Direct Intrinsics: Learning Albedo-Shading Decomposition by Convolutional Regression

code online

Takuya Narihira — *UC Berkeley / ICSI / Sony Corp.*

Michael Maire — *TTI Chicago*

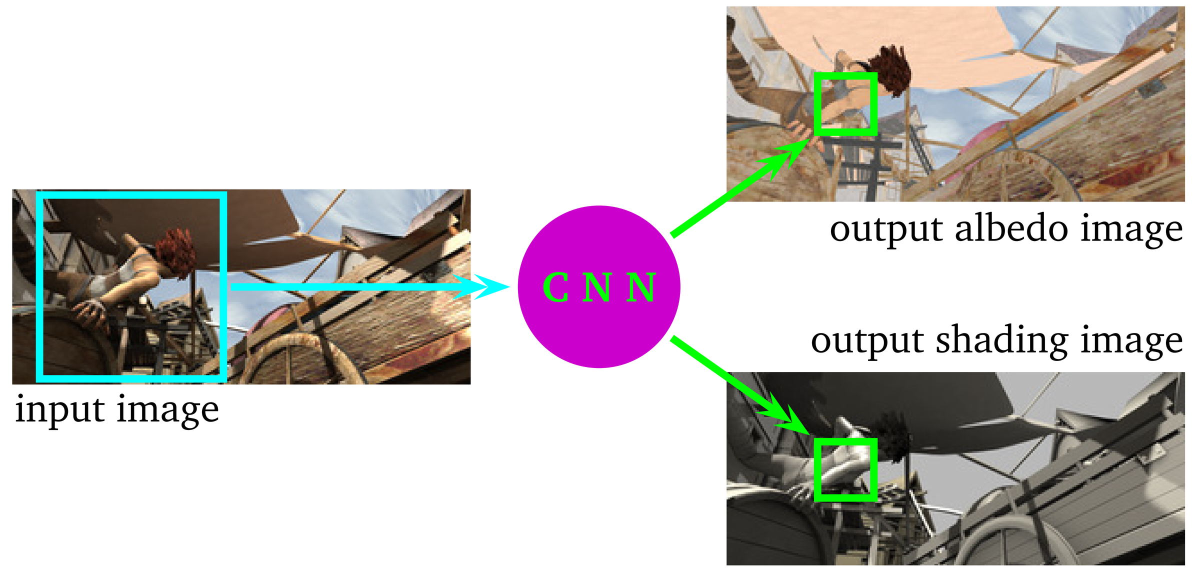
Stella X. Yu — *UC Berkeley / ICSI*

Overview

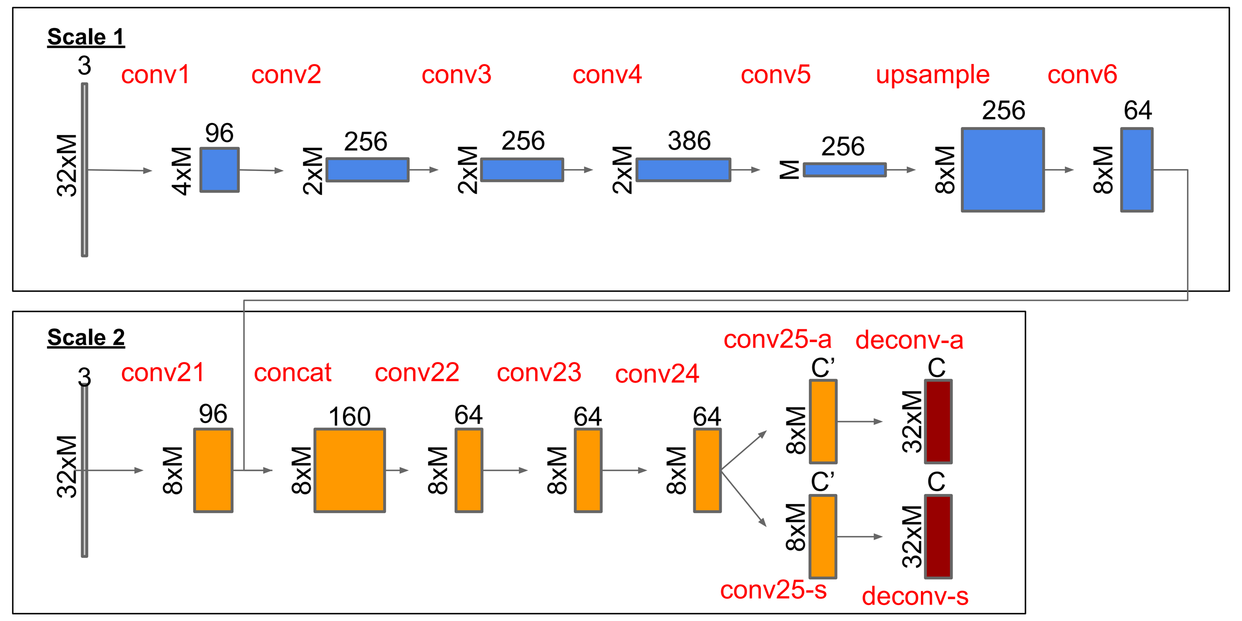
Goal: Decompose $I = A \cdot S$ (albedo, shading components)

