Scene Intrinsics and Depth from a Single Image

Evan Shelhamer
UC Berkeley
shelhamer@cs.berkeley.edu

Jonathan T. Barron
Google
barron@google.com

Trevor Darrell
UC Berkeley
trevor@cs.berkeley.edu

Abstract

Intrinsic image decomposition factorizes an observed image into its physical causes. This is most commonly framed as a decomposition into reflectance and shading, although recent progress has made full decompositions into shape, illumination, reflectance, and shading possible. However, existing factorization approaches require depth sensing to initialize the optimization of scene intrinsics. Rather than relying on depth sensors, we show that depth estimated purely from monocular appearance can provide sufficient cues for intrinsic image analysis. Our full intrinsic pipeline regresses depth by a fully convolutional network then jointly optimizes the intrinsic factorization to recover the input image. This combination yields full decompositions by uniting feature learning through deep network regression with physical modeling through statistical priors and random field regularization. This work demonstrates the first pipeline for full intrinsic decomposition of scenes from a single color image input alone.

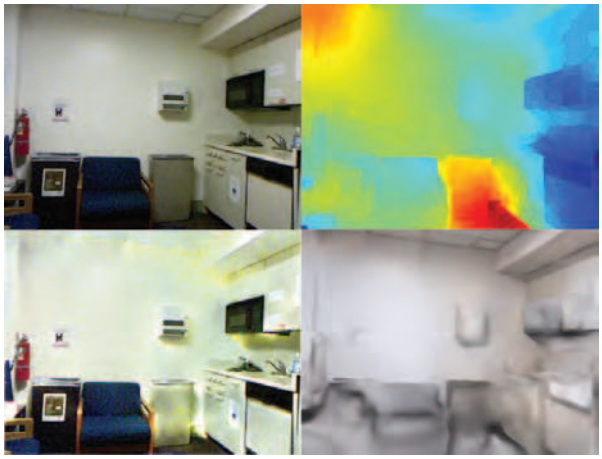

Figure 1. A full intrinsic decomposition of a scene from a single input image (top left) into depth (top right), reflectance (bottom left), and shading (bottom right). A fully convolutional network first estimates the depth, then a constrained intrinsic factorization stage jointly optimizes the intrinsic decomposition.

1. Introduction

Intrinsic image decomposition seeks to factorize an observed image into its physical causes such as reflectance, shape, and illumination, as shown in Figure 1. Each cause is represented as an "image" of a given physical quantity. Most successful approaches to this problem have focused on the decomposition of an image into a reflectance image and a shading image, the product of which should reproduce the image in question. The essence of these approaches is to express constraints corresponding to knowledge of the physics of scene formation, searching for an underlying pixel-wise factorization of observed brightness into underlying scene components subject to appropriate global constraints. The full recovery of scene intrinsics, in contrast to this more limited "shading and reflectance problem", is a significantly harder problem which has accordingly seen less progress. Due to the difficulty of this problem, recent methods have exploited observed depth information for a scene to recover an intrinsic factorization. While powerful, such methods are limited in that they can only be applied to scenes where depth observations are available, or alternatively, very strong prior constraints can be enforced. Such methods have not been applicable to scene-level intrinsic factorization of static scene images, and to our knowledge no method for reliably inferring scene intrinsics including scene depth (or surface normal information) from a single monocular image has previously been reported.

Monocular inference of depth via spatial appearance cues is a long-standing goal of computer vision. Many methods have been proposed with limited success, including methods which regress local patches across modalities or hallucinate depth textures by patch matching. These approaches were often limited either to effective local matching while ignoring the holistic content of a scene or could match the "gist" of a scene while missing the fine details. The recent advent of end-to-end, deep fully convolutional networks provides a new tool for this task, allowing training with large amounts of paired appearance and target images to learn a direct regressor to targets including semantic labels and depth [22, 21].

