

Unified Depth Prediction and Intrinsic Image Decomposition from a Single Image via Joint Convolutional Neural Fields

Seungryong Kim^{1*}, Kihong Park¹, Kwanghoon Sohn¹, and Stephen Lin²

¹Yonsei University, ²Microsoft Research
{srkim89,khpark7727,khsohn}@yonsei.ac.kr, stevelin@microsoft.com

Abstract. We present a method for jointly predicting a depth map and intrinsic images from single-image input. The two tasks are formulated in a synergistic manner through a joint conditional random field (CRF) that is solved using a novel convolutional neural network (CNN) architecture, called the joint convolutional neural field (JCNF) model. Tailored to our joint estimation problem, JCNF differs from previous CNNs in its sharing of convolutional activations and layers between networks for each task, its inference in the gradient domain where there exists greater correlation between depth and intrinsic images, and the incorporation of a gradient scale network that learns the confidence of estimated gradients in order to effectively balance them in the solution. This approach is shown to surpass state-of-the-art methods both on single-image depth estimation and on intrinsic image decomposition.

Keywords: single-image depth estimation, intrinsic image decomposition, conditional random field, convolutional neural networks

1 Introduction

Perceiving the physical properties of a scene undoubtedly plays a fundamental role in understanding real-world imagery. Such inherent properties include the 3-D geometric configuration, the illumination or shading, and the reflectance or albedo of each scene surface. Depth prediction and intrinsic image decomposition, which aims to recover shading and albedo, are thus two fundamental yet challenging tasks in computer vision. While they address different aspects of scene understanding, there exist strong consistencies among depth and intrinsic images, such that information about one provides valuable prior knowledge for recovering the other.

In the intrinsic image decomposition literature, several works have exploited measured depth information to make the decomposition problem more tractable [1–5]. These techniques have all demonstrated better performance than using RGB images alone. On the other hand, in the literature for single-image depth

* This work was done while Seungryong Kim was an intern at Microsoft Research.

prediction, illumination-invariant features have been utilized for greater robustness in depth inference [6, 7], and shading discontinuities have been used to detect surface boundaries [8], suggesting that intrinsic images can be employed to enhance depth prediction performance. Although the two tasks are mutually beneficial, most previous research have solved for them only in sequence, by using estimated intrinsic images to constrain depth prediction [8], or vice versa [9]. We propose in this paper to instead jointly predict depth and intrinsic images in a manner where the two complementary tasks can assist each other.

We address this joint prediction problem using convolutional neural networks (CNNs), which have yielded state-of-the-art performance for the individual problems of single-image depth prediction [6, 7] and intrinsic image decomposition [9–11], but are hampered by ambiguity issues that arise from limited training sets. In our work, the two tasks are formulated synergistically in a joint conditional random field (CRF) that is solved using a novel CNN architecture, called the joint convolutional neural field (JCNF) model. This architecture differs from previous CNNs in several ways tailored to our particular problem. One is the sharing of convolutional activations and layers between networks for each task, which allows each network to account for inferences made in other networks. Another is to perform learning in the gradient domain, where there exist stronger correlations between depth and intrinsic images than in the image value domain, which helps to deal with the ambiguity problem from limited training sets. A third is the incorporation of a gradient scale network which jointly learns the confidence of the estimated gradients, to more robustly balance them in the solution. These networks of the JCNF model are iteratively learned in a piece-wise manner using a unified energy function in a joint CRF.

Within this system, depth, shading and albedo are predicted in a coarse-to-fine manner that yields more globally consistent results. Our experiments show that this joint prediction outperforms existing depth prediction methods and intrinsic image decomposition techniques on various benchmarks.

2 Related Work

Depth Prediction from a Single Image Traditional methods for this task have formulated the depth prediction as a Markov random field (MRF) learning problem [12–14]. As exact MRF learning and inference are intractable in general, most of these approaches employ approximation methods, such as through linear regression of depth with image features [12], learning image-depth correlation with a non-linear kernel function [13], and training category-adaptive model parameters [14]. Although these parametric models infer plausible depth maps to some extent, they cannot estimate the depth of natural scenes reliably due to their limited learning capability.

By leveraging the availability of large RGB-D databases, data-driven approaches have been actively researched [15, 16]. Konrad *et al.* [15] proposed a depth fusion scheme to infer the depth map by retrieving the nearest images in the dataset, followed by an aggregation via weighted median filtering. Karsch *et al.* [16] presented the depth transfer (DT) approach which retrieves the nearest

similar images and warps their depth maps using dense SIFT flow. Inspired by this method, Choi *et al.* [17] proposed the depth analogy (DA) approach that transfers depth gradients from the nearest images, demonstrating the effectiveness of gradient domain learning. Although these methods can extract reliable depth for certain scenes, there exist many others for which the nearest images are dissimilar and unsuitable. Recently, Kong *et al.* [8] extended the DT approach [16] by using albedo and shading for image matching as well as for detecting contours at surface boundaries. In contrast to our approach, the intrinsic images are estimated independently from the depth prediction.

More recently, methods have been proposed based on CNNs. Eigen *et al.* [6] proposed multi-scale CNNs (MS-CNNs) for predicting depth maps directly from a single image. Other CNN models were later proposed for depth estimation [18], including a deep convolutional neural field (DCNF) by Fayao *et al.* [7] that estimates depth on each superpixel while enforcing smoothness within a CRF. CNN-based methods clearly outperform conventional techniques, and we aim to elevate the performance further by accounting for intrinsic image information.

Intrinsic Image Decomposition The notion of intrinsic images was first introduced in [19]. Conventional methods are largely based on Retinex theory [20–22], which attributes large image gradients to albedo changes, and smaller gradients to shading. More recent approaches have employed a variety of techniques, based on gradient distribution priors [23], dense CRFs [24], and hybrid L_2 - L_p optimization to separate albedo and shading gradients [25]. These single-image based methods, however, are inherently limited by the fundamental ill-posedness of the problem. To partially alleviate this limitation, several approaches have utilized additional input, such as multiple images [26–28], user interaction [29, 30], and measured depth maps [1–5]. The use of additional data such as measured depth clearly increases performance but reduces their applicability.

Related to our work is the method of Barron and Malik [31], which estimates object shape in addition to intrinsic images. To regularize the estimation, the method utilizes statistical priors on object shape and albedo which are not generally applicable to images of full scenes.

More recently, intrinsic image decomposition has been addressed using CNNs [9–11]. Zhou *et al.* [10] proposed a multi-stream CNN to predict the relative reflectance ordering between image patches from large-scale human annotations. Narihira *et al.* [11] learned a CNN that directly predicts albedo and shading from an RGB image patch. Shelhamer *et al.* [9] estimated depth through a fully convolutional network and used it to constrain the intrinsic image decomposition. Unlike our approach, the depth and intrinsic images are estimated sequentially.

