# Unsupervised Learning for Intrinsic Image Decomposition from a Single Image

Yunfei Liu[1] Yu Li[2] Shaodi You[3] Feng Lu[1, 4, *]
[1] State Key Laboratory of VR Technology and Systems, School of CSE, Beihang University
[2]Applied Research Center (ARC), Tencent PCG [3]University of Amsterdam, Amsterdam, Netherland
[4]Peng Cheng Laboratory, Shenzhen, China
{lyunfei,lufeng}@buaa.edu.cn ianyli@tencent.com s.you@uva.nl

## Abstract

*Intrinsic image decomposition, which is an essential task in computer vision, aims to infer the reflectance and shading of the scene. It is challenging since it needs to separate one image into two components. To tackle this, conventional methods introduce various priors to constrain the solution, yet with limited performance. Meanwhile, the problem is typically solved by supervised learning methods, which is actually not an ideal solution since obtaining ground truth reflectance and shading for massive general natural scenes is challenging and even impossible. In this paper, we propose a novel unsupervised intrinsic image decomposition framework, which relies on neither labelled training data nor hand-crafted priors. Instead, it directly learns the latent feature of reflectance and shading from unsupervised and uncorrelated data. To enable this, we explore the independence between reflectance and shading, the domain invariant content constraint and the physical constraint. Extensive experiments on both synthetic and real image datasets demonstrate consistently superior performance of the proposed method.*

## 1. Introduction

The appearance of a natural image depends on various factors, such as illumination, shape and material. Intrinsic image decomposition aims to decompose such a natural image into an illumination-invariant component and an illumination-variant component. Therefore, it can benefit a variety of high-level computer vision tasks such as texture editing [3], face appearance editing [5] and many others. In this paper, we follow the common practice [8, 18, 28] that assumes the ideal Lambertian surface. Then, a natural image $I$ can be decomposed as the pixel-wise product of the illumination invariance, the reflectance $R(I)$; and the illumination variance, the shading $S(I)$, *i.e.*,

$$I = R(I) \odot S(I). \tag{1}$$

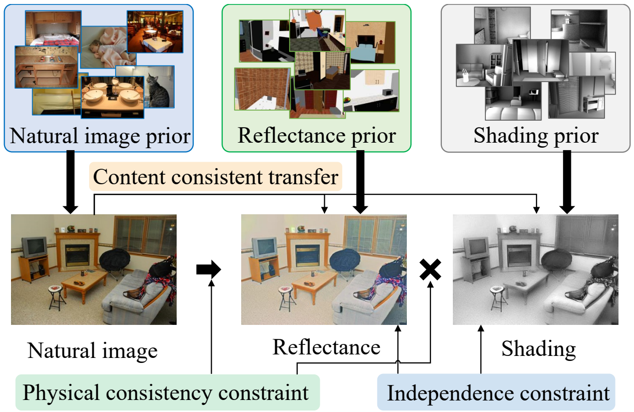

Figure 1. Our method learns intrinsic image decomposition in an unsupervised fashion where the ground truth reflectance and shading is not available in the training data. We learn the distribution priors from unlabeled and uncorrelated collections of natural image, reflectance and shading. Then we perform intrinsic image decomposition through content preserving image translation with independence constraint and physical consistency constraint.

Eq. (1) is ill-posed because there are twice the unknowns than the knowns. Conventional methods [3, 18] therefore explore physical priors as extra constraints, while recent researches tend to use deep neutral networks to directly learn such priors [8, 28, 33, 38].

Unlike high-level vision tasks, intrinsic image decomposition is obviously physics-based, and therefore designing a supervised learning method will heavily rely on high quality physical realistic ground truth. However, existing datasets are either created from a small set of manually painted objects [9], synthetic objects or scenes [4, 6, 19] or manual annotations [15, 31]. These datasets are either too small or far from natural images and therefore limit the performance of supervised learning.

[*]Corresponding Author.

This work is partially supported by the National Natural Science Foundation of China (NSFC) under Grant 61972012 and Grant 61732016, and Baidu academic collaboration program.

A few semi-supervised and unsupervised learning methods have been exploited very recently. Janner *et al*. [14] proposed self-supervised intrinsic image decomposition which relies on few labelled training data and then transfers to other unlabelled data. However, many other supervised information such as shape of the object need be involved. Li *et al*. [20] and Ma *et al*. [26] work on unlabelled image sequences where the scene requires to be fixed within the sequence and only the lighting and shading allow to change. Such settings are still very limited.

In this paper, we aim to explore single image unsupervised intrinsic image decomposition. The key idea is that the natural image, the reflectance and the shading all share the same content, which reflects the nature of the target object in the scene. Therefore, we consider estimating reflectance and shading from a natural image as transferring image style but remaining the image content. Based on such an idea, we can actually use unsupervised learning method to learn the style of natural image, reflectance and shading by collecting three unlabelled and un-correlated samples for each set. Then we apply auto-encoder and generative-adversarial network to transfer the natural image to the desired style while preserve the underlying content. Unlike naı"ve unsupervised style transfer methods which are from one domain to another, our method transfers from one domain to another two domains with explicit physical meanings. We therefore explicitly adopt three physical constraints into our proposed method which are 1) the physical consistent constraint as in Eq. (1), 2) domain invariant content constraint that natural image and its decomposition layers share the same object, layout, and geometry, 3) the physical independent constraint that reflectance is illumination-invariant and shading is illumination-variant.

Rigorous experiments show that our method can produce superior performance against state-of-the-art unsupervised methods on four benchmarks, namely ShapeNet, MPI Sintel benchmark, MIT intrinsic dataset and IIW. Our method also demonstrates comparable performance with state-of-the-art fully supervised methods [8, 33], and even outperforms some of them appeared in the recent years [28, 38].

The contributions of this work are threefold:

- To the best of our knowledge, we propose the first physics based *single image* unsupervised learning for intrinsic image decomposition. Specifically, we adopt three physical constraints: the physical consistency constraint, the domain invariant content constraint and the reflectance-shading independence.
- We propose and implement a completely unsupervised learning network architecture for single image intrinsic decomposition.
- The proposed method outperforms existing unsupervised methods and shows comparable results against fully supervised methods on different intrinsic image benchmarks.