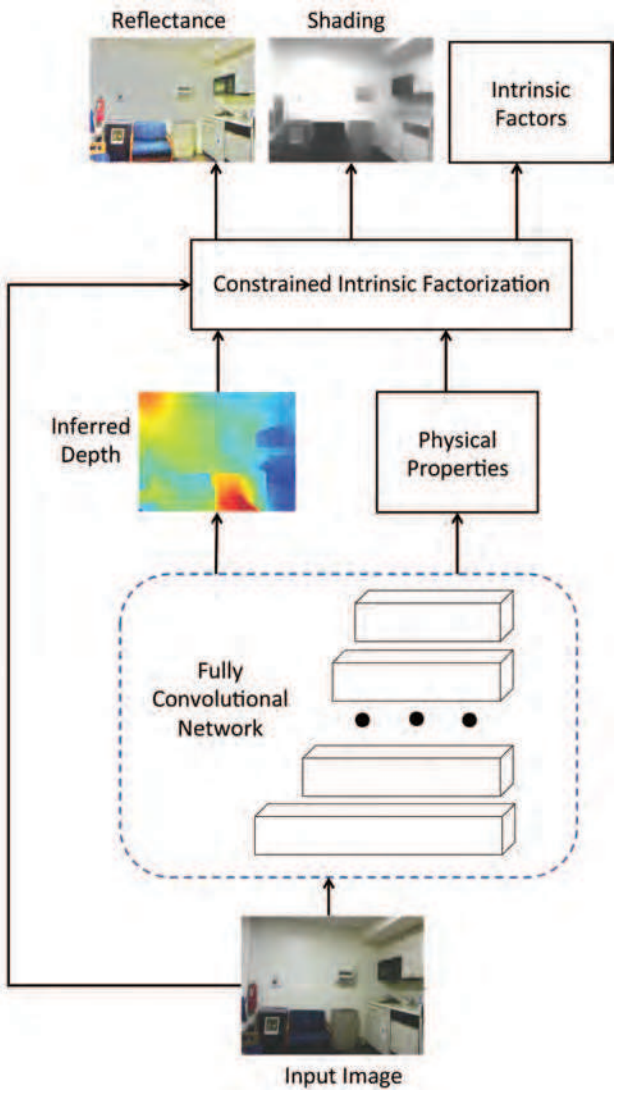

Figure 2. The full intrinsics pipeline. First, fully convolutional nets predict physical properties from the input image. The input image and FCN outputs are then fed into the constrained factorization stage. The decomposition is jointly optimized over reflectance, shading, and other intrinsic factors.

Our pipeline takes a single color image as input, infers monocular depth, and decomposes the image and the inferred depth into intrinsic reflectance and shading by joint optimization. The pipeline is illustrated in Figure 2 and described in detail in Section 3. For depth inference, we include an FCN for monocular depth regression based on the recent DCNF-FCSP network [21] as it is the state-of-the-art for indoor scenes. This network is trained end-to-end for depth estimation from a single image by fine-tuning from image classification pre-training, which demonstrates a role for recognition in transfer learning to reconstruction. Incorporating direct, data-driven depth regression by FCN into physics-constrained intrinsic factorization methods makes it possible to decompose intrinsics for a scene without depth input. For factorization, we experiment with approaches that previously required Kinect depth inputs to provide full intrinsic decomposition: the Scene-SIRFS method of Barron and Malik [2] and the albedo and component shading factorization method of Chen and Koltun [8].

The first stage leverages fully convolutional networks for the direct regression of physical properties of the scene such as depth, and then the factorization stage solves the decomposition of image appearance into intrinsic components constrained by inferred depth and other physics-based constraints.

We show that the inferred depth produced from a convnet using just a single monocular image is a strong enough cue to obtain surprisingly accurate initialization for these models, and that the resulting depth and intrinsic models are quite close to the quality obtained using ground truth Kinect observations. See Section 4 for implementation details and experimental results.

It remains an open question whether a pure convnet could learn how to completely factorize an image and invert image formation; while it may be possible, given the very limited amount of real-world training data available for many intrinsic image factors (shading, reflectance, illumination) we believe that physical modeling via the design of priors and regularization is currently necessary.

In the context of end-to-end learning, explicit factorization can be interpreted as a means to constrain model capacity. We view the combination of direct depth regression training and explicit factorization incorporating known physical constraints of image formation to be an opportune design choice which makes tractable the goal of monocular full intrinsic image recovery for general scenes.

Our results show that hallucinated depth suffices for monocular intrinsic image decomposition, and open the door for new classes of visual inference methods which now leverage full scene intrinsics—estimates of scene depth, reflectance, and illumination—for a variety of goals, including recognition, relighting, and animation. These tasks heretofore had been limited to scenes observed by a Kinect sensor or equivalent model; our pipeline allows relatively accurate estimates (compared to Kinect-based ground truth) to be obtained using a single forward pass through a fully convolutional network followed by a factorization optimization.