3 Formulation

3.1 Problem Statement and Model Architecture

Let us define a color image I such that $I_p : \mathcal{I} \rightarrow \mathbb{R}^3$ for pixel p , where $\mathcal{I} \subset \mathbb{N}^2$ is a discrete image domain. Similarly, depth, albedo and shading can be defined as $D_p : \mathcal{I} \rightarrow \mathbb{R}$ and $A_p, S_p : \mathcal{I} \rightarrow \mathbb{R}^3$. All of these image quantities are defined in

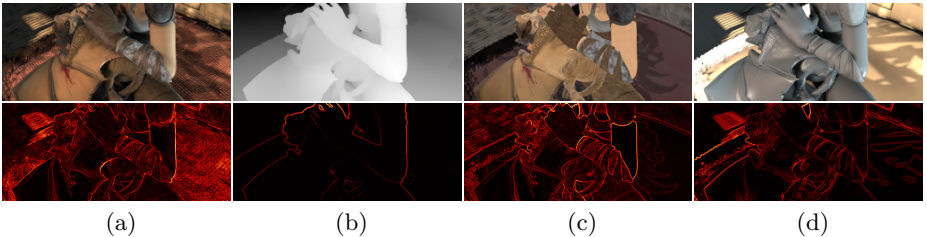


Fig. 1. For an example from the MPI-SINTEL dataset [32], its (a) color image I , (b) depth D , (c) albedo A , (d) shading S , and their corresponding gradient fields ∇I , ∇D , ∇A , and ∇S shown below. Compared to quantities in the value domain, correlations are stronger among gradient fields, such that estimates of one may help in learning others. Furthermore, the gradient consistency between ∇I , ∇D , ∇A , and ∇S can be used to estimate the confidence of each gradient.

the log domain. Given a training set of color, depth, albedo, and shading images denoted by $\mathcal{C} = \{(I^i, D^i, A^i, S^i) \mid i = 1, 2, \dots, \mathcal{N}_C\}$, where \mathcal{N}_C is the number of training images, we first aim to learn a prediction model that approximates depth D^i , albedo A^i , and shading S^i from each color image $I^i \in \mathcal{C}$. This prediction model will then be used to infer reliable depth D , albedo A , and shading S simultaneously from a single query image I .

We specifically learn the joint prediction model in the gradient domain, where depth and intrinsic images generally exhibit stronger correlation than in the value domain, as exemplified in Fig. 1. This greater correlation and reduced discrepancy among ∇D , ∇A , and ∇S facilitate joint learning of the two tasks by allowing them to better leverage information from each other¹. We therefore formulate our model to predict the depth, albedo, and shading gradient fields from the color image. Our method additionally learns the confidence of predicted gradients based on their consistency among one another in the training set.

We formulate this joint prediction using convolutional neural networks (CNNs) in a joint conditional random field (CRF). Our system architecture is structured as three cooperating networks, namely a depth prediction network, an intrinsic prediction network, and a gradient scale network. The depth prediction network is modeled by two feed-forward processes $\mathcal{F}(I^i; \mathbf{w}_{\mathcal{F}}^D)$ and $\mathcal{F}(I^i; \mathbf{w}_{\mathcal{F}}^{\nabla D})$, where $\mathbf{w}_{\mathcal{F}}^D$ and $\mathbf{w}_{\mathcal{F}}^{\nabla D}$ represent the network parameters for depth and depth gradients. The intrinsic prediction network is similarly modeled by feed-forward processes $\mathcal{F}(I^i; \mathbf{w}_{\mathcal{F}}^A)$ and $\mathcal{F}(I^i; \mathbf{w}_{\mathcal{F}}^{\nabla S})$, where $\mathbf{w}_{\mathcal{F}}^A$ and $\mathbf{w}_{\mathcal{F}}^{\nabla S}$ represent the network parameters for albedo gradients and shading gradients. The gradient scale network learns the confidence of depth, albedo and shading gradients using a feed-forward process for each, denoted by $\mathcal{G}(\nabla I^i, \nabla A^i, \nabla S^i; \mathbf{w}_{\mathcal{G}}^D)$, $\mathcal{G}(\nabla I^i, \nabla D^i, \nabla S^i; \mathbf{w}_{\mathcal{G}}^A)$, and $\mathcal{G}(\nabla I^i, \nabla D^i, \nabla A^i; \mathbf{w}_{\mathcal{G}}^S)$, where $\mathbf{w}_{\mathcal{G}}^D$, $\mathbf{w}_{\mathcal{G}}^A$, and $\mathbf{w}_{\mathcal{G}}^S$ are their respective network parameters. The three networks in our system are jointly learned in a manner where each can leverage information from the other networks.

¹ ∇ is a differential operator defined in the \mathbf{x} - and \mathbf{y} -direction such that $\nabla = [\nabla_{\mathbf{x}}, \nabla_{\mathbf{y}}]$.

3.2 Joint Conditional Random Field

The networks in our model are jointly learned by minimizing the energy function of a joint CRF. The joint CRF is formulated so that each task can leverage information from the other complementary task, leading to improved prediction in comparison to separate estimation models. Our energy function $\mathbf{E}(D, A, S|I)$ is defined as unary potentials \mathbf{E}_u and pairwise potentials \mathbf{E}_s for each task:

$$\begin{aligned} \mathbf{E}(D, A, S|I) = & \mathbf{E}_u(D|I) + \mathbf{E}_u(A, S|I) \\ & + \lambda_D \mathbf{E}_s(D|I, A, S) + \lambda_A \mathbf{E}_s(A|I, D, S) + \lambda_S \mathbf{E}_s(S|I, D, A), \end{aligned} \quad (1)$$

where λ_D , λ_A , and λ_S are weights for each pairwise potential. In the training procedure, this energy function is minimized over all the training images, *i.e.*, by minimizing $\sum_i \mathbf{E}(D^i, A^i, S^i|I^i)$. For testing, given a query image I and the learned network parameters, the final solutions of D , A , and S are estimated by minimizing the energy function $\mathbf{E}(D, A, S|I)$.

Unary Potentials The unary potentials consist of two energy functions, $\mathbf{E}_u(D|I)$ and $\mathbf{E}_u(A, S|I)$. The depth unary function $\mathbf{E}_u(D|I)$ is formulated as

$$\mathbf{E}_u(D|I) = \sum_p (D_p - \mathcal{F}(I_{\mathcal{P}}; \mathbf{w}_{\mathcal{F}}^D))^2, \quad (2)$$

which represents the squared differences between depths D_p and predicted depths from $\mathcal{F}(I_{\mathcal{P}}; \mathbf{w}_{\mathcal{F}}^D)$, where \mathcal{P} is the local neighborhood² for pixel p . It can be considered as a Dirichlet boundary condition for depth pairwise potentials, which will be described shortly.

