https://doi.org/10.2352/CIC.2022.30.1.33

# Image editing of light and color from a single image: a baseline framework

***Yixiong Yang* [1], *Hassan Ahmed Sial* [2], *Ramon Baldrich* [1], *Maria Vanrell* [1]**
[1] ***Computer Vision Center, Universitat Autònoma de Barcelona, Barcelona, Spain***
[2] ***ISGlobal, Barcelona, Spain***

## Abstract

*Smart image editing is drawing attention and a wide range of edit operations have been investigated. We address the problem of creating new image versions where light conditions and object colors can be altered while maintaining physical coherence across the scene. We propose a baseline framework comprised of a surreal dataset with a large Ground-Truth on light effects and a set of basic deep architectures relying on intrinsic decomposition. Our proposal is evaluated for image relighting and outperforms the state-of-the-art on the previous VIDIT dataset. The codes and datasets are available:* `https://github.com/liulisixin/ImageEditingSI`

## Introduction

Deep architectures provide a versatile tool for (a) estimating intrinsic scene properties from a single image, and (b) generating new versions of an input image to a given target. The image generation approach has rapidly evolved in recent years in the pursuit of smart image editing, from the early color transfer by Reinhard et al. [24] based on color space transformation, to the impressive generation of photo-realistic versions of images directly from text captions [7] in the last months. However, the generation of realistic image versions under new light conditions or with new object colors is still an ongoing challenge. Up to this point, image relighting, intrinsic decomposition, and material editing have a substantial body of work behind them. In this work, we aim to create a baseline framework for evaluating these types of edit operations. Figure 1 illustrates our scheme for light and color editing based on the intrinsic estimation of reflectance, shading, and input light.

Thus, the aim of the paper is twofold. Firstly, we propose a new dataset of synthetic images with a large variety of light conditions to train different deep models. The dataset is surreal since the aim is to learn the light reflection properties, no matter which is the scene semantic content, we introduce a high degree of texture variability and a single light condition to ensure strong cast shadows. Secondly, we propose three basic architectures, from a single encoder-decoder that directly generates the target version and two additional models that estimate the intrinsic light components that will provide the scheme with robust editing abilities.

We intend to prove that intrinsic decomposition clearly improves the relighting performance and enlarges the possibility to introduce multiple and consistent editing operations on the generated versions, as shown in Figure 1. We will establish a baseline evaluation based on the proposed dataset and a family of basic architectures that outperform the state-of-the-art methods on the VIDIT dataset [12, 13] that was designed for a similar purpose.

## Previous works

**AI-based Image Editing.** Creating new versions of a given image with different tasks in mind has been a goal in computer vision research. Considering the whole range of possible editing operations, we identify two different families of works. On one side, we have *Semantic image editing*, that pursues to manipulate the image content. It has evolved from the initial works focusing on altering facial appearance [27] to more recent and refined versions that allow to manipulate any image part based from its semantic segmentation [20]. On the other side, we have *Image style transfer*, that from the seminal work of Gatys et al. [10] pursues to create artistic imagery combining image content and specific style [15]. In between these two extremes, we can find a family of works that pursue to manipulate the image content, either by inserting or removing specific objects [37] or by altering their color [31] while keeping the consistency of the physically realistic light effects, such as shading or cast shadows. We review some of these works in the next section.

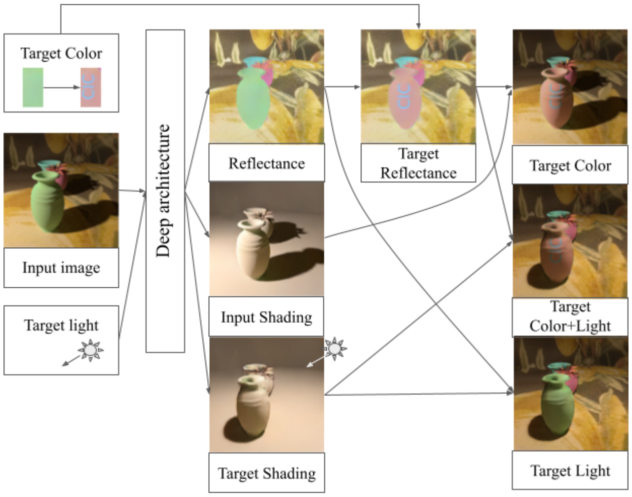


***Figure 1.*** *Image editing and relighting pipeline based on intrinsic decomposition. Except for the input image, the other images are all our predictions.*