## 2. Related Works

**Optimization based methods.** Intrinsic image decomposition is a typical image layer separation problem [16, 18, 22, 23] which has been studied for nearly fifty years. To handle the ill-posed problem, additional priors with an optimization framework have been applied. For instance, Land *et al*. [16] propose a seminal Retinex algorithm which assumes large image gradients corresponding to changes in reflectance, while smaller gradients are with shading. Subsequently, many priors for intrinsic image decomposition have been explored. Inspired by the large image gradients and piece-wise constant property, reflectance sparsity [30, 32] and low-rank reflectance [1] are taken as regularization term in object functions. There are also many constraints for shading such as the distribution difference in gradient domain [3, 18]. Recently, Chen *et al*. [7] use near-infrared image and propose near infrared prior to regularize the intrinsic image decomposition. Although these handcrafted priors are reasonable in small image set, they are not likely to cover complex scenes [15, 31]. Furthermore, the above mentioned methods assume Lambertian material and thus cannot be adopted to more complex situations with general non-Lambertian reflectance [24, 25].

**Supervised learning methods.** Many supervised learning based methods have been proposed in recent years, [8, 19, 28, 33, 36, 38]. These methods try to estimate the reflectance and shading on labelled training data with different network architectures. However, the training data from publicly available datasets have obvious shortcomings: Sintel [4], ShapeNet [6] and CGIntrinsics [19] are highly synthetic datasets, and networks trained on them cannot generalize well to real-world scenes. The MIT intrinsic dataset [9] consists of real images, but the number of these images are too limited since it contains just 20 objects with ground truth. As a result, these supervised methods often cannot generalize well on one dataset if they are trained on another dataset. Recently, human-labelled dataset IIW [31] and SAW [15] only contain sparse annotations. Not only that, it is difficult to collect such annotations at scale.

**Semi-supervised and unsupervised learning methods.** Involving other supervised information such as shape, illumination source of the object, Janner *et al*. [14] propose self-supervised intrinsic image decomposition, which relied on few labeled training data and then transferred to other unlabeled data. InverseRenderNet [37] uses a series of related images as input and adds multi-view stereo as supervised signal to intrinsic decomposition. Leveraging videos or image sequences, together with physical constraints, learning without intrinsic image has recently become an emerging topic of research. The most of existing unsupervised intrinsic image decomposition methods [20, 26] are mainly focus on training on images with fixed scene and varied illumination, then the trained model can be tested with single input. However, the scene-level real images for training are still limited to various scenes in most cases. Motivated by this, we propose an alternative unsupervised method, which doesn't rely on fixed structure image sequences to train.

**Image-to-image translation.** Image-to-image translation aims to learn the mapping from two image domains. Pix2pix [13] employs conditional GAN to learn the mapping, CycleGAN [39], UNIT [21] applies the cycle consistency to regularize the training. More recently, MUNIT [12] and DRIT [17] assume *partially shared latent space assumption* and make multi-modal image-to-image translation. However, there is still a great gap between unsupervised image-to-image translation between and intrinsic image decomposition because 1) image-to-image are fully statistics driven whereas intrinsic image decomposition is physic-based, 2) the translated image can be various of modalities while the intrinsic images of an input image are explicit. Thus, the image-to-image translation method is not directly adaptable to intrinsic image decomposition.

# 3. Unsupervised Single Input Intrinsic Image Decomposition

## 3.1. Problem formulation and assumptions

**Single input intrinsic image decomposition.** To begin with, we formulate the task with precise denotations. As illustrated in Fig. 1 and Eq. (1), the goal of single image intrinsic decomposition is to decompose a natural image, denoted as $I$, into two layers, illumination-invariance, namely the reflectance $R(I)$; and illumination-variance, namely the shading $S(I)$.

Eq. (1) has more 'unknowns' than 'knowns' and therefore is not directly solvable. Providing sufficient amount of data-samples with ground truths, *i.e*., the triplet samples $\{(I_i, R(I_i)^{GT}, S(I_i)^{GT})\}$, supervised learning based methods have also been explored[8, 19, 28, 33]. In previous sections, we have discussed the difficulty in obtaining the ground truth. We now focus on unsupervised learning.

**Unsupervised Intrinsic Image Decomposition.** In this section, we define the Unsupervised Single Image Intrinsic Image Decomposition (USI$^3$D) problem. Assuming we collect unlabelled and unrelated samples, we learn the appearance style of each collection. Say, we can learn the style of reflectance, the marginal distribution $p(R_j)$, by providing a set of unlabelled reflectance images: $\{R_j \in \mathcal{R}\}$; we learn the shading style, the marginal distribution $p(S_k)$, by providing a set of unlabelled shading images: $\{S_k \in \mathcal{S}\}$; and we learn the natural image style, marginal distribution $p(I_i)$, by providing a set of unlabelled natural images: $\{I_i \in \mathcal{I}\}$. Then, we aim to infer $R(I_i)$, $S(I_i)$ of $I_i$ from the marginal distributions.

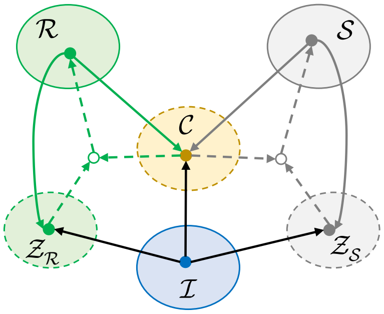

Figure 2. Content preserving translation among domains. $\mathcal{I}$ is the domain of natural image. $\mathcal{S}$ is the domain of shading and $\mathcal{R}$ is the domain of reflectance. For our unsupervised learning method, we learn a set of encoders which encode the appearance from each domain to domain-invariance latent space $\mathcal{C}$. We also learn the encoders to encode the appearance to domain-depended prior space for reflectance ($\mathcal{Z}_\mathcal{R}$) and shading ($\mathcal{Z}_\mathcal{S}$) correspondingly. Later the image style can be transferred from encoders (solid arrows) to generators (dash arrows).

