Separate from Observation: Unsupervised Single Image Layer Separation

Yunfei Liu, Feng Lu∗
State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China
{lyunfei, lufeng}@buaa.edu.cn

Abstract

Unsupervised single image layer separation aims at extracting two layers from an input image where these layers follow different distributions. This problem arises most notably in reflection inference removal and intrinsic image decomposition. Since there exist an infinite set of combinations that can construct the given input image, one could infer nothing about the solutions without additional assumptions. To address the problem, we make the shared information consistency assumption and separated layer independence assumption to constrain the solutions. In this end, we propose an unsupervised single image separation framework based on cycle GANs and self supervised learning. The proposed framework is applied for the reflection removal and intrinsic image problems. Numerical and visual results show that the proposed method achieves the state-of-the-art performance among unsupervised methods which require single image as input. Based on the slightly modified version of the presented framework, we also demonstrate the promising results of decomposing an image into three layer.

1 Introduction

Images of indoor/outdoor scenes are usually mixed by different meaningful information. Scenes contaminated with reflection, objects with different shadings are such phenomena, which due to illumination affected through different medium or materials in the environment. Specifically, the irradiance received by the camera from the scene point is blended with different cues along with the line of sight. Different factors make the image look real, but they make it more difficult for computer to understand the image [1,2,3]. Even worse, some factors make images degraded, decrease the visibility of scenes [4,5].

The separation of images into multiple layers is desired in both computational photography and various vision tasks like surface re-texturing, 3D object compositing [1], 3D point cloud processing [6]. In this regard, the single image separation intents to extract two independent layers from an image, in which the input image \( I \) can be constructed as the pixel-wise addition of an image \( L_1 \) and another image \( L_2 \), i.e.,

\[
I = L_1 + L_2.
\] (1)

Many traditional image separation problems can be formulated as Eqn. [1]. For instance, reflection interference often arises when a photo of a scene is taken behind a glass window. This is a typical image separation problem and can be expressed as a linear combination of a reflection layer \( L_R \) and the background scene \( L_B \), as \( I = L_R + L_B \). The intrinsic image model assumes an input image \( I \) is

∗ Corresponding Author
the pixel-wise product of an albedo (or reflectance) image $L_R$ and a shading image $L_S$. This can be reformulated into the form in Eqn. 1 by taking the log, i.e., $\log(I) = \log(L_R) + \log(L_S)$.

While obviously useful, estimating such layers is fundamentally ill-posed as there exist infinitely many feasible solutions to Eqn. 1. To constrain the space of feasible solutions, lots of prior information is instantiated through carefully tailored image filters or energy terms \[4, 7\]. For example, Li et al. \[4\] assume that one output layer is more smooth than the other layer. Based on this assumption, they proposed relative smoothness prior to separate image into two. However, when the scene goes complicated, such hand-crafted prior is no longer enough to describe the difference between these two output layers. On the other hand, given access to ground truth aligned dataset, deep convolutional neural networks (CNNs) provide a data-driven candidate for solving the ill-posed inverse problem with fewer potentially heuristic or hand-crafted assumptions. However, existing databases are limited to single image separation problem in various aspects: (1) It’s hard to transfer information among different datasets for the same task. Existing synthetic datasets vary from each other because there are different application scenarios \[8, 9, 10, 11\]. (2) Ground truth data of real images is extremely hard to acquire for training a general CNNs model \[11\]. Consequently, each existing dataset is limited in different ways, and thus far, supervised deep network models built using them likewise display a high degree of dataset-tailed architectural variance.

To this end, we proposed an USIS (Unpaired Single Image Separation) method which takes three image domains which without ground-truth data for training. Based on the difference of distributions in different image sets and cycle consistency, the USIS learns the relationship among the different domains via generative adversarial manner. The USIS separates the single input image into two meaningful components which are independent to each other, after learning the feature statics distribution from the given output image sets.

Experimental results show that our proposed framework could separate the input image into two desired images which following the distribution of output image distributes properly. The proposed architecture can be adopt to single image reflection removal, intrinsic image decomposition tasks without accessing ground-truth. We also extend the proposed method with slightly modification on a more challenging, single image three layer separation task. Results demonstrate that the proposed USIS can handle such problem properly.

