TrimBot2020 Deliverable D4.1
Intrinsic Image Decomposition

Principal Author: University of Amsterdam (UvA)
Contributors: University of Amsterdam (UvA)
Dissemination:

Abstract: This report describes and evaluates the algorithms that have been developed to perform intrinsic image decomposition for outdoor gardens.

Deliverable due: Month 36

1 Overview

The imaging conditions of gardens may vary significantly over, even small, periods of time influencing the appearance of garden objects and scene. For example, garden objects may contain shadow, specular highlights, shading and illumination changes. Therefore, intrinsic image/video decomposition is investigated to separate garden recordings into reflectance (i.e. albedo) and illumination artifacts. Intrinsic representations of garden proto-objects are important for robust 3D semantic segmentation and detection, measuring invariant appearance properties of proto-objects, the proper rendering of objects for visualization, and updating the 3D map of the garden.

Intrinsic image decomposition is the process of separating an image into its formation components such as reflectance and shading (illumination) [1]. Reflectance is the color of the object, invariant to camera viewpoint and illumination conditions, whereas shading, dependent on camera viewpoint and object geometry, consists of different illumination effects, such as shadows, shading and inter-reflections. Using intrinsic images, instead of the original *RGB* images, can be beneficial for many computer vision algorithms. For instance, for shape-from-shading algorithms, the shading images contain important visual cues to recover geometry, while for segmentation and detection algorithms, reflectance images can be beneficial as they are independent of confounding illumination effects.

2 Creating Synthetic Dataset for Garden-specialized Training

The availability of annotated large-scale datasets is key to the success of supervised deep learning methods. Because of the absence of large-scale datasets annotated with garden-specialized semantics and intrinsic properties which requires a good amount of person-hour effort but play a key role in deep learning algorithms, we decided to establish a synthetic design. A large set of synthetic gardens are created featuring plants and common garden objects. The dataset contains different species of vegetation such as trees and flowering plants with different types of terrains and landscapes under different lighting conditions. Furthermore, scenarios are created which involves human intervention such as the presence of bushes (like rectangular hedges or spherical topiaries), fences, flowerpots and planters, and etc. (16 classes in total). There is a substantial variety of object colors and geometry. The dataset is constructed by using the parametric tree models [2] (implemented as add-ons in Blender software), and several manually-designed models from the Internet that aim for realistic natural scenes and environments. Ambient lighting is provided by real HDR sky images with a parallel light source. Light source properties are designed to correspond to daytime lighting conditions such as clear sky, cloudy, sunset, twilight, etc. For each virtual garden, we capture the scene from different perspectives with motion blur effects. Scene are rendered with the physics-based Blender Cycles[1] engine respecting the design of the TrimBot's ring camera setup. The dataset consists of 400K images, depicting 130 various garden models (12 meters x 12 meters) under 6 lighting conditions. The dataset includes intrinsics, semantics, depth maps, surface normals, 3D point clouds, optical flow, light source properties and camera parameters. A number of samples are shown in Figure 1.

[1]https://www.blender.org/

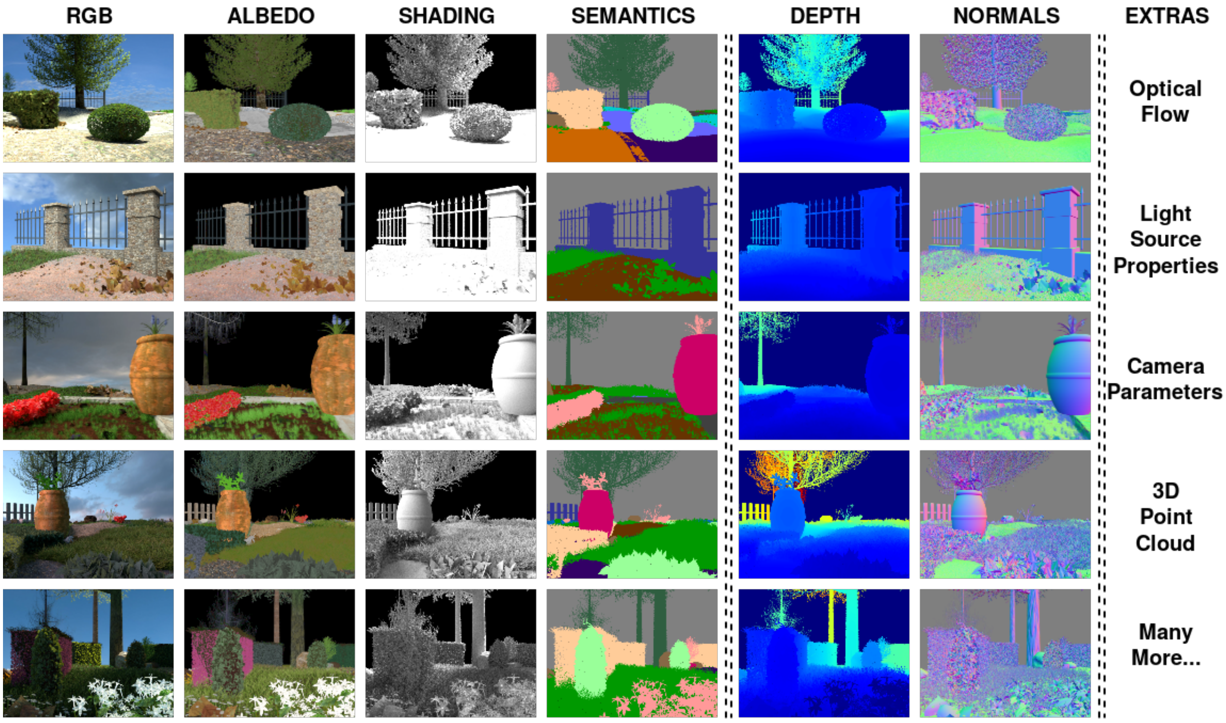

Figure 1: Sample images from the synthetic garden dataset.

3 Three for one and one for three: Flow, Segmentation, and Surface Normals

Optical flow, semantic segmentation, and surface normals represent different information modalities, yet together they bring better cues for scene understanding problems. In this section, we study the influence between the three modalities: how one impacts on the others and their efficiency in combination. To assist the training process we use the synthetic garden dataset mentioned in the previous section. This way we can also evaluate the synthetic garden dataset's quality on the results. As different information sources provide different cues to understand the world, they could also become complementary to each other. For example, certain objects have specific motion patterns (flow and semantics), an object's geometry provides specific cues about its category (surface normals and semantics), and object's boundary curves provide cues about motion boundaries (flow and surface normals). Details of the research can be found in our publication [3]. The main conclusions of this work for the intrinsic image decomposition task is (i) the benefit of surface normals and semantic segmentation tasks on other visual modalities, which will be discussed in Section 5 and Section 6, and (ii) the effectiveness of the generated synthetic garden dataset.

4 CNN based Learning using Reflection and Retinex Models for Intrinsic Image Decomposition

Since there are multiple unknowns and multiple solutions to recover the pixel intrinsics, intrinsic image decomposition is an ill-posed and under-constrained problem. Therefore, most of the traditional work derive priors about the scene characteristics and impose constraints on the reflectance and shading maps. Usually an optimization procedure is used enforcing imaging constraints for pixel-wise decomposition. In addition to the traditional work, more recent research focuses on using deep learning (e.g. CNN) models. However, these deep learning-based methods do not consider the well-established, traditional image formation process as the basis of their intrinsic learning process. Deep learning is used as in-and-out black box, which may lead to inadequate or restricted results. Furthermore, the contribution and physical interpretation of what the network learned is often difficult to interpret. As a consequence, although current deep learning approaches show superior performance when considering quantitative benchmark results, traditional approaches are still dominant in achieving high qualitative results.

Therefore, the aim of this research is to exploit the best of the two worlds and to get an insight on how successful will the deep learning models work for the intrinsic image decomposition task for TrimBot2020. As a result, a method is proposed that (1) is empowered by deep learning capabilities, (2) considers a physics-based reflection model to steer the learning process, and (3) exploits the traditional approach to obtain intrinsic images by exploiting reflectance and shading gradient information.

To this end, a physics-based convolutional neural network, *IntrinsicNet*, is proposed first. A standard CNN architecture is chosen to exploit the dichromatic reflection model [4] as a standard reflection model to steer the training process by introducing a physics-based loss function called the *image formation loss*, which takes into account the reconstructed image of the predicted reflectance and shading images:

The goal is to analyze the contribution of exploiting the image formation process as a constraining factor in a standard CNN architecture for intrinsic image decomposition. Then, we propose the *RetiNet*, which is a two-stage Retinex [5] inspired convolutional neural network which first learns to decompose (color) image gradients into intrinsic image gradients i.e. reflectance and shading gradients. Finally, these intrinsic gradients are used to train the CNN to decompose, at the pixel, the full image into its corresponding reflectance and shading images.

As mentioned in Section 2, the availability of annotated large-scale datasets is key to the success of supervised deep learning methods. However, the largest publicly available dataset with intrinsic image ground-truth has around a thousand of redundant images taken from an animated cartoon-like short film [6]. Nonetheless, at the time of this research was conducted our synthetic garden dataset was under construction. Therefore, to train our CNNs, we introduce another large scale dataset with only intrinsic ground-truth images: a synthetic dataset with man-made objects. The dataset consists of around 20,000 images. Rendered with different environment maps and viewpoints, the dataset provides a variety of possible images in indoor and outdoor scenes. A number of samples are shown in Figure 2.

4.1 Approach

In this section, the image formation model is described first. Then, we propose an encoder-decoder CNN, called *IntrinsicNet*, which is a convolutional neural network based on the reflection model by introducing the image formation loss. Finally, we propose a new CNN architecture, *RetiNet*, which is a Retinex-inspired scheme that exploits image gradients in combination with the image formation loss.

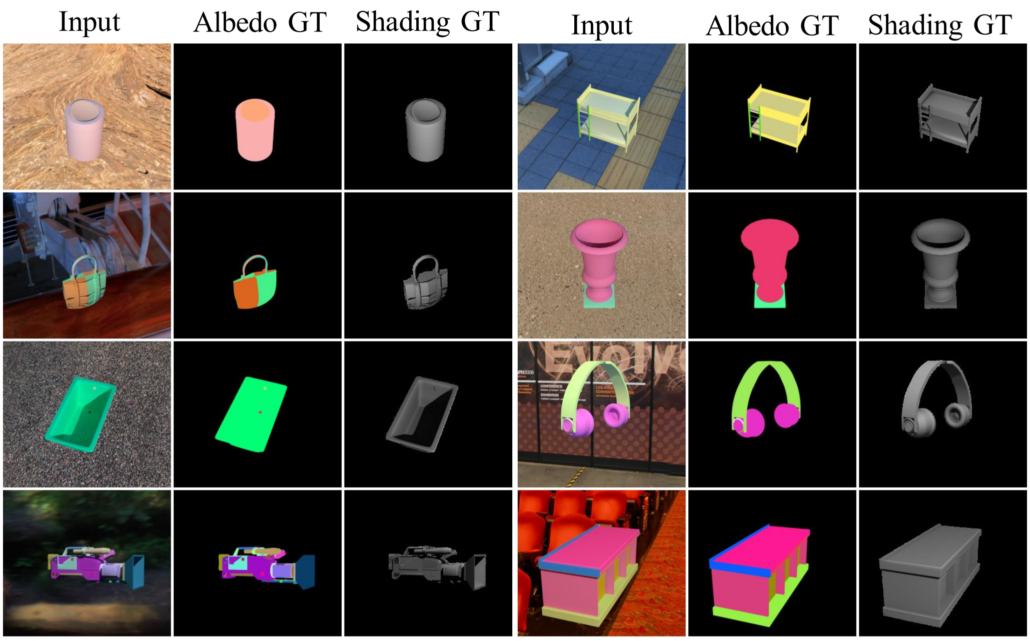

Figure 2: Overview of the synthetic dataset with man-made objects. Different environment maps are used to render the models for realistic appearance.

4.1.1 Image Formation Model

The dichromatic reflection model [4] describes a surface as a composition of the body I_b (diffuse) and specular I_s (interface) reflectance:

$$I = I_b + I_s \; . \tag{1}$$

Then, the pixel value, measured over the visible spectrum ω, is expressed by:

$$I = m_b(\vec{n},\vec{l}) \int_\omega f_c(\lambda)\; e(\lambda)\; \rho_b(\lambda)\; \mathrm{d}\lambda \;+\; m_s(\vec{n},\vec{l},\vec{v}) \int_\omega f_c(\lambda)\; e(\lambda)\; \rho_s(\lambda)\; \mathrm{d}\lambda \; , \tag{2}$$

where $\vec{n}$ is the surface normal, $\vec{l}$ is the light source direction, and $\vec{v}$ is the viewing/camera direction. m is a function of the geometric dependencies (e.g. Lambertian $\vec{n} \cdot \vec{l}$). Furthermore, λ is the wavelength, $f_c(\lambda)$ is the camera spectral sensitivity, $e(\lambda)$ defines the spectral power distribution of the illuminant, ρ_b characterizes the diffuse surface reflectance i.e. the albedo (intrinsic color), and ρ_s is the specular reflectance with Fresnel reflection.

Assuming a linear sensor response, a single light source and narrow band filters (λ_I), Equation (2) is as follows:

$$I = m_b(\vec{n},\vec{s})\; e(\lambda_I)\; \rho_b(\lambda_I) \;+\; m_s(\vec{n},\vec{s},\vec{v})\; e(\lambda_I)\; \rho_s(\lambda_I) \; . \tag{3}$$

Then, under the assumption of body (diffuse) reflection, the decomposition of the observed image $I(\vec{x})$ at position $\vec{x}$ can be approximated as the element-wise product of its reflectance $R(\vec{x})$ and shading $S(\vec{x})$ intrinsics:

$$I(\vec{x}) = R(\vec{x}) \times S(\vec{x}) \; . \tag{4}$$

In Equation (3), $e(\lambda_I)$ is modeled as a single, canonical light source. We can extend the model for a non-canonical light source as follows:

$$I(\vec{x}) = R(\vec{x}) \times S(\vec{x}) \times E(\vec{x}) \; , \tag{5}$$

where $E(\vec{x})$ describes the color of the light source at position $\vec{x}$. The model for a global, non-canonical light source is described by:

$$I(\vec{x}) = R(\vec{x}) \times S(\vec{x}) \times E \; . \tag{6}$$

Equation (4) is extended to non-diffuse reflection by adding the specular (surface) term $H(\vec{x})$:

$$I(\vec{x}) = R(\vec{x}) \times S(\vec{x}) + H(\vec{x}) \; , \tag{7}$$

and for a non-canonical light source by:

$$I(\vec{x}) = R(\vec{x}) \times S(\vec{x}) \times E(\vec{x}) + H(\vec{x}) \times E(\vec{x}) \; . \tag{8}$$

Finally, for a global, non-canonical light source we obtain:

$$I(\vec{x}) = R(\vec{x}) \times S(\vec{x}) \times E + H(\vec{x}) \times E \; . \tag{9}$$

In the next section, the reflection model is considered to introduce different image formation losses within an encoder-decoder CNN model for intrinsic image decomposition.

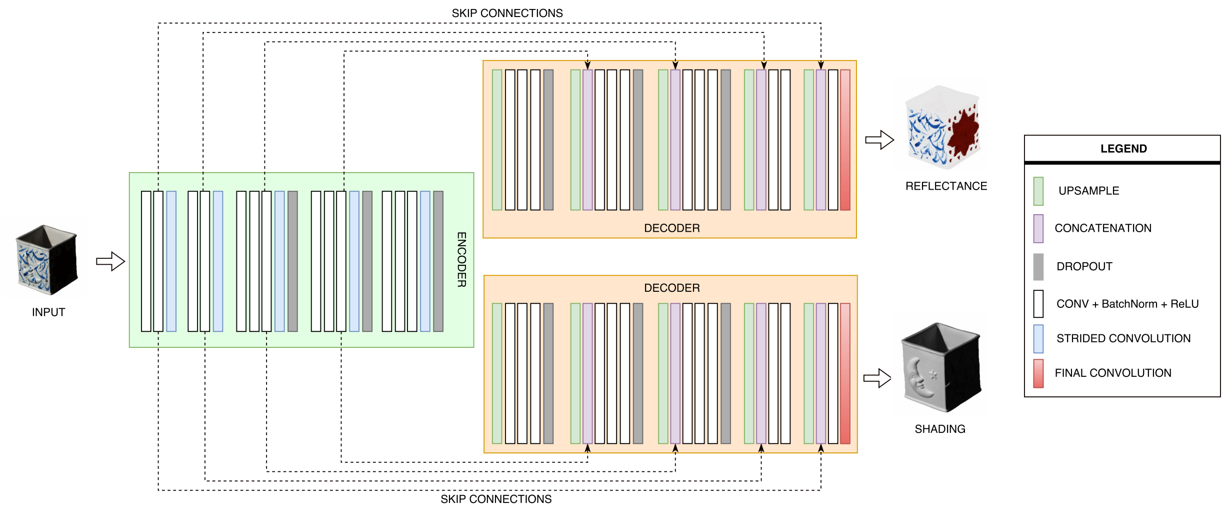

Figure 3: IntrinsicNet model architecture with one shared encoder and two separate decoders: one for shading and one for reflectance prediction. The encoder part contains both shading and reflectance characteristics. The decoder parts aim to disentangle those features.