**Image Relighting.** Editing light scene conditions has been mainly explored in two different scenarios: (a) portrait relighting [22] is the most popular, given an input image, a relit version is obtained in the output, and the target light condition is introduced in the bottleneck of an encoder-decoder network [30, 39, 21], some introduce GAN architectures [11] or intrinsic decomposition [36] steps to improve the final appearance, and (b) outdoor scene relighting, that has been tackled by Lalonde et al. [19] from a single image based on estimating the position of the light, and by Duchene et al. [6], which in this case is based on a multiview dataset. With the aim of setting a baseline for this topic, an image relighting challenge [13] was held on the VIDIT dataset *(Virtual Image Dataset for Illumination Transfer)* [12] formed by synthetic images with very high complexity, including densely packed scenes, with large dark areas and transparent surfaces. The challenge focus on three different approaches to the relighting problem: one-to-one relighting, estimation of illumination settings, and any-to-any relighting. The best results were obtained by [23] and [32]. The first work [23] proposes a wavelet decomposed RelightNet called WDRN which is an encoder-decoder network under a multi-resolution framework. In the second [32], the author carries out a single image relighting through a novel Deep Relighting Network (DRN) with three parts: scene reconversion, shadow prior estimation, and re-renderer.

**Intrinsic decomposition.** Since the seminal work of Barrow and Tenenbaum [2], a lot of methods have explored the problem of estimating intrinsic components[1]. Reflectance and shading have been the main components, whose pixel-wise product recovers the original image. Lately, light position and color have been estimated too [29]. Initial unsupervised approaches has been substituted by deep architectures that extend the U-Net paradigm to a one-to-two encoder-decoder version[28, 3, 34] or multi-stage trained architectures [8]. One of the main problems to deal with intrinsic decomposition has been the lack of adequate Ground-truth datasets. The small size of the datasets or the lack of physical coherence in the light conditions due to the use of environmental maps has been discussed in different works [4, 17, 28] and diverse synthetic datasets have been proposed.

## Relighting Surreal Dataset

We built a surreal synthetic dataset through the open source Blender rendering engine[5], which follows the methodology proposed in the research of intrinsic decomposition in Sial et al. [28]. Since our dataset contains both intrinsic decomposition and relighting, we call it ReLighting Surreal Intrinsic Dataset (RLSID). Some examples are displayed in Figure 2

RLSID has 10,077 scenes, and each scene has about 10 different light conditions, resulting in 100,242 images with intrinsic data for the whole image. The synthetic scenes are formed by 3D objects surrounded by walls. The 3D objects are randomly selected from various categories of the *ShapeNet* dataset [26] including electronics, pots, buses, cars, chairs, sofas, and airplanes. The roughness parameter is used to control how much light is reflected from the object's surface. Walls are set either to homogeneous colors or a rich set of textured patterns are randomly selected. As a result, the dataset presents significant reflectance and shading variations across the different objects and background surfaces. Each of the scenes is illuminated by a single small area color light source. Light properties include color and position. Light color is represented by one single parameter, which is color temperature. For a single scene, each image is generated using a random pan and tilt light positions represented on an upper semisphere of a radius size from 20 to 50 meters. To create a more natural environment, objects are always placed on the floor and around the center of the scene. The distance of the walls from the center of the scene is between 70 and 100 meters. The meter unit is only chosen in the Blender setting to capture high-quality images with less noise. The meter unit should not be considered as the real-world meter unit, only the scale matters.

## Methods

In this section, we explain the different architectures we want to evaluate on a single image relighting problem, after being trained on our RLSID dataset. We move from the simplest U-net (1 Encoder to 1 Decoder) architecture to different versions that increase the number of Decoders while constraining the training to the estimation of new intrinsic components. In Figure. 3 we plot a schema of our 3 proposed methods, which we refer to as 1-to-1 U-NET, 1-to-2 Intrinsic, and 1-to-3 Intrinsic.

### *Intrinsic decomposition constraints*

Intrinsic decomposition as proposed by Barrow and Tenenbaum [2] assumes that an image can be decomposed into the pixel-wise product of two components, reflectance, and shading:

$$I(x,y) = Ref(x,y) \cdot Sha(x,y) \tag{1}$$

where $I$ is the original image, $Ref$ is the reflectance component, $Sha$ is the shading component, and $(x,y)$ are pixel coordinates. This model assumes that the reflectance component is independent of the light position, which only affects the shading component. We use these assumptions as a physical constraint that forces the reflectance to be the same both for the input and target images, while shading varies accordingly with the input and target light position. This intrinsic decomposition is not going to be used in the 1-to-1 U-Net network.

### *Proposed network architectures*

**Basic 1-to-1 U-NET Architecture.** Considering the relationship between the input and relit images, the relighting task can be seen as an image-to-image translation problem. We propose a basic architecture that resembles the structure of the U-net network [25] described in Pix2pix [14], which is an encoder-decoder structure with skip connections, as shown in the left top of Figure. 3. We modify it to introduce in the U-Net bottleneck the target light condition as an input.