## 2 Previous Work

Unsupervised domain translation methods receive two sets of samples and learn a function that maps between a sample of one domain and the analogous sample of the other domain \[12, 13, 14, 15\]. However, due to the relationships among three sets of images fundamentally based on the physical model as shown in Sec. 1 previous unpaired image to image methods cannot adopt into image separate task directly.

There are various problems on image separation in computer vision area, different prior and physical models are applied in different sub-problems. For instance, relative smoothness \[4\], ghost cues \[7\] and layer independence prior are introduced for reflection removal problem. Although such hypothesis work in many cases, but these are all low-level prior, which is constructed with image gradient or color changing, such priors are not adaptable with complex scenes.

Recently, many deep learning based methods are carefully tailored for different datasets which contain ground-truth. In single image reflection removal task, full convolution neural networks with different guidance (like image gradient information \[8\], the face structure prior \[9\]) branches or losses (like perceptual losses \[10\]) designed. For intrinsic image decomposition, various U-net \[16\] like, encoder-decoder with skip connections are proposed to tackle the decomposition.

Due to it’s hard to collect real image database with ground truth label, unsupervised learning for single image separation is appealing because of inaccessibility of ground truth. However, the single image separation is even tough when the training image with unknown ground-truth. Michael et al. \[17\] proposed self-supervised intrinsic image decomposition by training on the few images with ground-truth data, then transfer the model to other unpaired images. However, their method require the training images to share the same reflectance layer in the same group. Ma et al. \[18\] and Li et al. \[19\] proposed unsupervised intrinsic image decomposition methods, but these methods need multiple input with same reflectance layer to train the model.
3 Unsupervised Single Image Layer Separation

Figure 1: (a) The shared information consistency assumption and layer independence assumption. Here $E_y, E_z$ are two encoding functions for sample $y$ and $z$, mapping images to latent codes. We assume a set of corresponding images $(x, y, z)$ in three domains $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, all the information in $y$ (or $z$) can be found in $x$ (green arrow). The separated samples $y, z$ are independent to each other in latent space, i.e., the joint probability density between latent distribution $F_y$ and $F_z$ is zero. (b) The proposed USIS framework. $f_y$ and $f_z$ are disentangled codes of $x$ in latent space, $f_y \in F_y$ and $f_z \in F_z$. $G_Y$ and $G_Z$ are two generation functions, mapping latent codes into corresponding images in domain $\mathcal{Y}$ and $\mathcal{Z}$. $\oplus$ is image blend function based on Eqn. 1. We represent $E_y, E_z, G_Y$ and $G_Z$ using CNNs. We implement the shared information consistency assumption using cycle consistency constraint, and implement layer independence assumption via self supervision learning. The $\sim$ denotes adversarial discriminators for the respective domain, in charge of evaluating whether the separated images are realistic or not.

3.1 Problem Formulation

Let $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ be three image domains. In supervised image-to-image translation, we are given samples $(x, y, z)$ drawn from a joint distribution $P_{X,Y,Z}(x, y, z)$. In unsupervised single image separation, we are given samples drawn from the marginal distributions $P_X(x), P_Y(y)$ and $P_Z(z)$. Besides, based on the Eqn. 1 $x = y + z$. However, as explained in Sec. 1 solving the problem is highly ill-posed. We can infer nothing about the joint distribution of the marginal samples without additional assumptions.

**Assumption 1.** Shared information consistency. $x$ is blended by $y$ and $z$, $x$ and $y$ (or $z$) share the same latent space, i.e., $\exists E(x), E(y) \in F_x$ and $\exists E(x), E(z) \in F_z$, where $E$ is a function for mapping image from color space to latent space, $F$ is the corresponding latent space.