To make the task tractable, we make the following three assumptions.

**Assumption-1. Domain invariant content.**
Physically, the natural appearance, the reflectance and the shading are all the appearance of a given object. As illustrated in Fig. 2, we assume such object property can be latent coded and shared amount domains. Following a style transfer terminology, we call such shared property as content, denoted as $c \in \mathcal{C}$. Also, we assume the content can be encoded from all three domains.

**Assumption-2. Reflectance-shading independence.**
Physically, reflectance is the invariance against lighting and orientation while shading is the variance. And therefore, to decompose these two components, we assume their conditional priors are independent and can be learned separately. As illustrated in Fig. 2, we denote the latent prior for reflectance as $z_R \in \mathcal{Z}_\mathcal{R}$ which can be encoded from both the reflectance domain and the natural image domains. Similarly, we define the latent prior for shading, $z_S \in \mathcal{Z}_\mathcal{S}$.

**Assumption-3. The latent code encoders are reversible.**
This assumption is widely used in image-to-image translation [12, 17]. In detail, it assumes an image can be encoded into the latent code, which can be decoded to image at the same time. This allows us to transfer style and contents among domains. Particularly, this allow us to transfer natural image to reflectance and shading.

## 3.2. Implementation

The detailed implementation of USI$^3$D network is illustrated in Fig. 3.

**The Content-sharing architecture.** As with Assumption-1, we design our content-sharing architecture. We use en-

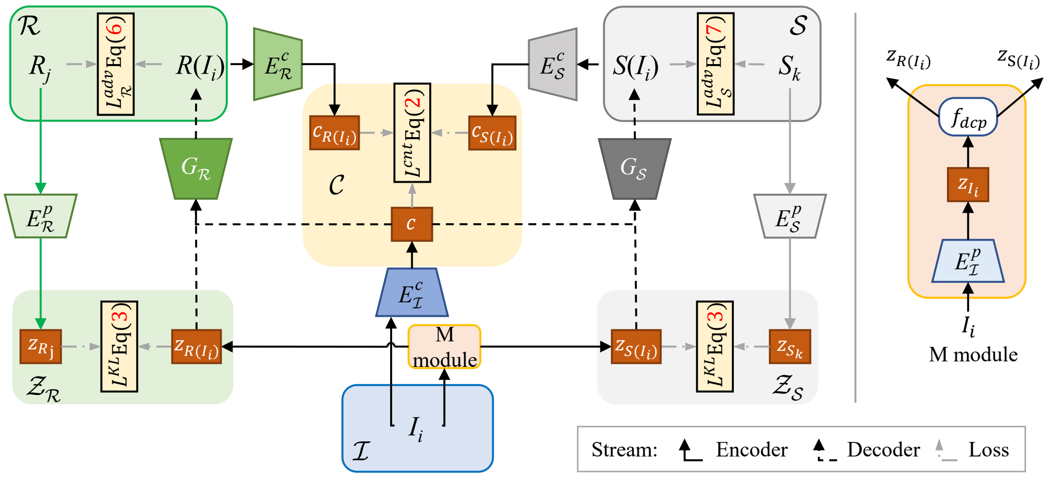

Figure 3. The proposed architecture of USI$^3$D. Our method decompose intrinsic images with unsupervised learning manner, which translate images from natural image domain $\mathcal{I}$ to reflectance domain $\mathcal{R}$ and shading domain $\mathcal{S}$.

coder $E^c_{\mathcal{I}}$ to extract the content code $c$ of the input image $I_i$, then $c$ is used to generate the decomposition layers $R(I_i)$ and $S(I_i)$ through generators $G_{\mathcal{R}}$ and $G_{\mathcal{S}}$, respectively. Next, we extract the content code $c_{R(I_i)}$ of $R(I_i)$ and the content code $c_{S(I_i)}$ of $S(I_i)$ by using $E^c_{\mathcal{R}}$, $E^c_{\mathcal{S}}$, respectively. Finally, we apply *content consistent loss* to make content encoders $E^c_{\mathcal{I}}$, $E^c_{\mathcal{R}}$ and $E^c_{\mathcal{S}}$ work correctly. In detail, we use the content consistency $\mathcal{L}^{cnt}$ to constrain the content code among input image $I_i$ and its predictions $R(I_i)$ and $S(I_i)$.

$$\mathcal{L}^{cnt} = |c_{R(I_i)} - c|_1 + |c_{S(I_i)} - c|_1, \quad (2)$$

where $|\,\boldsymbol{.}\,|_1$ is *L1* distance.

**Mapping module (M module).** Following the Assumption-2, the prior codes of reflectance and shading are domain-variant and independent to each other. Because we need to infer the prior code $z_{R(I_i)}$ and $z_{S(I_i)}$ from $I_i$, we design the Mapping module (M module), as shown in Fig. 3. Specifically, we first extract the natural image prior code $z_{I_i}$, then we designed a decomposition mapping $f_{dcp}$ to infer the prior code $z_{R(I_i)}$ and $z_{S(I_i)}$. To constrain the $z_{R(I_i)}$ in the reflectance prior domain $\mathcal{Z}_{\mathcal{R}}$, we use Kllback-Leibler Divergence (KLD) and other real prior $z_{R_j}$ which is sampled from $\mathcal{Z}_{\mathcal{R}}$. $Z_{S(I_i)}$ is generated and constrained in the similar way. The definition *KLD loss* is

$$\mathcal{L}^{KL} = \mathbb{E}[\log p(\hat{z}) - \log q(z)], \quad (3)$$

where the prior code $\hat{z}$ is extracted from *M module* and its real prior code $z$ is extracted from its real image. Here are two prior domains $\mathcal{Z}_{\mathcal{R}}$ and $\mathcal{Z}_{\mathcal{S}}$, so the total KLD loss is $\mathcal{L}^{KL}_t = \mathbb{E}[\log p(z_{R(I_i)}) - \log q(z_{R_j})] + \mathbb{E}[\log p(z_{S(I_i)}) - \log q(z_{S_k})]$.