4.1.2 IntrinsicNet: CNN driven by Reflection Models

In this section, a physics-based deep learning network, *IntrinsicNet*, is proposed. Figure 3 illustrates our model. We use a standard CNN architecture, VGG16 [7], to constrain the training process by introducing a physics-based loss. The reason of using a standard CNN architecture is to analyze whether it is beneficial to constrain the CNN by the reflection model. Therefore, an end-to-end trainable encoder-decoder CNN is considered. These type of CNNs yield good results in most of the pixel-wise dense prediction tasks [8, 9]. An architecture is adopted with one shared encoder and two separate decoders: one for shading prediction and one for reflectance prediction. The features learned by the encoder stage contain both shading and reflectance cues. The purpose of the decoder parts is to disentangle those features. Obviously, the architecture can be extended by considering more image formation factors (e.g. the light source or highlights) by adding the corresponding decoder blocks.

To train the model, we use the standard L_2 reconstruction loss. Let $\hat{J}$ be the ground-truth intrinsic image and J be the prediction of the network. Then, the reconstruction loss $\mathcal{L}_{RL}$ is given by:

$$\mathcal{L}_{RL}(J, \hat{J}) = \frac{1}{n} \sum_{\vec{x},c} ||\hat{J} - J||_2^2 \ , \tag{10}$$

where $\vec{x}$ denotes the image pixel, c the channel index and n is the total number of evaluated pixels. In our case, the final, combined loss $\mathcal{L}_{CL}$ is composed of 2 distinct loss functions, one for reflectance reconstruction $\mathcal{L}_{RL_R}$ and one for shading reconstruction $\mathcal{L}_{RL_S}$:

$$\mathcal{L}_{CL}(R, \hat{R}, S, \hat{S}) = \gamma_R \, \mathcal{L}_{RL_R}(R, \hat{R}) \ + \ \gamma_S \, \mathcal{L}_{RL_S}(S, \hat{S}) \ , \tag{11}$$

where the γs are the corresponding weights. In general, this type of network may generate color artifacts and blurry reflectance maps [8, 9]. The goal of the image formation loss is to increase the color reproduction quality because of the physics constraint.

More precisely, the image formation loss $\mathcal{L}_{IMF}$ takes into account the reconstructed image of the predicted reflectance and shading images. That is in addition to the RGB input image. Hence, this loss imposes the reflection model constraint of Equation 4:

$$\mathcal{L}_{IMF}(R, S, I) = \gamma_{IMF} \, \mathcal{L}_{RL_{IMF}}((R \times S), \ I) \ , \tag{12}$$

where I is the input image. Thus, the final loss of the *IntrinsicNet* becomes:

$$\mathcal{L}_{FL}(I, R, \hat{R}, S, \hat{S}) = \mathcal{L}_{CL}(R, \hat{R}, S, \hat{S}) \ + \ \mathcal{L}_{IMF}(R, S, I) \ . \tag{13}$$

Note that the image formation loss is not limited to Equation 4. Any intrinsic image Equation 4-9 can be used depending on the intrinsic problem at hand. For example, the loss function for the full reflection model $\mathcal{L}_{FRM}$ is as follows:

$$\begin{aligned} \mathcal{L}_{FRM}(*) = \gamma_R \, \mathcal{L}_{RL_R}(R, \hat{R}) \ + \gamma_S \, \mathcal{L}_{RL_S}(S, \hat{S}) \ + \gamma_H \, \mathcal{L}_{RL_H}(H, \hat{H}) \ + \\ \gamma_E \, \mathcal{L}_{RL_E}(E, \hat{E}) \ + \ \gamma_{IMF} \, \mathcal{L}_{RL_{IMF}}((R \times S \times E + H \times E), \ I) \ . \end{aligned} \tag{14}$$

The image formation loss function is designed to augment the color reproduction. To augment both *color reproduction* and *edge sharpness*, in the next section, a two-stage Retinex-inspired CNN architecture is described which uses intrinsic gradients (for edge sharpness) and the image formation loss (for color reproduction).

4.1.3 RetiNet

In this section, we exploit how a well-established, traditional approach such as Retinex [5] can be used to steer the design of a CNN architecture for intrinsic image decomposition. Therefore, we propose the *RetiNet* model. In fact, the *RetiNet* architecture is a two-stage Retinex-inspired CNN that exploits gradient information in combination with the image formation loss. Actually, most of the traditional approaches follow the successful Retinex findings of using gradient separation [10–15]. In contrast to threshold-driven gradient separation, the goal of our network is to learn intrinsic gradients directly from data avoiding hard-coded thresholds. Further, for the re-integration process, we propose a series of simple convolutions to efficiently compute the intrinsic images separately. That is in contrast to other methods which try to find, by complex computations, the pseudo-inverse of an unconstrained system of derivatives, or to solve the Poisson equation.

Image gradients are calculated by taking the intermediate difference between neighboring pixels; horizontal (G_x) and vertical (G_y) separately. Finally, the gradient magnitude (G) is given as the square root of the sum of squares of the horizontal and the vertical components of the gradient:

$$G = \sqrt{{G_x}^2 + {G_y}^2} \tag{15}$$

This operation is carried out for each color channel individually. Then, the input is formed by concatenating the RGB image with its gradients per color channel, resulting in a 6 channel input. In this way, the network is assisted by image gradients. Finally, the encoder-decoder network is trained to separate color image gradients to intrinsic image gradients by using Equation (11):

$$\mathcal{L}_{S1} = \mathcal{L}_{CAL}(\nabla R, \nabla \hat{R}, \nabla S, \nabla \hat{S}), \tag{16}$$

where ∇ denotes the image gradient. For the first stage, we use the *IntrinsicNet* architecture described in the previous section. For the second stage, the input image is concatenated with the predicted intrinsic gradients this time. The newly formed input is provided to a fully convolutional sub-network to perform the actual decomposition by using Equation 13 with the intrinsic loss. Figure 4 illustrates our *RetiNet* model.

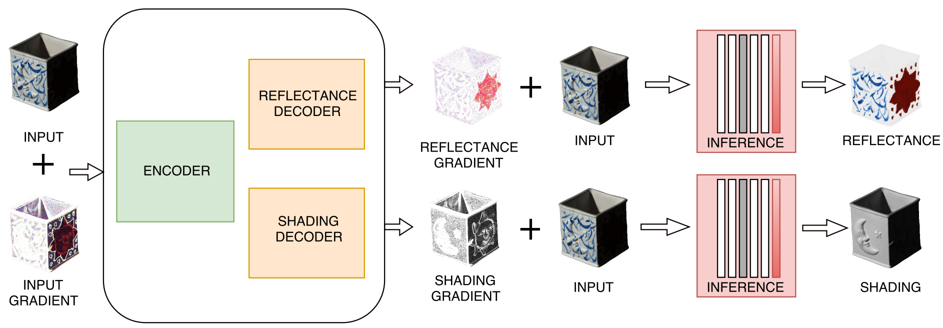

Figure 4: RetiNet model architecture. Refer to Figure 3 for layer types and encoder-decoder sub-network details. Instead of generating intrinsic image pixel values, the encoder-decoder network is trained to separate (color) image gradients into intrinsic image gradients. Then, for gradient re-integration part, the input image is concatenated with predicted intrinsic gradients and forwarded to a fully convolutional sub-network to perform the actual pixel-wise intrinsic image decomposition.

4.2 Experiments

4.2.1 New Synthetic Dataset of Man-made Objects

As mentioned in Section 4, for the experiments, a large scale synthetic dataset of man-made objects is created. We randomly sample around 20,000 3D models obtained from the ShapeNet dataset [16] for training. To create more variation and to decouple the correlation between image shape and texture, the texture of each component in a model is replaced by a random color. To enforce the lighting model, we apply a diffuse bidirectional scattering distribution function (BSDF) on the object surface with a random roughness parameter. The rendering is performed by the physics-based Blender Cycles[2]. Different environment maps are used to render the models at random viewpoints sampled from the upper hemisphere as conducted in [8]. To guarantee the relationship between reflectance and shading, the Cycles pipeline is modified to obtain the output image, its corresponding reflectance, and the shading map in high-dynamic range without gamma-correction. Since the images are taken from objects, the final dataset of around 20,000 images are object-centered. The object-centered dataset represent man-made objects. An overview of the datasets is given in Figure 2. Rendered with different environment maps and viewpoints, the dataset provides a variety of possible images in indoor and outdoor scenes.

4.2.2 Error Metrics

To evaluate our approach, metrics are chosen which are commonly used in the field. First, the results are evaluated in terms of the mean squared error (MSE) between the ground-truth intrinsic images and the measured ones. Following common practice, absolute brightness of each image is adjusted to minimize the error. Further, the local mean squared error (LMSE) [12] is chosen which is computed by aggregating the MSE scores over all local regions of size $k \times k$ with steps of $k/2$. Following the setup of [12], all the results in the evaluations use k = 20. The LMSE scores of the intrinsic images are averaged and normalized to make the maximum possible error equal to 1. Finally, to evaluate the perceptual visual quality of the results, the dissimilarity version of the structural similarity index (DSSIM) is taken, as done in [17].

4.3 Evaluation

4.3.1 Image Formation Loss

Figure 5 shows detailed views of a patch, demonstrating the benefits of the image formation loss. It can be derived that the image formation loss suppresses color artifacts and halo effects. Furthermore, Table 1 shows the quantitative evaluation results of our *IntrinsicNet* with and without the image formation loss ($\mathcal{L}_{IMF}$). The experiments on the MIT intrinsic benchmark [12] show that the image formation loss constrains the model to obtain improved color reproduction as expressed quantitatively by the DSSIM metric. In addition, the model with the image formation loss obtains better results for the MSE and LMSE metrics on average. On the ShapeNet test set, the model with the image formation loss achieves similar performance for MSE and LMSE. On DSSIM, it produces proper results for albedo prediction. Considering the generalization ability and the effect on a unseen real-world dataset, it can be observed that

[2]https://www.blender.org/

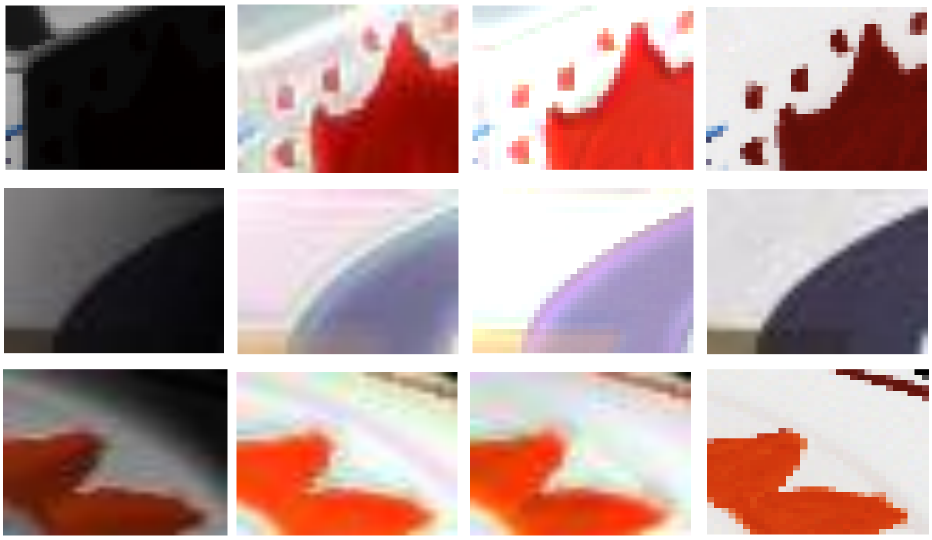

Figure 5: MIT intrinsic benchmark differentiated by the use of the image formation loss. IN (+/-) denotes the IntrinsicNet with/without the image formation loss. The image formation loss suppresses color artifacts and halo effects.

the network with image formation loss achieves best performance for all metrics. It shows the positive contribution of exploiting the image formation process as a constraining factor in a standard CNN approach for intrinsic image decomposition.

	MSE		LMSE		DSSIM	
	Albedo	Shading	Albedo	Shading	Albedo	Shading
*Without $\mathcal{L}_{IMF}$	**0.0045**	0.0062	0.0309	0.0326	0.0940	0.0704
*With $\mathcal{L}_{IMF}$	0.0051	**0.0029**	**0.0295**	**0.0157**	**0.0926**	**0.0441**
+Without $\mathcal{L}_{IMF}$	**0.0005**	**0.0007**	0.0300	**0.0498**	0.0075	**0.0082**
+With $\mathcal{L}_{IMF}$	**0.0005**	**0.0007**	**0.0297**	0.0505	**0.0072**	0.0084

Table 1: Evaluation results of the IntrinsicNet with and without image formation loss on the MIT intrinsic benchmark ($*$) and the ShapeNet test set ($+$). The image formation loss constrains the model to obtain better DSSIM performance. At the same time, it outperforms other models considering the MSE and LMSE metrics on real world images.

4.3.2 ShapeNet Dataset

We now test our models on the ShapeNet test partition. We follow the approach of [16] and randomly pick 1 image per test model, resulting in 7000 test images. For all experiments, the same test set is used. Table 2 provides the quantitative evaluation results of the synthetic test set of man-made objects. Figure 6 displays (visual) comparison results for a number of objects. Our proposed methods yield better results on the test set. Moreover, our *RetiNet* model outperforms all by a large margin. Visual comparison results show that all of our proposed models are capable of producing decent intrinsic image compositions on the test set.

	MSE		LMSE		DSSIM	
	Albedo	Shading	Albedo	Shading	Albedo	Shading
DirectIntrinsics [18]	0.1487	0.0505	0.6868	0.3386	0.0475	0.0361
ShapeNet [8]	0.0023	0.0037	0.0349	0.0608	0.0186	0.0171
IntrinsicNet	0.0005	0.0007	0.0297	0.0505	0.0072	0.0084
RetiNet	**0.0003**	**0.0004**	**0.0205**	**0.0253**	**0.0052**	**0.0064**

Table 2: Evaluation results on ShapeNet. Our proposed methods yield better results on the test set. Moreover, our RetiNet model outperforms all by a large margin.

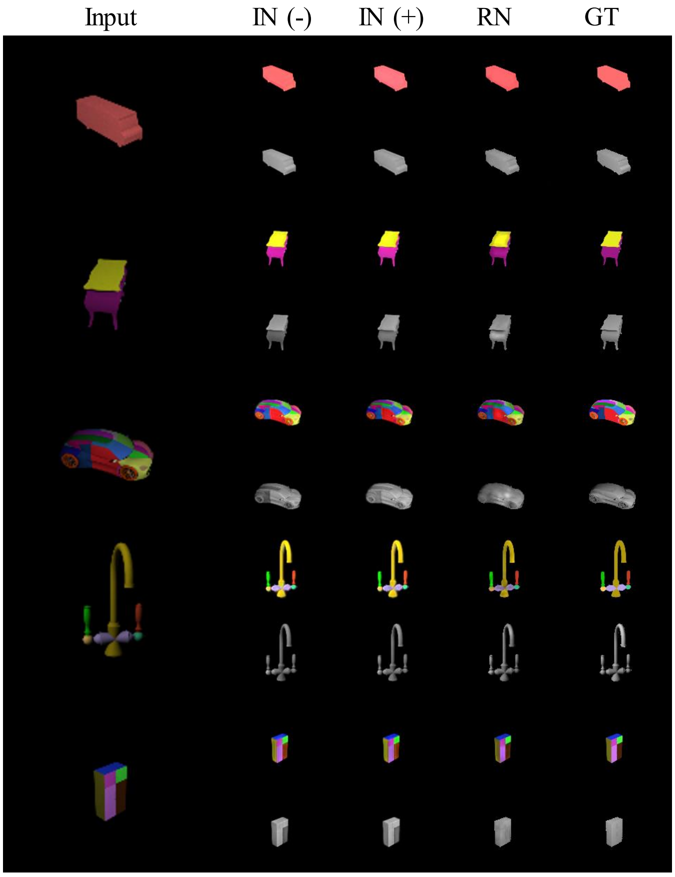

Figure 6: Evaluation results on the synthetic test set. All proposed models produce decent intrinsic image compositions. IN(+/-) denotes the IntrinsicNet with/without the image formation loss, and RN denotes the RetiNet model.