Computationally, to implement the information consistency assumption, the original image separation pipeline can be rewritten as:

$$x \rightarrow y, z \Rightarrow x \leftarrow f_y \xrightarrow{G_y} y$$

$$f_z \xrightarrow{G_z} z$$

where $f$ is the feature sample in latent space which is analogous to $x$, $f_y \in F_y, f_z \in F_z$. $G_Y$ and $G_Z$ are both mapping functions which are used to project the feature sample back to color space for reconstructing the analogous separated images $y$ and $z$. 

3
Assumption 2. Layer independence. The separated image \( y \) and \( z \) are independent to each other in latent space, i.e., samples from the same domain share more similar features in latent space than that between different domains.

Layer independence assumption can be implemented by minimizing distance between any two samples from the same domain in latent space, maximizing the distance between any two samples from different domain in latent space. Computationally, for any two samples \( f_y^a, f_y^b \) from latent space \( \mathcal{F}_y \), any other two samples \( f_z^a, f_z^b \) from latent space \( \mathcal{F}_z \), we have:

\[
d_{\phi} \gg d_{\phi}, \forall d_{\phi} \in \{|f_y^a - f_y^b|, |f_y^a - f_z^b|, |f_y^b - f_z^a|, |f_z^a - f_z^b|\}, \forall d_{\phi} \in \{|f_y^a - f_y^b|, |f_z^a - f_z^b|\},
\]

where \( d_{\phi} \) is distance between samples from different latent space, \( d_{\phi} \) is distance between two samples of the same latent space.

3.2 Unsupervised Single Image Separation Learning

Self Supervising (SS). Based on shared information consistence assumption, \( x \) contains all information constructing \( y \) and \( z \). Furthermore, features of \( x \) in latent space still contains all information of features of \( y \) and \( z \). As shown in Fig. 1(b), an ideal encoder \( E_y \) can encode \( x \) from RGB space to the analogous feature code \( f_y \) latent space \( \mathcal{F}_y \), which contains all information of \( y \). Add on layer independence assumption, \( f_y \) contains all information of \( y \) without information of \( z \), i.e., in this way, we can disentangle \( y \) and \( z \) from \( x \) through two encoders \( E_y \) and \( E_z \), into two independent analogous feature of \( y \) and \( z \) in latent space \( \mathcal{F}_y \) and \( \mathcal{F}_z \).

Following this idea, for \( E_y \), we minimize the L1 distance between \( E_y(x_i) \) and \( E_y(y_i) \), where \( x_i \) and \( y_i \) are unpaired samples from \( \mathcal{X} \) and \( \mathcal{Y} \). For \( E_z \), we minimize the L1 distance between \( E_z(x_j) \) and \( E_z(z_j) \), where \( x_j \) and \( z_j \) are unpaired samples from \( \mathcal{X} \) and \( \mathcal{Z} \). To make features of \( \mathcal{Y} \) and \( \mathcal{Z} \) more distinguishable, we also maximize the distance between features yielding from \( E_y \) and \( E_z \).

GANs. Note that the self-supervising constraint and same information consistency assumption do not guarantee that output corresponding images in two domains have different latent code are feasible to look like real image in domain \( \mathcal{Y} \) and \( \mathcal{Z} \). Hence, we adopt generative adversarial framework to make the output \( y \) and \( z \) look as real as samples from the corresponding domain \( \mathcal{Y} \) and \( \mathcal{Z} \) in perceptual manner.

There two parts of sub-networks in generative adversarial networks. The generator aims to solve the problem of single image separation, this part aims to get the mapping function \( S_{x \to y, z} \), the encoder-decoder based mapping technique [20], two encoders and two decoders are learned. We denote the encoder for mapping image color space to latent space, \( E_Y(\cdot), E_Z(\cdot) \) and the decoders of different output layers \( G_Y(\cdot), G_Z(\cdot) \). Specifically, we write \( S_{x \to y, z} = E_Y \circ G_Y, E_Z \circ G_Z \) as the generator of our framework. Whereas we use two discriminators \( D_Y \) and \( D_Z \) to discriminate whether the separated images belong to the domain \( \mathcal{X}, \mathcal{Y}, \mathcal{Z} \) respectively, e.g., for real images from domain \( \mathcal{Y}, D_Y \) should output true, while from the generator \( E_Y \circ G_Y, \) it should output false.