**Auto-encoders.** Per Assumption-3, we implement three auto-encoders. Left of Fig. 4 shows the detail of implementing the auto-encoder for natural image stream. The auto-encoder for reflectance and shading are implemented in a similar way where the detail are provided in the supplementary material. We follow the recent image-to-image translation methods [12, 17] and use *bidirectional reconstruction constraints* which enables reconstruction in both image → latent code → image and latent code → image → latent code directions. In detail:

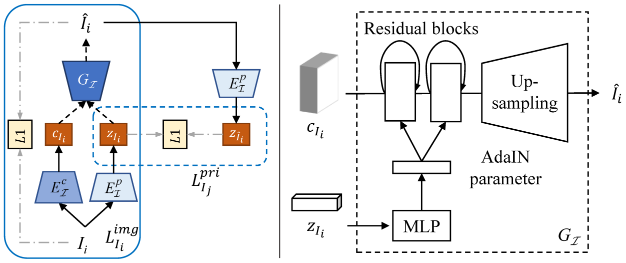

Figure 4. Left: Auto-encoder for natural image stream, the reflectance and shading streams are designed similarly. Right: Architecture of the generator.

*Image reconstruction loss.* Given an image sampled from the data distribution, we can reconstruct it after encoding and decoding.

$$\mathcal{L}^{img} = \sum_{x \in \{I_i\} or \{R_j\} or \{S_k\}} |G(E^c(x), E^p(x)) - x|_1. \quad (4)$$

*Prior code reconstruction loss.* Given a prior code sampled from the latent distribution at decomposition time, we should be able to reconstruct it after decoding and encoding. Different from Eq.(3), which is suitable to constrain the distributions of two samples, the constraint of the prior codes of the image and the reconstructed image should be

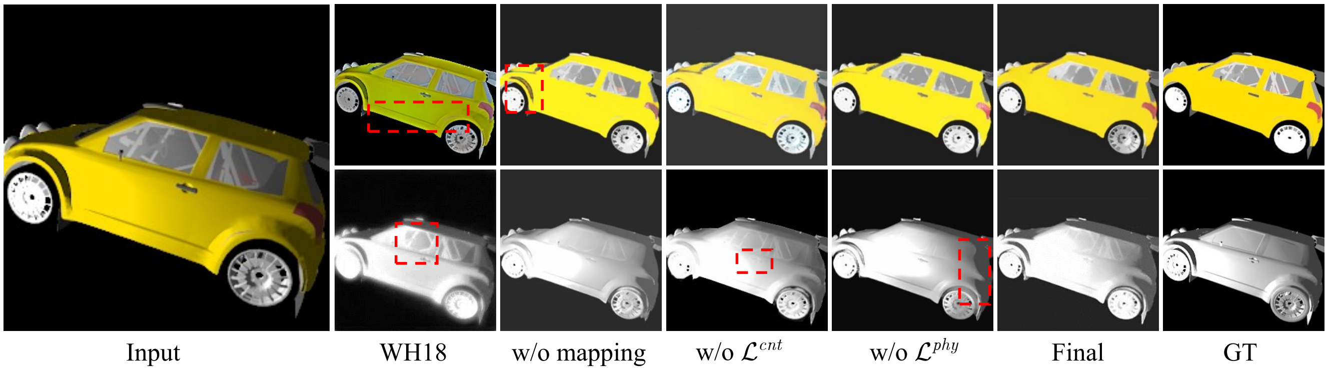

Figure 5. Visual comparison with state-of-the-art WH18 [26] and ablation study on ShapeNet dataset.

identical. Here we use $L_1$ for $\mathcal{L}^{pri}$.

$$\begin{aligned}\mathcal{L}^{pri} =& |E^p_{\mathcal{I}}(G_{\mathcal{I}}(c_{I_i}, z_{I_i})) - z_{I_i}|_1 \\ &+|E^p_{\mathcal{R}}(G_{\mathcal{R}}(c_{R_j}, z_{R_j})) - z_{R_j}|_1 \\ &+|E^p_{\mathcal{S}}(G_{\mathcal{S}}(c_{S_k}, z_{S_k})) - z_{S_k}|_1.\end{aligned} \tag{5}$$

To make the decomposed intrinsic image be indistinguishable from real image in the target domain, we use GANs to match the distribution of generated images to the target data distribution. The adversarial losses are defined as follow:

$$\mathcal{L}^{adv}_{\mathcal{R}} = \log(1 - D_{\mathcal{R}}(R(I_i)) + \log D_{\mathcal{R}}(R_j), \tag{6}$$

$$\mathcal{L}^{adv}_{\mathcal{S}} = \log(1 - D_{\mathcal{S}}(S(I_i))) + \log D_{\mathcal{S}}(S_k). \tag{7}$$

The total adversarial loss is $\mathcal{L}^{adv}_t = \mathcal{L}^{adv}_{\mathcal{R}} + \mathcal{L}^{adv}_{\mathcal{S}}$.

We notice that Eq. (1) means that image $I_i$ is equal to the pixel-product of its analogous $R(I_i)$ and $S(I_i)$, thus this *physical loss* can be employed to regularize our method.

$$\mathcal{L}^{phy} = |I_i - R(I_i) \odot S(I_i)|_1. \tag{8}$$

**Total loss.** By using the GAN scheme, we jointly train the encoders $E$, decoders $G$, mapping function $f$ and discriminators $D$ to optimize the weighted sum of the different loss terms.

$$\begin{aligned}\min_{E,G,f}\max_{D}(E,G,f,D) =& \mathcal{L}^{adv}_t + \lambda_1\mathcal{L}^{cnt} + \lambda_2\mathcal{L}^{KL} \\ &+ \lambda_3\mathcal{L}^{img} + \lambda_4\mathcal{L}^{pri} + \lambda_5\mathcal{L}^{phy},\end{aligned} \tag{9}$$

where $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$, and $\lambda_5$ are weights that are control the importance of different loss terms.

# 4. Experiments

## 4.1. Implementation details

**Encoder and generators.** The distribution prior encoder $E^p$ is constructed by several strided convolutional layers to downsample the input image, then followed by a global average pooling layer and a dense layer. Our content encoder $E^c$ includes several strided convolutional layers and several residual blocks [10] to downsample the input. All convolutional layers are followed by Instance Normalization [34]. The detailed architecture of encoders are provided in the supplementary materials.