2. Related Work

The "intrinsic images" problem, as originally outlined by Barrow and Tenenbaum [3], is the task of unraveling a single observed image into the constituent "images" which together conspired to create that observed image. This is a fundamental task in computer vision, as "un-rendering" an image requires reasoning about the shape and surface

orientation of the underlying scene, the colors and material properties of the objects in that scene, and the global illumination which lit that scene. And just as it is a fundamental problem, it is also an extremely difficult problem, which has led many researchers to pursue solutions to sub-problems of this broad task.

As a full decomposition into shape, reflectance, and shading is our goal, physics and learning-based approaches to intrinsic images are relevant to our method, as is monocular shape and depth estimation, and prior approaches to deep image factorization.

2.1. Classic approaches

One of the earliest such approaches to this task was shape-from-shading [13], well surveyed here: [30]. Shape-from-shading only addresses the subset of the intrinsic image problem in which that all objects assumed to be smooth and white, and the illumination is assumed to be known. Reasonable shape-from-shading algorithms can be derived from first-principles using just the physics of image formation, but more successful algorithms rely on smoothness assumptions [15], hinting that even this simplified problem is best addressed in a statistical or learning-based framework.

A parallel line of research into this problem as been the "shading vs reflectance" subset of the intrinsic image problem, often confusingly referred to as "intrinsic images". The first algorithm for this task was Land and McCann's "Retinex" algorithm [20], which was used effectively by Horn [14]. Despite it's age and simplicity, Retinex was the most effective intrinsic image algorithm for a considerable time [9].

2.2. Learning and optimization-based methods

Modern attempts at this problem have employed statistical and/or learning-based approaches [5, 26] with some success, and very recent work has seen significant progress through the use of modern optimization techniques and large, well-labeled datasets [6].

Several attempts have been made at subsets of the intrinsic image problem using related machinery. Yu et al. assume the geometry of the scene to be known, and try to recover illumination and reflectance [29]. Blanz and Vetter make the very strong assumption that the image contains a specific object category (such as a human head) and try to recovery shape by modeling the variation within that object category [7]. Barron and Malik's SIRFS model relaxes nearly all of the assumptions made in prior intrinsic image work, and from a single image (of a segmented object) solves for a complete model of shape, shading, reflectance, and illumination [2]. But SIRFS was found to perform poorly at the task of recovering coarse shape, which requires reasoning about more than simple statistical priors on shapes and reflectance. This issue can be ameliorated by incorporating some external knowledge of the shape of the scene, such as a depth map from some active depth sensor [1] but this is not a satisfying conclusion in our goal for a complete intrinsic image algorithm which requires as input only a single RGB image.

The work of [27] combines Scene-SIRFS [1] with semantic reasoning, and shows that improved reasoning about intrinsics can improve semantic segmentation. Recent work on an efficient optimization-based approach introduced a heterogeneous illumination model [8]; given a Kinect depth map, this method would recover shading and reflectance and would infer both indirect illumination and direct illumination, beyond the conventional model in [1] and others. We incorporate and experiment with the models of [1] and [8] in our framework as described below.

2.3. Shape inference methods

Coarse shape recovery from an image appears to be a fundamentally hard both for algorithms [4] and for humans [18]. This is because the local appearance of an image patch is indicative of just the local shape of that patch. Resolving a complete holistic model of the depth of a scene requires integrating many independent cues which are often hard to analytically model and difficult to combine, such as contours and occlusion [28, 17], object support relationships and size priors [12], and even semantic knowledge [10]. Furthermore, because of the interconnected nature of the three dimensional world and the projective nature of image formation, reasoning about shape must be done globally, and global inference problems are difficult to make tractable. Accordingly, most recent approaches to predicting coarse depth have framed the problem as one of learning and inference [11, 23, 16], relying on large datasets and learning machinery to abstract away the difficult analytical nature of this problem. As of late, given the resurgence of interest in convolutional networks, training a convnet to directly estimate per-pixel depth has been found to be effective [21]. The insight of our work is to combine the recent success of convnet-based depth-estimation algorithms with the also-recent success of intrinsic image algorithms at recovering intrinsic image measures other than depth. Combining these two formerly disparate lines of work gives us, for the first time, an effective method for complete intrinsic image estimation.

2.4. Deep network approaches

Deep learning models are trained end-to-end to capture structure in the input but generally lack explicit representations of this structure. Instead, factors of variation are cast as "nuisance" variables to discount instead of decompose. Exceptions that define factorized deep models show promising but preliminary decompositions for the constrained setting of faces. The deep Lambertian network [25] factorizes

identity, albedo, and illumination through a conditional restricted boltzmann machine (RBM) for a generative model of face appearance. The convolutional inverse graphics network [19] decomposes a face into pose, lighting, and other hidden factors by training a variational auto-encoder with a carefully constructed curriculum of training set transformations. These models are restricted to laboratory images of faces that are low resolution, closely cropped, and lack the complexity of real world imagery. However these models have potential extensions to natural images.