Approach:

- Direct prediction via convolutional neural network (CNN)
- RGB input only (depth channel not used)
- Utilize synthetic data (MPI Sintel) during training
- Test on real and synthetic RGB data; outperforms RGB+D work



CNN Architecture



Our multiscale CNN regression (MSCR) architecture extends prior designs [4]:

Multiscale for global/local context fusion

- Scale 1 coarse net for global context
- Scale 2 fine net for local information
- Arbitrary size input (fully convolutional)

PReLUs for better convergence

- Learn negative slopes a_i

$$g(x_i) = \begin{cases} x_i, & x_i \geq 0 \\ a_i x_i, & x_i < 0 \end{cases}$$

Deconvolution for finer output

- Learnable convolutional upsampling filters
- Apply at the end of network
- $C' = C = 3$ for upsampling (our baseline)
- $C' = 64,\ C = 3$ for deconvolution

Experimental variants

- Hypercolumns in scale 1 for cue fusion
- Training: alternative loss functions
- Training: data augmentation and synthesis

Training Loss Functions

Scale Invariant Loss [4]

$$\mathcal{L}_{\mathrm{SI}}(Y^*, Y) = \frac{1}{n}\sum_{i,j,c} y_{i,j,c}^2 - \lambda \frac{1}{n^2}\left(\sum_{i,j,c} y_{i,j,c}\right)^2$$

- Y^*: ground-truth in *log* space, Y: prediction map, $y = Y^* - Y$
- Imposed on both albedo and shading outputs

Gradient Loss

$$\mathcal{L}_{\mathrm{grad}}(Y^*, Y) = \frac{1}{n}\sum_{i,j,c}\left[\nabla_i y_{i,j,c}^2 + \nabla_j y_{i,j,c}^2\right]$$

- ∇_i, ∇_j: derivative operators in the i- and j-dimensions
- Optionally applied to albedo to account for piece-wise constancy

Training Data Augmentation

Random augmentation

- Cropping, horizontal mirroring (baseline), scaling and rotation (DA)
- Dropout with $p = 0.5$ for all conv layers except conv1-conv5