4.3.3 MIT Intrinsic Dataset

To assess our model on real world images, the MIT intrinsic image dataset [12] is used. The dataset consists of 20 object-centered images with a single canonical light source. Figure 8 displays (visual) results and Table 3 provides the numeric comparison to other state-of-the-art approaches. Our proposed methods yield better results compared with ShapeNet [8] and DirectIntrinsics [18] models. It can be derived that our proposed models properly recover the reflectance and shading information.

However, *IntrinsicNet* without the image formation loss generates color artifacts, and both

	MSE		LMSE		DSSIM	
	Albedo	Shading	Albedo	Shading	Albedo	Shading
Color Retinex [12]	**0.0032**	0.0348	0.0353	0.1027	0.1825	0.3987
DirectIntrinsics [18]	0.0277	0.0154	0.0585	0.0295	0.1526	0.1328
ShapeNet [8]	0.0468	0.0194	0.0752	0.0318	0.1825	0.1667
IntrinsicNet	0.0051	**0.0029**	**0.0295**	**0.0157**	0.0926	**0.0441**
RetiNet	0.0128	0.0107	0.0652	0.0746	0.0909	0.1054
RetiNet + GT∇	0.0072	0.0034	0.0429	0.0224	**0.0550**	0.0443

Table 3: Evaluation results on MIT intrinsic benchmark. Our proposed methods yield better results compared with other models. Experiment with intrinsic gradient ground-truths shows the benefits of exploiting them.

*IntrinsicNet*s create blurry results compared with *RetiNet*. In addition, if an image contains a strong shadow cast, as in the *deer* image, models struggle to eliminate it from the reflectance image. On the other hand, in *RetiNet* colors appear more vivid in the reflectance image and it suppresses most of the remaining color artifacts and blurriness that are present in *IntrinsicNet*s. Figure 7 displays a detailed analysis of *RetiNet*.

Finally, a small experiment is conducted to evaluate the edge performance of our methods. 100 patches were randomly selected around edges. These edge patches are tested by our two CNNs for patch level edge quality comparison. The results for the reflectance images are for *IntrinsicNet* vs. *RetiNet* MSE(0.0108 vs. **0.0093**), LMSE(0.0804 vs. **0.0595**), DSSIM(0.0977 vs. **0.0887**). That demonstrates the superior performance of *RetiNet* model on edge patches.

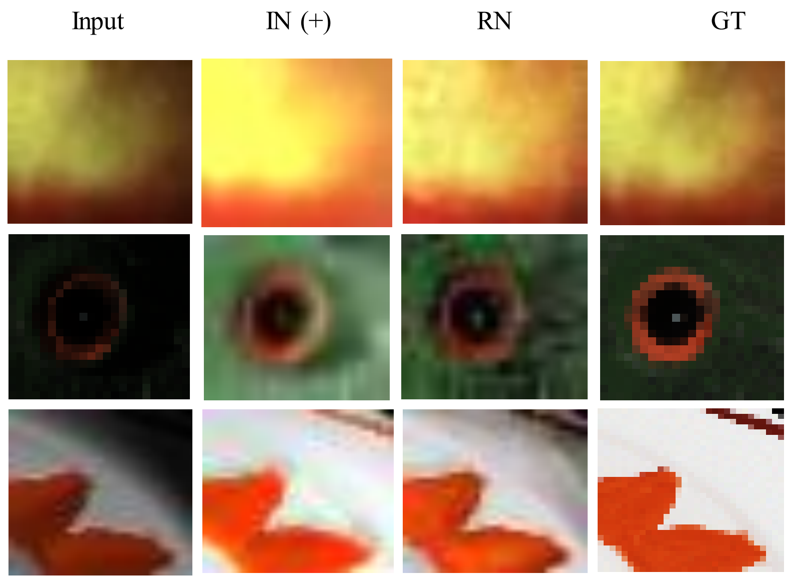

Figure 7: MIT intrinsic benchmark differentiated by the different models. IN(+) is the IntrinsicNet with the image formation loss, and RN denotes the RetiNet model (including the image formation loss). In RetiNet colors appear more vivid in the reflectance image and it suppresses most of the remaining color artifacts and blurriness that are present in IntrinsicNets.

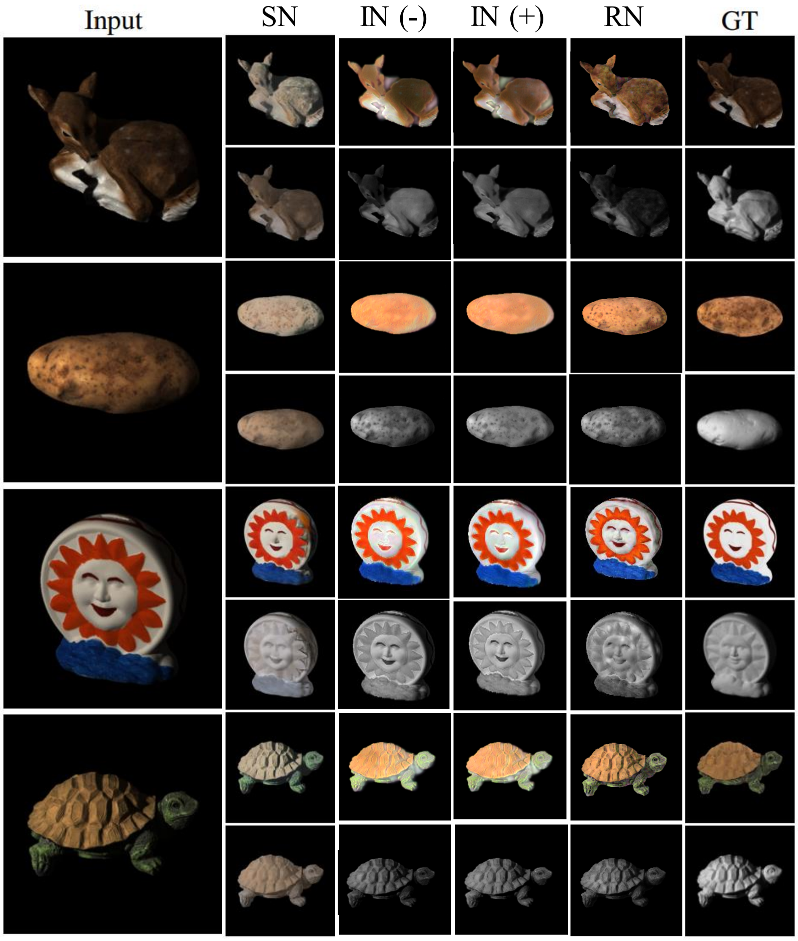

Figure 8: MIT intrinsic benchmark differentiated by the different models. SN is the ShapeNet model of [8], IN(+/-) denotes the IntrinsicNet with/without the image formation loss, and RN denotes the RetiNet model (including the image formation loss). Proposed models properly recover the reflectance and shading information. IntrinsicNet without the image formation loss generates color artifacts, and both IntrinsicNets create blurry results compared with RetiNet.

4.3.4 Real and In-the-wild Images

We also evaluate our *RetiNet* algorithm on real and in-the-wild images. Figure 9 shows the performance of our method for a number of images. The results show that it can capture proper reflectance image, free of shadings caused by geometry. Finally, we present the reconstructed input from its albedo and shading prediction to show that the decomposition is consistent.

4.4 Conclusion

We proposed two deep learning models considering a physics-based reflection model and gradient information to steer the learning process. The contributions of the research are as follows. 1: New is the physics-based image formation model in the design of the loss functions. 2: A novel, end-to-end solution is proposed to the well-known Retinex approach based on derivatives. 3: New is the gradient separation part of the *RetiNet* model in which albedo and shading gradients are learned using a CNN. 4: A (re)integration part is introduced where images are integrated

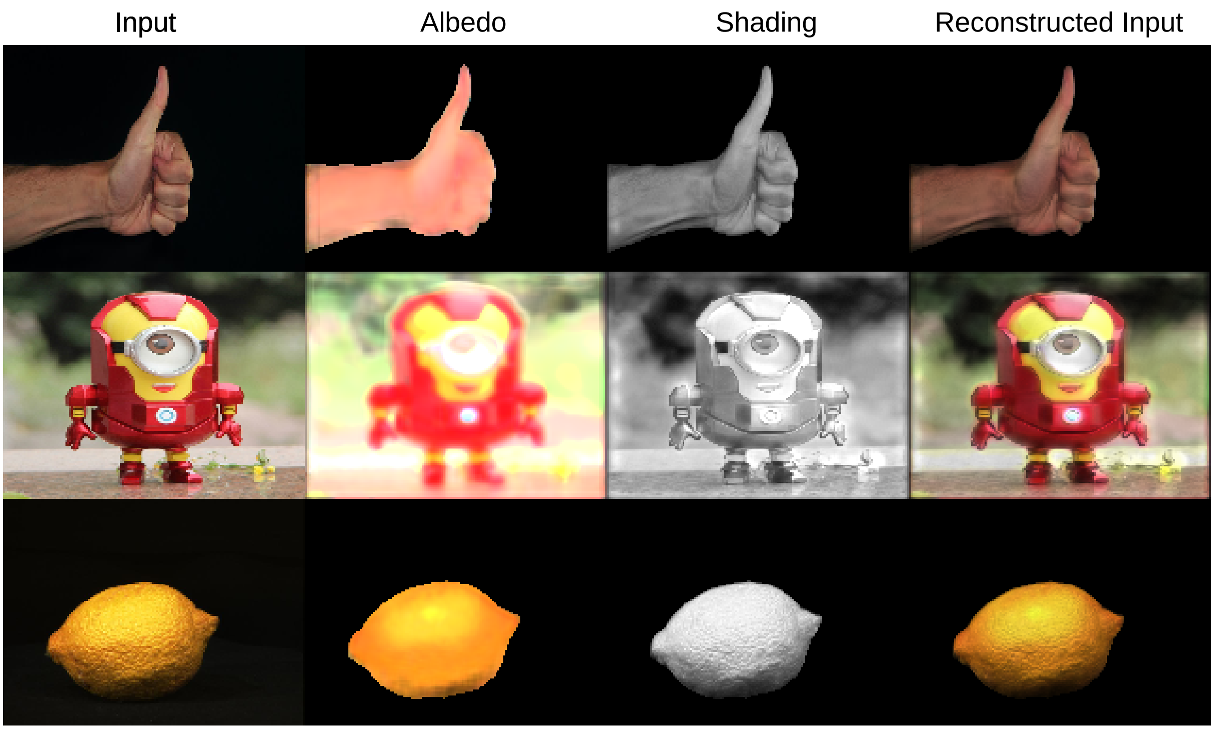

Figure 9: RetiNet applied on real images. It can capture proper albedo image, free of shadings due to geometry.

based on gradients by a set of simple convolutions. To train the models, an object centered large-scale synthetic dataset with intrinsic ground-truth images was created. Proposed models were evaluated on synthetic, real world and in-the-wild images. The evaluation results demonstrated that the new model outperforms existing methods. Furthermore, visual inspection showed that the image formation loss function augments color reproduction and the use of gradient information produces sharper edges. Additional details of the research can be found in our publication [19].

5 Joint Learning of Intrinsic Images and Semantic Segmentation

Semantic segmentation of outdoor scenes is a challenging problem in computer vision. Variations in imaging conditions may negatively influence the segmentation process. These varying conditions include shading, shadows, inter-reflections, illuminant color and its intensity. As image segmentation is the process of identifying and semantically grouping pixels, drastic changes in pixel values may hinder a successful segmentation. Current methods try to mitigate the effects of illumination artificially by hand crafted features. Therefore, they are limited in compensating for possible changes in photometry (i.e. illumination). Deep learning based methods may learn to accommodate photometric changes through data exploration. However, they are constrained by the amount of data such that it is not possible to cover all the variations caused by the illumination. On the other hand, albedo is invariant to all kinds of illumination effects. As a result, using albedo images for semantic segmentation task can be favorable, as they do not contain any illumination effect. Additionally, not only segmentation may benefit from reflectance, but also segmentation may be useful for reflectance computation. Information about an object reveals strong priors about its intrinsic properties. Each object label constrains the color distribution and is expected to reflect that property to class specific reflectance values. Therefore, distinct object labels provided by semantic segmentation can guide the intrinsic image decomposition process by yielding object specific color distributions per label. Furthermore, semantic segmentation process can act as an object boundary guidance map for intrinsic image decomposition by enhancing cues that differentiate between reflectance and occlusion edges in a scene. In addition, homogeneous regions (i.e. in terms of color) within an object segment should have similar reflectance values.

Therefore, in this research, the tasks of semantic segmentation and intrinsic image decomposition are considered as a combined process by exploring their mutual relationship in a joint fashion. Hence, instead of using narrow and specific invariant features, we focus on image formation invariance induced by a full intrinsic image decomposition.

To this end, we propose a supervised end-to-end CNN architecture to jointly learn intrinsic image decomposition *and* semantic segmentation. The joint learning includes an end-to-end trainable encoder-decoder CNN with one shared encoder and three separate decoders: one for reflectance prediction, one for shading prediction, and one for semantic segmentation prediction. In addition to joint learning, we explore new cascade CNN architectures to use reflectance to improve semantic segmentation, and semantic segmentation to steer the process of intrinsic image decomposition.

5.1 Approach

5.1.1 Image Formation Model

As done in the previous section, to formulate our intrinsic image decomposition Equation 4 is considered as the diffuse reflectance component. Additionally, when the light source is colored, it is also embedded in the shading component. Then, semantic segmentation is defined as the task of per pixel labeling of the *RGB* image.

