Self-Supervised Intrinsic Image Decomposition

Michael Janner
MIT
janner@mit.edu

Jiajun Wu
MIT
jiajunwu@mit.edu

Tejas D. Kulkarni
DeepMind
tejasdkulkarni@gmail.com

Ilker Yildirim
MIT
ilkery@mit.edu

Joshua B. Tenenbaum
MIT
jbt@mit.edu

Abstract

Intrinsic decomposition from a single image is a highly challenging task, due to its inherent ambiguity and the scarcity of training data. In contrast to traditional fully supervised learning approaches, in this paper we propose learning intrinsic image decomposition by explaining the input image. Our model, the Rendered Intrinsics Network (*RIN*), joins together an image decomposition pipeline, which predicts reflectance, shape, and lighting conditions given a single image, with a recombination function, a learned shading model used to recompose the original input based off of intrinsic image predictions. Our network can then use unsupervised reconstruction error as an additional signal to improve its intermediate representations. This allows large-scale unlabeled data to be useful during training, and also enables transferring learned knowledge to images of unseen object categories, lighting conditions, and shapes. Extensive experiments demonstrate that our method performs well on both intrinsic image decomposition and knowledge transfer.

1 Introduction

There has been remarkable progress in computer vision, particularly for answering questions such as *"what is where?"* given raw images. This progress has been possible due to large labeled training sets and representation learning techniques such as convolutional neural networks [LeCun et al., 2015]. However, the general problem of visual scene understanding will require algorithms that extract not only object identities and locations, but also their shape, reflectance, and interactions with light. Intuitively disentangling the contributions from these three components, or *intrinsic images*, is a major triumph of human vision and perception. Conferring this type of intuition to an algorithm, though, has proven a difficult task, constituting a major open problem in computer vision.

This problem is challenging in particular because it is fundamentally underconstrained. Consider the porcelain vase in Figure 1a. Most individuals would have no difficulty identifying the true colors and shape of the vase, along with estimating the lighting conditions and the resultant shading on the object, as those shown in 1b. However, the alternatives in 1c, which posits a flat shape, and 1d, with unnatural red lighting, are entirely consistent in that they compose to form the correct observed vase in 1a.

The task of finding appropriate intrinsic images for an object is then not a question of simply finding a valid answer, as there are countless factorizations that would be equivalent in terms of their rendered combination, but rather of finding the most probable answer. Roughly speaking, there are two methods of tackling such a problem: a model must either (1) employ handcrafted priors on the reflectance, shape, and lighting conditions found in the natural world in order to assign probabilities

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

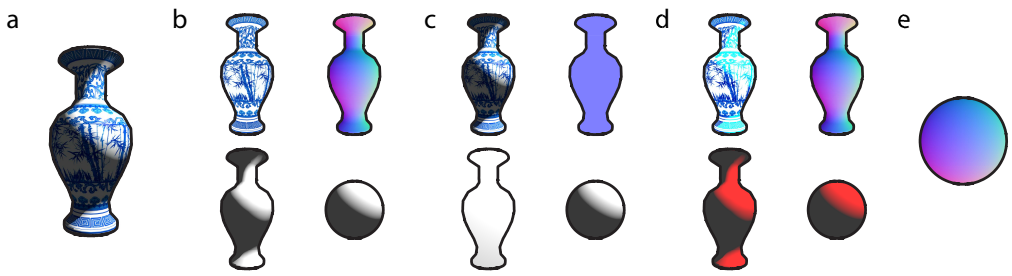

Figure 1: A porcelain vase (a) along with three predictions (b-d) for its underlying intrinsic images. The set in (c) assumes the contribution from shading is negligible by predicting a completely flat rather than rounded shape. The reflectance is therefore indistinguishable from the observed image. The set in (d) includes the correct shape but assumes red lighting and a much brighter blue color in the regions affected by shading. While the decomposition in (b) is much more inuitively pleasing than either of these alternatives, all of these options are valid in that they combine to exactly form the observed vase. (e) shows a sphere with our visualized normals map as a shape reference.

to intrinsic image proposals or (2) have access to a library of ground truth intrinsic images and their corresponding composite images.

Unfortunately, there are limitations to both methods. Although there has been success with the first route in the past [Barron and Malik, 2015], strong priors are often difficult to hand-tune in a generally useful fashion. On the other hand, requiring access to complete, high quality ground truth intrinsic images for real world scenes is also limiting, as creating such a training set requires an enormous amount of human effort and millions of crowd-sourced annotations [Bell et al., 2014].