The encoder is formed by a series of Convolution-BatchNorm-ReLU blocks. The output of this encoder is a latent space that further passes through one more convolution layer and a dense layer, yielding the light condition of the original image. After the original light condition is yielded, a new light condition is introduced to replace it. It is processed to create a new latent space, again formed by a dense layer and a transposed convolution layer, and reshaped to the same size as the output of the encoder. The actions of the decoder are like an invert of the encoder, and it takes the encoded message and the new light condition to predict a relit image.

The total loss function is defined as the sum of 3 losses as:

$$\begin{aligned}\mathscr{L}_{1to1} =& \omega_1 \mathscr{L}_{L_c}(L_c, \hat{L_c}) + \omega_2 \mathscr{L}_{L_p}(L_p, \hat{L_p}) \\ &+ \omega_3 \mathscr{L}_{RnS}(RnS, \hat{RnS})\end{aligned} \tag{2}$$

where $\hat{L}_p$ and $\hat{L}_c$ are denoting predictions for light position and color respectively, and $\hat{RnS}$ is the prediction of the relit image, thus $\mathscr{L}_{RnS}$ is the relit image loss. The different losses are combined with $\omega_i$ weights.

**1-to-2 Intrinsic Architecture.** In this second architecture, we introduce the intrinsic decomposition to additionally constrain the correct decomposition of reflectance and target shading with their GT versions. In this case, the relighting network is given in the scheme shown on the left bottom of Figure. 3. The new architecture presents the same encoder, but two decoders. One of them is used to predict the reflectance component, and the second one is used to yield the shading component under the target light condition. At the bottleneck, the output of the encoder is a latent representation that is transferred into the decoder to estimate the reflectance component. On the other side, the new light condition after being processed is regarded as the input of the decoder for the new shading. After the reflectance component and the new

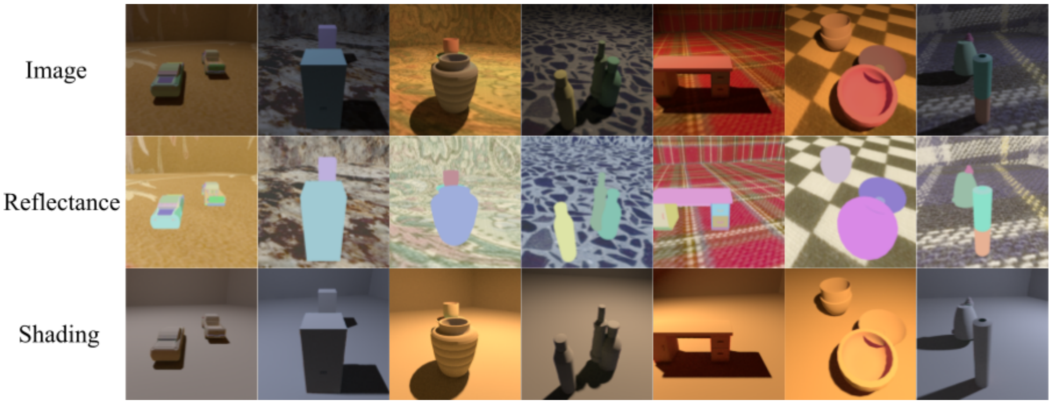


*Figure 2. Some examples of Images in RLSID dataset (first top row). Reflectance and Shading intrinsic components in second and third row, respectively.*

