# Joint Learning of Intrinsic Images and Semantic Segmentation

Anil S. Baslamisli[1], Thomas T. Groenestege[1,2], Partha Das[1,2], Hoang-An Le[1], Sezer Karaoglu[1,2], Theo Gevers[1,2]

[1]University of Amsterdam, [2]3DUniversum B.V.
{a.s.baslamisli, h.a.le, th.gevers}@uva.nl, s.karaoglu@3duniversum.com

**Abstract.** Semantic segmentation of outdoor scenes is problematic when there are variations in imaging conditions. It is known that albedo (reflectance) is invariant to all kinds of illumination effects. Thus, using reflectance images for semantic segmentation task can be favorable. Additionally, not only segmentation may benefit from reflectance, but also segmentation may be useful for reflectance computation. Therefore, in this paper, the tasks of semantic segmentation and intrinsic image decomposition are considered as a combined process by exploring their mutual relationship in a joint fashion. To that end, we propose a supervised end-to-end CNN architecture to jointly learn intrinsic image decomposition and semantic segmentation. We analyze the gains of addressing those two problems jointly. Moreover, new cascade CNN architectures for intrinsic-for-segmentation and segmentation-for-intrinsic are proposed as single tasks. Furthermore, a dataset of 35K synthetic images of natural environments is created with corresponding albedo and shading (intrinsics), as well as semantic labels (segmentation) assigned to each object/scene. The experiments show that joint learning of intrinsic image decomposition and semantic segmentation is beneficial for both tasks for natural scenes. Dataset and models are available at: https://ivi.fnwi.uva.nl/cv/intrinseg.

## 1 Introduction

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. To address this problem, several methods are proposed to mitigate the effects of illumination to obtain more robust image features to help semantic segmentation [1,2,3,4]. Unfortunately, these methods provide illumination invariance artificially by hand crafted features. Instead of using narrow and specific invariant features, in this paper, we focus on image formation invariance induced by a full intrinsic image decomposition.

Intrinsic image decomposition is the process of decomposing an image into its image formation components such as albedo (reflectance) and shading (illumination) [5]. The reflectance component contains the true color of objects in a scene. In fact, albedo is invariant to illumination, while the shading component heavily depends on object geometry and illumination conditions in a scene. As a result, using reflectance 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 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 paper, the tasks of semantic segmentation and intrinsic image decomposition are considered as a combined process by exploring their mutual relationship in a joint fashion.

To this end, we propose a supervised end-to-end convolutional neural network (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.

To train the proposed supervised network, a large dataset is needed with ground-truth images for both image semantic segmentation (i.e. class labels) and intrinsic properties (i.e. reflectance and shading). However, there is no such a dataset. Therefore, we have created a large-scale dataset featuring plants and objects under varying illumination conditions that are mostly found in natural environments. The dataset is at scene-level considering natural environments containing intrinsic image decomposition and semantic segmentation ground-truths. The dataset contains 35K synthetic images with corresponding albedo and shading (intrinsics), as well as semantic labels (segmentation) assigned to each object/scene.

Our contributions are: (1) a CNN architecture for joint learning of intrinsic image decomposition and semantic segmentation, (2) analysis on the gains of addressing those two problems jointly, (3) new cascade CNN architectures for intrinsic-for-segmentation and segmentation-for-intrinsic, and (4) a very large-scale dataset of synthetic images of natural environments with scene level intrinsic image decomposition and semantic segmentation ground-truths.

## 2 Related Work

**Intrinsic Image Decomposition.** Intrinsic image decomposition is an ill-posed and under-constrained problem since an infinite number of combinations of photometric and geometric properties of a scene can produce the same 2D image. Therefore, most of the work on intrinsic image decomposition considers priors about scene characteristics to constrain a pixel-wise optimization task. For instance, both [6] and [7] use non-local texture cues, whereas [8] and [9] constrain the problem with the assumption of sparsity of reflectance. In addition, the use of multiple images helps to resolve the ambiguity where the reflectance is constant and the illumination changes [10,11]. Nonetheless, with the success of supervised deep CNNs [12,13], more recent research on intrinsic image decomposition has shifted towards using deep learning. [14] is the first work that uses end-to-end trained CNNs to address the problem. They argue that the model should learn both local and global cues together with a multi-scale architecture. In addition, [15] proposes a model by introducing inter-links between decoder modules, based on the expectation that intrinsic components are correlated. Moreover, [16] demonstrates the capability of generative adversarial networks for the task. On the other hand, in more recent work, [17] considers an image formation loss together with gradient supervision to steer the learning process to achieve more vivid colors and sharper edges.

