produces reasonable decompositions. With further fine-tuning, it achieves the best results. For instance, our full model better recovers the albedo of the wheel cover of the car. For the motorcycle, it is capable of predicting the correct albedo of the wheel and the shading of the seat.

Table 4. Comparison against single image-based methods on MIT Dataset: Our unsupervised intrinsic model achieves comparable performance to fully supervised deep models. After fine-tuning, it is on par with the best performing method that exploits specialized priors.

Methods	Supervision	MSE			LMSE		
	Amounts	Albedo	Shading	Average	Albedo	Shading	Average
Barron *et al.* [3]	100%	**0.0147**	**0.0083**	**0.0115**	**0.0061**	**0.0039**	**0.0050**
Janner *et al.* [39]	100%	0.0336	0.0195	0.0265	0.0210	0.0103	0.0156
Shi *et al.* [48]	100%	0.0323	0.0156	0.0239	0.0132	0.0064	0.0098
Our method (U)	0%	0.0313	0.0207	0.0260	0.0116	0.0095	0.0105
Our method (F)	100%	0.0168	0.0093	0.0130	0.0074	0.0052	0.0063

(Semi-)supervised Intrinsic Learning. As mentioned in Sect. 4.3, our network can be easily extended to (semi-)supervised settings by exploiting ground truth images to directly supervise the intrinsic representations. To better understand how well our unsupervised representation is and exactly how much ground truth data we need in order to achieve comparable performance to previous methods, we gradually increase the degree of supervision during training and study the performance variation. The results on ShapeNet are plotted in Fig. 4. Our model is able to achieve state-of-the-art performance with only 50% of ground truth data. This suggests that our aggregation function is able to effectively constrain the solution space and capture the features that are not directly encoded in single images. In addition, we observe that our model has a larger performance gain with less ground truth data. The relative improvement gradually converges as the amount of supervision increases, showing our utility in low-data regimes.

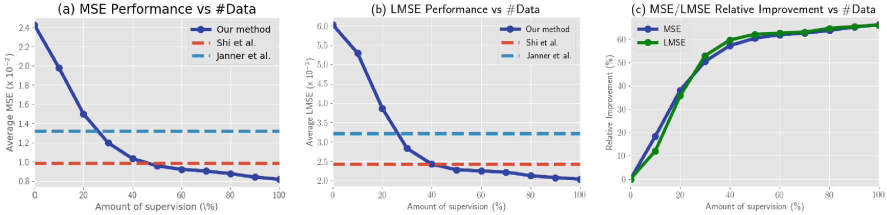

Fig. 4. Performance vs Supervision on ShapeNet: The performance of our model improves with the amount of supervision. (a) (b) Our results suggest that, with just 50% of ground truth, we can surpass the performance of other fully supervised models that used all of the labeled data. (c) The relative improvement is larger in cases with less labeled data, showing the effectiveness of our unsupervised objectives in low-data regimes.

5.4 Analysis

Ablation Study. To better understand the contribution of each component in our model, we visualize the output of the intrinsic network (*i.e.* $\hat{\mathcal{A}}$ and $\hat{\mathcal{S}}$) under different network configurations in Fig. 5. We start from the simple auto-encoder structure (*i.e.* using only $\mathcal{L}^{rec}$) and sequentially add other components back. At first, the model splits the image into arbitrary two components. This is expected since the representations are fully unconstrained as long as they satisfy $\mathcal{I} = \hat{\mathcal{A}} \odot \hat{\mathcal{S}}$. After adding the disentangle learning objective $\mathcal{L}^{dis}$, the albedo images becomes more "flat", suggesting that the model starts to learn that albedo components should be invariant of illumination. Finally, with the help of the Retinex loss $\mathcal{L}^{retinex}$, the network self-supervises the gradient images, and produces reasonable intrinsic representations without any supervision. The color is significantly improved due to the information lying in the gradient domain. The quantitative evaluations are shown in Table 5.

Table 5. Ablation studies: The performance of our model when employing different objectives.

Employed objectives			MSE		LMSE	
$\mathcal{L}_{rec}$	$\mathcal{L}_{dis}$	$\mathcal{L}_{retinex}$	Albedo	Shading	Albedo	Shading
✓			0.0362	0.0240	0.0158	0.0108
✓	✓		0.0346	0.0224	0.0141	0.0098
✓	✓	✓	**0.0313**	**0.0207**	**0.0116**	**0.0095**

Table 6. Degree of illumination invariance of the albedo image. Lower is better.