In this paper, we propose a deep structured autoencoder, the Rendered Instrinsics Network (*RIN*), that disentangles intrinsic image representations and uses them to reconstruct the input. The decomposition model consists of a shared convolutional encoder for the observation and three separate decoders for the reflectance, shape, and lighting. The shape and lighting predictions are used to train a differentiable shading function. The output of the shader is combined with the reflectance prediction to reproduce the observation. The minimal structure imposed in the model – namely, that intrinsic images provide a natural way of disentangling real images and that they provide enough information to be used as input to a graphics engine – makes RIN act as an autoencoder with useful intermediate representations.

The structure of RIN also exploits two natural sources of supervision: one applied to the intermediate representations themselves, and the other to the reconstructed image. This provides a way for RIN to improve its representations with unlabeled data. By avoiding the need for intrinsic image labels for all images in the dataset, RIN can adapt to new types of inputs even in the absence of ground truth data. We demonstrate the utility of this approach in three transfer experiments. RIN is first trained on a simple set of five geometric primitives in a supervised manner and then transferred to common computer vision test objects. Next, RIN is trained on a dataset with a skewed underlying lighting distribution and fills in the missing lighting conditions on the basis of unlabeled observations. Finally, RIN is trained on a single Shapenet category and then transferred to a separate, highly dissimilar category.

Our contributions are three-fold. First, we propose a novel formulation for intrinsic image decomposition, incorporating a differentiable, unsupervised reconstruction loss into the loop. Second, we instantiate this approach with the RIN, a new model that uses convolutional neural networks for both intrinsic image prediction and recombination via a learned shading function. This is also the first work to apply deep learning to the full decomposition into reflectance, shape, lights, and shading, as prior work has focused on the reflectance-shading decomposition. Finally, we show that RIN can make use of unlabeled data to improve its intermediate intrinsic image representations and transfer knowledge to new objects unseen during training.

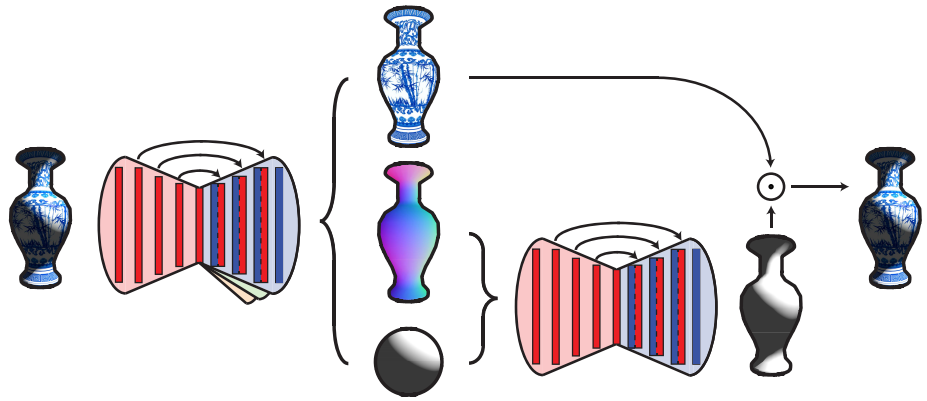

Figure 2: RIN contains two convolutional encoder-decoders, one used for predicting the intrinsic images from an input and another for predicting the shading stemming from a light source applied on a shape. The two networks together function as a larger structured autoencoder, forcing a specific type of intermediate representation in order to reconstruct the input image.

2 Related Work

Intrinsic images were introduced by Barrow and Tenenbaum as useful mid-level scene descriptors [Barrow and Tenenbaum, 1978]. The model posits that an image can be expressed as the pointwise product between contributions from the true colors of an object, or its reflectance, and contributions from the shading on that object:

$$\text{image } I = \text{reflectance } R \cdot \text{shading } S \tag{1}$$

Decomposing one step further, the shading is expressed as some function of an object's shape and the ambient lighting conditions. The exact nature of this shading function varies by implementation.

Early work on intrinsic image decomposition was based on insights from Land's Retinex Theory [Land and McCann, 1971]. Horn [1974] separated images into true colors and shading using the assumption that large image gradients tend to correspond to reflectance changes and small gradients to lighting changes. While this assumption works well for a hypothetical *Mondrian World* of flat colors, it does not always hold for natural images. In particular, Weiss [2001] found that this model of reflectance and lighting is rarely true for outdoor scenes.