Cycle Consistency (CC). Since the shared information consistency assumption and Eqn. 1 imply the cycle-consistency constraint, we also enforce the cycle-consistency constraint in the proposed framework to further regularizes the ill-posed unsupervised single image separation problem. Specifically, we get \( y' \) and \( z' \) through mapping function \( S \), then \( x' = y' + z' \). \( x' \) should consist to \( x \). Furthermore, \( y'' \) and \( z'' \) are generated by mapping function \( S \) again by given \( x', y'' \) (\( z'' \)) should consist to \( y' \) (\( z' \)).

Learning. We jointly solve the learning problems of self-supervising, GANs and cycle consistency for the image separation streams, the image reconstruction streams and the cycle-reconstruction streams:

\[
\min_{E_Y, E_Z, G_Y, G_Z} \max_{D_Y, D_Z} \mathcal{L}_{SS}(E_Y, E_Z) + \mathcal{L}_{GAN_Y}(E_Y, G_Y, D_Y) + \mathcal{L}_{GAN_Z}(E_Z, G_Z, D_Z) + \mathcal{L}_{CC_X}(E_Y, G_Y, E_Z, G_Z) + \mathcal{L}_{CC_Y}(E_Y, G_Y) + \mathcal{L}_{CC_Z}(E_Z, G_Z).
\]

(4)

Self-supervised training aims to split sample from \( \mathcal{X} \) in the latent space. Specifically, split the sample \( x \) into latent code \( f_x \) for the analogous sample in \( \mathcal{Y} \) and another latent code \( f_z \) for the analogous sample in \( \mathcal{Z} \). Based on Eqn. 3 the self-supervising object is:

\[
\mathcal{L}_{SS}(E_Y, E_Z) = \lambda_1 d_{\phi}(E_Y(y), E_Y(x)) + \lambda_2 d_{\phi}(E_Z(z), E_Z(x)) + \lambda_3 (1 - d_{\phi}(E_Y(x), E_Z(x))).
\]

(5)
where the hyper-parameters $\lambda_1$, $\lambda_2$ and $\lambda_3$ control the weights of different objective terms. We adopt $d_\phi(a, b) = |a - b|$ in practice, i.e., we compute the L1 distance of two inputs $a, b$. $d_\psi$ is the distance between the latent codes from $E_Y$ and $E_Z$, we use a modified sigmoid function as distance function:

$$d_\psi(a, b) = \frac{1}{1 + e^{g(a, b)}}, \quad \text{where} \quad g(a, b) = -\frac{|a - b| - \alpha e^\alpha}{\alpha^2},$$

(6)

where $\alpha$ controls the shape of distance curve, experimental results are provided in the top of Fig. 2 (b).

In Eqn. 4 the GAN objective functions are given by

$$\mathcal{L}_{GAN_y}(E_Y, G_Y, D_Y) = \lambda_0 \mathbb{E}_{y \sim p_y}[\log D_y(y)] + \lambda_0 \mathbb{E}_{f_1 \sim p_y(f_1|x)}[\log(1 - D_y(G_y(f_1)))]$$

$$\mathcal{L}_{GAN_z}(E_Z, G_Z, D_Z) = \lambda_0 \mathbb{E}_{z \sim p_z}[\log D_z(z)] + \lambda_0 \mathbb{E}_{f_2 \sim p_z(f_2|x)}[\log(1 - D_z(G_z(f_2)))]$$

(7)

The objective functions in Eqn. 7 are conditional GAN objective functions. They are used to ensure the separated images resembling images in the target domains, respectively. The hyper-parameter $\lambda_0$ controls the impact of the GAN objective functions.

We use the L1 difference function to model the cycle-consistency constraint, which is given by

$$\mathcal{L}_{CC_x} = \lambda_4 |y' + z' - x|,$$

$$\mathcal{L}_{CC_y} = \lambda_5 |G_Y(E_Y(y' + z')) - y'|, \mathcal{L}_{CC_z} = \lambda_6 |G_Z(E_Z(y' + z')) - z'|,$$

(8)

where $y' = G_Y(E_Y(x))$ and $z' = G_Z(E_Z(x))$. The hyper-parameters $\lambda_4$, $\lambda_5$ and $\lambda_6$ control the weights of these three different objective terms.