5.1.2 Baseline Model Architectures

Intrinsic Image Decomposition. We use the model proposed by [8], $ShapeNet$, without the specular highlight module. The model is shown in the dotted rectangle part of Figure 10. The model provides state-of-the results for intrinsic image decomposition task. Early features in the encoder block are connected with the corresponding decoder layers, which are called *mirror links*. That proves to be useful for keeping visual details and producing sharp outputs. Furthermore, the features across the decoders are linked to each other (*inter-connections*) to further strengthen the correlation between the components. Additionally, the network is lightweight and easy to train. To train the model for intrinsic image decomposition task, we use a combination of the standard L_2 reconstruction loss (MSE) with its scale invariant version (SMSE). Let J be the prediction of the network and $\hat{J}$ be the ground-truth intrinsic image. The standard L_2 reconstruction loss $\mathcal{L}_{MSE}$ is given by Equation 10. Then, SMSE scales J first and compares MSE with $\hat{J}$:

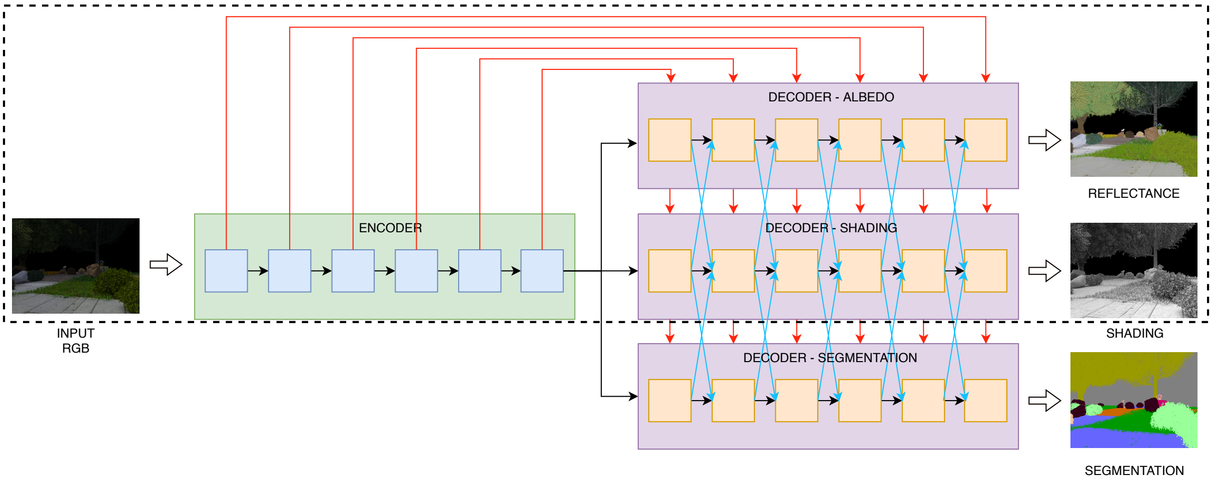

Figure 10: Model architecture for jointly solving intrinsic image decomposition and semantic segmentation with one shared encoder and three separate decoders: one for shading, one for reflectance, and one for semantic segmentation prediction. The part in the dotted rectangle denotes the baseline ShapeNet model of [8]. All the decoders are connected with each other.

$$\mathcal{L}_{SMSE}(J, \hat{J}) = \mathcal{L}_{MSE}(\alpha J, \hat{J}), \tag{17}$$

$$\alpha = argmin\ \mathcal{L}_{MSE}(\alpha J, \hat{J}). \tag{18}$$

Then, the combined loss $\mathcal{L}_{IL}$ for training an intrinsic component becomes:

$$\mathcal{L}_{IL}(J, \hat{J}) = \gamma_{SMSE}\ \mathcal{L}_{SMSE}(J, \hat{J}) + \gamma_{MSE}\ \mathcal{L}_{MSE}(J, \hat{J}), \tag{19}$$

where the γs are the corresponding loss weights. The final loss $\mathcal{L}_{FIL}$ for training the model for intrinsic image decomposition task becomes:

$$\mathcal{L}_{FIL}(R, \hat{R}, S, \hat{S}) = \gamma_R\ \mathcal{L}_{IL}(R, \hat{R}) + \gamma_S\ \mathcal{L}_{IL}(S, \hat{S}). \tag{20}$$

Semantic segmentation. The same architecture is used as the baseline for semantic segmentation task. However, one of the decoders is removed from the architecture, because there is only one task. As a consequence, inter-connection links are not used for the semantic segmentation task. Furthermore, as a second baseline, we train an off-the-shelf segmentation algorithm [20], $SegNet$, that is specifically engineered for semantic segmentation task.

To train the model for semantic segmentation, we use the cross entropy loss:

$$\mathcal{L}_{CE} = -\frac{1}{n}\sum_{\vec{x}}\sum_{L\in O_{\vec{x}}} log(p^L_{\vec{x}})\ , \tag{21}$$

where p is the output of the softmax function to compute the posterior probability of a given pixel $\vec{x}$ belonging to L^{th} class, where $L \in O_{\vec{x}}$ and $O_{\vec{x}} = \{0, 1, 2, \cdots, C\}$ as the category set for pixel level class label.

5.1.3 Joint Model Architecture

In this section, a new joint model architecture is proposed. It is an extension of the base model architecture for intrinsic image decomposition task, $ShapeNet$ [8], that combines the two tasks i.e. intrinsic image decomposition and semantic segmentation. We modify the baseline model architecture to have one encoder and three distinct decoders i.e. one for reflectance prediction, one for shading prediction, and one for semantic segmentation prediction. We maintain the mirror links and inter-connections. That allows for the network to be constrained with different outputs, and thus reinforce the learned features from different tasks. As a result, the network is forced to learn joint features for the two tasks at hand not only in the encoding phase, but also in the decoding phase. Both encoder and decoder parts contain both intrinsic properties and semantic segmentation characteristics. This setup is expected to be exploited by individual decoder blocks to learn extra cues for the task at hand. Figure 10 illustrates the joint model architecture. To train the model jointly, we combine the task specific loss functions by summing them together:

$$\mathcal{L}_{JL}(I, R, \hat{R}, S, \hat{S}) = \gamma_{CE}\, \mathcal{L}_{CE} + \gamma_{FIL}\, \mathcal{L}_{FIL}(R, \hat{R}, S, \hat{S}). \tag{22}$$

5.2 Experiments

5.2.1 The Synthetic Garden Dataset

For the experiments, we utilize the synthetic dataset introduced in Section 2 for garden-specialized training. We take a subset consisting of 35,000 images, which depicted 40 various gardens under 5 lighting conditions. For the experiments, the dataset is randomly split into 80% training and 20% testing (scene split).

5.2.2 Error Metrics

Error metrics introduced in Section 4.2.2 are used to evaluate intrinsic image decomposition tasks. For the semantic segmentation task, we report on global pixel accuracy, mean class accuracy and mean intersection over union (mIoU).

5.3 Evaluation

5.3.1 Influence of Reflectance on Semantic Segmentation

In this experiment, we evaluate the performance of reflectance and RGB color images as input for the semantic segmentation task. We train an off-the-shelf segmentation algorithm $SegNet$ [20] using (i) ground-truth reflectance ($Albedo - SegNet$) and (ii) RGB color images ($RGB - SegNet$); separately, and (iii) RGB + reflectance ($Comb. - SegNet$); together, as input. The results are summarized in Table 4 and illustrated in Figure 11. Further, confusion matrices for ($RGB - SegNet$) and ($Albedo - SegNet$) are provided in Figure 12.

The results show that semantic segmentation algorithm highly benefits from illumination invariant intrinsic properties (i.e. reflectance). The combination ($Comb. - SegNet$) outperforms single RGB input ($RGB - SegNet$). On the other hand, the results with reflectance as single input ($Albedo - SegNet$) are superior to the results with inputs including RGB color

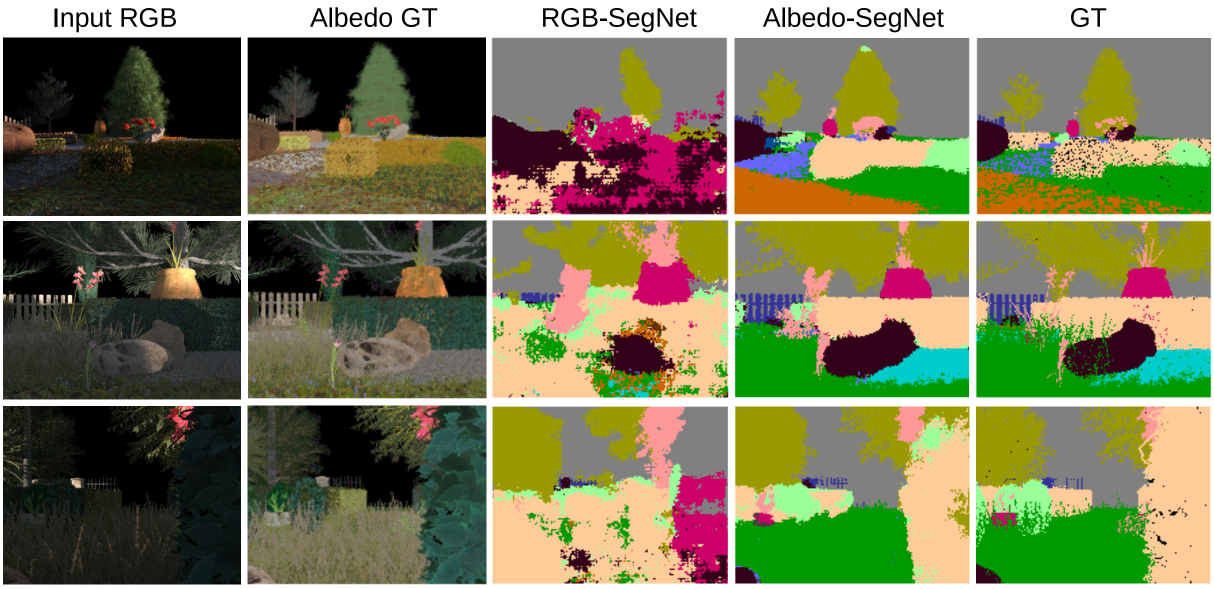

Figure 11: Qualitative evaluation of the influence of reflectance on semantic segmentation. The results show that the semantic segmentation algorithm highly benefits from illumination invariant intrinsic properties (i.e. reflectance).

Table 4: Semantic segmentation accuracy using albedo and RGB images as inputs. Using albedo images significantly outperforms RGB images.

Methodology	Global Pixel	Class Average	mIoU
$RGB - SegNet$	0.8743	0.6259	0.5217
$Comb. - SegNet$	0.8958	0.6607	0.5577
$Albedo - SegNet$	**0.9147**	**0.6739**	**0.5810**

images in all metrics. The combined input ($Comb. - SegNet$) is not better than using only reflectance ($Albedo - SegNet$), because the network may be negatively influenced by the varying photometric cues introduced by the RGB input. Although the CNN framework may learn, to a certain degree, illumination invariance, it is not possible to cover all the variations caused by the illumination. Therefore, a full illumination invariant representation (i.e. reflectance) helps the CNN to improve semantic segmentation performance. Moreover, the confusion matrices show that the network is unable to distinguish a number of classes based on RGB input. Using reflectance, the same network gains the ability to correctly classify the ground class, as well as making fewer mistakes with similar-looking box hedge and topiary classes.

5.3.2 Influence of Semantic Segmentation on Intrinsic Decomposition

In this experiment, we evaluate the performance of intrinsic image decomposition using ground-truth semantic segmentation labels as an extra source of information to the RGB images. We compare the performance of intrinsic image decomposition trained with RGB images (RGB) only as input and intrinsic decomposition trained with RGB images and ground-truth semantic segmentation labels ($RGB + SegGT$) together as their input. As for $RGB + SegGT$, four input channels (i.e. RGB color image and semantic segmentation labels) are provided

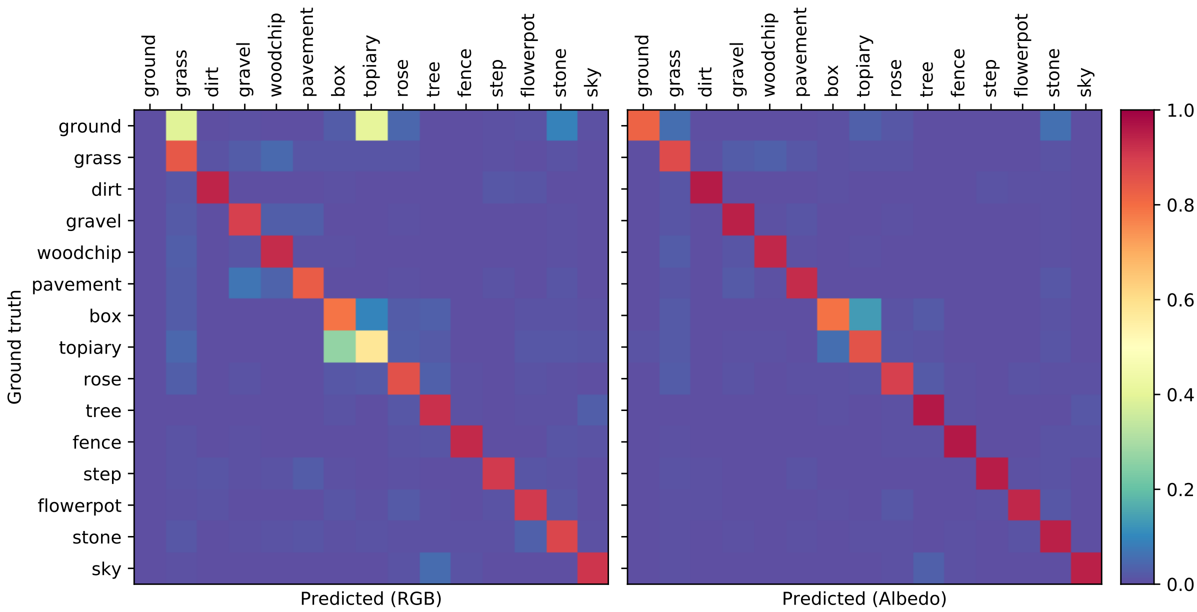

Figure 12: Confusion matrices for ($RGB-SegNet$) and ($Albedo-SegNet$). Using reflectance, the network gains the ability to correctly classify the ground class, as well as making fewer mistakes with similar-looking box and topiary classes.

Table 5: The influence of semantic segmentation on intrinsic property prediction. Providing segmentation as an additional input ($RGB + SegGT$) clearly outperforms the approach of using only RGB color images as their input.

	MSE		LMSE		DSSIM	
	Alb	Shad	Alb	Shad	Alb	Shad
RGB	0.0094 ± 0.008	0.0088 ± 0.0078	0.0679 ± 0.0412	0.0921 ± 0.0582	0.1310 ± 0.0535	**0.1303 ± 0.0495**
$RGB + SegGT$	**0.0076 ± 0.0063**	**0.0078 ± 0.0064**	**0.0620 ± 0.0384**	**0.0901 ± 0.0613**	**0.1141 ± 0.0472**	0.1312 ± 0.0523

as input. The results are summarized in Table 5. As shown in the table, intrinsic image decomposition clearly benefits from segmentation labels. $RGB+SegGT$ outperforms RGB in all metrics. DSSIM metric, accounting for the perceptual visual quality, shows the improvement on reflectance predictions, which indicates that the semantic segmentation process can act as an object boundary guidance map for reflectance prediction. A number of qualitative comparisons are shown for RGB and $RGB + SegGT$ in Figure 13.

5.3.3 Joint Learning of Semantic Segmentation and Intrinsic Decomposition

In this section, we evaluate the influence of joint learning on intrinsic image decomposition and semantic segmentation performances. We perform three experiments. First, we evaluate the effectiveness of joint learning of intrinsic properties and semantic segmentation considering semantic segmentation performance. Second, we evaluate the effectiveness of joint learning of intrinsic property and semantic segmentation to obtain intrinsic property prediction. Finally, we study the effects of the weights of the loss functions for the tasks.

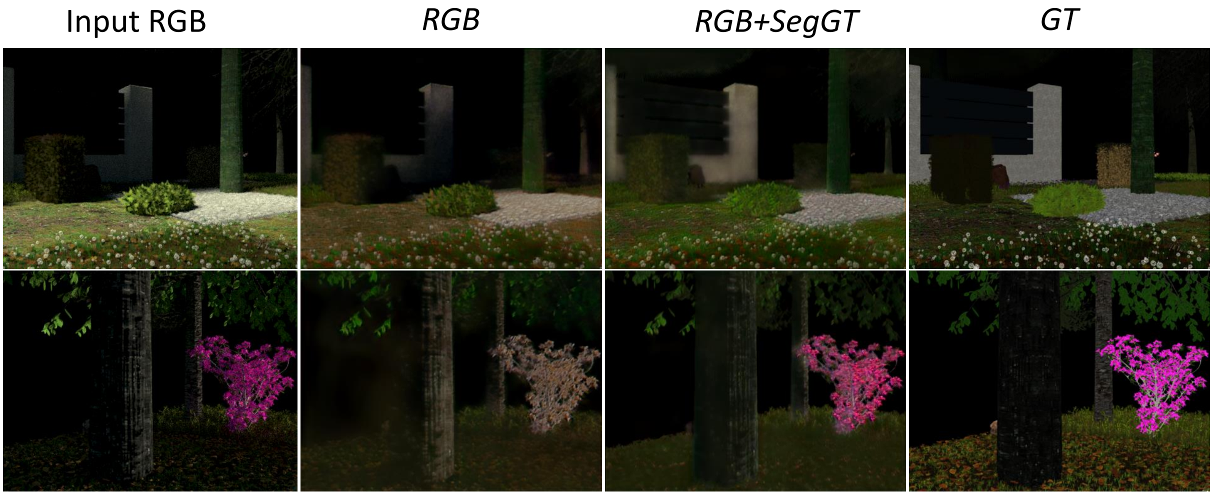

Figure 13: Columns 2 and 3 show that $RGB + SegGT$ is better in removing shadows and shading from the reflectance images, as well as preserving sharp object boundaries and vivid colors, and therefore is more similar to the ground truth.

Experiment I. In this experiment, we evaluate the performance of the proposed joint learning-based semantic segmentation algorithm ($Joint$), an off-the-shelf semantic segmentation algorithm [20] ($SegNet$) and the baseline of one encoder one decoder ShapeNet [8] ($Single$). All CNNs receive RGB color images as their input. $SegNet$ and $Single$ output only pixel level object class label predictions, whereas the proposed method predicts intrinsic property (i.e. reflectance and shading) in addition to the object class labels. We compare the accuracy of the models in Table 6.