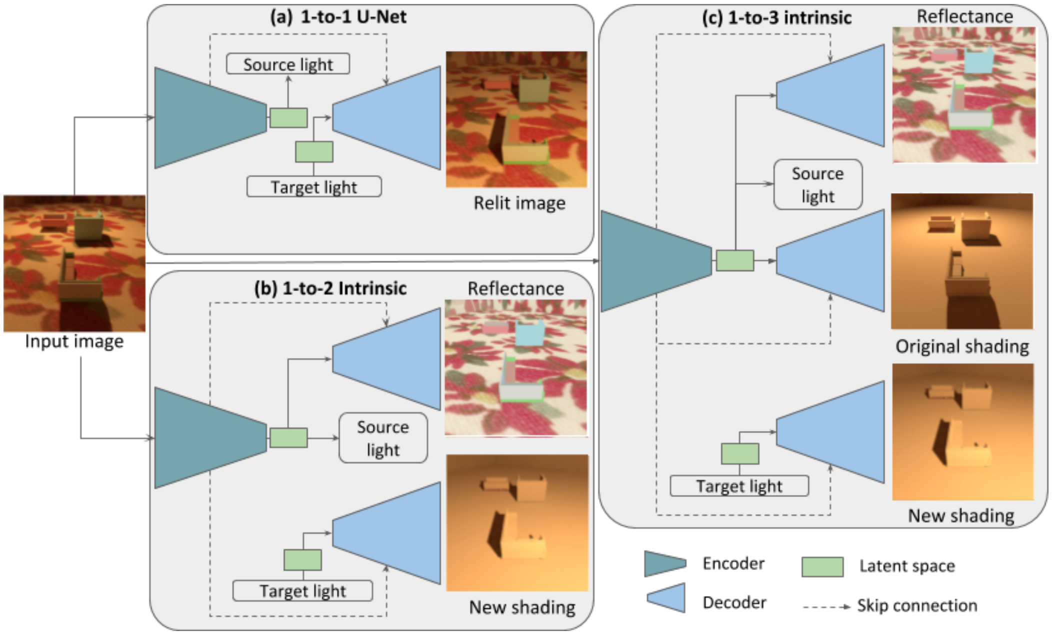


*Figure 3. Proposed network architectures: (a) 1-to-1 U-Net with one encoder and one decoder: (b) 1-to-2 intrinsic with one encoder and two decoders (c) 1-to-3 intrinsic with one encoder and three decoders.*

shading component are predicted, the relit image is yielded by the product given in Equ.1.

As a result, the loss function is derived as:

$$\begin{aligned}\mathcal{L}_{1to2} =& \omega_1 \mathcal{L}_{L_c}(L_c,\hat{L_c}) + \omega_2 \mathcal{L}_{L_p}(L_p,\hat{L_p}) \\ &+ \omega_3 \mathcal{L}_{RnS}(RnS,\hat{RnS}) + \omega_4 \mathcal{L}_R(R,\hat{R}) \\ &+ \omega_5 \mathcal{L}_{nS}(nS,\hat{nS})\end{aligned} \tag{3}$$

where $\hat{R}$ is the reflectance prediction. And $\hat{nS}$ is the new shading prediction, which is the shading under the target light.

**1-to-3 Intrinsic Architecture.** Furthermore, an architecture with three decoders has also been implemented as shown on the right side of Figure. 3. Compared with the previous schemes, it introduces one more decoder to also predict the shading of the original image. Likewise, the reflectance decoder, this new decoder only receives the encoded information from the encoder and has no connection with the new light. In other words, a full model for intrinsic decomposition is enclosed in this new architecture. With the output of the shading of the original light, the reconstruction of the input image can be generated by the product of reflectance and the original shading.

The estimation of the original shading requires the addition of two more losses, which results in:

$$\begin{aligned}\mathcal{L}_{1to3} =& \omega_1 \mathcal{L}_{L_c}(L_c,\hat{L_c}) + \omega_2 \mathcal{L}_{L_p}(L_p,\hat{L_p}) \\ &+ \omega_3 \mathcal{L}_{RnS}(RnS,\hat{RnS}) + \omega_4 \mathcal{L}_R(R,\hat{R}) \\ &+ \omega_5 \mathcal{L}_{nS}(nS,\hat{nS}) + \omega_6 \mathcal{L}_S(S,\hat{S}) \\ &+ \omega_7 \mathcal{L}_{RS}(RS,\hat{RS})\end{aligned} \tag{4}$$

where $\hat{S}$ is the prediction of the original shading, and $\hat{RS}$ is the prediction of the input image from the estimated original shading.

## *Implementation details and evaluation metrics*

To train on the RLSID dataset, we randomly divided the dataset into three sets: 80% for the training set, 5% for the validation set, and 15% for the test set. The training set has about 80,000 images with intrinsic data and we train all three proposed networks with identical settings and hyperparameters. For the optimization, we use Adam optimizer [16] with a learning rate of 0.0002 and a batch size of 96. The inputs and outputs including images and light conditions are all normalized to $[0,1]$. To constrain the images in Equ.2, Equ.3 and Equ.4, we use a sum of L1 loss, SSIM loss [33, 35] and LPIPS loss [38]. The light position is constrained by the angular error, and the light color is constrained by L1 loss.

Afterward, we fine-tune three proposed models on the VIDIT dataset [12] and some real images. For all these experiments, the resolution of input and relit images is 256 × 256.

Table 1: Estimation errors for relit images and light conditions on RLSID dataset

| Methods | Relit images | | | | | Estimation of light condition | |
|---|---|---|---|---|---|---|---|
| | MPS↑ | SSIM↑ | LPIPS↓ | PSNR↑ | MSE↓ | Light position | Light color |
| 1-to-1 U-Net | 0.9114 | 0.9106 | 0.0878 | 25.91 | 0.0038 | 12.95 | 2.1710 |
| 1-to-2 Intrinsic | 0.9168 | 0.9154 | 0.0818 | 26.50 | **0.0033** | 12.87 | 1.1645 |
| 1-to-3 Intrinsic | **0.9180** | **0.9167** | **0.0807** | **26.68** | **0.0033** | **12.72** | **1.1171** |