Inheriting from GAN, training of the proposed framework results in solving a min-max problem where the optimization aims to find a saddle point. To make a training process stable, we apply gradient update scheme similar to the one described in [21] and gradient penalization to solve Eqn. 4. Specifically, we first apply a gradient descent step to update $E_Y$, $E_Z$, $G_Y$, $G_Z$ with $D_Y$ and $D_Z$ fixed. We then apply a gradient ascent step to update $D_Y$ and $D_Z$ with $E_Y$, $E_Z$, $G_Y$, $G_Z$ fixed.

4 Network Architecture

Our model USIS follows the pipeline of generative adversarial network introduced by [21] which consists of one generator and two discriminators. The generator aims to separate the input sample $x$ into the analogous $y$ and $z$. Discriminators share the same architecture with different parameters, discriminate whether the sample $y$ (or $z$) is real or fake, i.e., discriminate whether the sample belongs to the distribution of real images.

Generator. As illustrated in Fig. 1(b), there are two types of generators in the USIS.

The first generator is called separate generator which is constructed by two convolutional encoder-decoder networks $\{E_Y, G_Y, E_Z, G_Z\}$ with skip connections. These two generators share the same structure but different parameters. According to the difference of experiment task, we designed two types of generators for different tasks:

Toy problem task. See Fig. 2 (a) for for the task description. Both networks employ mirror-link connections introduced by [16], which connect layers of the encoder and decoder of the same size. These connections yield sharper results than the blurred outputs characteristic of many deconvolutional models. The encoder has 5 convolutional layers with $\{16, 32, 64, 128, 256\}$ filters of size $4 \times 4$ and stride of 2. Batch normalization [22] and leaky ReLU activation are applied after every convolutional layer. The layers in these two decoders have the same number of features as the encoder but in reverse order plus a final layer with 3 channels.

Reflection removal & Intrinsic image decomposition task. We adopt pretrained VGG-19 [23]. Selecting ‘conv1\_2’, ‘conv2\_2’, and ‘conv3\_2’ as skip connected features, shown to be successful for image synthesis and enhancement [10]. The first 4 blocks in decoder are cascade convolution layer and upsample operations to fuse features from encoder. The next contextual block is a fully convolutional network with 64 filters of size $3 \times 3$, stride of 1 and dilation rate of $\{2, 4, 8, 16, 32, 1\}$, followed an output layer, which is a convolution layer with 3 filters of size $1 \times 1$. Instance normalization [24] and leaky ReLU activation are applied after every convolutional layer.
The other generator is called the combine generator, with no learnable parameters, is illustrated by $\oplus$. Combine generator aims to combine the predicted analogous $y'$ and $z'$ together based on the Eqn 1.

**Discriminator.** We adopt multi-scale discriminator as UINT [15] to distinguish the real and fake images. For each discriminator network, which is constructed by multi-branches of sub-networks, distinguishes the real and fake images in different scales. For toy problem task, the number of branches is 1, otherwise, we set it to 3. Specifically, each sub-branch has 4 convolutional layers with $[32, 64, 128, 32]$ filters of size 4×4 and stride of 2. Instance normalization [24] and LeakyReLU activation are applied after every convolutional layer. For $i$-th branch, image is down sampled by $i - 1$ times via average pooling operation as input. Finally, the features yielded from different branches are fused together and followed with sigmoid activation.

5 Experiments

We first analyze different components of the proposed framework based on a toy problem. We then present visual and numerical results on real image separation tasks. Finally, we extend our framework to a more challenging single image separation task.