The unary function $\mathbf{E}_u(A, S|I)$ for intrinsic images is used in minimizing the reconstruction errors of color image I from albedo A and shading S :

$$\mathbf{E}_u(A, S|I) = \sum_p (L_p(I_p - A_p - S_p))^2, \quad (3)$$

where $L_p = \text{lum}(I_p) + \varepsilon$, and $\text{lum}(I)$ denotes the luminance of I with $\varepsilon = 0.001$. It has been noted that processing of luminance balances out the influence of the unary potential across the image [1, 28], and that treating the image formation equation (*i.e.*, $I_p = A_p + S_p$) as a soft constraint can bring greater stability in optimization [25], especially for dark pixels whose chromaticity can be greatly distorted by sensor noise.

Pairwise Potentials The pairwise potentials, which include $\mathbf{E}_s(D|I, A, S)$, $\mathbf{E}_s(A|I, D, S)$, and $\mathbf{E}_s(S|I, D, A)$, represent differences between gradients and estimated gradients in the depth, albedo, and shading images. The pairwise potential $\mathbf{E}_s(D|I, A, S)$ for depth gradients is defined as

$$\mathbf{E}_s(D|I, A, S) = \sum_p \|\nabla D_p - \mathcal{G}(\nabla I_{\mathcal{P}}, \nabla A_{\mathcal{P}}, \nabla S_{\mathcal{P}}; \mathbf{w}_{\mathcal{G}}^{\nabla D}) \circ \mathcal{F}(I_{\mathcal{P}}; \mathbf{w}_{\mathcal{F}}^{\nabla D})\|^2, \quad (4)$$

² It is defined as the receptive field through the CNNs for pixel p [33].

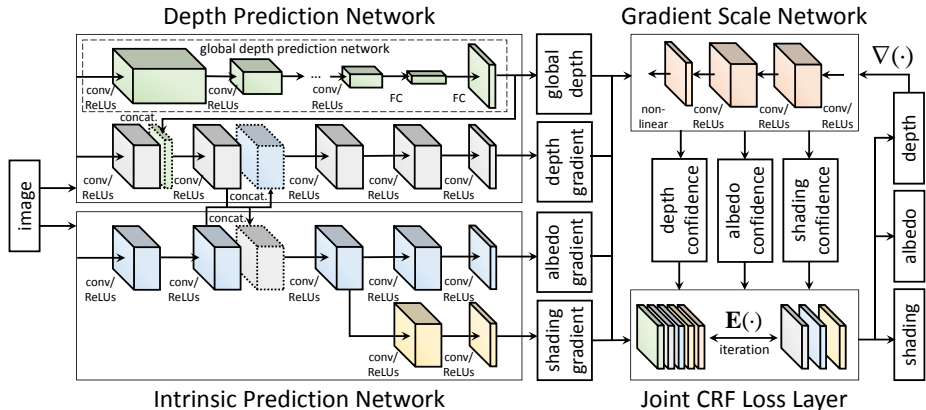


Fig. 2. Network architecture of the JCNF model. It consists of a depth prediction network, an intrinsic prediction network, and a gradient scale network. These networks are learned by minimizing a joint CRF loss function.

where \circ denotes the Hadamard product, and the estimated depth gradients of $\mathcal{F}(I_p; \mathbf{w}_{\mathcal{F}}^{\nabla D})$ provide a guidance gradient field for depth, similar to a Poisson equation [34, 35]. They are weighted by a confidence factor $\mathcal{G}(\nabla I_p, \nabla A_p, \nabla S_p; \mathbf{w}_{\mathcal{G}}^{\nabla D})$ learned in the gradient scale network to reduce the impact of erroneous gradients. This gradient scale is similar to the derivative-level confidence employed in [36] for image restoration, except that our gradient scale is learned non-locally with CNNs and different types of guidance images, as later described in Sec. 3.4. The pairwise potentials for albedo gradients $\mathbf{E}_s(A|I, D, S)$ and shading gradients $\mathbf{E}_s(S|I, D, A)$ are defined in the same manner. Since the gradient scales are jointly estimated with each other task, these pairwise potentials are computed within an iterative solver, which will be described in Sec. 4.1.

3.3 Joint Depth and Intrinsic Prediction Network

Our joint depth and intrinsic prediction network utilizes the aforementioned energy function to predict D , ∇D , ∇A , and ∇S from a single image I . The joint network consists of a depth prediction network for D and ∇D , and an intrinsic prediction network for ∇A and ∇S . In contrast to previous methods for single-image depth prediction [6, 37, 11], our system jointly estimates the gradient fields ∇D , ∇A , and ∇S , which are used to reduce ambiguity in the solution and obtain more edge-preserved results. To allow the different estimation tasks to leverage information from one another, we design the depth and intrinsic networks to share concatenated convolutional activations, and share convolutional layers between albedo and shading networks, as illustrated in Fig. 2.

Depth Prediction Network The depth prediction network consists of a global depth network and a depth gradient network. For the global depth network, we learn its parameters $\mathbf{w}_{\mathcal{F}}^D$ for predicting an overall depth map from the entire image structure. Similar to [6, 37, 11], it provides coarse, spatially-varying depth

that may be lacking in fine detail. This coarse depth will later be refined using the output of the depth gradient network.

The global depth network consists of five convolutional layers, three pooling layers, six non-linear activation layers, and two fully-connected (FC) layers. For the first five layers, the pre-trained parameters from the AlexNet architecture [38] are employed, and fine-tuning for the dataset is done. Rectified linear units (ReLUs) are used for the non-linear layers, and the pooling layers employ max pooling. The first FC layer encodes the network responses into fixed-dimensional features, and the second FC layer infers a coarse global depth map at 1/16-scale of the original depth map.

The depth gradient network predicts fine-detail depth gradients for each pixel. Its parameters $\mathbf{w}_{\mathcal{F}}^{\nabla D}$ are learned using an end-to-end patch-level scheme inspired by [39, 35], where the network input is an image patch and the output is a depth gradient patch. For inference of depth gradients at the pixel level, the depth gradient network consists of five convolutional networks followed by ReLUs, without stride convolutions or pooling layers. The first convolutional layer is identical to the first convolutional layer in the AlexNet architecture [38]. Four additional convolutional layers are also used as shown in Fig. 2. The depth gradient patches that are output by this network will be used for depth reconstruction in Sec. 4.2. Note that in the testing procedure, the depth gradient network is applied to overlapping patches over the entire image, which are aggregated in the last convolutional layer to yield the full gradient field.