Table 6: Comparison of the semantic segmentation accuracy. The proposed joint learning framework outperforms the single task frameworks in all metrics.

Methodology	Global Pixel	Class Average	mIoU
$Single$	0.8022	0.4584	0.3659
$SegNet$	0.8743	0.6259	0.5217
$Joint$	**0.9302**	**0.7055**	**0.6332**

As shown in Table 6, the proposed joint learning framework outperforms the single task frameworks in all metrics. Further, visual comparison between $SegNet$ and the proposed $joint$ framework is provided in Figure 14. By analyzing the 3rd and 4th row of the figure, it can be derived that unusual lighting conditions negatively influence the results of the $SegNet$. In contrast, our proposed method is not effected by varying illumination due to the joint learning scheme. Furthermore, our method preserves object shapes and boundaries when compared to the $SegNet$ model (rows 1, 2 and 5). Note that the joint network does not perform any additional fine-tuning operations (e.g. CRF etc.). Additionally, $SegNet$ architecture is deeper than our proposed model. However, our method still outperforms $SegNet$. Finally, the joint network outperforms the single task cascade network; for mIoU 0.6332 vs. 0.5810, see Table 4 and Table 6, as the joint scheme enforces to augment joint features. Finally, confusion matrices are provided in Figure 15. Confusion matrices show that the ability to distinguish close-color classes under different lighting conditions is further improved by joint learning. Similar to the case with using albedo as input for SegNet architecture, joint learning also improves the semantic segmentation performance significantly with certain classes. For the ground class, confusion is reduced remarkably by also learning intrinsics. Likewise, similar looking (in terms of shape and color) box and topiary classes are also better distinguished. In addition, most of the small confusions are eliminated.

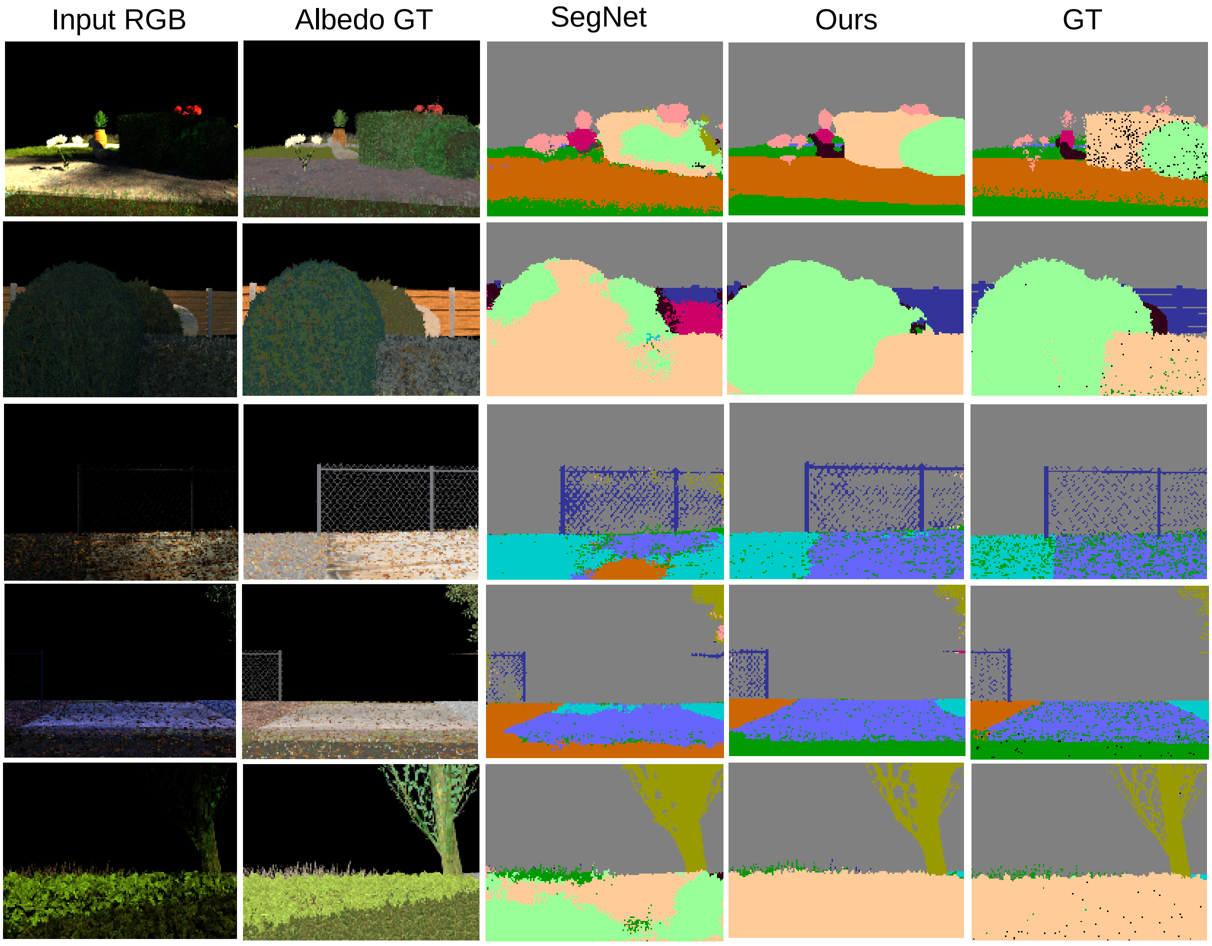

Figure 14: Proposed joint learning framework outperforms single task framework $SegNet$. Our method preserves the object shapes and boundaries better and is robust against varying lighting conditions

Experiment II. In this experiment, we evaluate the performance of the proposed joint learning-based and the state-of-the-art intrinsic image decomposition algorithms [8] ($ShapeNet$). Both CNNs receive RGB color images as input. $ShapeNet$ outputs only intrinsic properties (i.e. reflectance and shading), whereas the proposed method predicts pixel level object class labels as well as intrinsic properties. We train $ShapeNet$ and the proposed method using ground-truth reflectance and shading labels on the training set of the proposed dataset. We compare the accuracy of $ShapeNet$ and the proposed method in Table 7.

As shown in Table 7, the performance of the proposed joint learning framework outperforms single task learning ($ShapeNet$) in all the metrics for reflectance (albedo) and shading estimation. Further, our joint model obtains lower standard deviation values. To give more insight on reflectance prediction performances, a number of visual comparisons between $ShapeNet$

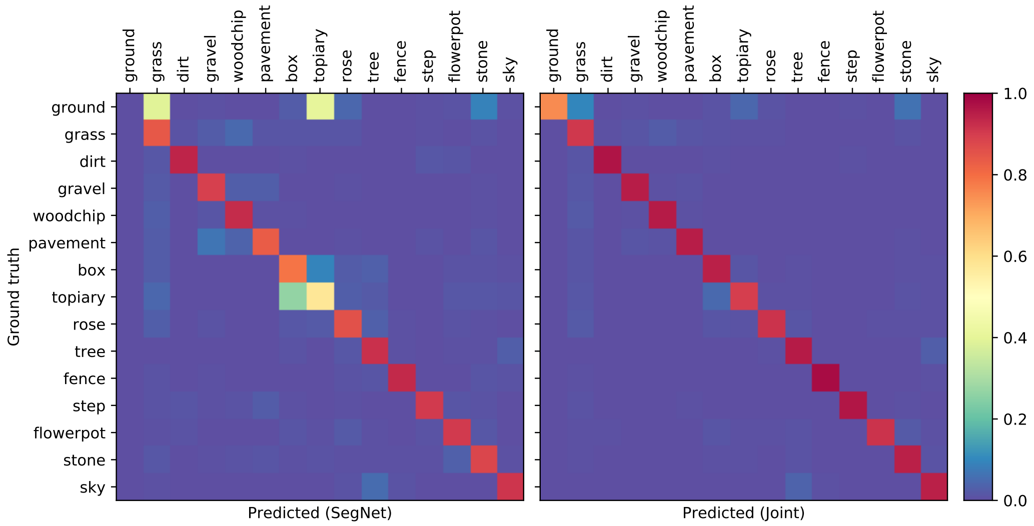

Figure 15: Confusion matrices for $SegNet$ and proposed $Joint$ model. Results suggest that close-color classes under different lighting conditions is further improved by joint learning and most of the small confusions are eliminated.

Table 7: Influence of joint learning on intrinsic property prediction

	MSE		LMSE		DSSIM	
	Alb	Shad	Alb	Shad	Alb	Shad
$ShapeNet$	0.0094 $\pm$ 0.0080	0.0088 $\pm$ 0.0078	0.0679 $\pm$ 0.0412	0.0921 $\pm$ 0.0582	0.1310 $\pm$ 0.0535	0.1303 $\pm$ 0.0495
Int.-Seg. Joint	**0.0030** $\pm$ **0.0040**	**0.0030** $\pm$ **0.0024**	**0.0373** $\pm$ **0.0356**	**0.0509** $\pm$ **0.0395**	**0.0753** $\pm$ **0.0399**	**0.0830** $\pm$ **0.0381**

and the proposed $joint$ framework are given in Figure 16. In the figure, (the first two columns) it can be derived that the semantic segmentation process acts as an object boundary guidance map for the intrinsic image decomposition task by enhancing cues to differentiate between reflectance and occlusion edges in a scene. Hence, object boundaries are better preserved by the proposed method (e.g. the separation between pavement and ground in the first image and the space between fences in the second image). In addition, information about an object reveals strong priors about it's intrinsic properties. Each object label adopts to a constrained color distribution. That can be observed in third and fourth columns. Semantic segmentation guides intrinsic image decomposition process by yielding the trees to be closer to green and flowers to be closer to pink. Moreover, for class-level intrinsics, the best improvement (3.3 times better) is obtained by *concrete step blocks*, which have achromatic colors. Finally, as in segmentation, the joint network outperforms the single task cascade network, see Table 5 and Table 7.

Experiment III. In this experiment, we study the effects of the weightings of the loss functions. As the cross entropy loss is an order of magnitude higher than the SMSE loss, we first normalize them by multiplying the intrinsic loss by 100. Then, we evaluate different weights on top of the normalization ($SMSE \times 100 \times w$). See Table 8 for the results. If higher weights are assigned to intrinsics, they both jointly increase. However, weights which are too high negatively influence the mIoU values. Therefore, $w = 2$ appears to be the proper setting for both tasks.

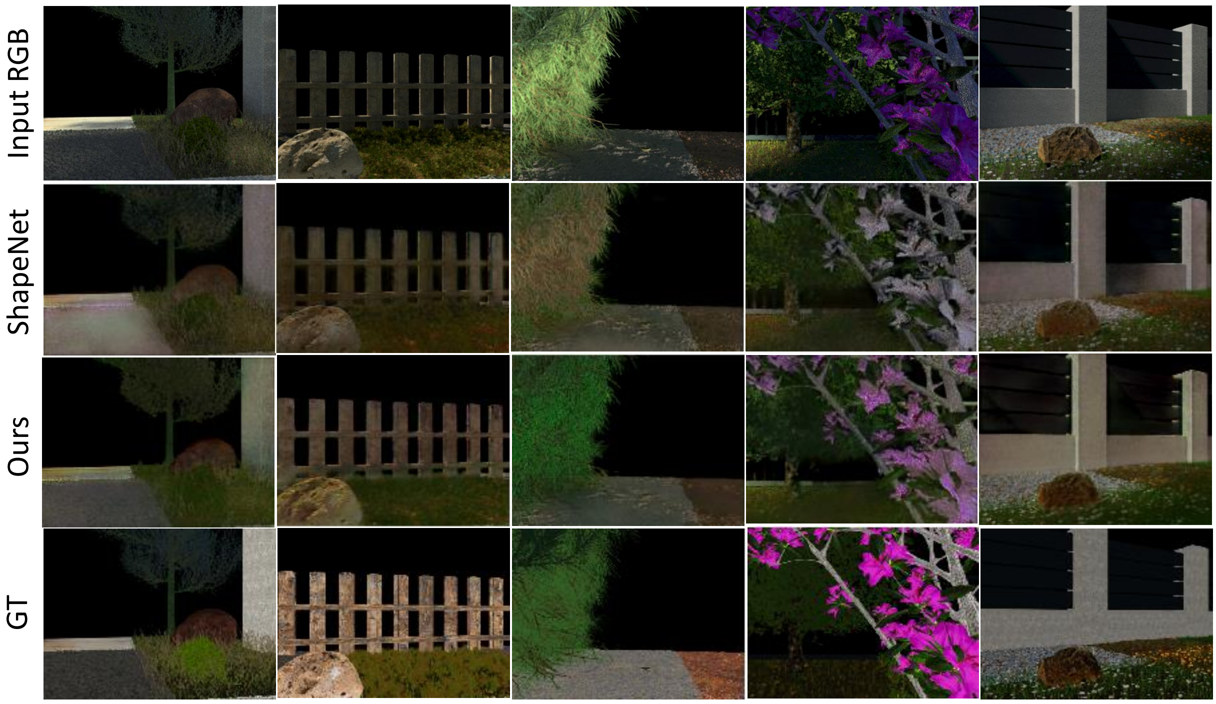

Figure 16: The first two columns illustrate that the proposed method provides sharper outputs especially at object boundaries than $ShapeNet$. The 3rd and 4th columns show that the proposed method predicts colours that are closer to the ground truth reflectance. The last column shows that the proposed method handles sharp cast shadows better than $ShapeNet$.

Table 8: Influence of the weighting of the loss functions. SMSE loss is weighted by ($SMSE \times 100 \times w$). $w = 2$ appears to be the proper setting for both tasks.

ω	Segmentation		MSE		LMSE		DSSIM	
	Global	mIoU	Alb	Shad	Alb	Shad	Alb	Shad
0.01	0.9179	0.567	0.0083 ± 0.0068	0.0083 ± 0.0072	0.0650 ± 0.0412	0.0920 ± 0.0611	0.1224 ± 0.0498	0.1343 ± 0.0545
0.5	0.7038	0.512	0.0038 ± 0.0037	0.0035 ± 0.0027	0.0398 ± 0.0311	0.0550 ± 0.0416	0.1633 ± 0.0538	0.1353 ± 0.0497
1	0.9048	0.533	0.0044 ± 0.0041	0.0044 ± 0.0036	0.0477 ± 0.0352	0.0655 ± 0.0474	0.0926 ± 0.0445	0.1040 ± 0.0421
2	0.9302	**0.633**	0.0030 ± 0.0040	0.0030 ± 0.0024	0.0373 ± 0.0356	0.0509 ± **0.0395**	0.0753 ± 0.0399	0.0830 ± 0.0381
4	**0.9334**	0.611	**0.0028** ± **0.3300**	**0.0028** ± **0.0023**	**0.0356** ± **0.0300**	**0.0491** ± 0.04081	**0.0716** ± **0.0380**	**0.0695** ± **0.0357**

5.3.4 Real and In-the-wild Images

Finally, our model is evaluated on real world garden images provided by the *3D Reconstruction meets Semantics challenge* [21]. The images are captured by a robot driving through a semantically-rich garden (Wageningen Trimbot test garden) with fine geometric details. Results of [8] are provided as a visual comparison on the performance in Figure 17. It shows that our method generates better results on real images with sharper reflectance images having more vivid and realistic colors. Moreover, our method mitigates sharp shadow effects better. Note that our model is trained fully on synthetic images and still provides satisfactory results on real, natural scenes. For semantic segmentation comparison, we fine-tuned SegNet [20] and our approach on the real world dataset after pre-training on the garden dataset. Since we only have the ground-truth for segmentation, we (only) unfreeze the segmentation branch. Results show that SegNet and our approach obtain 0.54 and 0.54 for mIoU and a global pixel accuracy of 0.85 and 0.88 respectively. Results are shown in Figure 18. For segmentation, the joint learning performs comparable to the baseline, yet we achieve sharper results. Note that our model is much smaller in size and predicts the intrinsics together with the segmentation.

Figure 17: Evaluation on real world garden images. We observe that our proposed method capture better colors and sharper outputs compared with [8].