More recently, Barron and Malik [2015] developed an iterative algorithm called SIRFS that maximizes the likelihood of intrinsic image proposals under priors derived from regularities in natural images. SIRFS proposes shape and lighting estimates and combines them via a spherical harmonics renderer to produce a shading image. Lombardi and Nishino [2012, 2016] and Oxholm and Nishino [2016] proposed a Bayesian formulation of such an optimization procedure, also formulating priors based on the distribution of material properties and the physics of lighting in the real world. Researchers have also explored reconstructing full 3D shapes through intrinsic images by making use of richer generative models [Kar et al., 2015, Wu et al., 2017].

Tang et al. [2012] combined Lambertian reflectance assumptions with Deep Belief Networks to learn a prior over the reflectance of greyscale images and applied their *Deep Lambertian Network* to one-shot face recognition. Narihira et al. [2015b] applied deep learning to intrinsic images first using human judgments on real images and later in the context of animated movie frames [Narihira et al., 2015a]. Rematas et al. [2016] and Hold-Geoffroy et al. [2017] also used convolutional neural networks to estimate reflectance maps and illumination parameters, respectively, in unconstrained outdoor settings.

Innamorati et al. [2017] generalized the intrinsic image decomposition by considering the contributions of specularity and occlusion in a direction-dependent model. Shi et al. [2017] found improved performance in the full decomposition by incorporating skip layer connections [He et al., 2016] in the network architecture, which were used to generate much crisper images. Our work can be seen

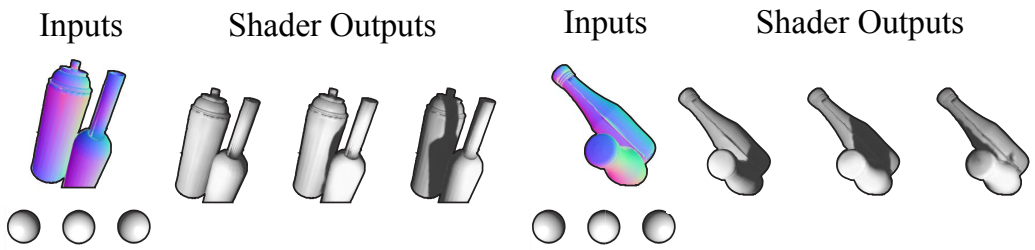

Figure 3: In contrast to simple Lambertian shading techniques, our learned shading model can handle shadows cast between objects. Inputs to the shader are shape and lighting parameter pairs.

as a further extension of these models which aims to relax the need for a complete set of ground truth data by modeling the image combination process, as in Nalbach et al. [2017].

Incorporating a domain-specific decoder to reconstruct input images has been explored by Hinton *et al.* in their *transforming autoencoders* [Hinton et al., 2011], which also learned natural representations of images in use by the vision community. Our work differs in the type of representation in question, namely images rather than descriptors like affine transformations or positions. Kulkarni et al. [2015] were also interested in learning disentangled representations in an autoencoder, which they achieved by selective gradient updates during training. Similarly, Chen et al. [2016] showed that a mutual information objective could drive disentanglement of a deep network's intermediate representation.

3 Model

3.1 Use of Reconstruction

RIN differs most strongly with past work in its use of the reconstructed input. Other approaches have fallen into roughly two groups in this regard:

1. Those that solve for one of the intrinsic images to match the observed image. SIRFS, for example, predicts shading and then solves equation 1 for reflectance given its prediction and the input [Barron and Malik, 2015]. This ensures that the intrinsic image estimations combine to form exactly the observed image, but also deprives the model of any reconstruction error.
2. Data-driven techniques that rely solely on ground truth labelings [Narihira et al., 2015a, Shi et al., 2017]. These approaches assume access to ground truth labels for all inputs and do not explicitly model the reconstruction of the input image based on intrinsic image predictions.

Making use of the reconstruction for this task has been previously unexplored because such an error signal can be difficult to interpret. Just as the erroneous intrinsic images in Fig 1c-d combine to reconstruct the input exactly, one cannot assume that low reconstruction error implies accurate intrinsic images. An even simpler degenerate solution that yields zero reconstruction error is:

$$\hat{R} = I \quad \text{and} \quad \hat{S} = \mathbf{1}\mathbf{1}^T, \tag{2}$$

where $\hat{S}$ is the all-ones matrix. It is necessary to further constrain the predictions such that the model does not converge to such explanations.