We implement the distribution prior mapping function $f_{dcp}$ via multi-layer perceptron (MLP). More specifically, the $f_{dcp}$ takes $z_{I_i}$ as input, and output the concatenation of $z_{R(I_i)}$ and $z_{S(I_i)}$.

The generator reconstructs the image from its content feature and distribution prior. As illustrated in the right of Fig. 4, it processes the content feature by several unsampling and convolutional layers. Inspired by recent unsupervised image-to-image translation works that use affine transformation parameters in normalization layers to represent image styles [11, 12], hence we use image style to represent intrinsic image distribution priors. To this end, we equip the Adaptive Instance Normalization (AdaIN) [11] after each of convolutional layers in residual blocks. The parameters of AdaIN are dynamically generated by a MLP from the distribution prior.

$$\text{AdaIN}(m, \gamma, \beta) = \gamma(\frac{m - \mu(m)}{\sigma(m)}) + \beta, \tag{10}$$

where $m$ is the activation of the previous convolutional layer, $\mu$ and $\sigma$ are channel wise mean and standard deviation. $\gamma$ and $\beta$ are parameters generated by the MLP.

**Discriminator.** We use multi-scale discriminators [35] to guide the generators to generate high quality images in different scales including correct global structure and realistic details. We employ the LSGAN [27] as the objective.

**Weights for loss terms.** In the total objective function Eq. (9), We follow [12] and set $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$ as 10.0, 0.1, 10.0 and 0.1, respectively. Based on the start value and convergence of $\mathcal{L}^{phy}$, we set $\lambda_5$ as 5.0 empirically.

## 4.2. Experimental setup

We have compared three types of intrinsic decomposition algorithms in the experiments. 1) *Unsupervised algo-*

*rithms.* Retinex [16], Barron *et al*. [2], LM14 [18], and $L_1$ flattening that tackle image intrinsic decomposition though optimization without deep learning. LS18 [20], WH18 [26] are trained on image sequences in which each sequence include fixed reflectance and variant shadings. 2) *Image-to-image translation method.* MUNIT [12] is one of the state-of-the-art image-to-image translation methods. Because the framework of MUNIT can only translate images between two domains, so we train this method twice for input ↔ reflectance and input ↔ shading for generating the reflectance and shading of the input image. 3) *Supervised methods.* We have also provided results of state-of-the-art supervised methods like Zhou *et al*. [38], MSCR [28] and FY18 [8] as references.

### 4.3. Qualitative and quantitative results

We evaluate the proposed USI$^3$D on four public intrinsic image benchmarks which are commonly used in image intrinsic decomposition.

Table 1. Numerical comparison and ablation study on ShapeNet intrinsic dataset.

| | MSE | | | LMSE |
|---|---|---|---|---|
| Method | Reflectance | Shading | Avg. | Total |
| LM14 [18] | 0.0338 | 0.0296 | 0.0317 | 0.0623 |
| FY18 [8] | 0.0302 | 0.0315 | 0.0309 | 0.0717 |
| WH18 [26] | 0.0284 | 0.0262 | 0.0273 | 0.0544 |
| Ours w/o mapping | 0.0211 | 0.0130 | 0.0171 | 0.0509 |
| Ours w/o $\mathcal{L}^{cnt}$ | 0.0239 | 0.0167 | 0.0203 | 0.0529 |
| Ours w/o $\mathcal{L}^{phy}$ | 0.0196 | 0.0151 | 0.0174 | 0.0531 |
| Ours final | **0.0185** | **0.0108** | **0.0147** | **0.0465** |

**ShapeNet intrinsic dataset.** We first make a comparison among exist methods and then perform an ablation study on this dataset. We use 8979 input images, 8978 reflectance images and 8978 shadings to train USI$^3$D , each set is shuffled before training. We use the other 2245 images for evaluation.

We make a comparison with one of the optimization-based methods LM14 [18], one of the state-of-the-art supervised methods FY18 [8] and one of the state-of-the-art unsupervised method WH18 [26]. We follow settings in [8, 28] and employ mean square error (MSE), Local MSE (LMSE) [28]. The numerical results are listed in Table 4.3. The visual results are shown in Fig. 5.

We employ three ablation studies to evaluate the USI$^3$D. 1) The effectiveness of mapping $f_{dcp}$. $f_{dcp}$ is aiming to map the prior code of input image into reflectance prior code and shading prior code. In this ablation study, we remove it and replace the decomposed prior code with a random vector. 2) The effectiveness of loss term $\mathcal{L}^{cnt}$. $\mathcal{L}^{cnt}$ is aiming to constrain the predicted reflectance and shading share the same content code with the input. We set this loss term's weight to 0 in this ablation study. 3) The effectiveness of loss term $\mathcal{L}^{phy}$. $\mathcal{L}^{phy}$ is used to enforce the predicted reflectance, shading and the input have relationship as described in Eq. (1). We set this loss term's weight to 0 in this ablation study. As shown in Fig. 5, The third column shows that without mapping, reflectance and shading aren't decomposed properly in some regions. The fourth column shows that without $\mathcal{L}^{cnt}$, some details are lost in shading. The fifth column shows without $\mathcal{L}^{phy}$, although shading looks fine while some area are incorrect.

To further explore how different volume of training data affect the final performance of our method, we reduce the volume of reflectance and shadings in training set and train our algorithm under the same condition. The results are illustrated in Fig. 6. As shown in Fig. 6, our method can still out-perform existing methods by using as few as 20% of training sample.

**MPI-Sintel benchmark.** MPI-Sintel benchmark [4] is a synthesized dataset, which includes 890 images from 18 scenes with 50 frames each (except for one that contains 40 images). We follow FY18[8] and make data argumentation, after that, 8900 patches of images are generated. Then, we adopt two-fold cross validation to obtain all 890 test results. In the training set, we randomly select half of the input image as input samples, the reflectance and shading with remain file names as our reflectance samples and shading samples, respectively.