3. Intrinsics Pipeline

Our pipeline connects a deep architecture for local physical prediction with joint intrinsic optimization. This combines fully convolutional networks (FCNs, [21, 22]) to estimate physical properties and a constrained intrinsic factorization (CIF) stage to optimize these estimates based on image formation factorization principles. While the physical FCNs can make efficient, direct predictions of intrinsics joint optimization is needed for a consistent decomposition. The constrained intrinsic factorization stage takes the observed input and FCN-inferred physical properties and outputs a set of intrinsic factors that reproduce the input image. These intrinsic factors may include refinements of the initial physical predictions that result from joint optimization.

This pipeline merges end-to-end feature learning with explicit factorization via physics-based vision. The fully convolutional network stage learns and infers pixel-to-pixel mappings of physical properties such as depth from sensed pixels. The constrained intrinsic factorization stage encodes visual knowledge through natural image priors and physics-inspired regularization. The quality and suitability of the inferred intrinsics is the focus of this work, but we note that in principle end-to-end learning is possible as the operations of both the FCN and CIF stages of our architecture are differentiable.

3.1. Fully Convolutional Physical Prediction

A fully convolutional network (FCN) is a model designed for spatial prediction problems. Every layer in an FCN computes a local operation on relative spatial coordinates. In this way, an FCN can take an input of any size and produce an output of corresponding dimensions. This model family is a natural choice for pixel-to-pixel mappings: for intrinsic image decomposition this could take the form of inferring depth, shadow, or specularity maps from the input image. In this usage the FCN mapping is $P = f_d\left(f_{\ldots}\left(f_1\left(I\right)\right)\right)$ where $P \in \mathbb{R}^{m \times n \times o}$ is the output map of physical properties and $I \in \mathbb{R}^{m \times n \times k}$ is the input image; these predictions P can serve as intrinsic factors F in Equation 1.

FCN inference and learning are performed on a whole image at a time by dense feedforward computation and backpropagation. The network is trained by presenting paired input and truth images for the pixel-to-pixel mapping, such as a registered color image and depth map pair from a sensor like the Kinect. Correlated properties can be jointly predicted by learning a single network that regresses different physical quantities from a shared feature representation. Although this approach requires sensed or annotated ground truth for training, there is no need for hand-engineer low-level features or structures since the model is learned end-to-end. Feedforward inference produces fast initial physical estimates.

While FCNs can provide maps of physical properties, these inferences may be noisy or inconsistent. Direct prediction is not enough to guarantee a coherent decomposition.

3.2. Constrained Intrinsic Factorization

The constrained intrinsic factorization (CIF) stage reconciles the input image with physical predictions through explicit factorization. Solving the joint optimization over intrinsic factors attempts to find a globally consistent image decomposition. The output of this network is constrained to reproduce the input image. The intrinsic factors are induced by explicit priors and regularization. Designing this stage is counter to the standard philosophy of end-to-end learning of deep networks. However, we argue that there is no need to re-discover relationships already known to natural image statistics and optics. The general form of the CIF is

$$\begin{aligned} \min_{F_1,\ldots,F_n} \quad & \sum_{i=1}^{n} c_i(F_i) \\ \text{s.t.} \quad & I = r(F_1, \ldots, F_n) \end{aligned} \tag{1}$$

for intrinsic factors F, cost functions c, and a "rendering" function r that composes the intrinsic factors to reproduce the input image I. A product or logspace sum are standard choices for r.

What costs and rendering should this net compute? We have our choice of factors F from the rich literature on intrinsic decompositions. The essential constraint is that the the reflectance and shading multiply to reproduce the image, and this can be accomplished by a hard or soft constraint. The hard formulation is to solve for reflectance or shading alone then obtain the other by dividing the image. The soft formulation is to define an energy that measures the departure of the reflectance-shading product from the observed image. Either can be instantiated in a network by a deterministic element-wise arithmetic layer or difference and norm layers respectively. With this constraint addressed, further terms may be defined to bias the decomposition to different properties such as smoothness of illumination.

4. Experiments

We evaluate our full reflectance, shading, and depth decompositions on the standard NYUDv2 dataset [24] of RGBD images. For the experiments reported here we employ a single depth regression network for the fully convolutional physical prediction network stage, and consider two different intrinsics approaches for the constrained intrinsic factorization stage.