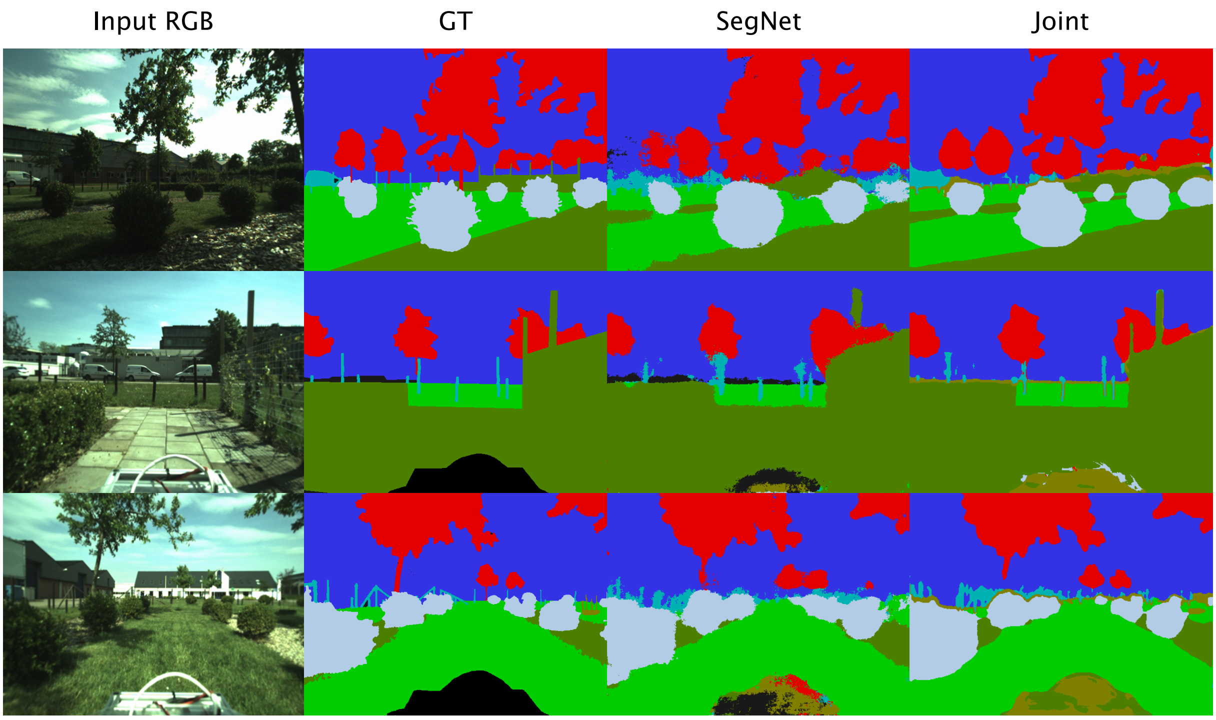

Figure 18: Semantic segmentation evaluation on real world garden images. The joint learning performs comparable to the baseline, yet achieves sharper results.

5.4 Conclusion

Our approach jointly learns intrinsic image decomposition and semantic segmentation. New CNN architectures are proposed for joint learning, and single intrinsic-for-segmentation and segmentation-for-intrinsic learning. The experiments show joint performance benefit when performing the two tasks (intrinsics and semantics) in a joint manner for natural scenes. Additional details of the research can be found in our publication [22].

6 ShadingNet: Image Intrinsics by Fine-Grained Shading Decomposition

In general, existing (traditional and new) intrinsic image decomposition algorithms assume that strong image variations are due to albedo changes and that smooth image variations are caused by shading. However, this assumption does not often hold for real images as they may suffer from strong photometric changes due to environmental conditions such as cast shadows and inter-reflections. As a consequence, existing methods may fail to correctly distinguish strong (cast) shadows from albedo variations. Confusing strong shadows with reflectance variations may negatively influence the quality of the resulting intrinsic image decomposition. For example, the *deer* image mentioned earlier in Section 4.3.3.

Therefore, instead of decomposing images into shading and reflectance only, we propose to decompose the shading component into three separate components to represent the different photometric effects.

To this end, the body term of the dichromatic reflection model [4] is extended to decompose the shading component into direct (light source) and indirect light conditions (ambient light and cast shadows). Using the fine-grained model, two different end-to-end deep convolutional neural networks (CNNs) are proposed. Further, to steer the deep learning models, surface normals are considered as an extra source of information. Surface normals are expected to assist (1) the shading prediction as they are part of the shading formation process and (2) the reflectance prediction as they are invariant to photometric effects.

Additionally, we extend a subset of the synthetic garden dataset to generate direct shading (shading due to surface geometry and illumination conditions), cast shadows, and ambient light (inter-reflections) maps.

6.1 Approach

6.1.1 Image Formation Model

We again use the diffuse (Lambertian) component of the dichromatic reflection model [4] as the basis of our image formation model, Equation 2. However, we redefine Equation 4 with new annotations which will be useful for further derivations:

$$I = \rho s_u \ , \tag{23}$$

where an image *I* can be modelled by a product of its unified shading $s_u = e(\vec{n}\cdot\vec{l})$ and reflectance ρ components. If the light source e is colored, then the color information is embedded in the illumination (shading) component. In general, in the context of intrinsic image decomposition, the shading component s_u is only defined for *direct* light (i.e. no occlusion) as follows:

$$s_d = e_d(\vec{n} \cdot \vec{l}) \ , \tag{24}$$

where e_d is the intensity of the light source. Obviously, Equation 24 does not include photometric effects such as ambient light or cast shadows. However, this assumption is often violated for real images. To compute intrinsic images, modelling these photometric effects may help to correctly distinguish strong (cast) shadows from albedo variations.

6.1.2 Image Formation Model with Composite Shading

To incorporate the photometric effects of ambient light and (cast) shadows, two terms are added to Equation 23:

$$I = \rho e_d(\vec{n} \cdot \vec{l}) + \rho e_a^+(\vec{n} \cdot \vec{l}) + \rho e_a^-(\vec{n} \cdot \vec{l}) \ , \tag{25}$$

where e_d is the intensity of direct light as defined by Equation 24. The indirect light $e_i = e_a^+ + e_a^-$ consists of ambient light, denoted by (e_a^+), resulting in an additive term. Shadows are modelled by a negative term e_a^-. Then, we obtain the composite shading model:

$$I = \rho s \ , \tag{26}$$

where the fine-grained shading component s distinguishes the three photometric effects:

$$s = e(\vec{n} \cdot \vec{l}) = (e_d + e_a^+ + e_a^-)(\vec{n} \cdot \vec{l}) \ , \tag{27}$$

where, $e = e_d + e_a^+ + e_a^-$ is the intensity of the composite lighting effects, e_d is the intensity of the light source (i.e. direct, non-occluded light), and indirect light (where direct light is occluded) is modelled by a combination of e_a^+ denoting the intensity of ambient light (e.g. inter-reflections) and e_a^- modeling the negative value of shadows. Ambient light (e_a^+) causes objects to appear brighter, whereas shadows (e_a^-) cause objects to appear dimmer.

6.1.3 ShadingNet

Using the image formation model of Equation 26, we propose two different modifications that can be applied to any regular encoder-decoder type CNN architecture that is designed for the standard intrinsic image decomposition task (simultaneous estimation of the intrinsics of Equation 23). Both modifications include end-to-end trainable encoder-decoder CNN models, *ShadingNet*s. First, we extend the shading decoder to contain multiple outputs for the photometric effects (intrinsic modification). Secondly, we extend the entire architecture by adding extra decoder blocks for each photometric effect (extrinsic modification). Figure 19 illustrates the different models. We show the effectiveness of the extensions by modifying $ShapeNet$ [8] which is a state-of-the-art architecture specifically engineered for intrinsic image decomposition. The model is shown in the dotted rectangle part of Figure 10.

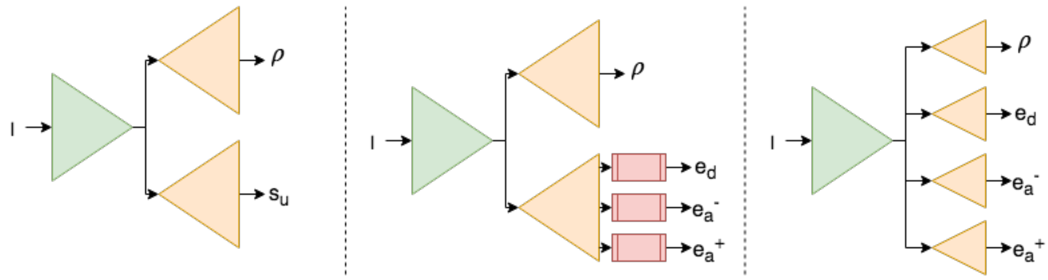

Figure 19: On the left, a standard encoder-decoder architecture (Equation 23), in the middle, intrinsic modification with $Squeeze - and - Excitation$ blocks [23], on the right, extrinsic modification with extra decoders. e_d denotes direct shading, e_a^- is for cast shadows and e_a^+ is for ambient light, s_u is for unified shading, and ρ is for albedo.

Intrinsic Modification. We extend the shading decoder to generate multiple outputs for the photometric effects. To that end, we extend the shading decoder module with *Squeeze-and-Excitation* blocks (SE) [23]. The motivation is that for the standard intrinsic image decomposition task, shading is taken as a single, unified component including all photometric effects. The shading decoder includes all shading features which can be further decomposed into the photometric effects that it is composed of (feature sharing). Therefore, we integrate SE blocks at the end of the shading decoder to perform feature re-calibration. By using SE blocks, predictions of the photometric cues are conditioned on one unified shading decoder that in return enhances feature discriminability.

Extrinsic Modification. We extend the entire architecture by adding extra decoder blocks per photometric effect. As a result, the architecture has 1 encoder and 4 distinct decoders; for albedo, direct shading, cast shadows and ambient light predictions. Unlike intrinsic modification, shading features are not shared within one decoder. In this way, the gradient flow from separate decoder blocks will boost the feature discriminability. Furthermore, we follow the design of *ShapeNet* and interconnect all the decoder blocks with each other. As a result, joint learning of features is reinforced.

6.1.4 Influence of Surface Normals

We use surface normals as an extra source of information. The surface normals are expected to assist (1) the shading prediction as they are part of the shading formation process and (2) the reflectance prediction as they are invariant to photometric effects. We explore two different ways to utilize surface normals as an input to a network. First, we use a single encoder with 6-channel input; RGB color image and surface normal ground-truths are concatenated and fed to the network (early fusion) for learning. Second, we use one separate encoder per input source. Then, the latent representations of both branches are combined to create a joint representation of the image (intermediate fusion).

6.2 Experiments

6.2.1 The Synthetic Garden Dataset

To train our models and baselines, we extend a subset of the synthetic garden dataset, around 30,000 images, to generate direct shading (shading due to surface geometry and illumination conditions), cast shadows, and ambient light (inter-reflections) ground-truth images. For the experiments, the dataset is randomly (scene) split resulting 15 gardens for training, around 25k images, and 3 gardens for testing, around 6k images. Figure 20 illustrates samples from the extended dataset.

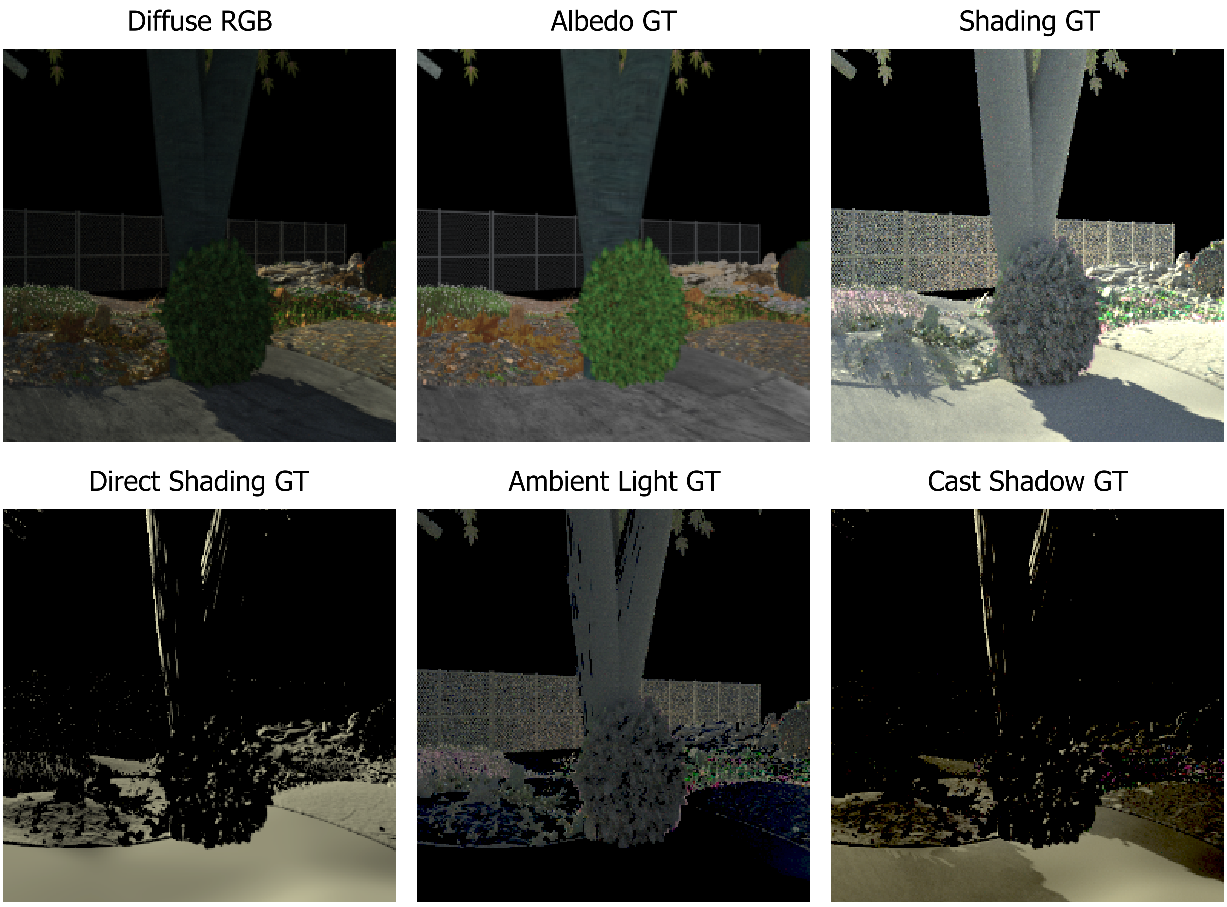

Figure 20: Sample images from the extended synthetic garden dataset with ground-truth intrinsics and fine-grained shading components.

6.2.2 Evaluation Metrics

Error metrics introduced in Section 4.2.2 are used to evaluate intrinsic image decomposition qualities; MSE, LMSE, and DSSIM. Moreover, we provide weighted human disagreement rate (WHDR) for IIW [24] evaluations.

6.3 Evaluation

6.3.1 Influence of Surface Normals

In this experiment, we evaluate the influence of surface normals as an extra source of information for intrinsic image decomposition. We train *ShapeNet* [8] and *IntrinsicNet* [19] architectures with modifications to the input branch by using NED. First, we use a single encoder with a 6-channel input; RGB color image and surface normal ground-truth image are concatenated and fed to the network (early fusion). Second, we use one separate encoder per input source. Then, the latent representations of both branches are combined to create a joint representation of the image (intermediate fusion). In this case, skip connections [25] are used for both encoders. We compare the results with the baselines that only use one branch encoder with RGB color images as input. All networks produce one reflectance map and one unified shading map. For the experiments, ground-truth surface normals are used. The results are summarized in Table 9.

ShapeNet [8]	MSE		LMSE		DSSIM	
	Albedo	Shading	Albedo	Shading	Albedo	Shading
RGB only	0.0053	0.0050	0.0597	0.0910	0.2516	**0.2186**
Early Fusion	0.0044	0.0056	0.0596	0.1102	0.328	0.3285
Intermediate Fusion	**0.0043**	**0.0035**	**0.0581**	**0.0854**	**0.2502**	0.2500

IntrinsicNet [19]	MSE		LMSE		DSSIM	
	Albedo	Shading	Albedo	Shading	Albedo	Shading
RGB only	0.0035	0.0037	0.0449	0.0791	0.2367	0.2110
Early Fusion	0.0027	0.0030	0.0358	0.0575	0.0967	0.1010
Intermediate Fusion	**0.0025**	**0.0027**	**0.0329**	**0.0528**	**0.0916**	**0.0883**

Table 9: Influence of surface normals on the intrinsic image decomposition predictions. Providing surface normals as additional inputs clearly outperforms the approach of using a single RGB color input. Intermediate fusing strategy appears to be the best approach.