3.2 Shading Engine

RIN decomposes an observation into reflectance, shape, and lighting conditions. As opposed to models which estimate only reflectance and shading, which may make direct use of Equation 1 to generate a reconstruction, we must employ a function that transforms our shape and lighting predictions into a shading estimate. Linear Lambertian assumptions could reduce such a function to a straightforward dot product, but would produce a shading function incapable of modeling lighting conditions that drastically change across an image or ray-tracing for the purposes of casting shadows.

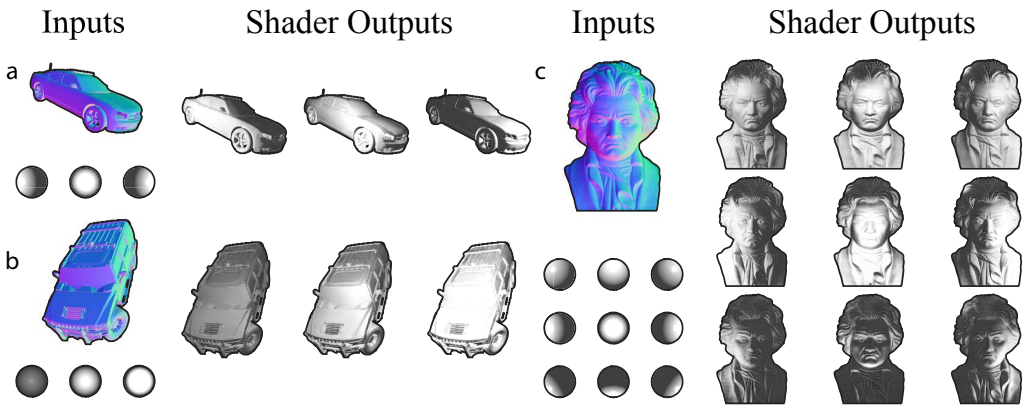

Figure 4: Our shading model's outputs after training only on synthetic car models from the ShapeNet dataset [Chang et al., 2015]. (a) shows the effect of panning the light horizontally and (b) shows the effect of changing the intensity of the light. The input lights are visualized by rendering them onto a sphere. Even though the shader was trained only on synthetic data, it generalizes well to real shapes with no further training. The shape input to (c) is an estimated normals map of a Beethoven bust [Quéau and Durou, 2015].

Instead, we opt to learn a shading model. Such a model is not limited in the way that a pre-defined shading function would be, as evidenced by shadows cast between objects in Fig 3. Learning a shader also has the benefit of allowing for different representations of lighting conditions. In our experiments, lights are defined by a position in three-dimensional space and a magnitude, but alternate representations such as the radius, orientation, and color of a spotlight could be just as easily adopted. For work that employs the shading engine from SIRFS [Barron and Malik, 2015] instead of learning a shader in a similar disentanglement context, see Shu et al. [2017]. The SIRFS engine represents lights as spherical harmonics coefficient vectors.

3.3 Architecture

Our model consists of two convolutional encoder-decoder networks, the first of which predicts intrinsic images from an observed image, and the second of which approximates the shading process of a rendering engine. Both networks employ mirror-link connections introduced by Shi et al. [2017], which connect layers of the encoder and decoder of the same size. These connections yield sharper results than the blurred outputs characteristic of many deconvolutional models.

The first network has a single encoder for the observation and three separate decoders for the reflectance, lighting, and shape. Unlike Shi et al. [2017], we do not link layers between the decoders so that it is possible to update the weights of one of the decoders without substantially affecting the others, as is useful in the transfer learning experiments. The encoder has 5 convolutional layers with $\{16, 32, 64, 128, 256\}$ filters of size 3×3 and stride of 2. Batch normalization [Ioffe and Szegedy, 2015] and ReLU activation are applied after every convolutional layer. The layers in the reflectance and shape decoders have the same number of features as the encoder but in reverse order plus a final layer with 3 output channels. Spatial upsampling is applied after the convolutional layers in the decoders. The lighting decoder is a simple linear layer with an output dimension of four (corresponding to a position in three-dimensional space and an intensity of the light).

The shape is passed as input to the shading encoder directly. The lighting estimate is passed to a fully-connected layer with output dimensionality matching that of the shading encoder's output, which is concatenated to the encoded shading representation. The shading decoder architecture is the same as that of the first network. The final component of RIN, with no learnable parameters, is a componentwise multiplication between the output of the shading network and the predicted reflectance.