In contrast, our proposed method jointly learns intrinsic properties and segmentation. Additionally, the success of supervised deep CNNs not only depends on a successful model, but also on the availability of annotated data. Generating ground-truth intrinsic images is only possible in a fully-controlled setup and it requires enormous effort and time [18]. To that end, the most popular real-world dataset for intrinsic image decomposition includes only 20 object-centered images with their ground-truth intrinsics [18], which alone is not feasible for deep learning. On the other hand, [19] presents scene-level real world relative reflectance comparisons over point pairs of indoor scenes. However, it does not include ground-truth intrinsic images. The most frequently used scene-level synthetic dataset for intrinsic image decomposition is the MPI Sintel Dataset [20]. It provides around a thousand of cartoon-like images with their ground-truth intrinsics. Therefore, a new dataset is created consisting of 35K synthetic (outdoor) images with 16 distinct object types/scenes which are recorded under different illumination conditions. The dataset contains intrinsic properties and object segmentation ground-truth labels. The dataset is described in detail in the experimental section.

**Semantic Segmentation.** Traditional semantic segmentation methods design hand-crated features to achieve per-pixel classification with the use of an external classifier such as support vector machines [21,22,23]. On the other hand, contemporary semantic segmentation methods such as [24,25,26] benefit from the powerful CNN models and large-scale datasets such as [27,28]. A detailed review on deep learning techniques applied to semantic segmentation task can be found in [29].

Photometric changes, which are due to varying illumination conditions, cause changes in the appearance of objects. Consequently, these appearance changes create problems for the semantic segmentation task. Therefore, several methods are proposed to mitigate the effects of varying illumination to accomplish a more robust semantic segmentation by incorporating illumination invariance in their algorithms [1,2,3,4]. However, these methods provide invariance 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. In this paper, we propose to use the intrinsic reflectance property (i.e. fully illumination invariance) to be used for semantic segmentation.

**Joint Learning.** Semantic segmentation has been used for joint learning tasks as it provides useful cues about objects and scenes. For instance, [30,31,32] propose joint depth prediction and semantic segmentation models. Joint semantic segmentation and 3D scene reconstruction is proposed by [33]. Furthermore, [34] formulates dense stereo reconstruction and semantic segmentation in a joint framework.

For intrinsic image decomposition, [35] introduces the first unified model for recovering shape, reflectance, and chromatic illumination in a joint optimization framework. Other works [36,37], jointly predict depth and intrinsic property. Finally, [38] exploits the relation between the intrinsic property and objects (i.e. attributes and segments). The authors propose to address these problems in a joint optimization framework. Using hand crafted priors, [38] designs energy terms per component and combines them in one global energy to be minimized. In contrast to previous methods, our proposed method is an end-to-end solution and does not rely on any hand crafted priors. Additionally, [38] does not optimize their energy function for each component separately. Therefore, the analysis on the influence of intrinsic image decomposition on semantic segmentation is omitted. In this paper, an in-depth analysis for each component is given.

## 3 Approach

### 3.1 Image Formation Model

To formulate our intrinsic image decomposition, the diffuse reflectance component is considered [39]. Then, an $RGB$ image, $I$, over the visible spectrum $\omega$, is defined by:

$$I = m_b(\boldsymbol{n}, \boldsymbol{s}) \int_{\omega} f_c(\lambda)\, e(\lambda)\, \rho_b(\lambda)\, \mathrm{d}\lambda. \tag{1}$$

In the equation, $\boldsymbol{n}$ denotes the surface normal, whereas $\boldsymbol{s}$ is the light source direction; together forming the geometric dependencies $m$, which in return forms the shading component $S(\boldsymbol{x})$ under white light. Additionally, $\lambda$ represents the wavelength, $f_c(\lambda)$ is the camera spectral sensitivity, $e(\lambda)$ specifies the spectral