Table 2: Estimation errors of the intrinsic components on RLSID dataset

| Components | Methods | MPS↑ | SSIM↑ | LPIPS↓ | PSNR↑ | MSE↓ |
|---|---|---|---|---|---|---|
| Reflectance | 1-to-2 Intrinsic | 0.9537 | 0.9552 | 0.0479 | 28.8062 | 0.0018 |
| | 1-to-3 Intrinsic | **0.9545** | **0.9559** | **0.0468** | **29.1480** | **0.0016** |
| New shading | 1-to-2 Intrinsic | 0.8878 | 0.8944 | 0.1188 | 22.5385 | 0.0081 |
| | 1-to-3 Intrinsic | **0.8891** | **0.8955** | **0.1173** | **22.7268** | **0.0079** |
| Original shading | 1-to-3 Intrinsic | 0.9767 | 0.9797 | 0.0262 | 31.5787 | 0.0011 |

Table 3: Quantitative results on VIDIT dataset

| Experiments | MPS↑ | SSIM↑ | LPIPS↓ | PSNR↑ |
|---|---|---|---|---|
| CET_SP[23] | 0.6452 | 0.6310 | 0.3405 | 17.07 |
| CET_CVLAB[23] | 0.6451 | 0.6362 | 0.3460 | 16.89 |
| Lyl[13] | 0.6436 | 0.6301 | 0.3430 | 16.68 |
| YorkU[13] | 0.6216 | 0.6091 | 0.3659 | 16.81 |
| IPCV_IITM[13] | 0.5897 | 0.5298 | 0.3505 | 17.05 |
| DeepRelight[32] | 0.5892 | 0.5928 | 0.4144 | 17.42 |
| Hertz[13] | 0.5339 | 0.5666 | 0.4989 | 16.92 |
| Image Lab[13] | 0.3746 | 0.3769 | 0.6278 | 16.89 |
| SILT[18] | n/a | 0.6060 | n/a | 17.00 |
| Pix2pix [14] | 0.5928 | 0.5825 | 0.3970 | 17.35 |
| 1-to-1 U-Net | 0.7192 | 0.6815 | 0.2431 | **17.88** |
| 1-to-2 Intrinsic | 0.7298 | **0.6948** | 0.2353 | 17.38 |
| 1-to-3 Intrinsic | **0.7306** | 0.6915 | **0.2303** | 17.72 |

When fine-tuning on other datasets and images, we have no ground truth of intrinsic components. In order to utilize the physical constraint, we introduce a reflectance consistency loss instead. After getting the relit image, we feed the relit image to the network again, thereby obtaining the reflectance of the relit image. The reflectance consistency constrains the reflectance of the input image and relit image to be the same.

The metrics used to evaluate our predictions include mean square error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [33, 35], Learned Perceptual Image Patch Similarity (LPIPS) [38], and Mean Perceptual Score (MPS) [13]. Mean Perceptual Score (MPS) is the average of SSIM and LPIPS scores, and it is used in the ranking score of the AIM 2020 relighting challenge [13]. We use the angular error both for the position (derived from pan and tilt) and for the color (derived from the 3D color vector) of the light.

## Results and Discussion

In this section, we first show the results obtained on our RLSID dataset, both a qualitative and a quantitative evaluation. Second, we present the results of our trained networks after fine-tuning on the VIDIT dataset and we compare our results with previous works. Finally, we test our networks on some real images to observe their generalization capabilities.

### *Results on RLSID Dataset*

**Quantitative Evaluation.** We perform a quantitative evaluation of our network on a testing dataset with more than 1500 scenes. The results on the relit images and the estimations of light conditions are displayed in Table 1. From these results, we can conclude that 1-to-3 intrinsic network achieve the best results with all the metrics. And 1-to-2 intrinsic network presents clear advantages over the 1-to-1 U-Net. As a result, these make us conclude again that the intrinsic decomposition is helping in the relighting task. Table 2 shows the quantitative results of the intrinsic decomposition of 1-to-2 and 1-to-3 networks. It can be seen that predicting and constraining the original shading provides 1-to-3 network with a clear advantage that makes it clearly overcomes the results of 1-to-2 in the estimation of reflectance and the target shading.

We observe that the error between the input and relit images varies with the position of the target light. In Figure 4 we show the PSNR values depending on the range of angles between the input and the target light position. In addition to the PSNR of our prediction, we also plot the difference between the input and relit images (GT) as a reference value (in blue). We can observe that the error is highest when the relative angles of the light positions are in the range of 80∼100 degrees, which means that at such degrees the relighting is harder. The figure illustrates that although PSNR of our model drops slightly at these angle intervals, it remains over 25 for all the models. Additionally, in most cases, the 1-to-3 is superior to the 1-to-2 intrinsic network, and 1-to-2 is superior to the 1-to-1 U-Net.