As listed in Table 2, we employ MSE, LMSE and dissimilarity structural similarity index measure (DSSIM) as evaluation metrics. We show the best of performance among unsupervised methods and even comparable with supervised method MSCR in many evaluation matrices such as LMSE and DSSIM. In qualitative comparison, we show one of the qualitative results in Fig. 7. The proposed USI$^3$D can produce much more correct reflectance and shading than unsupervised method LS18 [20]. Furthermore, our results still competitive than supervised method MSCR [28] since the results of MSCR with many blurry regions. The supervised method FY18 [8] shows the best visual results because they are trained across many other datasets and involves additional other guidance like boundary, domain filter, *etc*.

**MIT intrinsic dataset.** To test performance on real images, we use the 220 images in the MIT intrinsic dataset [32] as in [28]. This data contains only 20 different objects, each of which has 11 images. To compare with previous methods, we finetune our model using 10 objects via the split from [8], and evaluate the results using the remaining objects.

As shown in Table 4.3, our method gets the best performance among unsupervised methods and provides comparable results among supervised methods. We show two visual comparison samples in Fig. 8. Compared to the un-

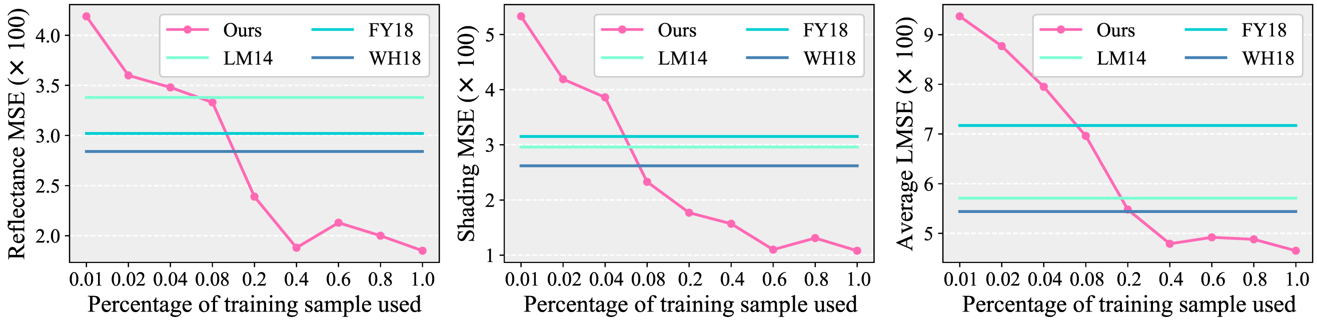

Figure 6. Performance of our unsupervised method on ShapeNet when using fewer training samples.

Table 2. Numerical comparison on MPI Sintel dataset. Sup. denotes that MSCR [28] and FY18 [8] are fully-supervised methods, which are trained with ground truth data, and thus their results can only serve as reference.

| | | MSE | | | LMSE | | | DSSIM | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Method | Reflectance | Shading | Avg. | Reflectance | Shading | Avg. | Reflectance | Shading | Avg. |
| Sup. (Ref.) | MSCR [28] | 0.0100 | 0.0092 | 0.0096 | 0.0083 | 0.0085 | 0.0084 | 0.2014 | 0.1505 | 0.1760 |
| | FY18 [8] | 0.0069 | 0.0059 | 0.0064 | 0.0044 | 0.0042 | 0.0043 | 0.1194 | 0.0822 | 0.1008 |
| Unsup. | Retinex [9] | 0.0606 | 0.0727 | 0.0667 | 0.0366 | 0.0419 | 0.0393 | 0.2270 | 0.2400 | 0.2335 |
| | Barron *et al.* [2] | 0.0420 | 0.0436 | 0.0428 | 0.0298 | 0.0264 | 0.0281 | 0.2100 | 0.2060 | 0.2080 |
| | LS18 [20] | 0.0215 | 0.0251 | 0.0233 | 0.0095 | 0.0118 | 0.0107 | 0.2070 | 0.1827 | 0.1949 |
| | MUNIT [12] | 0.0291 | 0.0232 | 0.0262 | 0.0207 | 0.0182 | 0.0195 | 0.2073 | 0.1893 | 0.1983 |
| | Ours | **0.0159** | **0.0148** | **0.0154** | **0.0087** | **0.0081** | **0.0084** | **0.1797** | **0.1474** | **0.1635** |

Table 3. Numerical comparison on MIT intrinsic dataset. Results of supervised methods are also provided for reference.

| | | MSE | | | LMSE |
|---|---|---|---|---|---|
| | Method | Reflectance | Shading | Avg. | Total |
| Sup. (Ref.) | Zhou *et al.* [38] | 0.0252 | 0.0229 | 0.0240 | 0.0319 |
| | Shi *et al.* [33] | 0.0216 | 0.0135 | 0.0175 | 0.0271 |
| | MSCR [28] | 0.0207 | 0.0124 | 0.0165 | 0.0239 |
| | FY18 [8] | **0.0134** | **0.0089** | **0.0111** | **0.0203** |
| Unsup. | LM14 [18] | 0.0286 | 0.0227 | 0.0255 | 0.0366 |
| | WH18 [26] | 0.0232 | 0.0166 | 0.0197 | 0.0379 |
| | MUNIT [12] | 0.0197 | 0.0170 | 0.0184 | 0.0302 |
| | Ours | **0.0157** | **0.0135** | **0.0146** | **0.0231** |

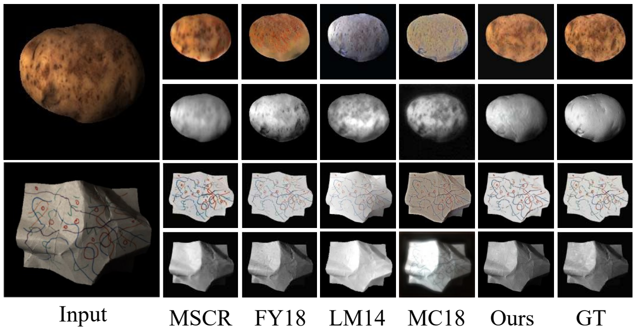

Figure 8. Visual results on MIT intrinsic image benchmark. The column 2 - 3 are supervised methods' results for reference.