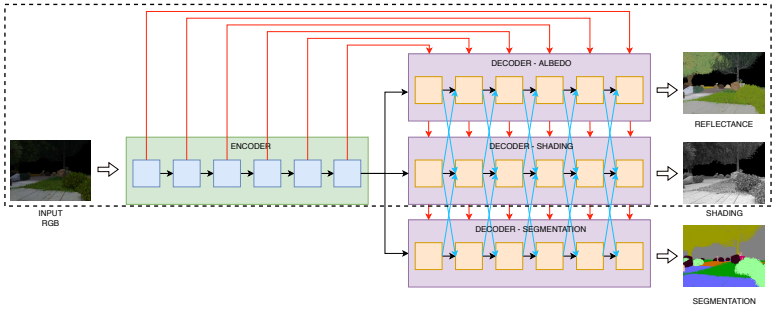


**Fig. 1.** Model architecture for jointly solving intrinsic image decomposition and semantic segmentation with one shared encoder and three separate decoders: one for shading, one for reflectance, and one for semantic segmentation prediction. The part in the dotted rectangle denotes the baseline ShapeNet model of [15].

power distribution of the illuminant, and $\rho_b$ represents the diffuse surface reflectance $R(\boldsymbol{x})$. Then, using narrow band filters and considering a linear sensor response under white light, intrinsic image decomposition can be formulated as:

$$I(\boldsymbol{x}) = R(\boldsymbol{x}) \times S(\boldsymbol{x}). \tag{2}$$

Then, for a position $\boldsymbol{x}$, $I(\boldsymbol{x})$ can be approximated by the element-wise product of its intrinsic components. When the light source is colored, it is also included in the shading component.

### 3.2 Baseline Model Architectures

**Intrinsic Image Decomposition.** We use the model proposed by [15], *ShapeNet*, without the specular highlight module. The model is shown in the dotted rectangle part of Figure 1. 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.

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. Then, the standard $L_2$ reconstruction loss $\mathcal{L}_{MSE}$ is given by:

$$\mathcal{L}_{MSE}(J, \hat{J}) = \frac{1}{n} \sum_{\boldsymbol{x},c} ||\hat{J} - J||_2^2, \tag{3}$$

where $\boldsymbol{x}$ denotes the pixel coordinate, $c$ is the color channel index and $n$ is the total number of evaluated pixels. Then, SMSE scales $J$ first and compares MSE with $\hat{J}$:

$$\mathcal{L}_{SMSE}(J, \hat{J}) = \mathcal{L}_{MSE}(\alpha J, \hat{J}), \tag{4}$$

$$\alpha = argmin\ \mathcal{L}_{MSE}(\alpha J, \hat{J}). \tag{5}$$

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

$$\mathcal{L}_{CL}(J, \hat{J}) = \gamma_{SMSE}\ \mathcal{L}_{SMSE}(J, \hat{J}) + \gamma_{MSE}\ \mathcal{L}_{MSE}(J, \hat{J}), \tag{6}$$

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

$$\mathcal{L}_{IL}(R, \hat{R}, S, \hat{S}) = \gamma_R\ \mathcal{L}_{CL}(R, \hat{R}) + \gamma_S\ \mathcal{L}_{CL}(S, \hat{S}). \tag{7}$$

**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 [24], *SegNet*, that is specifically engineered for semantic segmentation task.

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

$$\mathcal{L}_{CE} = -\frac{1}{n} \sum_{\boldsymbol{x}} \sum_{L \in O_{\boldsymbol{x}}} log(p^L_{\boldsymbol{x}})\ , \tag{8}$$

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

### 3.3 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* [15], 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 1 illustrates

the joint model architecture. To train the model jointly, we combine the task specific loss functions by summing them together:

$$\mathcal{L}_{JL}(I, R, \hat{R}, S, \hat{S}) = \gamma_{CE}\, \mathcal{L}_{CE} + \gamma_{IL}\, \mathcal{L}_{IL}(R, \hat{R}, S, \hat{S}). \tag{9}$$

The effect of the gamma parameters of Equation 6 and more implementation details can be found in the supplementary materials.