Intrinsic Prediction Network The intrinsic prediction network has a structure similar to the depth gradient prediction network. The network parameters $\mathbf{w}_{\mathcal{F}}^{\nabla A}$ and $\mathbf{w}_{\mathcal{F}}^{\nabla S}$ are learned for predicting the albedo and shading gradients at each pixel. To jointly infer the depth and intrinsic image gradients, the second convolutional activations for each task are concatenated and passed to their third convolutional layers as shown in Fig. 2. In the training procedure, the depth and intrinsic networks are iteratively learned, which enables each task to benefit from each other’s activations to provide more reliable estimates. Furthermore, similar to [11], the albedo and shading gradient networks share their first three convolutional layers, while the last two are separate. Since the albedo and shading images have related properties, these shared convolutional layers benefit their estimation. Details on kernel sizes and the number of channels for each layer are provided in the supplemental material for all the networks.

3.4 Gradient Scale Network

The estimated gradients from the depth and intrinsic prediction networks might contain errors due to the ill-posed nature of their problems. To help in identifying such errors, our system additionally learns the confidence of estimated gradients, specifically, whether a gradient exists at a particular location or not. The basic idea is to learn from the training data about the consistencies that exist among the different types of gradients given their local neighborhood \mathcal{P} . From this, we can determine the confidence of a gradient (*e.g.*, a depth gradient), based on

the other estimated gradients (*e.g.*, the albedo, shading, and image gradients). This confidence is modeled as a gradient scale that is similar to the scale map used in [36] to model derivative-level confidence for image restoration. It can be noted that in some depth and intrinsic image decomposition methods [1, 4, 7], the solutions are filtered with fixed parameters using the color image as guidance. Our system instead learns a network for defining the parameters, using not only a color image but also depth and intrinsic images as guidance.

The gradient scale network consists of three convolutional layers and one non-linear activation layer. For the case of depth gradients, the output of the gradient scale network $\mathcal{G}(\nabla I_{\mathcal{P}}, \nabla A_{\mathcal{P}}, \nabla S_{\mathcal{P}}; \mathbf{w}_{\mathcal{G}}^{\nabla D})$ is estimated as the convolution between $\mathbf{w}_{\mathcal{G}}^{\nabla D}$ and $(|\nabla I_{\mathcal{P}}|^2, |\nabla A_{\mathcal{P}}|^2, |\nabla S_{\mathcal{P}}|^2)$, followed by a non-linear activation *i.e.*, $f(\cdot) = (1 - \exp(1 - \cdot)) / (1 + \exp(1 - \cdot))$, which is defined within $[-1, 1]$. Here, $|\cdot|^2$ for a vector of gradients denotes a vector of the gradient magnitudes. Thus, in the gradient scale network, the network parameters are convolved with the gradient magnitudes. With the learned parameters $\mathbf{w}_{\mathcal{G}}^{\nabla D}$, the confidence of ∇D_p is estimated from $\nabla I_{\mathcal{P}}, \nabla A_{\mathcal{P}}, \nabla S_{\mathcal{P}}$. This can alternatively be viewed as a guidance filtering weight for D with guidance images I, A , and S . $\mathcal{G}(\nabla I_{\mathcal{P}}, \nabla D_p, \nabla S_{\mathcal{P}}; \mathbf{w}_{\mathcal{G}}^{\nabla A})$ and $\mathcal{G}(\nabla I_{\mathcal{P}}, \nabla D_p, \nabla A_{\mathcal{P}}; \mathbf{w}_{\mathcal{G}}^{\nabla S})$ are also similarly defined.

Some properties of gradient scales are as follows. A gradient scale can be either positive or negative. A large positive value indicates high confidence in the presence of a gradient. A large negative value also indicates high confidence, but for the reversed gradient direction. In addition, when a gradient field contains extra erroneous regions, gradient scales of value 0 can help to disregard them.

4 Unified Depth and Intrinsic Image Prediction

4.1 Training

The energy function $\mathbf{E}(D, A, S|I)$ from (1) is used to simultaneously learn the depth and intrinsic network parameters ($\mathbf{w}_{\mathcal{F}}^D, \mathbf{w}_{\mathcal{F}}^{\nabla D}, \mathbf{w}_{\mathcal{F}}^{\nabla A}, \mathbf{w}_{\mathcal{F}}^{\nabla S}$) and the gradient scale network parameters ($\mathbf{w}_{\mathcal{G}}^{\nabla D}, \mathbf{w}_{\mathcal{G}}^{\nabla A}, \mathbf{w}_{\mathcal{G}}^{\nabla S}$). Although the overall form of the energy is non-quadratic, it has a quadratic form with respect to each of its terms. The energy function can thus be minimized by alternating among its terms.

Loss Functions For the global depth unary potential of (2), the global depth network parameters $\mathbf{w}_{\mathcal{F}}^D$ can be solved by minimizing the following loss function

$$\mathcal{L}(\mathbf{w}_{\mathcal{F}}^D) = \sum_{\{i,p\}} (D_p^i - \mathcal{F}(I_{\mathcal{P}}^i; \mathbf{w}_{\mathcal{F}}^D))^2. \quad (5)$$

We note that the intrinsic image unary term does not contain network parameters to be learned, so it is used only in the testing procedure.

The pairwise potentials each incorporate two networks, namely the gradient prediction network and gradient scale network, so they are iteratively trained. The loss function for the depth gradient pairwise potential of (4) is defined as

$$\mathcal{L}(\mathbf{w}_{\mathcal{G}}^{\nabla D}, \mathbf{w}_{\mathcal{F}}^{\nabla D}) = \sum_{\{i,p\}} \|\nabla D_p^i - \mathcal{G}(\nabla I_{\mathcal{P}}^i, \nabla A_{\mathcal{P}}^i, \nabla S_{\mathcal{P}}^i; \mathbf{w}_{\mathcal{G}}^{\nabla D}) \circ \mathcal{F}(I_{\mathcal{P}}^i; \mathbf{w}_{\mathcal{F}}^{\nabla D})\|^2. \quad (6)$$

The loss functions for the pairwise potentials of the albedo gradients $\mathcal{L}(\mathbf{w}_G^{\nabla A}, \mathbf{w}_F^{\nabla A})$ and shading gradients $\mathcal{L}(\mathbf{w}_G^{\nabla S}, \mathbf{w}_F^{\nabla S})$ are similarly defined.

These loss functions are minimized using stochastic gradient descent with the standard back-propagation [40]. First, \mathbf{w}_F^D is estimated through $\partial\mathcal{L}(\mathbf{w}_F^D)/\partial\mathbf{w}_F^D$. Then $\mathbf{w}_G^{\nabla D}$ and $\mathbf{w}_F^{\nabla D}$ are iteratively estimated through $\partial\mathcal{L}(\mathbf{w}_G^{\nabla D}, \mathbf{w}_F^{\nabla D})/\partial\mathbf{w}_G^{\nabla D}$ and $\partial\mathcal{L}(\mathbf{w}_G^{\nabla D}, \mathbf{w}_F^{\nabla D})/\partial\mathbf{w}_F^{\nabla D}$. In each iteration, the loss functions are differently defined according to the other network outputs, where the network parameters are initialized with the values obtained from the previous iteration. In this way, the networks account for the improving outputs of the other networks.