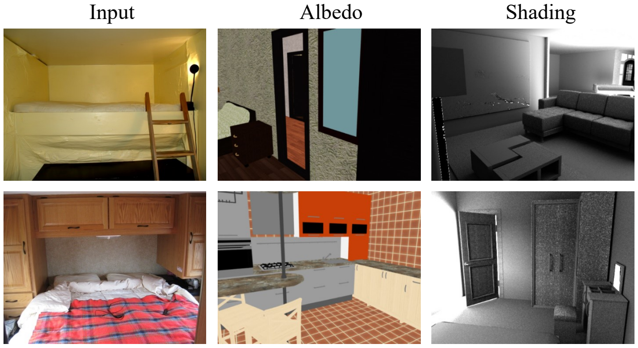

Figure 10. Samples from mixed dataset. The input images is collected from IIW, while the reflectance and shading are collected from CGIntrinsic dataset.

supervised methods, we show the best visual performance both in reflectance and shading.

**IIW benchmark.** The Intrinsic Images in the Wild (IIW) benchmark [31] contains 5,230 real images of mostly indoor scenes, combined with a total of 872,161 human judgments regarding the relative reflectance between pairs of points sparsely selected throughout the images. We split the input images of IIW into a training set (4184) and test set in the same way as [8, 19]. Because the IIW dataset contains no reflectance and shading images, we employ unpaired reflectance and shading from the rendered dataset CGIntrinsics [19]. To make a fair training, we choose the first 4000 consecutive reflectance sorted by the image ID and the last

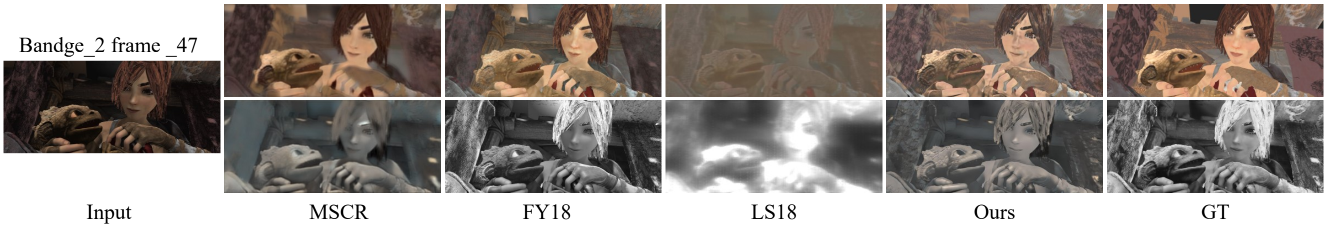

Figure 7. Visual results on MPI Sintel benchmark. Compared with state-of-the-art unsupervised method LS18 [20]. Supervised methods MSCR [28] and FY18 [8] are provided for reference.

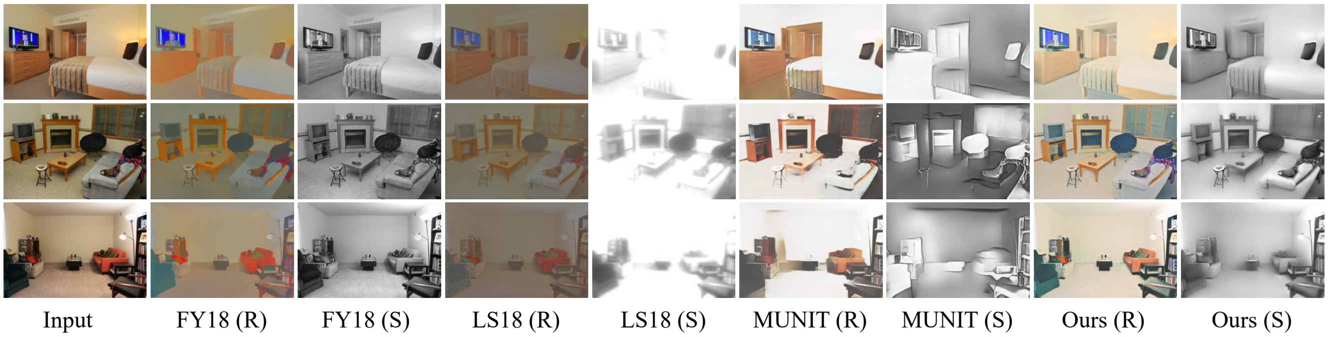

Figure 9. Qualitative comparison on the IIW test sets. FY18 [8] is supervised method. LS18 [20] and MUNIT [12] are unsupervised.

4000 shading images for training. Samples of the training set are illustrated in Fig. 10.

Table 4. Numerical comparison on IIW benchmark dataset using weighted human disagreement rate (WHDR) [29]. Lower is better.

| | Method | WHDR%(mean) |
|---|---|---|
| | Baseline (const reflectance) | 36.54 |
| | Baseline (const shading) | 51.37 |
| Sup. (Ref.) | Narihira *et al.* [29] | 18.10 |
| | Zhou *et al.* [38] | 15.70 |
| | CGIntrinsic [19] | 14.80 |
| | FY18 [8] | **14.45** |
| Unsup. | Retinex (color) [9] | 26.89 |
| | Retinex (gray) [9] | 26.84 |
| | WH18 [26] | 28.04 |
| | MUNIT [12] | 25.23 |
| | $L_1$ flattening [3] | 20.94 |
| | Bell *et al.* [31] | 20.64 |
| | LS18 [20] | 20.30 |
| | Ours | **18.69** |

For domain adjustment, we follow CGIntrinsics and apply reflectance smoothness $\mathcal{L}_R^{smooth}$ term to encourage reflectance predictions to be piece-wise constant.

$$\mathcal{L}_R^{smooth} = \sum_{i=1}^{N} \sum_{j \in \mathcal{N}(i)} v_{i,j} |\log x_R^i - \log x_R^j|_1, \tag{11}$$

where $\mathcal{N}(i)$ denotes the 8-connected neighbourhood of the pixel at position $i$. The reflectance weight $v_{i,j} = \exp(-\frac{1}{2}(\mathbf{f}_i - \mathbf{f}_j)^T \Sigma^{-1} (\mathbf{f}_i - \mathbf{f}_j))$, and the feature vector $\mathbf{f}_i$ is defined as $[\mathbf{p}_i, \mathbf{x}_i, r_i^1, r_i^2]$, where $\mathbf{p}_i$ and $\mathbf{x}_i$ are the spatial position and image intensity respectively, and $r_i^1$ and $r_i^2$ are the first two elements of chromaticity. $\Sigma$ is a covariance matrix which defines the distance between two feature vectors. We set this loss term with weight of 1.0 in this experiment.