## 4 Experiments

### 4.1 New Synthetic Dataset of Natural Environments

A large set of synthetic images is created featuring plants and objects that are mostly found in natural environments such as parks and gardens. 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 [40] (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 park/garden, we captured the scene from different perspectives with motion blur effects. Scene are rendered with the physics-based Blender Cycles[1] engine. To obtain annotations, the rendering pipeline is modified to output $RGB$ images, their corresponding albedo and shading profiles (intrinsics) and semantic labels (segmentation). The dataset consists of 35K images, depicted 40 various parks/gardens under 5 lighting conditions. A number of samples are shown in Figure 2. For the experiments, the dataset is randomly split into 80% training and 20% testing (scene split).

### 4.2 Error Metrics

To evaluate our method for intrinsic image decomposition task, we report on mean squared error (MSE), its scale invariant version (SMSE), local mean squared error (LMSE), and dissimilarity version of the structural similarity index (DSSIM). DSSIM accounts for the perceptual visual quality of the results. Following [18], for MSE, the absolute brightness of each image is adjusted to minimize the error. Further, $k = 20$ is used for the window size of LMSE. For semantic segmentation task, we report on global pixel accuracy, mean class accuracy and mean intersection over union (mIoU).

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

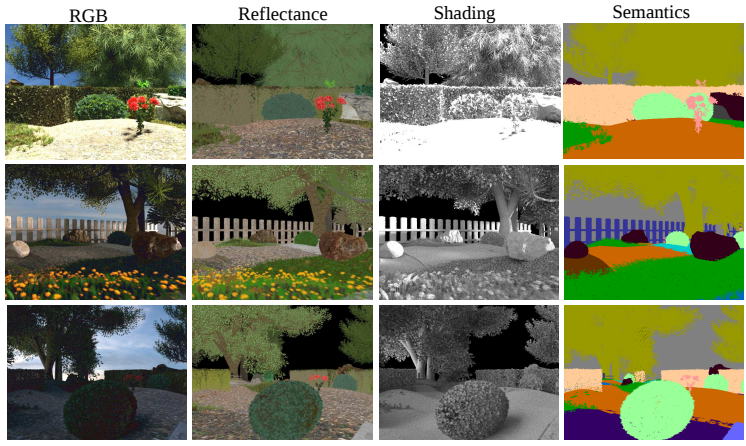


**Fig. 2.** Sample images from the Natural Environment Dataset (NED) featuring plants and objects under varying illumination conditions with ground-truth components

## 5 Evaluation

### 5.1 Influence of Reflectance on Semantic Segmentation

In this experiment, we evaluate the performance of reflectance and RGB color images as input for semantic segmentation task. We train an off-the-shelf segmentation algorithm *SegNet* [24] 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 1 and illustrated in Figure 3. Further, confusion matrices for ($RGB - SegNet$) and ($Albedo - SegNet$) are provided in Figure 4.

**Table 1.** 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** |

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 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

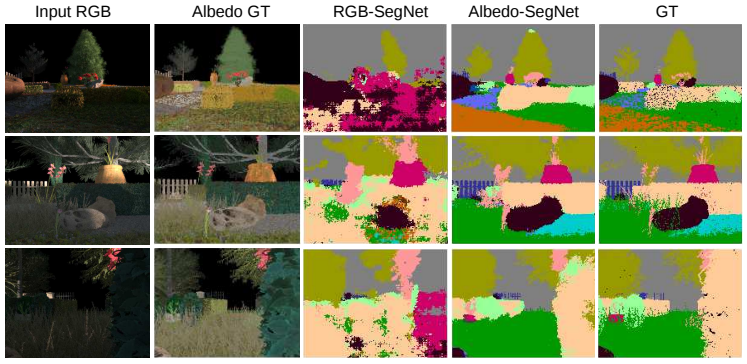


**Fig. 3.** Qualitative evaluation of the influence of reflectance on semantic segmentation. The results show that the semantic segmentation algorithm highly benefits from illumination invariant intrinsic properties (i.e. reflectance)

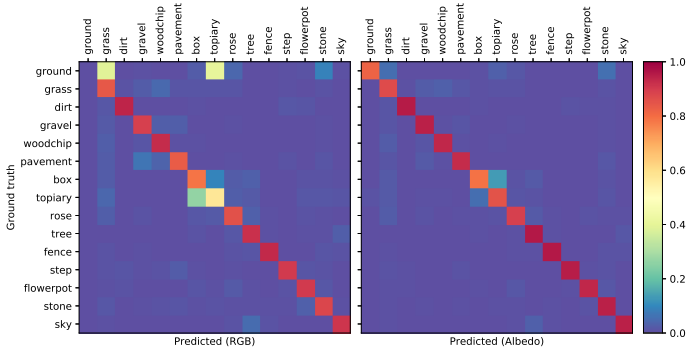


**Fig. 4.** Confusion matrices for $(RGB - SegNet)$ and $(Albedo - SegNet)$

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 and topiary classes.