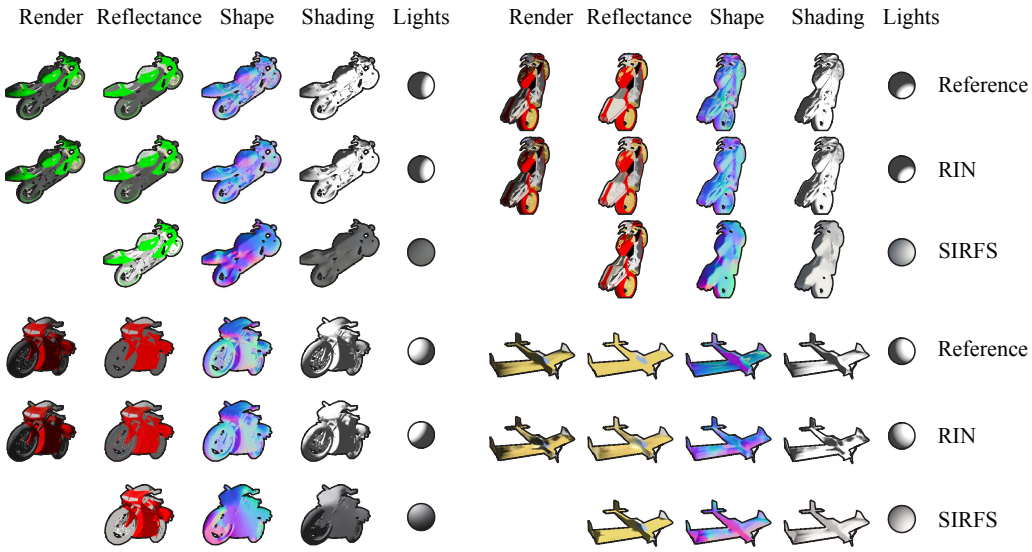

Figure 5: Intrinsic image prediction from our model on objects from the training category (motorbikes) as well as an example from outside this category (an airplane). The quality of the airplane intrinsic images is significantly lower, which is reflected in the reconstruction (labeled "Render" in the RIN rows). This allows reconstruction to drive the improvement of the intermediate intrinsic image representations. Predictions from SIRFS are shown for comparison. Note that the reflectance in SIRFS is defined based on the difference between the observation and shading prediction, so there is not an analogous reconstruction.

	Motorbike (Train)			**Airplane (Transfer)**		
	Reflectance	Shape	Lights	Reflectance	Shape	Lights
RIN	0.0021	0.0044	0.1398	0.0042	0.0119	0.4873
SIRFS	0.0059	0.0094	–	0.0054	0.0080	–

Table 1: MSE of our model and SIRFS on a test set of ShapeNet motorbikes, the category used to train RIN, and airplanes, a held-out class. The lighting representation of SIRFS (a vector with 27 components) is sufficiently different from that of our model that we do not attempt to compare performance here directly. Instead, see the visualization of lights in Fig 5.

4 Experiments

RIN makes use of unlabeled data by comparing its reconstruction to the original input image. Because our shading model is fully differentiable, as opposed to most shaders that involve ray-tracing, the reconstruction error may be backpropagated to the intrinsic image predictions and optimized via a standard coordinate ascent algorithm. RIN has one shared encoder for the intrinsic images but three separate decoders, so the appropriate decoder can be updated while the others are held fixed.

In the following experiments, we first train RIN (including the shading model) on a dataset with ground truth labels for intrinsic images. This is treated as a standard supervised learning problem using mean squared error on the intrinsic image predictions as a loss. The model is then trained further on an additional set of *unlabeled* data using only reconstruction loss as an error signal. We refer to this as the self-supervised transfer. For both modes of learning, we optimize using Adam [Kingma and Ba, 2015].