We experiment with factorization approaches that previously required Kinect depth inputs to provide full intrinsic factorization: the Scene-SIRFS method of Barron and Malik [2] and the heterogeneous factorization method of Chen and Koltun [8]. Our model takes a single color image as input, infers monocular depth, and decomposes the image and the inferred depth into intrinsic reflectance and shading by joint optimization. The intrinsic factors produced depend on the choice of factorization method.

4.1. Dataset and experimental protocol

We evaluate our methods on NYU Depth version 2 dataset [24] by quantitative comparison with the prior methods and by visual inspection. This dataset has 1,449 curated RGB-D images captured with the Microsoft Kinect as video then post-processed. All parameters are tuned and hyperparameters selected on the 795 image trainval split. Results here are reported over a randomly selected 100 image subset of the 654 image test split for reasons of computational convenience and limited space.[1] Direct evaluation of intrinsics on this dataset is not possible without ground truth reflectance and shading; we resort to an oracle comparison to results with Kinect input. [2]

4.2. FCN depth regression implementation

We include a fully convolutional network for monocular depth in our architecture based on the recent DCNF-FCSP network [21] as it is the state-of-the-art for indoor scenes on the NYUDv2 [24] dataset. This net is trained end-to-end for depth estimation from a single image. Deep convolutional neural fields (DCNF) fuse convolutional networks and CRFs in a jointly learned architecture with unary and pairwise potentials. The fully convolutional superpixel (FCSP) variant of the model incorporates superpixel pooling into a fully convolutional network for fast inference that respects local image structure. We utilize the FCSP model in the experiments reported here, following the exact training protocols in [21].

[1] Additional images are shown in the supplemental material file, including comparisons on the set of 16 images shown in [1]'s supplementary material.

[2] The dataset in [6] is appealing in that it has ground truth human estimates of reflectance gradients, but is limited in that it does not include depth ground truth. Nonetheless, further experimentation using our DIN framework on [6] is an appealing avenue of future work.

Table 1. FCN depth regression in the DIN improves intrinsic prediction. The mean-squared-error of intrinsics from our DIN model relative to that obtained with oracle Kinect depth is low, and outperforms models not using a FCN. The mean methods are scene-SIRFS and Chen & Koltun given the mean depth image over the trainval set as input. DIN-SIRFS and DIN-CK are instantiations of our model for the corresponding constrained intrinsic factorization methods. See text for details.

	r-MSE	*s*-MSE	*rs*-MSE	Avg.
SIRFS, mean	0.0168	0.0197	0.0081	0.0139
FCN-SIRFS	0.0130	0.0109	0.0086	**0.0107**
CK, mean	0.0027	0.0031	0.0025	0.0027
DIN-CK	0.0018	0.0020	0.0023	**0.0020**

4.3. CIF implementation

For constrained intrinsic factorization we examine state-of-the-art approaches that require depth input: the Scene-SIRFS method of Barron and Malik [2] and the albedo and component shading factorization method of Chen and Koltun [8]. We refer to the resulting regression and factorization pipelines as FCN-SIRFS and FCN-CK respectively.

FCN-SIRFS The SIRFS model [2] formulates a joint optimization of shape, illumination, and reflectance from shading that is solved by multi-scale L-BFGS. The image is reproduced by absorbing the hard constraint that $I = R + S$ by absorbing R into the objective and removing it as a free parameter. This model is limited to uncluttered views of an object under a single global illumination. Naïve application of the model to scenes gives degenerate solutions with implausible shape and shading. Scene-SIRFS extends the model to natural images and scenes by introducing mixtures over shapes and lights that compose to explain the whole scene. Scene-SIRFS optimizes a full decomposition over shape, illumination, and reflectance; however, it requires a depth input to initialize the shape.

FCN-CK The Chen & Koltun model [8] decomposes an image into reflectance and a factorization of shading into direct irradiance, indirect irradiance, and illumination color. Further decomposing shading into these constituents leads to simpler regularizers for each factor.

These regularizers are nonlocal CRFs defined on different neighborhooods according to the reflectance or shading factor. The image is reproduced by a soft constraint that defines a cost for differences between the result and the observed pixels. Altogether this formulation defines a simple energy that can be minimized by least squares optimization.