Table 9 shows that intrinsic image decomposition highly benefits from surface normals as an additional input to our CNN models. Both early and intermediate fusion strategies outperform the single RGB color input. Assigning each input source with its own encoder (with skip connections) and own bottleneck appears to be a better option. This is because the network extracts and feed-forwards features independently to the decoder blocks. In this way, the decoder blocks can decide which cues are related to their specific task.

6.3.2 Intrinsic vs. Extrinsic Modification

In this experiment, we evaluate the architectural modifications. We train the *ShadingNet* architectures with; the *intrinsic modification*, where a single shading decoder contains multiple outputs for the photometric effects with SE blocks, and, the *extrinsic modification*, where photometric effects are estimated via individual decoder blocks. Ultimately, modifications with surface normal inputs using individual encoder blocks are presented. The results are summarized in Table 10.

	MSE		LMSE		DSSIM	
	Albedo	Shading	Albedo	Shading	Albedo	Shading
ShapeNet [8]	0.0053	0.0050	0.0597	0.0910	0.2516	**0.2186**
ShapeNet [8] + SN	0.0043	0.0035	0.0581	**0.0854**	0.2502	0.2500
Intrinsic M.	0.0048	0.0054	0.0585	0.1179	0.2757	0.3188
Extrinsic M.	0.0049	0.0049	0.0592	0.1157	0.2774	0.2732
Intrinsic M. + SN	0.0027	0.0048	0.0405	0.1211	0.1019	0.2361
Extrinsic M. + SN	**0.0018**	**0.0029**	**0.0365**	0.0890	**0.0812**	0.2200

Table 10: Influence of intrinsic and extrinsic modifications. SN denotes the model with surface normal encoder. Both modifications improve albedo estimations due to the presence of extra photometric cues.

The results show that albedo estimations are improved by the multi-varied shading model. Both intrinsic and extrinsic modifications improve the results on average. Thus, intrinsic image decomposition highly benefits from the fine-grained shading decomposition. Assigning each photometric effect to its own decoder appears to be the better option. Further, the contribution of surface normals is more significant than the Lambertian model. However, to form the shading component, the 3 photometric effects are combined. Thus, the quality of the shading component depends on the quality of the photometric effects combined. For the Lambertian model there is only one shading component estimated, which may explain the drop in shading results. In addition, the quality of the photometric effects and the influence of the surface normal are presented for both modifications in Table 11.

	$\text{MSE}_{intrinsic}$		$\text{MSE}_{extrinsic}$	
	SN (+)	SN (-)	SN (+)	SN (-)
Cast Shadow	**0.0217**	0.0274	0.0441	0.0455
Ambient Light	**0.0018**	0.0025	0.0036	0.0040
Direct Shading	0.0290	0.0381	**0.0141**	0.0302

Table 11: Quality of the direct and indirect shading effects. SN (+) denotes the model with surface normal encoder.

Table 11 shows that surface normals are essential and influence positively the quality of the fine-grained shading intrinsics. The intrinsic model appears to be the better option for photometric effects as indirect shading components are conditioned on one unified shading decoder. On the other hand, the extrinsic model with individual decoders are better for direct shading and reflectance predictions and appears to differentiate intrinsic cues better. We also provide additional qualitative results. First of all, we provide qualitative results of ShadingNet with intrinsic modification (IM) and extrinsic modification (EM) for fine-grained shading components. Provided results are from networks that use surface normal encoders. Figure 21 shows visual results for cast shadow predictions, Figure 22 provides visual results for direct shading predictions, and Figure 23 illustrates ambient light predictions.

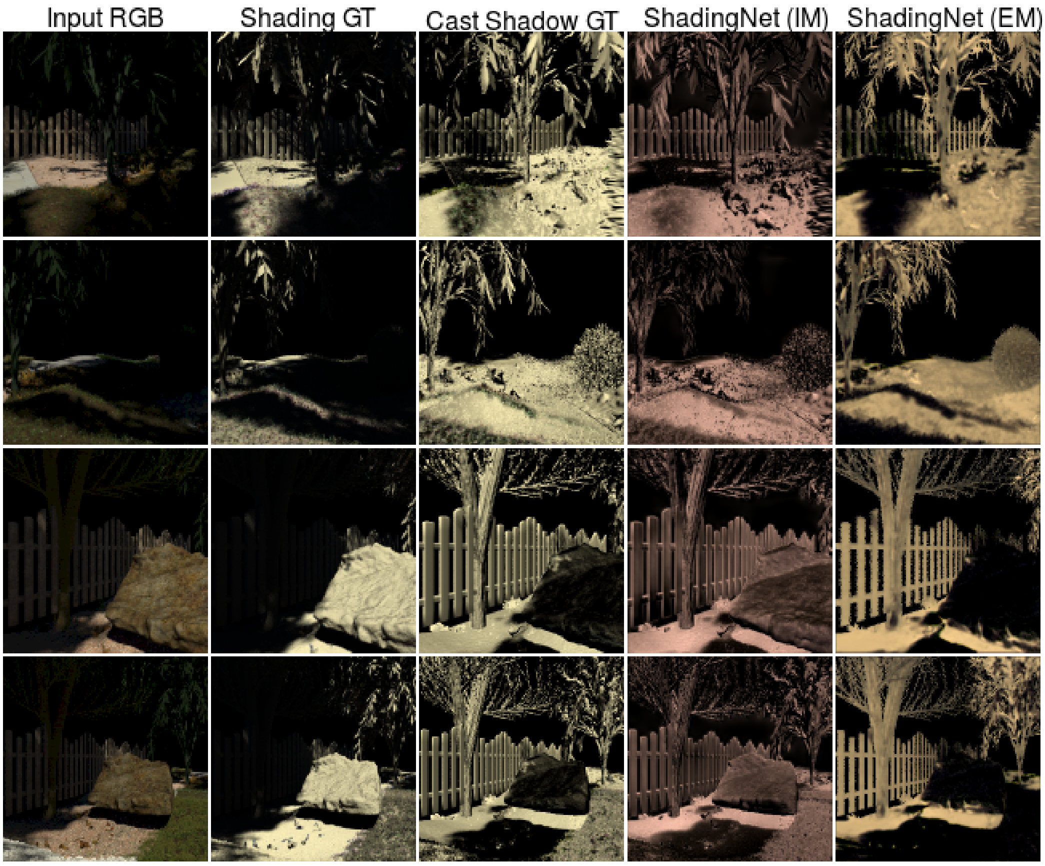

Figure 21: Qualitative results for cast shadow estimations. IM is for intrinsic modification, EM is for extrinsic modification. Both models can differentiate shadow cues.

6.3.3 Influence of Fine-Grained Shading

In this experiment, we evaluate the influence of the photometric components (indirect light) of the fine-grained shading term on the quality of intrinsic images. Following the promising results of the extrinsic modification (individual decoders per output), we train different versions of the *ShadingNet* architecture with modifications to the output decoder branches. First, we evaluate the photometric effects (e_i) as one unified component containing both ambient light (e_a^+) and cast shadows (e_a^-). Second, we train the *ShadingNet* architecture with the multiplicative image formation model of [17]. Both setups have 3 decoders; reflectance, direct shading (e_d) and composite indirect shading (e_i). Then, we provide results of the proposed model by having decoupled indirect shading terms. The results are summarized in Table 12.

	MSE		LMSE		DSSIM	
e_i	Albedo	Shading	Albedo	Shading	Albedo	Shading
Multiplicative [17]	0.0057	0.0292	0.0791	0.6989	**0.2195**	0.3943
Coupled	0.0060	0.0083	0.0638	0.1810	0.2877	0.2743
Decoupled	**0.0049**	**0.0049**	**0.0592**	**0.1157**	0.2774	**0.2732**

Table 12: Influence of fine-grained shading with extrinsic modification (extra decoders).

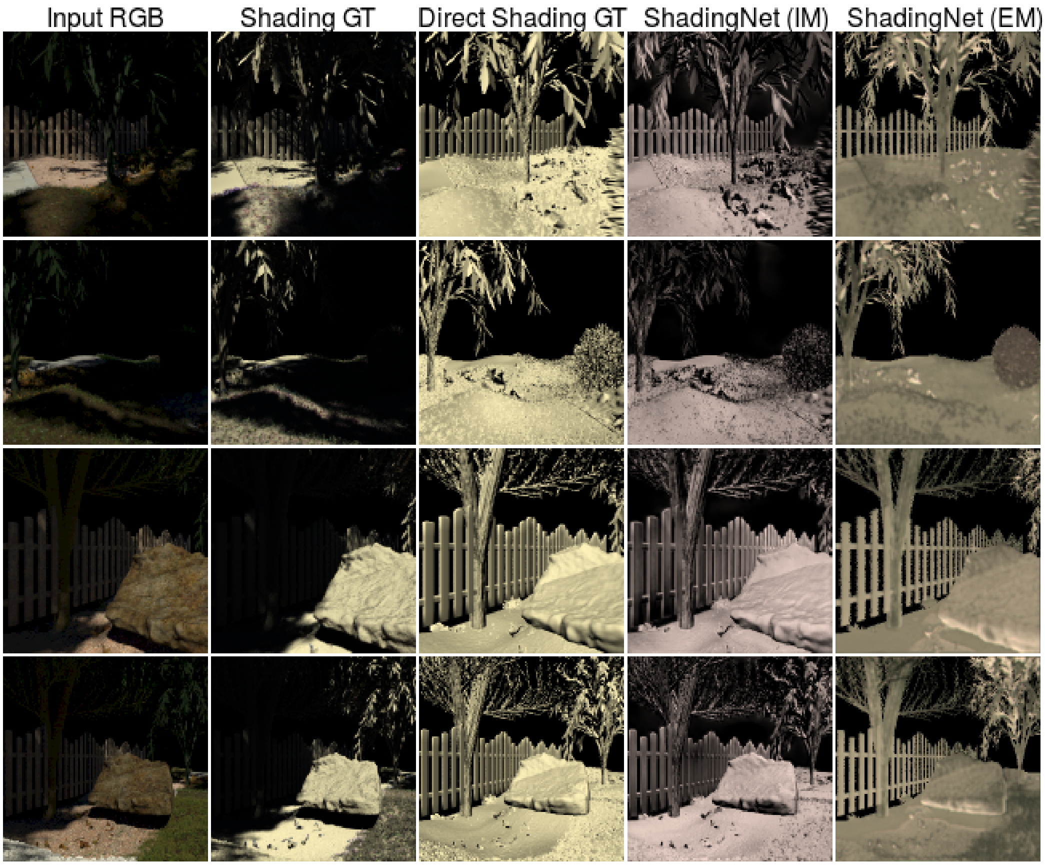

Figure 22: Qualitative results for direct shading estimations. IM is for intrinsic modification, EM is for extrinsic modification. Both models can differentiate direct shading cues. EM is closer to the ground-truth, while IM produce sharper results.

The results show that further decomposition of the indirect light (e_i) into ambient light (e_a^+) and cast shadows (e_a^-) shows to have great benefits over the coupled versions in all metrics. Further, our (additive) model yields better results than the (multiplicative) model of [17]. The reason their shading estimates are too noisy is that their multiplicative modeling yields unstable indirect shading values when direct shading values are very small. We do not further investigate the issue as it is out of the scope of this work.

6.4 Synthetic Outdoor Images

In this section, we compare our method with the color version of Retinex [12] a threshold-based traditional method, IIW [24] using optimization-based dense CRF, DirectIntrinsics [18] the pioneering coarse-to-fine multi-scale CNN network, IntrinsicNet [19] a standard CNN for the task with image formation loss and ShapeNet [8] with interconnections (all layers are connected to promote correlation between components). Further, we train our final model including the image formation loss (IMF), which also involves the shading formation process. All the models are trained on the same dataset (NED). Table 13 shows the quantitative evaluation results (6000 images) and Figure 24 displays visual comparison results for NED. In addition, Table 14 shows quantitative evaluation results (890 images) and for the MPI Sintel dataset. For MPI Sintel, we

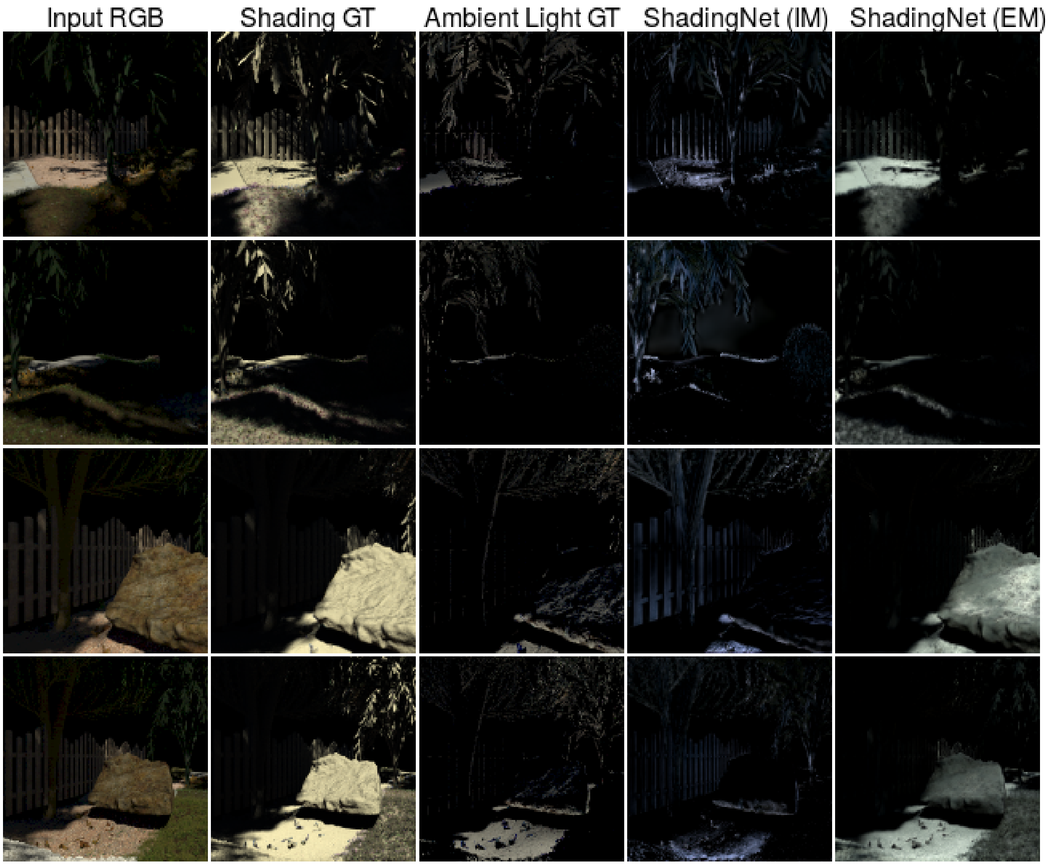

Figure 23: Qualitative results for ambient light estimations. IM is for intrinsic modification, EM is for extrinsic modification. Both models can differentiate ambient light cues from other photometric effects.

use surface normals generated by MarrRevisited [26].

Table 13 shows that our proposed models significantly outperforms other baselines on all metrics on the synthetic garden dataset. Table 14 demonstrates the good generalization ability of our models. Moreover, our models obtain better albedo results. For some metrics our models are on par or slightly worse, yet they predict fine-grained shading intrinsics with the albedo. Although (noisy) surface normals are estimated with an external network, it still helps to achieve better results. In addition, we support the findings of [19] such that the IMF loss constrains the model to obtain better results. Moreover, we show the contribution of colored shading by training and testing our final model on gray-scale shading, see the last row of Table 13. It can be derived that if the shading component is colored, it provides more cues to the decoders. Further, Figure 24 shows that ShadingNet with extrinsic modification (extra decoders) with surface normal encoder and image formation loss obtains better reflectance images. Moreover, our models removes cast shadows and shading leakage is minimal in the reflectance images.