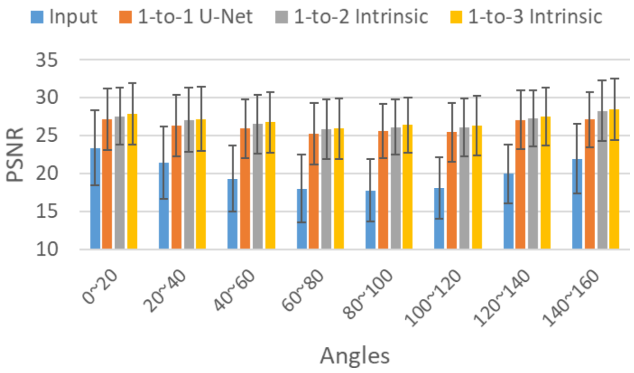
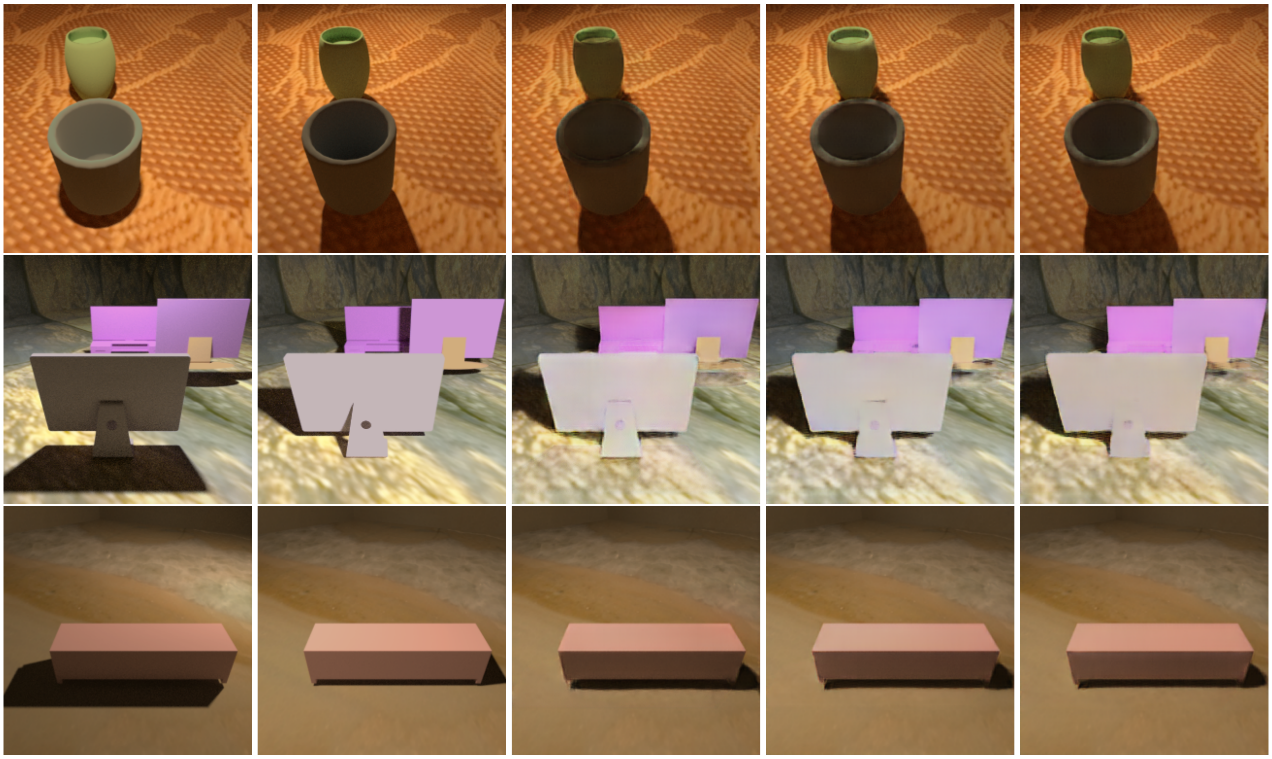