5.1 Performance Analysis

![Figure 2: (a) Toy problem data samples. Top row: input sample $x$ (blended by circle and square), middle row: $y$ (only square), bottom row $z$ (only circle). We separate the blended sample $x$ into $y$ and $z$, and measure the MSE scores achieved by different configurations of the proposed framework. (b) Top: Curve of how different $\alpha$ effect the distance loss $\alpha = 1.4$. Bottom: On evaluation set, separated layers’ MSE versus different internal parameters in self-supervision loss. (c) Impact of self-supervision and cycle consistency constraints on image separation accuracy.](a) (b) (c)

We used ADAM [25] for training where the learning rate was set to 0.0001 and momentums were set to 0.0 and 0.9. Each mini-batch consisted of an image from the domain $\mathcal{X}$, an image from the domain $\mathcal{Y}$ and an image from domain $\mathcal{Z}$. Our framework had several hyper-parameters as shown in Eqn. 5-8. The default values were $\lambda_0 = 5.0$, $\lambda_1 = 0.5$, $\lambda_2 = 0.5$, $\lambda_3 = 1.0$, $\lambda_4 = 1.0$, $\lambda_5 = 1.0$, $\lambda_6 = 1.0$, $\alpha = 1.4$.

We introduce a toy problem (visualized in Fig. 2(a)), in which: domain $\mathcal{Y}$ contains gray images with solution of 128×128, with squares in different {lightness, position, size} in each image. Shapes in domain $\mathcal{Z}$ are circles and each image in domain $\mathcal{X}$ are generated via $x = y + z$. Based on this toy problem, we generate a dataset which contains 5K sets of images in three domains (blended $x$, rectangle $y$ and circle $z$), which is convenient for quantitative evaluation. For the toy problem, the goal is to separate the blended image into the image containing only rectangle and the residual image which contains only circle. Here we use the unsupervised scheme, where we choose 4K images in each domain randomly without grouping the analogs images for training. We train for 200 epochs and use the final model, specifically, test the separate generator on the test set. We then compare the difference between separated images and the corresponding ground truth images via mean square error (MSE). Note that image pixel values are in [0, 255] in our experiment.
The hyperparameters analysis of self supervision loss and ablation study results are illustrated in bottom of Fig.2(b) and Fig.2(c).

5.2 Qualitative and quantitative results

Fig.3, 4 show the result of the proposed framework on two image separation tasks against state-of-the-art unsupervised methods with single input.

Reflection removal. We apply the proposed framework on single image reflection removal task. Most previous reflection removal works use synthesis data to train their CNN model [8, 10] and the published real data with ground truth have too limited size (Zhang et al. [10] proposed a small dataset which contains 110 pair of images without ground truth reflection (input, background)). Wan et al. [26] proposed a real dataset which contains 454 image sets with the corresponding background and reflection.

In this experiment, we train our model on the benchmark [26] directly. Noted that we use 400 images sets for training and the remain images for evaluation. In the each iteration of training, we choose non-corresponding sample x, y and z from the training set randomly. We train the network to separate reflection contaminated images of size 256x256 by random cropping patches from the image. We show the effectiveness of the encoders $E_Y$ and $E_Z$ for separating images in the latent space through clustering the features from encoders in Fig.3(a), then following several separate results of qualitative comparisons. Cycle GAN and UNIT only provide the prediction of y by given x.

Intrinsic image decomposition. We use the 220 images in the MIT intrinsic dataset [27]. The dataset is extended by [28]. This data contains only 20 different objects, each of which has 11 images. We train the network to decompose intrinsic images of size 256x256. The cluster result, visual comparisons are illuminated in Fig.4. The numerical comparison results are shown in Table 1.

5.3 Multi-layers separation

In this section we designed another toy but more challenging problem to evaluate the proposed USIS framework: separating one single image into three analogous layers, see Fig.5(a) for the visualization of the task. Same as separating an image into two layers, as shown in Fig.5(a), the task here aims to separate the image into three layers: square, circle and triangle. We generate 5K image
Figure 4: Visual results of intrinsic image decomposition. (a) Cluster result of features from encoders in USIS. (b) Qualitative comparison against two previous unsupervised methods. Here Cyc-GAN [14] and UINT [15] provide the input and reflectance translation.