6.4.1 In-the-wild Real World Indoor Images

Our method is also evaluated on in-the-wild real world images of IIW [24]. The dataset includes images of indoor scenes with complex lighting conditions. Figure 25 shows the performance of *ShadingNet* for a number of images. Table 15 show the mean WHDR results for IIW (randomly

	MSE		LMSE		DSSIM	
	Albedo	Shading	Albedo	Shading	Albedo	Shading
Color Retinex [12]	0.0114	0.0193	0.1204	0.2334	0.328	0.3515
IIW [24]	0.0095	0.0111	0.1343	0.1861	0.2098	0.3511
DirectIntrinsics [18]	0.0073	0.0065	0.1205	0.1798	0.3756	0.3843
IntrinsicNet [19]	0.0035	0.0037	0.0449	**0.0791**	0.2367	0.2110
ShapeNet [8]	0.0053	0.0050	0.0597	0.0910	0.2516	0.2186
ShadingNet (IM)	0.0048	0.0054	0.0585	0.1179	0.2757	0.3188
ShadingNet (EM)	0.0049	0.0049	0.0592	0.1157	0.2774	0.2732
ShadingNet (IM) + SN	0.0027	0.0048	0.0405	0.1211	0.1019	0.2361
ShadingNet (EM) + SN	0.0018	0.0029	0.0365	0.0890	0.0812	0.2200
ShadingNet (EM) + SN + IMF	**0.0016**	**0.0023**	**0.0326**	0.0840	**0.0751**	**0.1109**
ShadingNet (EM) + SN + IMF + GS	0.0027	0.0039	0.0407	0.1085	0.1010	0.2103

Table 13: Quantitative results for NED. SN denotes the model with surface normal encoder, IM is for intrinsic modification, EM is for extrinsic modification, IMF is for image formation loss, and GS is for gray-scale shading.

	MSE		LMSE		DSSIM	
	Albedo	Shading	Albedo	Shading	Albedo	Shading
Color Retinex [12]	0.0537	0.0617	0.0719	0.0665	0.2999	0.2646
IIW [24]	0.0371	0.0388	0.0720	0.0656	0.2673	0.2360
DirectIntrinsics [18]	0.0269	0.0315	0.0607	0.0943	0.3140	0.2895
IntrinsicNet [19]	0.0261	0.0334	0.0548	**0.0627**	**0.2047**	0.2112
ShapeNet [8]	0.0201	0.0371	0.0447	0.0767	0.2816	0.3132
ShadingNet (IM)	0.0207	0.0393	0.0459	0.0825	0.2711	0.2987
ShadingNet (EM)	0.0207	0.0393	0.0480	0.0960	0.2900	0.2465
ShadingNet (IM) + SN	**0.0194**	0.0300	0.0459	0.0656	0.2182	0.2277
ShadingNet (EM) + SN	0.0219	0.0314	0.0478	0.0665	0.2261	0.2010
ShadingNet (EM) + SN + IMF	0.0209	**0.0297**	**0.0458**	0.0706	0.2170	**0.1956**

Table 14: Quantitative results for Sintel. SN denotes the model with surface normal encoder, IM is for intrinsic modification and EM is for extrinsic modification, and IMF is for image formation loss.

picked 500 images). IIW dataset does not provide surface normals or depth data.

Figure 25 shows that despite the fact that our models are trained solely on synthetic outdoor data with single light source, they can capture proper reflectance image with smoothed out light effects. Table 15 shows that extrinsic modification is better than intrinsic modification and other baselines that predict unified shading. Note that the dataset is extremely challenging and very different from our settings as there are multiple light sources present in scenes that are also close in range (not point distant). Finally, we did not observe any improvement when surface normals are estimated by a network as in the case for IIW. The reason might be that, along with the complex light setting, the surface normal characteristics of indoor and outdoor are quite different. The current state-of-the-art method [27] of the dataset achieves 15% WHDR performance by combining 3 indoor datasets (synthetic and real, including IIW itself) with 8 different loss functions specifically engineered for the problem. Moreover, they post-process the results with a guided filter. On the other hand, we use a single outdoor synthetic dataset

Figure 24: Results for the synthetic garden dataset. IM is for intrinsic modification, EM is for extrinsic modification with surface normal encoder. Proposed models provide better reflectance components. EM produce better reflectance images with minimal shadow leakage.

with a single reconstruction loss without any post processing or fine-tuning.

	WHDR
DirectIntrinsics [18]	42 %
IntrinsicNet [19]	40 %
ShapeNet [8]	41 %
ShadingNet (IM)	43 %
ShadingNet (EM)	**37 %**

Table 15: Quantitative results for IIW. The lower the better for WHDR. SN denotes the model with surface normal encoder, IM is for intrinsic modification and EM is for extrinsic modification, and IMF is for image formation loss.

6.4.2 In-the-wild Real World Outdoor Images

Finally, our model is evaluated on real world garden images provided by the *3D Reconstruction meets Semantics challenge* [21] (Wageningen Trimbot test garden). The results are provided

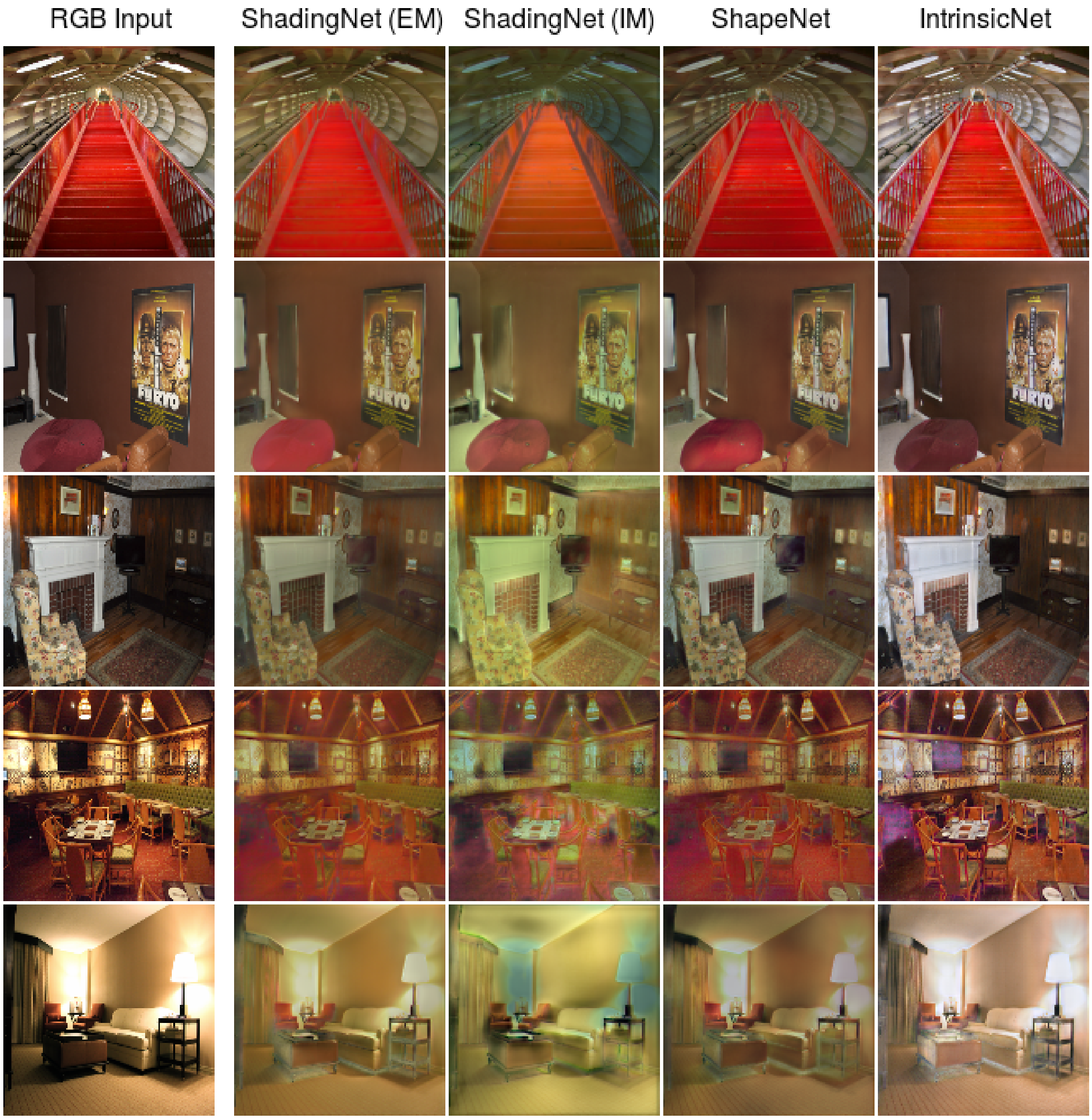

Figure 25: Reflectance prediction results for IIW dataset. Although our models are trained solely on synthetic outdoor data with a single light source, they properly capture the reflectance image.

in Figure 26. Results show that both intrinsic and extrinsic modifications can differentiate photometric cues and the shadow leakage is minimal.

6.5 Conclusion

We proposed to separate shading into different photometric effects such as shading caused by direct shading (object geometry) and indirect shading (shadows and ambient light) to improve intrinsic image decomposition results. Two end-to-end supervised CNN models, *ShadingNets*, were utilized to exploit the fine-grained shading model. Further, surface normals were considered as an input to the models and their contributions to the task were analyzed. The proposed models were evaluated on synthetic and real world in-the-wild images. The evaluation results show that intrinsic image decomposition highly benefits from (1) surface normals as an input to a CNN model and (2) the proposed fine-grained shading model. Our approaches outperform the existing unified shading methods. Moreover, visual inspection shows that the proposed method reduces the leakage of photometric effects in reflectance images.

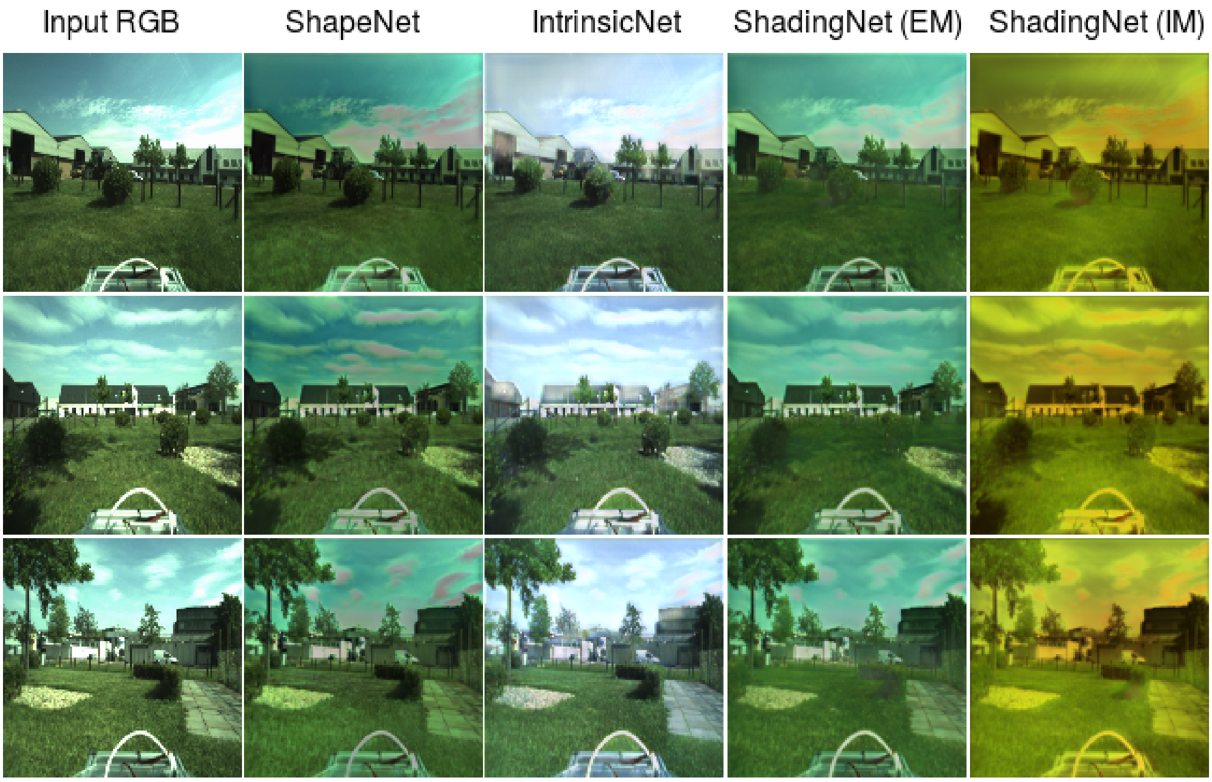

Figure 26: Qualitative results on real world garden images. IM is for intrinsic modification, EM is for extrinsic modification. EM produce better reflectance images with minimal shadow leakage.

7 Conclusion

In this report, we have described the intrinsic image decomposition algorithms that are currently used and developed as part of the Trimbot2020 project. We have shown algorithms based on image gradients and physics-based image formation models, joint learning with semantic segmentation, and fine-grained shading decompositions. All models can achieve proper intrinsic image decomposition and can directly be generalized to real world garden scenes without any fine-tuning.

References

[1] Barrow, H.G., Tenenbaum, J.M.: Recovering intrinsic scene characteristics from images. Computer Vision Systems (1978) 3–26

[2] Weber, J., Penn, J.: Creation and rendering of realistic trees. In: Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). (1995)

[3] Le, H.A., Baslamisli, A.S., Mensink, T., Gevers, T.: Three for one and one for three: Flow, segmentation, and surface normals. In: British Machine Vision Conference. (2018)

[4] Shafer, S.: Using color to separate reflection components. Color research and applications (1985) 210–218

[5] Land, E.H., McCann, J.J.: Lightness and retinex theory. Journal of Optical Society of America (1971) 1–11

[6] Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: European Conference on Computer Vision. (2012)

[7] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations. (2015)

[8] Shi, J., Dong, Y., Su, H., Yu, S.X.: Learning non-lambertian object intrinsics across shapenet categories. In: IEEE Conferance on Computer Vision and Pattern Recognition. (2017)

[9] Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T.: Demon: Depth and motion network for learning monocular stereo. In: IEEE Conferance on Computer Vision and Pattern Recognition. (2017)

[10] Bell, M., Freeman, W.T.: Learning local evidence for shading and reflectance. In: IEEE International Conference on Computer Vision. (2001)

[11] Gehler, P.V., Rother, C., Kiefel, M., Zhang, L., Schlkopf, B.: Recovering intrinsic images with a global sparsity prior on reflectance. In: Advances in Neural Information Processing Systems. (2011)

[12] Grosse, R., Johnson, M.K., Adelson, E.H., Freeman, W.T.: Ground truth dataset and baseline evaluations for intrinsic image algorithms. In: IEEE International Conference on Computer Vision. (2009)

[13] Shen, L., Tan, P., Lin, S.: Intrinsic image decomposition with non-local texture cues. In: IEEE Conferance on Computer Vision and Pattern Recognition. (2008)

[14] Tappen, M.F., Adelson, E.H., Freeman, W.T.: Estimating intrinsic component images using non-linear regression. In: IEEE Conferance on Computer Vision and Pattern Recognition. (2006)

[15] Tappen, M.F., Freeman, W.T., Adelson, E.H.: Recovering intrinsic images from a single image. In: Advances in Neural Information Processing Systems. (2003)

[16] Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)

[17] Chen, Q., Koltun, V.: A simple model for intrinsic image decomposition with depth cues. In: IEEE International Conference on Computer Vision. (2013)

[18] Narihira, T., Maire, M., Yu, S.X.: Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In: IEEE International Conference on Computer Vision. (2015)

[19] Baslamisli, A.S., Le, H.A., Gevers, T.: Cnn based learning using reflection and retinex models for intrinsic image decomposition. In: IEEE Conference on Computer Vision and Pattern Recognition. (2018)

[20] Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence (2017)

[21] Sattler, T., Tylecek, R., Brok, T., Pollefeys, M., Fisher, R.B.: 3d reconstruction meets semantics - reconstruction challange 2017. In: IEEE International Conference on Computer Vision Workshop. (2017)

[22] Baslamisli, A.S., Groenestege, T.T., Das, P., Le, H.A., Karaoglu, S., Gevers, T.: Joint learning of intrinsic images and semantic segmentation. In: European Conference on Computer Vision. (2018)

[23] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE Conference on Computer Vision and Pattern Recognition. (2018)

[24] Bell, S., Bala, K., Snavely, N.: Intrinsic images in the wild. ACM Trans. on Graphics (TOG) (2014)

[25] Mao, X., Shen, C., Yang, Y.: Image restoration using very deep fully convolutional encoder-decoder networks with symmetric skip connections. In: Advances in Neural Information Processing Systems. (2016)

[26] Bansal, A., Russell, B., Gupta, A.: Marr revisited: 2d-3d model alignment via surface normal prediction. In: IEEE Conference on Computer Vision and Pattern Recognition. (2016)

[27] Li, Z., Snavely, N.: Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In: European Conference on Computer Vision. (2018)