Methods	MPRE ($\times 10^{-4}$)
Barron *et al.* [3]	2.6233
Janner *et al.* [39]	4.8372
Shi *et al.* [48]	5.1589
Our method (U)	3.2341
Our method (F)	**2.4151**

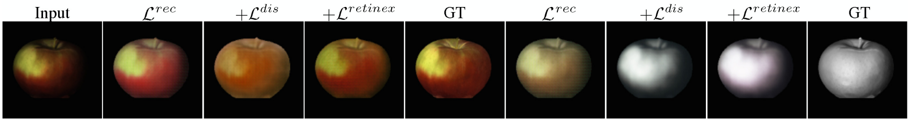

Fig. 5. Contributions of each objectives: Initially the model separates the image into two arbitrary components. After adding the disentangled loss $\mathcal{L}^{dis}$, the network learns to exclude illumination variation from albedo. Finally, with the help of the Retinex loss $\mathcal{L}^{retinex}$, the albedo color becomes more saturated.

Natural Image Disentangling. To demonstrate the generalizability of our model, we also evaluate on natural images in the wild. Specifically, we use our full model on MIT Dataset and the images provided by Barron *et al.* [3]. The images are taken by a iPhone and span a variety of categories. Despite our model is trained purely on laboratory images and have never seen other objects/scenes before, it still produces good quality results (see Fig. 6). For instance, our model successfully infers the intrinsic properties of the banana and the plants. One limitation of our model is that it cannot handle the specularity in the image. As we ignore the specular component when formulating the task, the specular parts got treated as sharp material changes and are classified as albedo. We plan to incorporate the idea of [48] to address this issue in the future.

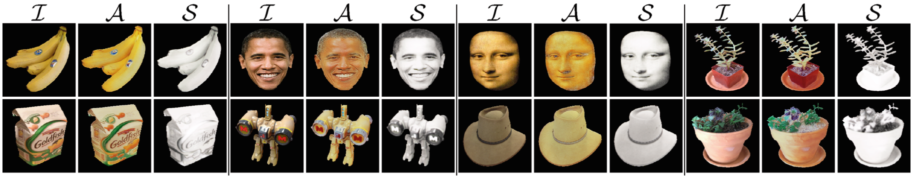

Fig. 6. Decomposing unseen natural images: Despite being trained on laboratory images, our model generalizes well to real images that it has never seen before.

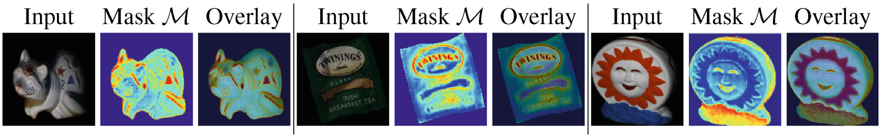

Fig. 7. Network interpretation: To understand how our model sees an edge in the input image, we visualize the soft assignment mask $\mathcal{M}$ predicted by the intrinsic network. An edge has a higher probability to be assigned to albedo when there is a drastic color change. (Color figure online)

Robustness to Illumination Variation. Another way to evaluate the effectiveness of our approach is to measure the degree of illumination invariance of our albedo model. Following Zhou *et al.* [61], we compute the MSE between the input image $\mathcal{I}_1$ and the disentangled reconstruction $\hat{\mathcal{I}}_1^{dis}$ to evaluate the illumination invariance. Since our model explicitly takes into account the disentangled objective $\mathcal{L}^{dis}$, we achieve the best performance. Results on MIT Dataset are shown in Table 6.