### 5.2 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 as input. The results are summarized in Table 2.

**Table 2.** 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 shown in Table 2, 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 Fig. 5.

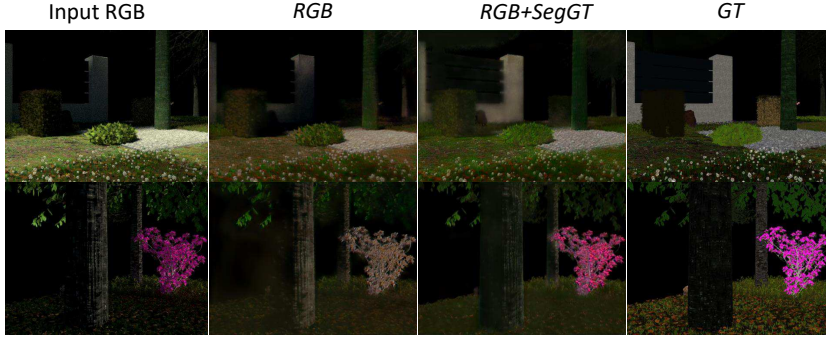


**Fig. 5.** Columns 2 and 3 show that $RGB + SegGT$ is better in removing shadows and shading from the reflectance images, as well as preserving sharp object boundaries and vivid colors, and therefore is more similar to the ground truth

### 5.3 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.

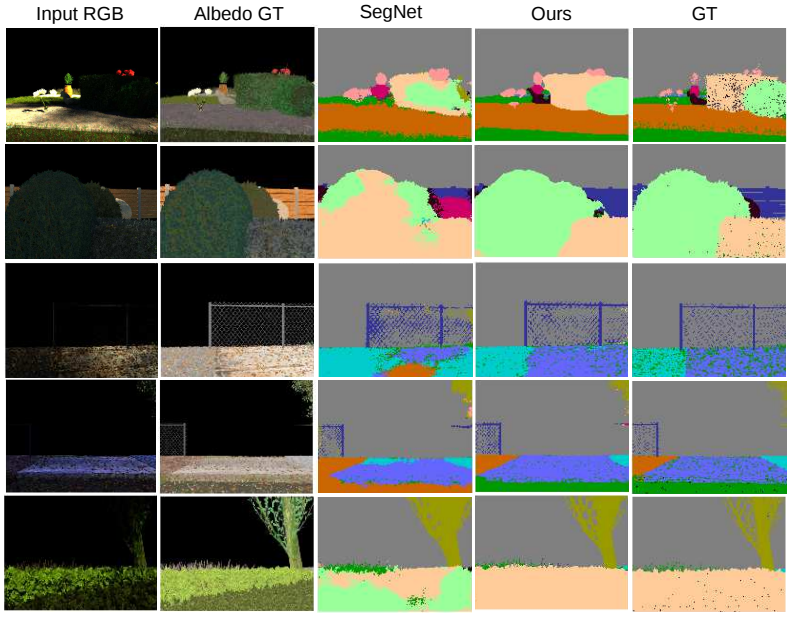


**Fig. 6.** Proposed joint learning framework outperforms single task framework *SegNet*. Our method preserves the object shapes and boundaries better and is robust against varying lighting conditions

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.

**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 [24] (*SegNet*) and the baseline of one encoder one decoder ShapeNet [15] (*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 3. As shown in Table 3, 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 Fig. 6. In addition, confusion matrices are provided in the supplementary material.

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,

**Table 3.** 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** |

our method still outperforms *SegNet*. Finally, the joint network outperforms the single task cascade network; for mIoU 0.6332 vs. 0.5810, see Table 1 and Table 3, as the joint scheme enforces to augment joint features.

**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 [15] (*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 4.

**Table 4.** Influence of joint learning on intrinsic property prediction

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

As shown in Table 4, 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* and the proposed *joint* framework are given in Fig. 7. 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 2 and Table 4.