During transfer, one half of a minibatch will consist of the unlabeled transfer data the other half will come from the labeled data. This ensures that the representations do not shift too far from those learned during the initial supervised phase, as the underconstrained nature of the problem can drive

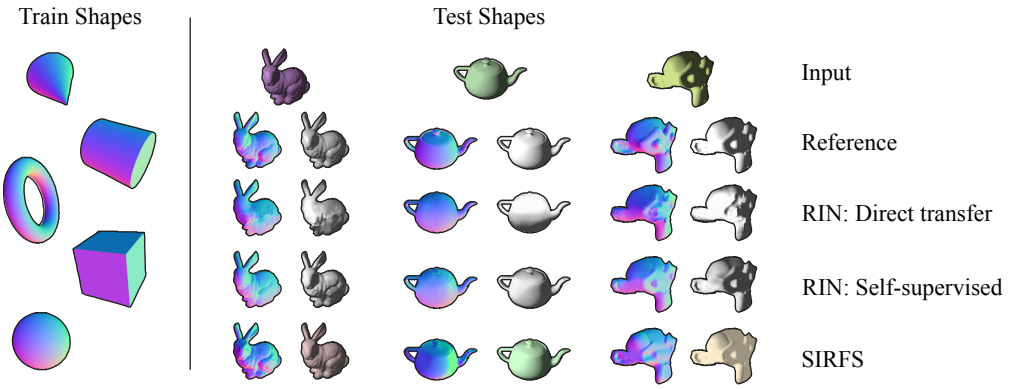

Figure 6: Predictions of RIN before ("Direct transfer") and after ("Self-supervised") it adapts to new shapes on the basis of unlabeled data.

	Stanford Bunny		**Utah Teapot**		**Blender Suzanne**	
	Shape	Shading	Shape	Shading	Shape	Shading
Direct transfer	0.074	0.071	0.036	0.043	0.086	0.104
Self-supervised	0.048	0.005	0.029	0.003	0.058	0.007

Table 2: MSE of RIN trained on five geometric primitives before and after self-supervised learning of more complicated shapes.

the model to degenerate solutions. When evaluating our model on test data, we use the outputs of the three decoders and the learned shader directly; we do not enforce that the predictions must explain the input exactly.

Below, we demonstrate that our model can effectively transfer to different shapes, lighting conditions, and object categories without ground truth intrinsic images. However, for this unsupervised transfer to yield benefits, there must be a sufficient number of examples of the new, unlabeled data. For example, the MIT Intrinsic Images dataset [Grosse et al., 2009], containing twenty real-world images, is not large enough for the unsupervised learning to affect the representations of our model. In the absence of any unsupervised training, our model is similar to that of Shi et al. [2017] adapted to predict the full set of intrinsic images.

4.1 Supervised training

Data The majority of data was generated from ShapeNet [Chang et al., 2015] objects rendered in Blender. For the labeled datasets, the rendered composite images were accompanied by the object's reflectance, a map of the surface normals at each point, and the parameters of the lamp used to light the scene. Surface normals are visualized by mapping the XYZ components of normals to appropriate RGB ranges. For the following supervised learning experiments, we used a dataset size of 40,000 images.

Intrinsic image decomposition The model in Fig 5 was trained on ShapeNet motorbikes. Although it accurately predicts the intrinsic images of the train class, its performance drops when tested on other classes. In particular, the shape predictions suffer the most, as they are the most dissimilar from anything seen in the training set. Crucially, the poor intrinsic image predictions are reflected in the reconstruction of the input image. This motivates the use of reconstruction error to drive improvement of intrinsic images when there is no ground truth data.

Shading model In contrast with the intrinsic image decomposition, shading prediction generalized well outside of the training set. The shader was trained on the shapes and lights from the same set of rendered synthetic cars as above. Even though this represents only a narrow distribution over

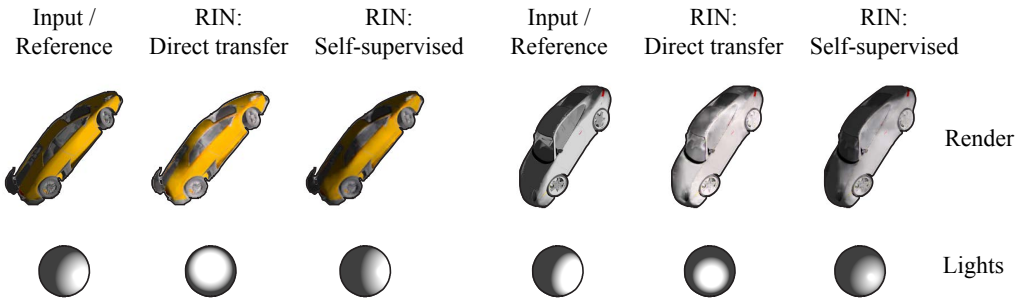

Figure 7: Predictions of RIN trained on left-lit images before and after self-supervised learning on right-lit images. RIN uncovers the updated lighting distribution without external supervision or ground truth data.

shapes, we found that the shader produced plausible predictions for even real-world objects (Fig 4). Because the shader generalized without any further effort, its parameters were never updated during self-supervised training. Freezing the parameters of the shader prevents our model from producing nonsensical shading images.