Interpreting the Soft Assignment Mask. The soft assignment mask predicts the probability that a certain edge belongs to albedo. It not only enables the self-supervised Retinex loss, but can also serve as a probe to our model, helping us interpret the results. By visualizing the predicted soft assignment mask $\mathcal{M}$, we can understand how the network *sees* an edge—an edge caused by albedo change or variation of shading. Some visualization results of our unsupervised intrinsic network are shown in Fig. 7. The network believes that drastic color changes are most of the time due to albedo edges. Sometimes it mistakenly classify the edges, *e.g.* the variation of the blue paint on the sun should be due to shading. This mistake is consistent with the sun albedo result in Fig. 3, yet it provides another intuition of why it happens. As there is no ground truth to directly evaluate the performance of the predicted assignment map, we instead measure the pixel-wise difference between the ground truth gradient images $\nabla\mathcal{A}, \nabla\mathcal{S}$ and the "pseudo" ground truths $\nabla\mathcal{I} \odot \mathcal{M}, \nabla\mathcal{I} \odot (1 - \mathcal{M})$ that we used for self-supervision. Results show that our data-driven assignment mask (1.7×10^{-4}) better explains the real world images than traditional Retinex algorithm (2.6×10^{-4}).

6 Conclusion

An accurate estimate of intrinsic properties not only provides better understanding of the real world, but also enables various applications. In this paper, we present a novel method to disentangle the factors of variations in the image. With the carefully designed architecture and objectives, our model automatically learns reasonable intrinsic representations without any supervision. We believe it is an interesting direction for intrinsic learning and we hope our model can facilitate further research in this path.

References

1. Adelson, E.H., Pentland, A.P.: The perception of shading and reflectance. In: Perception as Bayesian Inference. Cambridge University Press, New York (1996)
2. Barron, J.T., Malik, J.: Intrinsic scene properties from a single RGB-D image. In: CVPR (2013)
3. Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. In: PAMI (2015)
4. Barrow, H., Tenenbaum, J.: Recovering intrinsic scene characteristics from images. Comput. Vis. Syst. **2**, 3–26 (1978)

5. Bell, M., Freeman, E.: Learning local evidence for shading and reflectance. In: ICCV (2001)
6. Bell, S., Bala, K., Snavely, N.: Intrinsic images in the wild. TOG **33**(4), 159 (2014)
7. Bonneel, N., Sunkavalli, K., Tompkin, J., Sun, D., Paris, S., Pfister, H.: Interactive intrinsic video editing. TOG **33**(6), 197 (2014)
8. Bousseau, A., Paris, S., Durand, F.: User-assisted intrinsic images. TOG **28**(5), 130 (2009)
9. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 611–625. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33783-3_44
10. Chang, A.X., et al.: ShapeNet: an information-rich 3D model repository. arXiv (2015)
11. Chen, Q., Koltun, V.: A simple model for intrinsic image decomposition with depth cues. In: ICCV (2013)
12. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In: NIPS (2016)
13. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: NIPS (2014)
14. Finlayson, G.D., Hordley, S.D., Drew, M.S.: Removing shadows from images using retinex. In: Color and Imaging Conference (2002)
15. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR (2016)
16. Grosse, R., Johnson, M.K., Adelson, E.H., Freeman, W.T.: Ground truth dataset and baseline evaluations for intrinsic image algorithms. In: ICCV (2009)
17. Hauagge, D., Wehrwein, S., Bala, K., Snavely, N.: Photometric ambient occlusion. In: CVPR (2013)
18. Hauagge, D.C., Wehrwein, S., Upchurch, P., Bala, K., Snavely, N.: Reasoning about photo collections using models of outdoor illumination. In: BMVC (2014)
19. Horn, B.: Robot Vision. Springer, Heidelberg (1986). https://doi.org/10.1007/978-3-662-09771-7
20. Hui, Z., Sankaranarayanan, A.C., Sunkavalli, K., Hadap, S.: White balance under mixed illumination using flash photography. In: ICCP (2016)
21. Janner, M., Wu, J., Kulkarni, T.D., Yildirim, I., Tenenbaum, J.: Self-supervised intrinsic image decomposition. In: NIPS (2017)
22. Jayaraman, D., Grauman, K.: Learning image representations tied to ego-motion. In: ICCV (2015)
23. Jeon, J., Cho, S., Tong, X., Lee, S.: Intrinsic image decomposition using structure-texture separation and surface normals. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 218–233. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10584-0_15
24. Karsch, K., Hedau, V., Forsyth, D., Hoiem, D.: Rendering synthetic objects into legacy photographs. TOG **30**(6), 157 (2011)
25. Kim, S., Park, K., Sohn, K., Lin, S.: Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 143–159. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_9
26. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv (2014)
27. Kong, N., Black, M.J.: Intrinsic depth: improving depth transfer with intrinsic images. In: ICCV (2015)