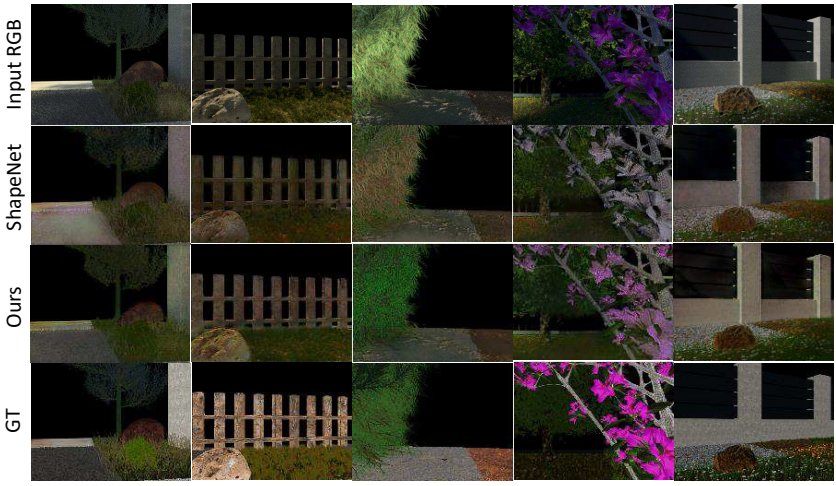


**Fig. 7.** The first two columns illustrate that the proposed method provides sharper outputs especially at object boundaries than *ShapeNet*. The 3rd and 4th columns show that the proposed method predicts colours that are closer to the ground truth reflectance. The last column shows that the proposed method handles sharp cast shadows better than *ShapeNet*

**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 5 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.

**Table 5.** 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

| $\omega$ | 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 ± 0.02997 | 0.0491 ± 0.04081 | 0.0716 ± 0.03804 | 0.0695 ± 0.0357 |

### 5.4 Real World Outdoor Dataset

Finally, our model is evaluated on real world garden images provided by the *3D Reconstruction meets Semantics challenge* [41]. The images are captured by


**Fig. 8.** Evaluation on real world garden images. We observe that our proposed method capture better colors and sharper outputs compared with [15]

a robot driving through a semantically-rich garden with fine geometric details. Results of [15] are provided as a visual comparison on the performance in Fig. 8. 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 [24] 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. Note that our model is much smaller in size and predicts the intrinsics together with the segmentation. More results are provided in the supplementary material.

## 6 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. A dataset of 35K synthetic images of natural environments has been created with corresponding albedo and shading (intrinsics), and semantic labels (segmentation). The experiments show joint performance benefit when performing the two tasks (intrinsics and semantics) in joint manner for natural scenes.

**Acknowledgements:** This project was funded by the EU Horizon 2020 program No. 688007 (TrimBot2020). We thank Gjorgji Strezoski for his contributions to the website.

## References


1. Upcroft, B., McManus, C., Churchill, W., Maddern, W., Newman, P.: Lighting invariant urban street classification. In: IEEE International Conference on Robotics and Automations. (2014) 1, 4
2. Wang, C., Tang, Y., Zou, X., Situ, W., Feng, W.: A robust fruit image segmentation algorithm against varying illumination for vision system of fruit harvesting robot. Optik-International Journal for Light and Electron Optics 131 (2017) 626–631 1, 4
3. Suh, H.K., Hofstee, J.W., van Henten, E.J.: Shadow-resistant segmentation based on illumination invariant image transformation. In: International Conference of Agricultural Engineering. (2014) 1, 4
4. Ramakrishnan, R., Nieto, J., Scheding, S.: Shadow compensation for outdoor perception. In: IEEE International Conference on Robotics and Automation. (2015) 1, 4
5. Land, E.H., McCann, J.J.: Lightness and retinex theory. Journal of Optical Society of America (1971) 1–11 2
6. Shen, L., Tan, P., Lin, S.: Intrinsic image decomposition with non-local texture cues. In: IEEE Conferance on Computer Vision and Pattern Recognition. (2008) 3
7. Zhao, Q., Tan, P., Dai, Q., Shen, L., Wu, E., Lin, S.: A closed-form solution to retinex with nonlocal texture constraints. IEEE Trans. on Pattern Analysis and Machine Intelligence (2012) 1437–1444 3
8. 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) 3
9. Shen, L., Yeo, C.: Intrinsic images decomposition using a local and global sparse representation of reflectance. In: IEEE Conferance on Computer Vision and Pattern Recognition. (2011) 3
10. Weiss, Y.: Deriving intrinsic images from image sequences. In: IEEE International Conference on Computer Vision. (2001) 3
11. Matsushita, Y., Lin, S., Kang, S.B., Shum, H.Y.: Estimating intrinsic images from image sequences with biased illumination. In: European Conference on Computer Vision. (2004) 3
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations. (2015) 3
13. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conferance on Computer Vision and Pattern Recognition. (2014) 3
14. Narihira, T., Maire, M., Yu, S.X.: Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In: IEEE International Conference on Computer Vision. (2015) 3
15. 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) 3, 5, 6, 11, 12, 14
16. Lettry, L., Vanhoey, K., Gool, L.V.: Darn: a deep adversarial residual network for intrinsic image decomposition. In: IEEE Winter Conference on Applications of Computer Vision. (2018) 3