We use the same test split with compared method. Because there is no pixel-wise ground truth, we use the weighted human disagreement rate (WHDR) which is introduced in the dataset [31]. The numerical results are listed in Table 4.3, lower is better. The qualitative results are illustrated in Fig. 9. Our proposed method generates better results than the existing methods.

## 5. Conclusion

In this paper, we propose an unsupervised single image intrinsic decomposition (USI$^3$D) method. We introduce three assumptions about distributions of different image domains, *i.e.*, domain invariant content, reflectance-shading independence and the latent code encoders are reversible. We implemented our pipeline based on these assumptions. Experimental results on four intrinsic image benchmarks show that our method outperforms existing state-of-the-art unsupervised methods and even some state-of-the-art supervised methods.

## References

[1] Bousseau Adrien, Paris Sylvain, and Durand Frédo. User-assisted intrinsic images. *TOG*, 2009. 2

[2] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *TPAMI*, 2015. 6, 7

[3] Sai Bi, Xiaoguang Han, and Yizhou Yu. An l1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. *TOG*, 2015. 1, 2, 8

[4] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012. 1, 2, 6

[5] Chong Cao, Feng Lu, Chen Li, Stephen Lin, and Xukun Shen. Makeup removal via bidirectional tunable de-makeup network. *IEEE TMM*, 2019. 1

[6] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, and Hao Su. Shapenet: An information-rich 3d model repository. In *arXiv preprint arXiv:1512.03012*, 2015. 1, 2

[7] Ziang Cheng, Yinqiang Zheng, Shaodi You, and Imari Sato. Non-local intrinsic decomposition with near-infrared priors. In *ICCV*, October 2019. 2

[8] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. Revisiting deep intrinsic image decompositions. In *CVPR*, 2018. 1, 2, 3, 6, 7, 8

[9] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, 2011. 1, 2, 7, 8

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[11] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 5

[12] Xun Huang, Ming-Yu Liu, Serge Belongie, and Kautz Jan. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 3, 4, 5, 6, 7, 8

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 3

[14] Michael Janner, Jiajun Wu, Tejas Kulkarni, D., Ilker Yildirim, and Joshua Tenenbaum, B. Self-supervised intrinsic image decomposition. In *NIPS*, 2017. 2

[15] Balazs Kovacs, Sean Bell, Noah Snavely, and Kavita Bala. Shading annotations in the wild. In *CVPR*, 2017. 1, 2

[16] Edwin H Land and John J McCann. Lightness and retinex theory. *Josa*, 2008. 2, 6

[17] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 3, 4

[18] Yu Li and Michael S. Brown. Single image layer separation using relative smoothness. In *CVPR*, 2014. 1, 2, 6, 7

[19] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *ECCV*, 2018. 1, 2, 3, 7, 8

[20] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *CVPR*, 2018. 2, 3, 6, 7, 8

[21] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017. 3

[22] Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. Semantic guided single image reflection removal. In *arXiv preprint arXiv:1907.11912*, 2019. 2

[23] Yunfei Liu and Feng Lu. Separate in latent space: Unsupervised single image layer separation. In *AAAI*, 2020. 2

[24] Feng Lu, Xiaowu Chen, Imari Sato, and Yoichi Sato. Symps: Brdf symmetry guided photometric stereo for shape and light source estimation. *IEEE T-PAMI*, 2017. 2

[25] Feng Lu, Yasuyuki Matsushita, Imari Sato, Takahiro Okabe, and Yoichi Sato. From intensity profile to surface normal: photometric stereo for unknown light sources and isotropic reflectances. *IEEE T-PAMI*, 2015. 2

[26] Wei-Chiu Ma, Hang Chu, Bolei Zhou, Raquel Urtasun, and Antonio Torralba. Single image intrinsic decomposition without a single intrinsic image. In *ECCV*, 2018. 2, 3, 5, 6, 7, 8

[27] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, Oct 2017. 5

[28] Takuya Narihira, Michael Maire, and Stella X Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *ICCV*, 2015. 1, 2, 3, 6, 7, 8

[29] Takuya Narihira, Michael Maire, and Stella X Yu. Learning lightness from human judgement on relative reflectance. In *CVPR*, 2015. 8

[30] Carsten Rother, Martin Kiefel, Lumin Zhang, Bernhard Schölkopf, and Peter V Gehler. Recovering intrinsic images with a global sparsity prior on reflectance. In *NIPS*, 2011. 2

[31] Bell Sean, Bala Kavita, and Snavely Noah. Intrinsic images in the wild. In *SIGGRAPH*, 2014. 1, 2, 7, 8

[32] Li Shen and Chuohao Yeo. Intrinsic images decomposition using a local and global sparse representation of reflectance. In *CVPR*, 2011. 2, 6

[33] Jian Shi, Yue Dong, Hao Su, and Stella X Yu. Learning non-lambertian object intrinsics across shapenet categories. In *CVPR*, 2017. 1, 2, 3, 7

[34] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. In *arXiv preprint arXiv:1607.08022*, 2016. 5

[35] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Highresolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 5

[36] Zhongji Wang and Feng Lu. Single image intrinsic decomposition with discriminative feature encoding. In *ICCVW*, 2019. 2

[37] Ye Yu and William AP Smith. Inverserendernet: Learning single image inverse rendering. In *CVPR*, 2019. 2

[38] Tinghui Zhou, Philipp Krahenbuhl, and Alexei A. Efros. Learning data driven reflectance priors for intrinsic image decomposition. In *ICCV*, 2015. 1, 2, 6, 7, 8

[39] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 3