4.2 Shape transfer

Data We generated a dataset of five shape primitives (cubes, spheres, cones, cylinders, and toruses) viewed at random orientations using the Blender rendering engine. These images are used for supervised training. Three common reference shapes (Stanford bunny, Utah teapot, and Blender's Suzanne) are used as the unlabeled transfer class. To isolate the effects of shape mismatch in the labeled versus unlabeled data, all eight shapes were rendered with random monochromatic materials and a uniform distribution over lighting positions within a contained region of space in front of the object. The datasets consisted of each shape rendered with 500 different colors, with each colored shape being viewed at 10 orientations.

Results By only updating weights for the shape decoder during self-supervised transfer, the predictions for held-out shapes improves by 29% (averaged across the three shapes). Because a shape only affects a rendered image via shading, the improvement in shapes comes alongside an improvement in shading predictions as well. Shape-specific results are given in Table 2 and visualized in Fig 6.

4.3 Lighting transfer

Data Cars from the ShapeNet 3D model repository were rendered at random orientations and scales. In the labeled data, they were lit only from the left side. In the unlabeled data, they were lit from both the left and right.

Results Before self-supervised training on the unlabeled data, the model's distribution over lighting predictions mirrored that of the labeled training set. When tested on images lit from the right, then, it tended to predict centered lighting. After updating the lighting decoder based on reconstruction error from these right-lit images though, the model's lighting predictions more accurately reflected the new distribution and lighting mean-squared error reduces by 18%. Lighting predictions, along with reconstructions, for right-lit images are shown in Fig 7.

4.4 Category transfer

In the previous transfer experiments, only one intrinsic image was mismatched between the labeled and unlabeled data, so only one of RIN's decoders needed updating during transfer. When transferring between object categories, though, there is not such a guarantee. Although it might be expected that a model trained on sufficiently many object categories would learn a generally-useful distribution over reflectances, it is difficult to ensure that this is the case. We are interested in these sorts of

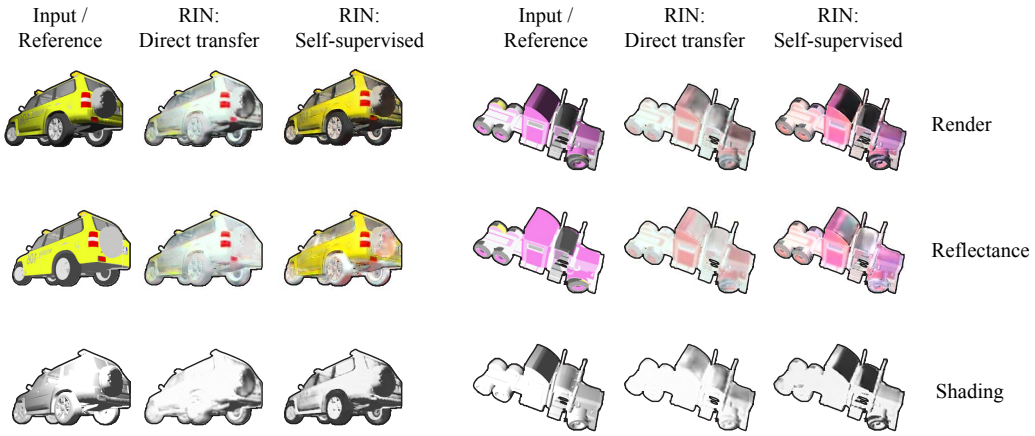

Figure 8: RIN was first trained on ShapeNet airplanes and then tested on cars. Because most of the airplanes were white, the reflectance predictions were washed out even for colorful cars. RIN fixed the mismatch between datasets without any ground truth intrinsic images of cars.

	Reflectance	Shape	Lights	Shading	Render
Direct transfer	0.019	0.014	0.584	0.065	0.035
Self-supervised	0.015	0.014	0.572	0.044	0.006

Table 3: MSE of RIN trained on ShapeNet airplanes before and after self-supervised transfer to cars. Although RIN improves its shading predictions, these are not necessarily driven by an improvement in shape prediction.

scenarios to determine how well self-supervised transfer works when more than one decoder needs to be updated to account for unlabeled data.