17. 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) 3
18. 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) 3, 7
19. Bell, S., Bala, K., Snavely, N.: Intrinsic images in the wild. In: ACM Trans. on Graphics (TOG). (2014) 3
20. 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) 3
21. Fulkerson, B., Vedaldi, A., Soatto, S.: Class segmentation and object localization with superpixel neighborhoods. In: IEEE International Conference on Computer Vision. (2009) 3
22. Csurka, G., Perronnin, F.: An efficient approach to semantic segmentation. International Journal of Computer Vision, 95(2) (2011) 198–212 3
23. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. International Journal of Computer Vision, 95(2) (2009) 2–23 3
24. 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) 3, 6, 8, 11, 14
25. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conferance on Computer Vision and Pattern Recognition. (2015) 3
26. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. arXiv preprint arXiv:1606.00915 (2016) 3
27. Everingham, M., Eslami, S.M.A., van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision, 111(1) (2015) 98–136 3
28. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: IEEE Conferance on Computer Vision and Pattern Recognition. (2016) 3
29. Garcia-Garcia, A., Orts-Escolano, S., Oprea, S.O., Villena-Martinez, V., Garcia-Rodriguez, J.: A survey on deep learning techniques for image and video semantic segmentation. Applied Soft Computing, 70 (2018) 41–65 3
30. Jafari, O.H., Groth, O., Kirillov, A., Yang, M.Y., Rother, C.: Analyzing modular cnn architectures for joint depth prediction and semantic segmentation. In: IEEE International Conference on Robotics and Automation. (2017) 4
31. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: IEEE International Conference on Computer Vision. (2015) 4
32. Mousavian, A., Pirsiavash, H., Kosecka, J.: Joint semantic segmentation and depth estimation with deep convolutional networks. In: IEEE International Conference on 3D Vision. (2016) 4
33. Kundu, A., Li, Y., Dellaert, F., Li, F., Rehg, J.M.: Joint semantic segmentation and 3d reconstruction from monocular video. In: European Conference on Computer Vision. (2014) 4

34. Ladicky, L., Sturgess, P., Russell, C., Sengupta, S., Bastanlar, Y., Clocksin, W., Torr, P.H.S.: Joint optimization for object class segmentation and dense stereo reconstruction. International journal of computer vision, 100(2) (2012) 4
35. Barron, J.T., Malik, J.: Color constancy, intrinsic images,and shape estimation. In: European Conference on Computer Vision. (2012) 4
36. Kim, S., Park, K., Sohn, K., Lin, S.: Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In: IEEE Conference on Computer Vision and Pattern Recognition. (2016) 4
37. Shelhamer, E., Barron, J.T., Darrell, T.: Scene intrinsics and depth from a single image. In: IEEE International Conference on Computer Vision Workshop. (2015) 4
38. Vineet, V., Rother, C., Torr, P.H.S.: Higher order priors for joint intrinsic image, objects, and attributes estimation. In: Advances in Neural Information Processing Systems. (2013) 4
39. Shafer, S.: Using color to separate reflection components. Color research and applications (1985) 210–218 4
40. 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) 7
41. 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) 13