4.4. Results

For a visual comparison of FCN-SIRFS and FCN-CK decompositions with the previous methods given ground truth depth results see the figures on the following pages. For quantitative comparison, see Table 1 for measures of the relative error of FCN and factorization outputs vs. in-

trinsics from mean depth image input scored against the oracle results of previous methods given the ground truth depth from post-processed Kinect depth recordings. The relative mean-squared-error of intrinsics from our FCN and factorization model vs. performance without a deep network but using mean scene depth input to the underlying factorization method is shown. Performance is reported relative to oracle intrinsics using ground truth Kinect depth inputs to SIRFS [1] and CK [8] decompositions. *r*-MSE, *s*-MSE, and *rs*-MSE are metrics for reflectance and shading error from [1] and *Avg.* is their geometric average.

5. Conclusion

Hallucinated depth suffices for monocular intrinsic image decomposition. The intrinsic outputs from ground truth and hallucinated depth are visually similar and oracle experiments verify that the hallucinated depth improves the intrinsic results relative to mean depth input. To the best of our knowledge this is the first report of a full scene intrinsics decomposition from a single input image without further information. These results serve as a new baseline for scene intrinsics.

References

[1] J. T. Barron and J. Malik. Intrinsic scene properties from a single rgb-d image. *CVPR*, 2013. 3, 5, 6, 7, 8

[2] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *TPAMI*, 2015. 2, 3, 5

[3] H. Barrow and J. Tenenbaum. Recovering intrinsic scene characteristics from images. In *Computer Vision Systems*, 1978. 2

[4] P. Belhumeur, D. Kriegman, and A. Yuille. The Bas-Relief Ambiguity. *IJCV*, 1999. 3

[5] M. Bell and W. T. Freeman. Learning local evidence for shading and reflectance. *ICCV*, 2001. 3

[6] S. Bell, K. Bala, and N. Snavely. Intrinsic images in the wild. *ACM Trans. on Graphics (SIGGRAPH)*, 33(4), 2014. 3, 5

[7] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. *SIGGRAPH*, 1999. 3

[8] Q. Chen and V. Koltun. A simple model for intrinsic image decomposition with depth cues. In *ICCV*, December 2013. 2, 3, 5, 6, 7, 8

[9] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman. Ground-truth dataset and baseline evaluations for intrinsic image algorithms. *ICCV*, 2009. 3

[10] S. Gupta, P. A. Arbeláez, R. B. Girshick, and J. Malik. Aligning 3D models to RGB-D images of cluttered scenes. *CVPR*, 2015. 3

[11] D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. *ICCV*, 1:654–661, 2005. 3

[12] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *CVPR*, 2006. 3

[13] B. K. P. Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. Technical report, MIT, 1970. 3

[14] B. K. P. Horn. Determining lightness from an image. *Computer Graphics and Image Processing*, 1974. 3

[15] K. Ikeuchi and B. Horn. Numerical shape from shading and occluding boundaries. *Artificial Intelligence*, 1981. 3

[16] K. Karsch, C. Liu, and S. B. Kang. Depthtransfer: Depth extraction from video using non-parametric sampling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2014. 3

[17] J. Koenderink. What does the occluding contour tell us about solid shape? *Perception*, 1984. 3

[18] J. Koenderink, A. Van Doorn, C. Christou, and J. Lappin. Perturbation study of shading in pictures. *Perception*, 1996. 3

[19] T. D. Kulkarni, W. Whitney, P. Kohli, and J. B. Tenenbaum. Deep convolutional inverse graphics network. *arXiv preprint arXiv:1503.03167*, 2015. 4

[20] E. H. Land and J. J. McCann. Lightness and retinex theory. *JOSA*, 1971. 3

[21] F. Liu, C. Shen, G. Lin, and I. Reid. Learning depth from single monocular images using deep convolutional neural fields. *arXiv preprint arXiv:1212.5701*, 2015. 1, 2, 3, 4, 5

[22] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014. 1, 4

[23] A. Saxena, M. Sun, and A. Ng. Make3d: learning 3d scene structure from a single still image. *TPAMI*, 2008. 3

[24] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 5

[25] Y. Tang, R. Salakhutdinov, and G. Hinton. Deep lambertian networks. In *ICML*, 2012. 3

[26] M. F. Tappen, W. T. Freeman, and E. H. Adelson. Recovering intrinsic images from a single image. *TPAMI*, 2005. 3

[27] V. Vineet, C. Rother, and P. Torr. Higher order priors for joint intrinsic image, objects, and attributes estimation. *NIPS*, 2013. 3

[28] A. Witkin. Shape from contour. *AITR*, 1980. 3

[29] Y. Yu, P. Debevec, J. Malik, and T. Hawkins. Inverse global illumination: recovering reflectance models of real scenes from photographs. *SIGGRAPH*, 1999. 3