***Figure 4.*** *Error variation depending on the angle between input and target light.*

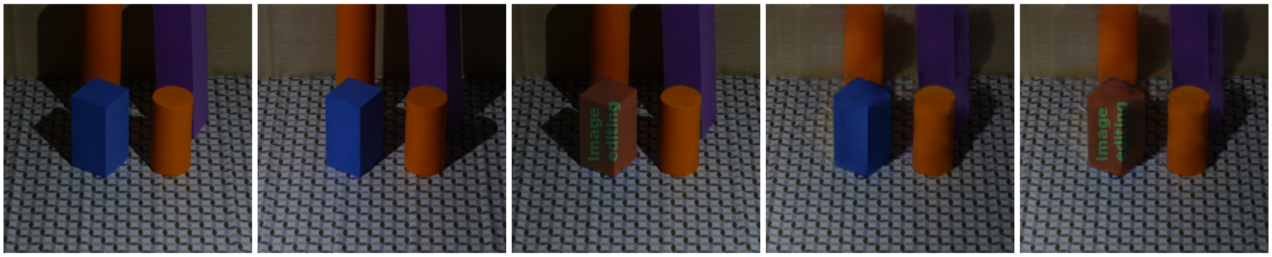
**Qualitative Evaluation.**

Figure 1 show some qualitative result on manipulating both light and color. In Figure 5 we show more qualitative results of relighting on the testing dataset. The angular distance between the source and target light positions is obvious in these cases. The first row shows that our models can generate a clear shadow that casts at the front of the object. In the second row, we demonstrate that our models can remove the cast shadow and shading



Input Relit (GT) Relit (1-to-1 U-Net) Relit (1-to-2 Intrinsic) Relit (1-to-3 Intrinsic)

***Figure 5.*** *Qualitative results of relit images on the RLSID dataset.*



Input Relit(GT) Target color Target light Target color + light

***Figure 6.*** *Results of editing color and light on real images.*

on the facade of the object, and then relight it properly. The third row illustrates that the networks can move the cast shadow from left to right. In general, the image quality obtained by the 1-to-3 Intrinsic network is better than the other two. For example, in the first example, the upper mug interior is properly illuminated, whereas the other two are not. Another advantage of 1-to-2 Intrinsic and 1-to-3 Intrinsic over 1-to-1 U-Net is they present less noise. In general, we can say that intrinsic decomposition qualitatively helps the image relighting process.

## Results on VIDIT Dataset

Table 3 shows the quantitative comparison with other methods on VIDIT Dataset. The top rows are the results from the original challenge on VIDIT [13]. The subsequent rows in Table 3 are the results from [18]. Since VIDIT dataset did not release the ground truth for the test, [18] and [9] split the released part with a rate of 80:10:10 for the train:validation:test, respectively. However, [18] did not share the result of the split, so we did our own split with the same rate. As a result, the splits in Table 3 are not all the same. Based on our split, we train from scratch and evaluate the model modified from [14] as a baseline, which only adds a module to embed the light conditions. The results are shown in the bottom rows of Table 3, where we apply our three models that are pre-trained on our RLSID dataset and fine-tune them on the VIDIT dataset. The 1-to-3 intrinsic network achieves the best results on MPS and LPIPS. In particular, MPS is the final metric used in the challenge. 1-to-2 intrinsic network and 1-to-1 U-Net achieve the best results with SSIM and PSNR respectively. These results confirm the abilities of our network and our RLSID dataset in generalizing for other scenarios.

## Results on Real images

To test on real images, we fine-tune our 1-to-3 Intrinsic network (only trained by the RLSID dataset) on some real scenes (40 images) we captured in lab conditions. In Figure 6 we show the results of editing on a different testing image. The image in the third column shows that our model can change the color of any object and add some text by editing the predicted reflectance component. In the fourth column, we present a relit version of the input image which is built with the predicted target shading and the estimated reflectance. Finally, the image in the fifth column depicts the combined editing results.

## Conclusion

In this paper we have explored the problem of editing the light and color of an image while keeping the physical coherence of the light across the scene.

We build a large dataset of surreal scenes with consistent light conditions, a comprehensible number of objects, and a sufficient level of diversity to approach the problem step by step. The dataset will be used to train some deep architectures to establish a baseline framework, but it can easily be enlarged with more complex light conditions in further research.

We evaluate the performance of three deep architectures from a basic encoder-decoder, to some extensions that introduce physical constraints derived from the intrinsic decomposition model, and which are used to facilitate the image editing task.

**Limitations and Future work:** The scenes we study in this work are limited in the following ways: (a) the number and shape of the objects, as well as the shape of the background; (b) the object surfaces only present diffuse reflection; (c) the light source only considers a single small area light source. In future research, we plan to overcome all these limitations.

## Acknowledgments

This work was supported by Grant PID2021-128178OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by ERDF ”A way of making Europe”. The author Yixiong Yang is also supported by China Scholarship Council.

## References

[1] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1670–1687, 2014.

[2] H. Barrow, J. Tenenbaum, A. Hanson, and E. Riseman. Recovering intrinsic scene characteristics. *Comput. Vis. Syst*, 2(3-26):2, 1978.

[3] A. S. Baslamisli, H.-A. Le, and T. Gevers. Cnn based learning using reflection and retinex models for intrinsic image decomposition. In *CVPR*, pages 6674–6683, 2018.