4.2 Testing

Iterative Joint Prediction In the testing procedure, the outputs D , ∇D , ∇A and ∇S for a given input image I are predicted by minimizing the energy function $\mathbf{E}(D, A, S|I)$ from (1) with constraints from the estimates computed using the learned network parameters and forward-propagation. Similar to the training procedure, we minimize $\mathbf{E}(D, A, S|I)$ with an iterative scheme due to its non-quadratic form, where $\mathbf{E}(D|I)$ and $\mathbf{E}(A, S|I)$ are minimized in alternation.

For the depth prediction, $\mathbf{E}(D|I)$ is defined as a data term for global depth and a pairwise term for depth gradients:

$$\mathbf{E}(D|I) = \sum_p (D_p - D_p^*)^2 + \lambda_D \sum_p \|\nabla D_p - C(\nabla D_p^*) \circ \nabla D_p^*\|^2, \quad (7)$$

where $*$ denotes network outputs, and $C(\nabla D_p^*)$ is the gradient scale of ∇D_p^* derived from $\mathcal{G}(\nabla I_p^*, \nabla A_p^*, \nabla S_p^*; \mathbf{w}_G^{\nabla D})$. We note that since $C(\nabla D_p^*)$ is computed with ∇I_p^* , ∇A_p^* , and ∇S_p^* , all of the predictions need to be iteratively estimated.

For the intrinsic prediction, $\mathbf{E}(A, S|I)$ is also defined as data and pairwise terms, with the image formation equation and the albedo and shading gradients:

$$\begin{aligned} \mathbf{E}(A, S|I) = & \sum_p (L_p(I_p - A_p - S_p))^2 \\ & + \sum_p \lambda_A \|\nabla A_p - C(\nabla A_p^*) \circ \nabla A_p^*\|^2 + \lambda_S \|\nabla S_p - C(\nabla S_p^*) \circ \nabla S_p^*\|^2, \end{aligned} \quad (8)$$

where $C(\nabla A_p^*)$ and $C(\nabla S_p^*)$ are defined similarly to $C(\nabla D_p^*)$. This energy function can be optimized with an existing linear solver [1]. These two energy functions $\mathbf{E}(D|I)$ and $\mathbf{E}(A, S|I)$ are iteratively minimized while providing information in the form of depth, albedo, and shading gradients to each other.

Coarse-to-Fine Joint Prediction In estimating depth and intrinsic images, enforcing a degree of global consistency can lead to performance gains [1, 5]. For greater global consistency, we apply our joint prediction model in a coarse-to-fine manner, where color images I^l are constructed at \mathcal{N}_L image pyramid levels $l = \{1, \dots, \mathcal{N}_L\}$, and the depth D^l and intrinsic images A^l and S^l are predicted from I^l . Coarser scale results are then used as guidance for finer levels.

Specifically, we reformulate $\mathbf{E}(D|I)$ as $\mathbf{E}(D^l|I^l, I^{l-1})$:

$$\begin{aligned} \mathbf{E}(D^l|I^l, I^{l-1}) = & \sum_p (D_p^l - D_p^{l,*})^2 + \sum_p (D_p^l - D_p^{l-1})^2 \\ & + \lambda_D \sum_p \|\nabla D_p^l - C(\nabla D_p^{l,*}) \circ \nabla D_p^{l,*}\|^2. \end{aligned} \quad (9)$$

Similarly, $\mathbf{E}(A, S|I)$ is reformulated as $E(A^l, S^l|I^l, I^{l-1})$:

$$\begin{aligned}
 E(A^l, S^l|I^l, I^{l-1}) &= \sum_p (L_p^l(I_p^l - A_p^l - S_p^l))^2 + (A_p^l - A_p^{l-1})^2 + (S_p^l - S_p^{l-1})^2 \\
 &+ \sum_p \lambda_A \|\nabla A_p^l - C(\nabla A_p^{l,*}) \circ \nabla A_p^{l,*}\|^2 + \lambda_S \|\nabla S_p^l - C(\nabla S_p^{l,*}) \circ \nabla S_p^{l,*}\|^2,
 \end{aligned} \tag{10}$$

where the multi-scale unary functions lead to more reliable solutions and faster convergence. The high-level algorithm for the training and testing procedures is provided in the supplemental material.

5 Experimental Results

For our experiments, we implemented the JCNF model using the VLFeat MatConvNet toolbox [40]. The energy function weights were set to $\{\lambda_D, \lambda_A, \lambda_S\} = \{1, 0.1, 0.1\}$ by cross-validation. The filter weights of each network layer were initialized by drawing randomly from a Gaussian distribution with zero mean and a standard deviation of 0.001. The network learning rates were set to 10^{-4} , except for the final layer of the gradient networks where it was set to 10^{-5} .

We additionally augmented the training data by applying random transforms to it, including scalings in the range $[0.8, 1.2]$, in-plane rotations in the range $[-15, 15]$, translations, RGB scalings, image flips, and different gammas.

In the following, we evaluated our system through comparisons to state-of-the-art depth prediction and intrinsic image decomposition methods on the MPI SINTEL [41], NYU v2 [42], and Make3D [43] benchmarks. We additionally examined the performance contributions of the joint network learning (wo/jnl), the gradient scale network (wo/gsn), and the coarse-to-fine scheme (wo/ctf). The experimental details are provided in the supplemental material.

5.1 MPI SINTEL Benchmark

We evaluated our JCNF model on both depth prediction and intrinsic image decomposition on the MPI SINTEL benchmark [41], which consists of 890 images from 18 scenes with 50 frames each. For a fair evaluation, we followed the same experimental protocol as in [1, 11], with their two-fold cross-validation and training/testing image splits. Fig. 3 and Fig. 4 exhibit predicted depth and intrinsic images from a single image, respectively. Table 1 and Table 2 are quantitative evaluations for both tasks using a variety of metrics, including average relative difference (rel), average \log_{10} error (\log_{10}), root-mean-squared error (rms), its log version (rms_{\log}), and accuracy with thresholds $\delta = \{1.25, 1.25^2, 1.25^3\}$ [7]. For quantitatively evaluating intrinsic image decomposition performance, we used mean-squared error (MSE), local mean-squared error (LMSE), and the dissimilarity version of the structural similarity index (DSSIM) [11].

For the depth prediction task, data-driven approaches (DT [16] and DA [17]) provided limited performance due to their low learning capacity. CNN-based depth prediction (DCNF-FCSP [7]) using a pre-trained model from NYU v2

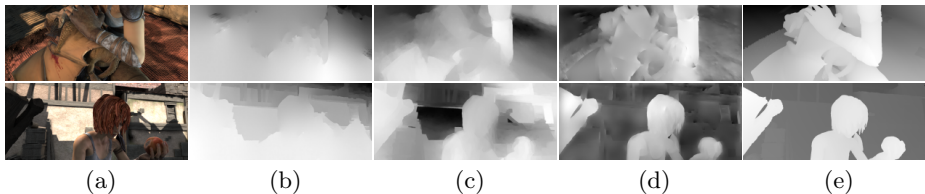


Fig. 3. Qualitative results on MPI SINTEL [41] for depth prediction. (a) color image, (b) DA [17], (c) DCNF-FCSP(NYU) [7], (d) JCNF, and (e) ground truth.