28. Kong, N., Gehler, P.V., Black, M.J.: Intrinsic video. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 360–375. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10605-2_24
29. Laffont, P.Y., Bazin, J.C.: Intrinsic decomposition of image sequences from local temporal variations. In: ICCV (2015)
30. Laffont, P.Y., Bousseau, A., Drettakis, G.: Rich intrinsic image decomposition of outdoor scenes from multiple views. In: TVCG (2013)
31. Land, E.H., McCann, J.J.: Lightness and retinex theory. J. Opt. Soc. Am. **61**(1), 1–11 (1971)
32. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 577–593. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_35
33. Li, Z., Snavely, N.: Learning intrinsic image decomposition from watching the world. In: CVPR (2018)
34. Liu, X., Jiang, L., Wong, T.T., Fu, C.W.: Statistical invariance for texture synthesis. In: TVCG (2012)
35. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
36. Matsushita, Y., Nishino, K., Ikeuchi, K., Sakauchi, M.: Illumination normalization with time-dependent intrinsic images for video surveillance. In: PAMI (2004)
37. Meka, A., Maximov, M., Zollhöfer, M., Chatterjee, A., Richardt, C., Theobalt, C.: Live intrinsic material estimation. arXiv (2018)
38. Meka, A., Zollhöfer, M., Richardt, C., Theobalt, C.: Live intrinsic video. TOG **35**(4), 109 (2016)
39. Narihira, T., Maire, M., Yu, S.X.: Direct intrinsics: learning Albedo-shading decomposition by convolutional regression. In: ICCV (2015)
40. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
41. Oh, B.M., Chen, M., Dorsey, J., Durand, F.: Image-based modeling and photo editing. In: Computer Graphics and Interactive Techniques (2001)
42. Omer, I., Werman, M.: Color lines: image specific color representation. In: CVPR (2004)
43. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: feature learning by inpainting. In: CVPR (2016)
44. Rezende, D.J., Eslami, S.A., Mohamed, S., Battaglia, P., Jaderberg, M., Heess, N.: Unsupervised learning of 3D structure from images. In: NIPS (2016)
45. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: MIC-CAI (2015)
46. Rother, C., Kiefel, M., Zhang, L., Schölkopf, B., Gehler, P.V.: Recovering intrinsic images with a global sparsity prior on reflectance. In: NIPS (2011)
47. Shen, J., Yang, X., Jia, Y., Li, X.: Intrinsic images using optimization. In: CVPR (2011)
48. Shi, J., Dong, Y., Su, H., Yu, S.X.: Learning non-lambertian object intrinsics across shapenet categories (2017)
49. Shu, Z., Yumer, E., Hadap, S., Sunkavalli, K., Shechtman, E., Samaras, D.: Neural face editing with intrinsic image disentangling. In: CVPR (2017)
50. Tappen, M.F., Freeman, W.T., Adelson, E.H.: Recovering intrinsic images from a single image. In: NIPS (2003)

51. Tung, H.Y., Tung, H.W., Yumer, E., Fragkiadaki, K.: Self-supervised learning of motion capture. In: NIPS (2017)
52. Tung, H.Y.F., Harley, A.W., Seto, W., Fragkiadaki, K.: Adversarial inverse graphics networks: learning 2D-to-3D lifting and image-to-image translation from unpaired supervision. In: ICCV (2017)
53. Vijayanarasimhan, S., Ricco, S., Schmid, C., Sukthankar, R., Fragkiadaki, K.: SFM-Net: learning of structure and motion from video. arXiv (2017)
54. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: ICCV (2015)
55. Weiss, Y.: Deriving intrinsic images from image sequences. In: ICCV (2001)
56. Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H.: Perspective transformer nets: learning single-view 3D object reconstruction without 3D supervision. In: NIPS (2016)
57. Yang, J., Reed, S.E., Yang, M.H., Lee, H.: Weakly-supervised disentangling with recurrent transformations for 3D view synthesis. In: NIPS (2015)
58. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 649–666. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_40
59. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. arXiv (2018)
60. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: CVPR (2017)
61. Zhou, T., Krahenbuhl, P., Efros, A.A.: Learning data-driven reflectance priors for intrinsic image decomposition. In: ICCV (2015)
62. Zoran, D., Isola, P., Krishnan, D., Freeman, W.T.: Learning ordinal relationships for mid-level vision. In: ICCV (2015)