[30] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah. Shape-from-shading: a survey. *TPAMI*, 2002. 3

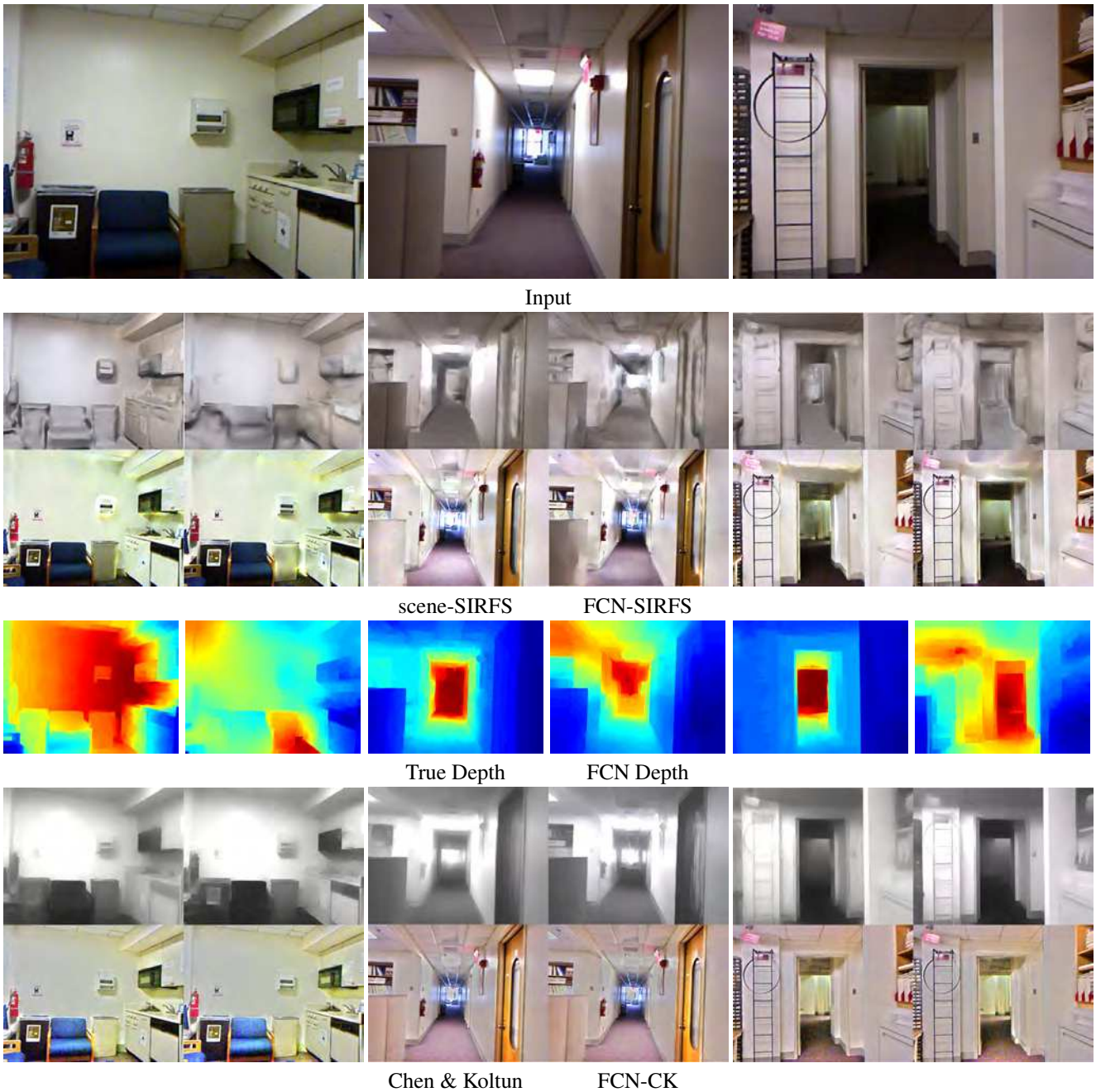

Figure 3. Comparison of intrinsic decompositions given real Kinect depth sensor input and our network and factorization pipeline with one input per column. The rows from top to bottom are: the input image, the SIRFS (left) and FCN-SIRFS (right) reflectance and shading, the true (left) and hallucinated (right) depth, and the Chen & Koltun (left) and FCN-CK (right) reflectance and shading. The FCN-SIRFS and FCN-CK outputs are similar to the respective decompositions of Barron & Malik [1] and Chen & Koltun [8] on real depth.