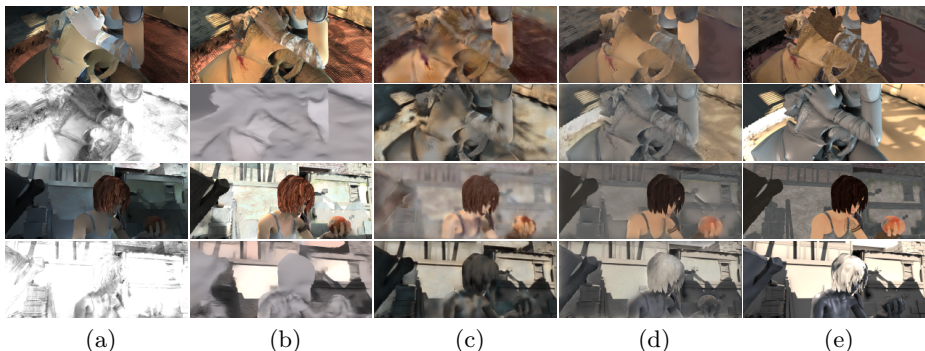


Fig. 4. Qualitative results on MPI SINTEL [41] for intrinsic decomposition of Fig. 3. (a) Shen *et al.* [30], (b) SIRFS [31], (c) MSCR [11], (d) JCNF, and (e) ground truth.

[42] showed better performance, but is restricted by depth ambiguity problems. Our JCNF model achieved the best results both quantitatively and qualitatively, whether pre-trained using MPI SINTEL or NYU v2 datasets. Furthermore, it is shown that omitting the gradient scale network, coarse-to-fine processing, or joint learning significantly reduced depth prediction performances.

In intrinsic image decomposition, existing single-image based methods [44, 23, 30, 22, 24] produced the lowest quality results as they do not benefit from any additional information. RGB-D based methods [1, 5, 4] performed better with measured depth as input. CNN-based intrinsic decomposition [11] surpassed RGB-D based techniques even without having depth as an input, but its results exhibit some blur, likely due to ambiguity from limited training datasets. Thanks to its gradient domain learning and leverage of estimated depth information, our JCNF model provides more accurate and edge-preserved results, with the best qualitative and quantitative performance.

5.2 NYU v2 RGB-D Benchmark

For further evaluation, we obtained a set of RGB, depth, and intrinsic images by applying RGB-D based intrinsic image decomposition methods [1, 4] on the NYU v2 RGB-D database [42]. Of its 1449 RGB-D images of indoor scenes, we used 795 for training and 654 for testing, which is the standard training/testing split for the dataset.

Methods	Error				Accuracy		
	rel	log ₁₀	rms	rms _{log}	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Depth Transfer [16]	0.448	0.193	9.242	3.121	0.524	0.712	0.735
Depth Analogy [17]	0.432	0.167	8.421	2.741	0.621	0.799	0.812
DCNF-FCSP(NYU) [7]	0.424	0.164	8.112	2.421	0.652	0.782	0.824
JCNF(NYU)	0.293	0.131	7.421	1.812	0.715	0.812	0.831
JCNF wo/jnl	0.292	0.138	7.471	1.973	0.714	0.783	0.839
JCNF wo/gsn	0.271	0.119	7.451	1.921	0.724	0.793	0.893
JCNF wo/ctf	0.252	0.101	7.233	1.622	0.729	0.812	0.878
JCNF	0.183	0.097	6.118	1.037	0.823	0.834	0.902

Table 1. Quantitative results on MPI SINTEL [41] for depth prediction. DCNF-FCSP (NYU) [7] and JCNF(NYU) predict the depth by pre-training on NYU v2 [42].

Methods	MSE			LMSE			DSSIM		
	albedo	shading	avg.	albedo	shading	avg.	albedo	shading	avg.
Retinex [44]	0.053	0.049	0.051	0.033	0.028	0.031	0.214	0.206	0.210
Li <i>et al.</i> [23]	0.042	0.041	0.037	0.024	0.031	0.034	0.242	0.224	0.194
Shen <i>et al.</i> [30]	0.043	0.039	0.048	0.028	0.027	0.032	0.221	0.210	0.232
Zhao <i>et al.</i> [22]	0.047	0.041	0.031	0.028	0.029	0.031	0.210	0.257	0.214
IIW [24]	0.041	0.032	0.041	0.032	0.031	0.027	0.281	0.241	0.284
SIRFS [31]	0.042	0.047	0.043	0.029	0.026	0.028	0.210	0.206	0.208
Jeon <i>et al.</i> [4]	0.042	0.033	0.032	0.021	0.021	0.023	0.204	0.181	0.193
Chen <i>et al.</i> [1]	0.031	0.028	0.029	0.019	0.019	0.019	0.196	0.165	0.181
MSCR [11]	0.020	0.017	0.021	0.016	0.011	0.011	0.201	0.150	0.176
JCNF wo/jnl	0.012	0.015	0.016	0.014	0.010	0.010	0.149	0.123	0.141
JCNF wo/gsn	0.008	0.011	0.011	0.010	0.009	0.008	0.146	0.112	0.132
JCNF wo/ctf	0.008	0.012	0.010	0.009	0.008	0.008	0.127	0.110	0.119
JCNF	0.007	0.009	0.007	0.006	0.007	0.007	0.092	0.101	0.097

Table 2. Quantitative results on MPI SINTEL [41] for intrinsic decomposition using methods based on single images, RGB-D, CNNs, and our JCNF model.

For depth prediction, comparisons are made to the ground truth depth in Fig. 5 and Table 2 using the same experimental settings as in [7]. The state-of-the-art CNN-based methods [6, 7] clearly outperformed other previous methods. The performance of our JCNF model was even higher, with pre-training on either MPI SINTEL or NYU v2. Our depth prediction network is similar to [6], but it additionally predicts depth gradients and leverages intrinsic image estimates to elevate performance.

In intrinsic image decomposition of Fig. 6, RGB-D based methods [1, 4] are used as ground truth for training. It is seen that our JCNF more closely resembles that assumed ground truth than single-image based methods [23, 24].