[4] S. Bell, K. Bala, and N. Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):1–12, 2014.

[5] B. O. Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2022.

[6] S. Duchêne, C. Riant, G. Chaurasia, J. Lopez-Moreno, P.-Y. Laffont, S. Popov, A. Bousseau, and G. Drettakis. Multi-view intrinsic images of outdoors scenes with an application to relighting. 2015.

[7] C. S. et al. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[8] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf. Revisiting deep intrinsic image decompositions. In *CVPR*, pages 8944–8952, 2018.

[9] P. Gafton and E. Maraz. 2d image relighting with image-to-image translation. *arXiv preprint arXiv:2006.07816*, 2020.

[10] L. A. Gatys, A. S. Ecker, and M. Bethg. A neural algorithm of artistic style. *Journal of Vision*, 2015.

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

[12] M. E. Helou, R. Zhou, J. Barthas, and S. Süsstrunk. Vidit: Virtual image dataset for illumination transfer. *arXiv preprint arXiv:2005.05460*, 2020.

[13] M. E. Helou, R. Zhou, S. Süsstrunk, R. Timofte, M. Afifi, M. S. Brown, K. Xu, H. Cai, Y. Liu, L.-W. Wang, et al. Aim 2020: Scene relighting and illumination estimation challenge. *arXiv preprint arXiv:2009.12798*, 2020.

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017.

[15] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song. Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics*, 26(11):3365–3385, 2020.

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] B. Kovacs, S. Bell, N. Snavely, and K. Bala. Shading annotations in the wild. In *CVPR*, pages 6998–7007, 2017.

[18] N. Kubiak, A. Mustafa, G. Phillipson, S. Jolly, and S. Hadfield. Silt: Self-supervised lighting transfer using implicit image decomposition. In *BMVC*, 2021.

[19] J.-F. Lalonde, A. A. Efros, and S. G. Narasimhan. Estimating natural illumination from a single outdoor image. In *ICCV*, pages 183–190. IEEE, 2009.

[20] H. Ling, K. Kreis, D. Li, S. W. Kim, A. Torralba, and S. Fidler. Editgan: High-precision semantic image editing, 2021.

[21] T. Nestmeyer, J.-F. Lalonde, I. Matthews, and A. Lehrmann. Learning physics-guided face relighting under directional light. In *CVPR*, pages 5124–5133, 2020.

[22] R. Pandey, S. O. Escolano, C. Legendre, C. Haene, S. Bouaziz, C. Rhemann, P. Debevec, and S. Fanello. Total relighting: learning to relight portraits for background replacement. *ACM TOG*, 40(4):1–21, 2021.

[23] D. Puthussery, M. Kuriakose, J. C V, et al. Wdrn: A wavelet decomposed relightnet for image relighting. *arXiv preprint arXiv:2009.06678*, 2020.

[24] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

[25] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[26] J. Shi, Y. Dong, H. Su, and S. X. Yu. Learning non-lambertian object intrinsics across shapenet categories. *CVPR*, pages 5844–5853, 2017.

[27] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, pages 5444–5453, 2017.

[28] H. A. Sial, R. Baldrich, and M. Vanrell. Deep intrinsic decomposition trained on surreal scenes yet with realistic light effects. *J. Opt. Soc. Am. A*, 37(1):1–15, Jan 2020.

[29] H. A. Sial, R. Baldrich, M. Vanrell, and D. Samaras. Light direction and color estimation from single image with deep regression. In *IS&T London Imaging Conference*, 2020.

[30] T. Sun, J. T. Barron, Y.-T. Tsai, Z. Xu, X. Yu, G. Fyffe, C. Rhemann, J. Busch, P. E. Debevec, and R. Ramamoorthi. Single image portrait relighting. *ACM Trans. Graph.*, 38(4):79–1, 2019.

[31] P. Vitoria, L. Raad, and C. Ballester. Chromagan: Adversarial picture colorization with semantic class distribution. In *WACV*, pages 2434–2443, 2020.

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

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[34] Z. Wang and F. Lu. Single image intrinsic decomposition with discriminative feature encoding. In *ICCV Workshops*, pages 0–0, 2019.

[35] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[36] Z. Wang, X. Yu, M. Lu, Q. Wang, C. Qian, and F. Xu. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Transactions on Graphics (TOG)*, 39(6):1–13, 2020.

[37] E. Zhang, R. Martin-Brualla, J. Kontkanen, and B. L. Curless. No shadow left behind: Removing objects and their shadows using approximate lighting and geometry. In *CVPR*, pages 16397–16406, 2021.

[38] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

[39] H. Zhou, S. Hadap, K. Sunkavalli, and D. W. Jacobs. Deep single-image portrait relighting. In *ICCV*, pages 7194–7202, 2019.