Table 1: Numerical comparison against two other unsupervised methods on reflection removal task and intrinsic image decomposition task. We compare MSE and SSIM [30] between predicted $y$ and ground truth $y$ in two tasks.

| Method                  | Cyc-GAN [14] | UNIT [15] | USIS (proposed) |
|-------------------------|--------------|-----------|-----------------|
|                         | SSIM    | MSE    | SSIM    | MSE    | SSIM    | MSE    |
| Reflection removal      | 0.622   | 97.04  | 0.738   | 68.15  | 0.842   | 51.14  |
| Intrinsic decomposition | 0.572   | 48.95  | 0.821   | 37.41  | 0.893   | 30.10  |

Figure 5: The extend experiment on single image to three layers separation task. (a) The extend toy problem data samples. The input sample and three different samples from three different domains. (b) Two visual results of proposed USIS. From left to right: input and separated layers.

6 Conclusion

In this paper, we propose an Unsupervised Single Image Separation network for single image separation task. We show that by learning both image consistency and independence of distributions in different layers, USIS can make use of the information of different layer distributions to separate single image into analogous different layers. USIS allows unlabeled data to be used in training. Experimental results show that the proposed framework can make unsupervised single image reflection removal and single intrinsic image decomposition properly.
References

[1] Sai Bi, Xiaoguang Han, and Yizhou Yu. An l1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. *ACM Transactions on Graphics (TOG)*, 34(4):78, 2015.

[2] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.

[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[4] Yu Li and Michael S. Brown. Single image layer separation using relative smoothness. In *CVPR*, 2014.

[5] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2341–2353, 2011.

[6] Jae-Young Sim Jae-Seong Yun. Reflection removal for large-scale 3d point clouds. In *CVPR*, 2018.

[7] YiChang Shih, Dilip Krishnan, Fredo Durand, and William T. Freeman. Reflection removal using ghosting cues. In *CVPR*, 2015.

[8] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *ICCV*, 2017.

[9] Haoliang Li Ling-Yu Duan Alex C. Kot Renjie Wan, Boxin Shi. Face image reflection removal. In *CVPR*, 2019.

[10] Xuuner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *CVPR*, 2018.

[11] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. Revisiting deep intrinsic image decompositions. In *CVPR*, pages 8944–8952, 2018.

[12] L. Metz A. Radford and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

[13] Zhang H. Tan P. Gong-M. Yi, Z. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017.

[14] Park T. Isola P. Efros-A.A. Zhu, J.Y. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *CVPR*, 2017.

[15] Breuel T. Kautz J. Liu, M.Y. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems 30*, 2017.

[16] Fischer P. Brox T Ronneberger, O. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[17] Tejas D. Kulkarni-Ilker Yildirim Joshua B. Tenenbaum Michael Janner, Jiajun Wu. Self-supervised intrinsic image decomposition. In *Advances in Neural Information Processing Systems*, 2017.

[18] Bolei Zhou-Raquel Urtasun Wei-Chiu Ma, Hang Chu and Antonio Torralba. Single image intrinsic decomposition without a single intrinsic image. In *ECCV*, 2018.

[19] Noah Snavely Zhengqi Li. Learning intrinsic image decomposition from watching the world. In *CVPR*, 2018.

[20] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[21] M. Mirza-B. Xu D. Warde-Farley S. Ozair A. Courville I. Goodfellow, J. Pouget-Abadie and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

[22] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[23] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
[24] Andrea Vedaldi Dmitry Ulyanov and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. In arXiv preprint arXiv:1607.08022, 2016.

[25] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In ICLR, 2015.

[26] Renjie Wan, Boxin Shi, Ling Yu Duan, Ah Hwee Tan, and Alex C. Kot. Benchmarking single-image reflection removal algorithms. In IEEE ICCV, 2017.

[27] E. H. Adelson R. Grosse, M. K. Johnson and W. T. Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In ICCV, 2011.

[28] M. M. Takuya Narihira and S. X. Yu. Direct intrinsics:learning albedo-shading decomposition by convolutional regression. In CVPR, 2015.

[29] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(12):2579–2605, 2008.

[30] Wang Zhou, Bovik Alan Conrad, Sheikh Hamid Rahim, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process, 13(4):600–612, 2004.