Data Datasets of ShapeNet cars and airplanes were created analogously to those in Section 4.1. The airplanes had a completely different color distribution than the cars as they were mostly white, whereas the cars had a more varied reflectance distribution. The airplanes were used as the labeled category to ensure a mismatch between the train and transfer data.

Results To transfer to the new category, we allowed updates to all three of the RIN decoders. (The shader was left fixed as usual.) There were pronounced improvements in the shading predictions (32%) accompanied by modest improvements in reflectances (21%). The shading predictions were not always caused by improved shape estimates. Because there is a many-to-one mapping from shape to shading (conditioned on a lighting condition), it is possible for the shape predictions to worsen in order to improve the shading estimates. The lighting predictions also remained largely unchanged, although for the opposite reason: because no lighting region were intentionally left out of the training data, the lighting predictions were adequate on the transfer classes even without self-supervised learning.

5 Conclusion

In this paper, we proposed the Rendered Intrinsics Network for intrinsic image prediction. We showed that by learning both the image decomposition and recombination functions, RIN can make use of reconstruction loss to improve its intermediate representations. This allowed unlabeled data to be used during training, which we demonstrated with a variety of transfer tasks driven solely by self-supervision. When there existed a mismatch between the underlying intrinsic images of the labeled and unlabeled data, RIN could also adapt its predictions in order to better explain the unlabeled examples.

Acknowledgements

This work is supported by ONR MURI N00014-16-1-2007, the Center for Brain, Minds and Machines (NSF #1231216), Toyota Research Institute, and Samsung.

References

Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE TPAMI*, 37 (8):1670–1687, 2015.

H.G. Barrow and J.M. Tenenbaum. Recovering intrinsic scene characteristics from images. *Computer Vision Systems*, 1978.

Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM TOG*, 33(4):159, 2014.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

Xi Chen, Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016.

Roger Grosse, Micah K. Johnson, Edward H. Adelson, and William T. Freeman. Ground-truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *ICANN*, 2011.

Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-Francois Lalonde. Deep outdoor illumination estimation. In *CVPR*, 2017.

Berthold K.P. Horn. Determining lightness from an image. *Computer Graphics and Image Processing*, 3: 277–299, 1974.

Carlo Innamorati, Tobias Ritschel, Tim Weyrich, and Niloy J. Mitra. Decomposing single images for layered photo retouching. *Computer Graphics Forum*, 36:15–25, 07 2017.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *CVPR*, 2015.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *NIPS*, 2015.

Edwin H. Land and John J. McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61:1–11, 1971.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Stephen Lombardi and Ko Nishino. Single image multimaterial estimation. In *CVPR*, 2012.

Stephen Lombardi and Ko Nishino. Reflectance and illumination recovery in the wild. *IEEE TPAMI*, 38(1): 129–141, 2016.

Oliver Nalbach, Elena Arabadzhiyska, Dushyant Mehta, Hans-Peter Seidel, and Tobias Ritschel. Deep shading: Convolutional neural networks for screen-space shading. *Computer Graphics Forum*, 36(4), 2017.

Takuya Narihira, Michael Maire, and Stella X. Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *ICCV*, 2015a.

Takuya Narihira, Michael Maire, and Stella X. Yu. Learning lightness from human judgement on relative reflectance. In *CVPR*, 2015b.

Geoffrey Oxholm and Ko Nishino. Shape and reflectance estimation in the wild. *IEEE TPAMI*, 38(2):376–389, 2016.

Yvain Quéau and Jean-Denis Durou. Edge-preserving integration of a normal field: Weighted least-squares, tv and L1 approaches. In *International Conference on Scale Space and Variational Methods in Computer Vision*, 2015.

Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Efstratios Gavves, and Tinne Tuytelaars. Deep reflectance maps. In *CVPR*, June 2016.

Jian Shi, Yue Dong, Hao Su, and Stella X. Yu. Learning non-lambertian object intrinsics across shapenet categories. In *CVPR*, 2017.

Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, July 2017.

Yichuan Tang, Ruslan Salakhutdinov, and Geoffrey Hinton. Deep lambertian networks. In *ICML*, 2012.

Yair Weiss. Deriving intrinsic images from image sequences. In *ICCV*, 2001.

Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, William T Freeman, and Joshua B Tenenbaum. Marrnet: 3d shape reconstruction via 2.5d sketches. In *NIPS*, 2017.