5.3 Make3D RGB-D Benchmark

We also evaluated our JCNF model on the Make3D dataset [43], which contains 534 images depicting outdoor scenes (with 400 used for training and 134 for testing). To account for a limitation of this dataset [12, 45, 7], we calculate depth errors in two ways [45, 7]: on only regions with ground truth depth less than

Methods	Error				Accuracy		
	rel	log ₁₀	rms	rms _{log}	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Make3D [12]	0.349	-	1.214	0.409	0.447	0.745	0.897
Depth Transfer [16]	0.350	0.134	1.1	0.378	0.460	0.742	0.893
Depth Analogy [17]	0.328	0.132	1.31	0.392	0.471	0.799	0.891
MS-CNNs [6]	0.228	-	0.901	0.293	0.611	0.873	0.961
DCNF-FCSP [7]	0.221	0.095	0.760	0.281	0.604	0.885	0.974
JCNF(MPI)	0.214	0.093	0.716	0.241	0.677	0.879	0.927
JCNF wo/jnl	0.216	0.101	0.753	0.241	0.625	0.896	0.925
JCNF wo/gsn	0.210	0.091	0.728	0.254	0.621	0.890	0.975
JCNF wo/ctf	0.208	0.106	0.708	0.237	0.681	0.901	0.972
JCNF	0.201	0.077	0.711	0.212	0.690	0.910	0.979

Table 3. Quantitative results on the NYU v2 dataset [42] for depth prediction.

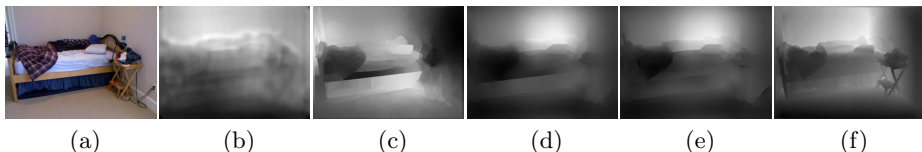


Fig. 5. Qualitative results on NYU v2 [42] for depth prediction. (a) color image, (b) MS-CNNs [6], (c) DCNF-FCSP [7], (d) JCNF(MPI), (e) JCNF, and (f) ground truth.

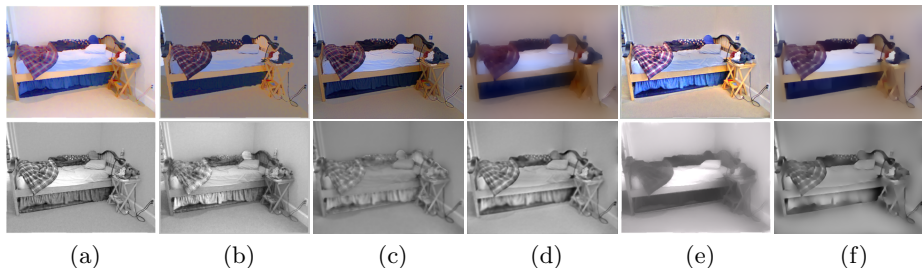


Fig. 6. Qualitative results on NYU v2 [42] for intrinsic decomposition of Fig. 5. (a) Li *et al.* [23], (b) IIW [24], (c) Jeon *et al.* [4], (d) JCNF learned using [4], (e) Chen *et al.* [1], and (f) JCNF learned using [1].

70 meters (denoted by C1), and over the entire image (C2). From the depth prediction results in Fig. 7 and Table 4, our JCNF model is found to yield the highest accuracy, even when pretrained on MPI SINTEL [41] or NYU v2 [42] (*i.e.*, JCNF(MPI) and JCNF(NYU)). For the intrinsic image decomposition results given in Fig. 8, JCNF also outperforms the comparison techniques.

6 Conclusion

We presented Joint Convolutional Neural Fields (JCNF) for jointly predicting depth, albedo and shading maps from a single input image. Its high performance can be attributed to its sharing network architecture, its gradient domain inference, and the incorporation of gradient scale network. It is shown through extensive experimentation that synergistically solving for these physical scene properties through the JCNF leads to state-of-the-art results in both single-image

Methods	Error (C1)				Error (C2)			
	rel	log ₁₀	rms	rms _{log}	rel	log ₁₀	rms	rms _{log}
Make3D [12]	0.412	0.165	11.1	0.451	0.407	0.155	16.1	0.486
Depth Transfer [16]	0.355	0.127	9.20	0.421	0.438	0.161	14.81	0.461
Depth Analogy [17]	0.371	0.121	8.11	0.381	0.410	0.144	14.52	0.479
DCNF-FCSP [7]	0.331	0.119	8.60	0.392	0.307	0.125	12.89	0.412
JCNF(MPI)	0.273	0.110	7.70	0.351	0.263	0.117	8.62	0.347
JCNF(NYU)	0.274	0.097	7.22	0.352	0.287	0.127	8.22	0.341
JCNF	0.262	0.092	6.61	0.321	0.243	0.091	6.34	0.302

Table 4. Quantitative results on the Make3D dataset [43] for depth prediction.

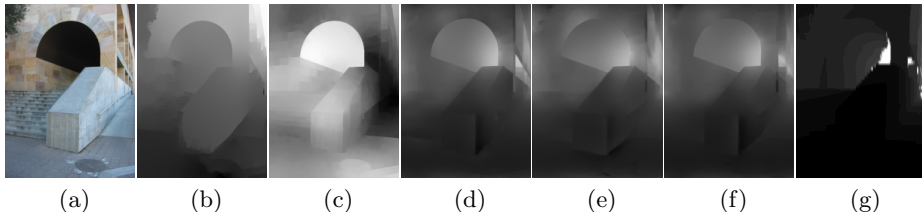


Fig. 7. Qualitative results on Make3D [42] for depth prediction. (a) color image, (b) DA [17], (c) DCNF-FCSP [7], (d) JCNF(MPI), (e) JCNF(NYU), (f) JCNF, and (g) ground truth.

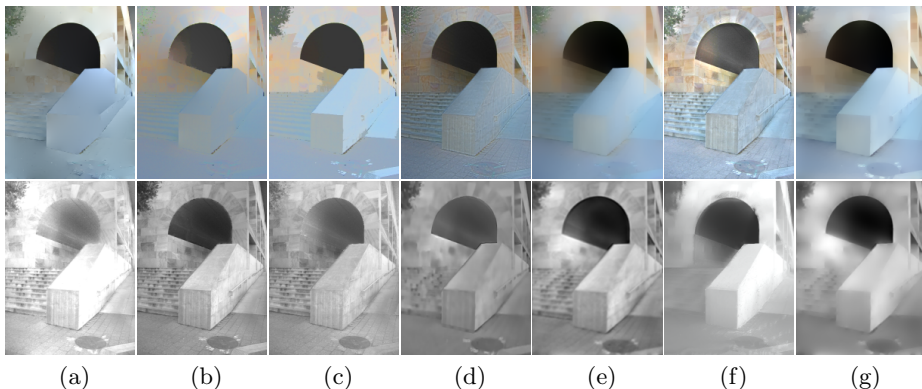


Fig. 8. Qualitative results on Make3D [43] for intrinsic decomposition of Fig. 7. (a) Li *et al.* [23], (b) Zhao *et al.* [22], (c) IHW [24], (d) Jeon *et al.* [4], (e) JCNF learned using [4], (f) Chen *et al.* [1], and (g) JCNF learned using [1].

depth prediction and intrinsic image decomposition. In future work, JCNF can potentially benefit shape refinement and image relighting from a single image.