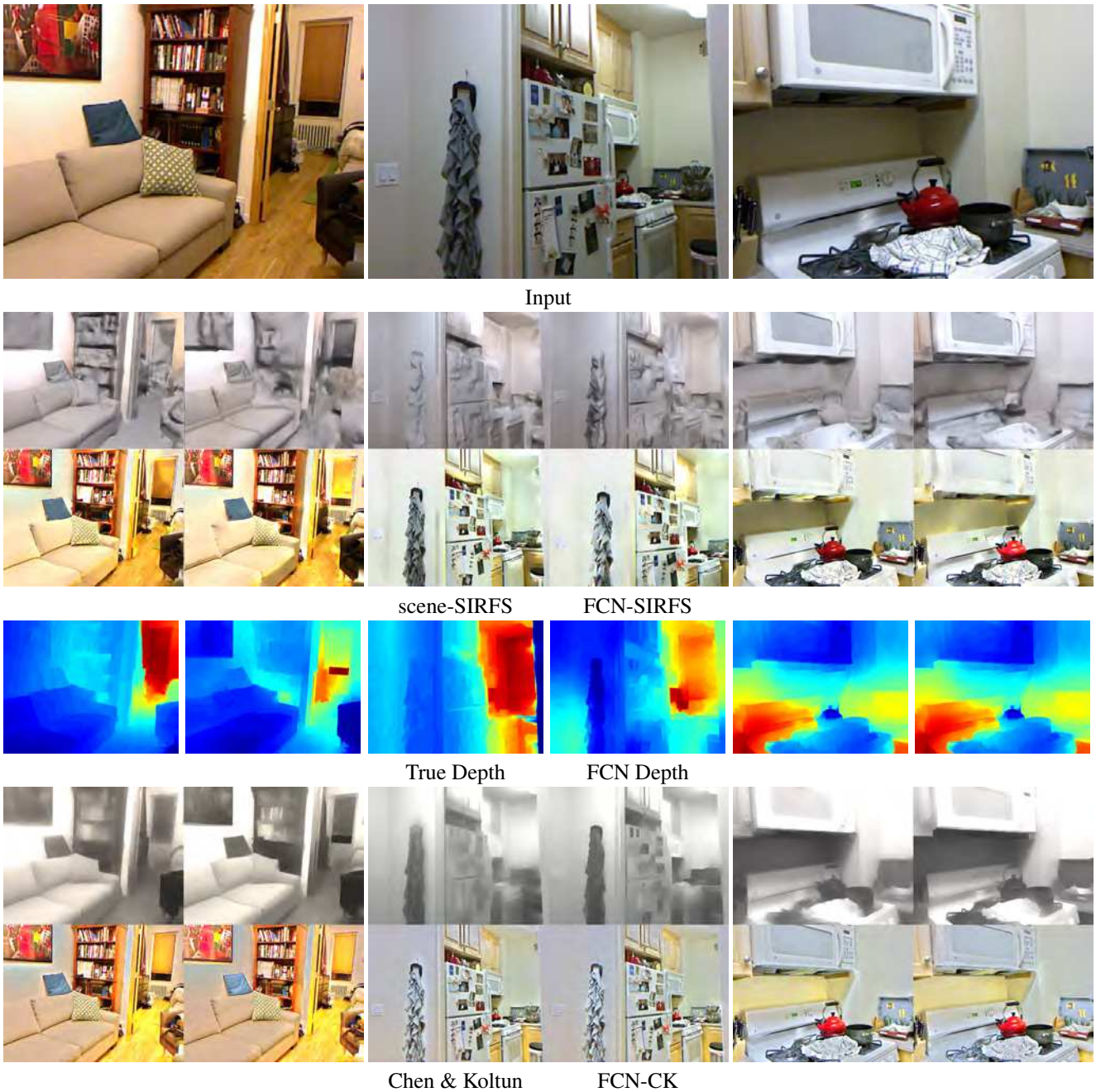

Figure 4. Comparison of intrinsic decompositions given real Kinect depth sensor input and our network and factorization pipeline with one input per column. The rows from top to bottom are: the input image, the SIRFS (left) and FCN-SIRFS (right) reflectance and shading, the true (left) and hallucinated (right) depth, and the Chen & Koltun (left) and FCN-CK (right) reflectance and shading. The FCN-SIRFS and FCN-CK outputs are similar to the respective decompositions of Barron & Malik [1] and Chen & Koltun [8] on real depth.