Acknowledgement. This research was supported by the MSIP (The Ministry of Science, ICT and Future Planning), Korea and Microsoft Research, under ICT/SW Creative research program supervised by the IITP(Institute for Information & Communications Technology Promotion) (IITP-2015-R2212-15-0008).

References

1. Chen, Q., Koltun, V.: A simple model for intrinsic image decomposition with depth cues. *ICCV* (2013)
2. Laffont, P.Y., Bousseau, A., Paris, S., Durand, F., Drettakis, G.: Coherent intrinsic images from photo collections. *ACM TOG* **31**(6) (2012) 1–11
3. Lee, K.J., Zhao, Q., Tong, X., Gong, M., Izadi, S., L.S.U., Tan, P., Lin, S.: Estimation of intrinsic image sequences from image + depth video. *ECCV* (2012)
4. Jeon, J., Cho, S., Tong, X., Lee, S.: Intrinsic image decomposition using structure-texture separation and surface normals. *ECCV* (2014)
5. Barron, J.T., Malik, J.: Intrinsic scene properties from a single rgb-d image. *CVPR* (2013)
6. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. *NIPS* (2014)
7. Fayao, L., Chunhua, S., Guosheng, L.: Deep convolutional neural fields for depth estimation from a single images. *CVPR* (2015)
8. Kong, N., Black, M.J.: Intrinsic depth: Improving depth transfer with intrinsic images. *ICCV* (2015)
9. Shelhamer, E., Barron, J., Darrell, T.: Scene intrinsics and depth from a single image. *ICCV workshop* (2015)
10. Zhou, T., Krahenbuhl, P., Efors, A.A.: Learning data-driven reflectance priors for intrinsic image decomposition. *ICCV* (2015)
11. Narihira, T., Maire, M., Yu, S.X.: Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. *ICCV* (2015)
12. Saxena, A., Sun, M., Andrew, Y.: Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. PAMI* **31**(5) (2009) 824–840
13. Wang, Y., Wang, R., Dai, Q.: A parametric model for describing the correlation between single color images and depth maps. *IEEE SPL* **21**(7) (2014) 800–803
14. Xiu, L., Hongwei, Q., Yangang, W., Yongbing, Z., Qionghai, D.: Dept: Depth estimation by parameter transfer for single still images. *CVPR* (2014)
15. Konrad, J., Wang, M., Ishwar, P., Wu, C., Mukherjee, D.: Learning-based, automatic 2d-to-3d image and video conversion. *IEEE Trans. IP* **22**(9) (2013) 3485–3496
16. Karsch, K., Liu, C., Kang, S.B.: Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE Trans. PAMI* **32**(11) (2014) 2144–2158
17. Choi, S., Min, D., Ham, B., Kim, Y., Oh, C., Sohn, K.: Depth analogy: Data-driven approach for single image depth estimation using gradient samples. *IEEE Trans. IP* **24**(12) (2015) 5953–5966
18. Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.: Towards unified depth and semantic prediction from a single image. *CVPR* (2015)
19. Barrow, H.G., Tenenbaum, J.M.: Recovering intrinsic scene characteristics from images. *CVS* (1978)
20. Land, E.H., McCann, J.J.: Lightness and retinex theory. *JOSA* **61**(1) (1971) 1–11
21. Shen, J., Tan, P., Lin, S.: Intrinsic image decomposition with non-local texture cues. *CVPR* (2008)
22. Zhao, Q., Tan, P., Dai, Q., Shen, L., Wu, E., Lin, S.: A closed-form solution to retinex with non-local texture constraints. *IEEE Trans. PAMI* **34**(7) (2012) 1437–1444
23. Li, Y., Brown, M.S.: Single image layer separation using relative smoothness. *CVPR* (2004)

24. Bell, S., Bala, K., Snavely, N.: Intrinsic images in the wild. *ACM TOG* **33**(4) (2014)
25. Bonneel, N., Sunkavalli, K., Tompkin, J., Sun, D., Paris, S., Pfister, H.: Interactive intrinsic video editing. *ACM Trans. Graphics (SIGGRAPH ASIA)* (2014)
26. Wiess, Y.: Deriving intrinsic images from image sequences. *ICCV* (2001)
27. Laffont, P.Y., Bousseau, A., Drettakis, G.: Rich intrinsic image decomposition of outdoor scenes from multiple views. *IEEE TVCG* **19**(2) (2013) 1–11
28. Kong, N., Gehler, P.V., Black, M.J.: Intrinsic video. *ECCV* (2014)
29. Bousseau, A., Paris, S., Durand, F.: User-assisted intrinsic images. *ACM TOG* **28**(5) (2009) 1–11
30. Shen, J., Yang, X., Jia, Y.: Intrinsic image using optimization. *CVPR* (2011)
31. Barron, J., Malik, J.: Shape, albedo, and illumination from a single image of an unknown object. *CVPR* (2012)
32. Butler, D., Wulff, J., Stanley, G., Black, M.: A naturalistic open source movie for optical flow evaluation. *ECCV* (2012)
33. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. PAMI* **37**(9) (2015) 1904–1916
34. Perez, P., Gangnet, M., Blake, A.: Poisson image editing. *ACM TOG* **22**(3) (2003)
35. Xu, L., Ren, J., Yan, Q., Liao, R., Jia, J.: Deep edge-aware filters. *ICML* (2015)
36. Shen, X., Yan, Q., Xu, L., Ma, L., Jia, J.: Multispectral joint image restoration via optimizing a scale map. *IEEE Trans. PAMI* **31**(9) (2015) 1582–1599
37. Eigen, D., R, F.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *ICCV* (2015)
38. Alex, K., Ilya, S., E, H.: Imagenet classification with deep convolutional neural networks. *NIPS* (2012)
39. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. *IEEE Trans. PAMI* **37**(3) (2015) 597–610
40. Online.: <http://www.vlfeat.org/matconvnet/>.
41. Online.: <http://sintel.is.tue.mpg.de/>.
42. Online.: <http://cs.nyu.edu/silberman/datasets/>.
43. Online.: <http://make3d.cs.cornell.edu/>.
44. Grosse, R., Johnson, M.K., Adelson, E.H., Freeman, W.T.: Ground truth and baseline evaluations for intrinsic image algorithms. *ICCV* (2009)
45. Liu, M., Salzmann, M., He, X.: Discrete-continuous depth estimation from a single image. *CVPR* (2014)