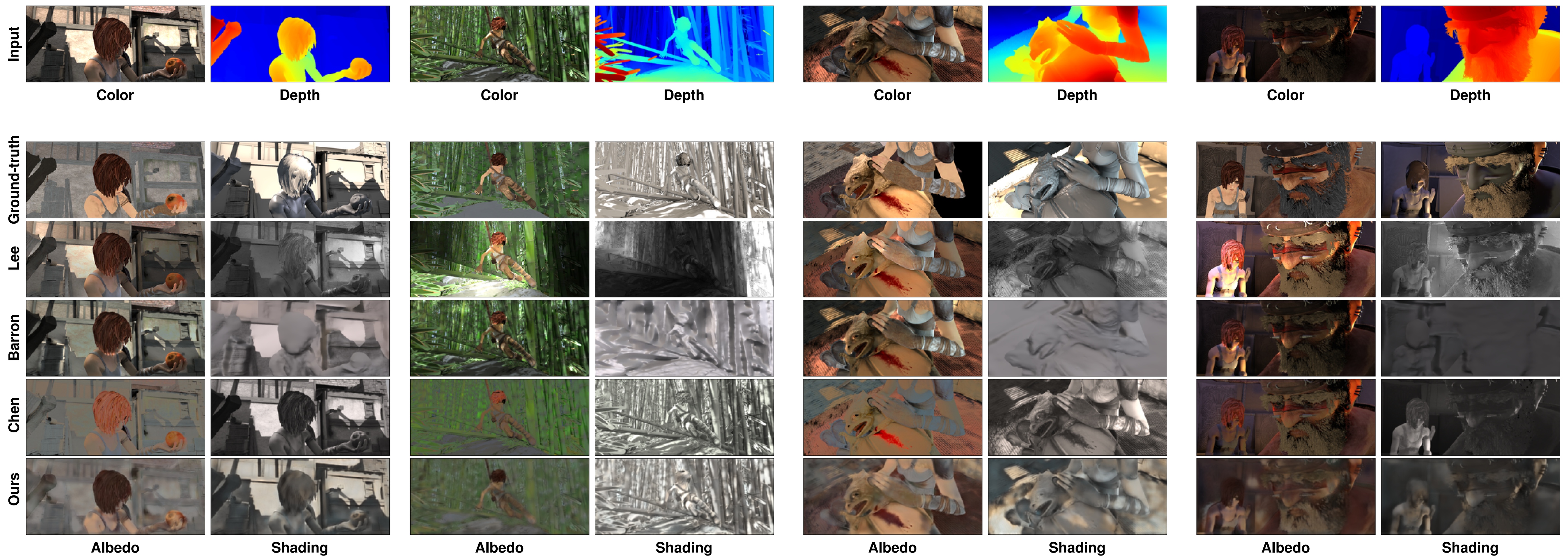
Generated MIT shading (GenMIT)

- Each object has only one ground-truth shading example (the original light source image) in MIT dataset
- Generate more shading examples from ground-truth albedo and 10 additional diffuse images by $S = \alpha I / A$

Resynthesize Sintel for adaptation to MIT training (ResynthSintel)

- Rendered Sintel ground-truth does not satisfy $I = \alpha A \cdot S$
- Generate resynthesized Sintel images by $I' = A \cdot S$

Our Model on RGB Outperforms the State-of-the-Art on RGB+Depth



Best Performance on Sintel

Sintel Training & Testing: Image Split	MSE		LMSE		DSSIM	
	Albedo	Shading	Albedo	Shading	Albedo	Shading
Baseline: Shading Constant	0.0531	0.0488	0.0326	0.0284	0.2140	0.2060
Baseline: Albedo Constant	0.0369	0.0378	0.0240	0.0303	0.2280	0.1870
Retinex [5]	0.0606	0.0727	0.0366	0.0419	0.2270	0.2400
Lee *et al.* [6]	0.0463	0.0507	0.0224	0.0192	0.1990	0.1770
Barron *et al.* [1]	0.0420	0.0436	0.0298	0.0264	0.2100	0.2060
Chen and Koltun [3]	0.0307	0.0277	0.0185	0.0190	**0.1960**	0.1650
MSCR+dropout+GL	**0.0100**	**0.0092**	**0.0083**	**0.0085**	0.2014	**0.1505**

Sintel Training & Testing: Scene Split	MSE		LMSE		DSSIM	
	Albedo	Shading	Albedo	Shading	Albedo	Shading
MSCR (scale 2 only)	0.0255	0.0269	0.0171	0.0186	0.2293	0.1882
MSCR	0.0238	0.0250	0.0155	0.0172	0.2226	0.1816
MSCR+dropout	0.0228	0.0240	0.0147	0.0168	0.2192	0.1746
MSCR+dropout+HC	0.0231	0.0247	0.0147	0.0167	0.2187	0.1750
MSCR+dropout+GL	0.0219	0.0242	0.0143	0.0166	0.2163	0.1737
MSCR+dropout+deconv+DA	0.0209	**0.0221**	0.0135	**0.0144**	0.2081	0.1608
*MSCR+dropout+deconv+DA+GenMIT	**0.0201**	0.0224	**0.0131**	0.0148	**0.2073**	**0.1594**

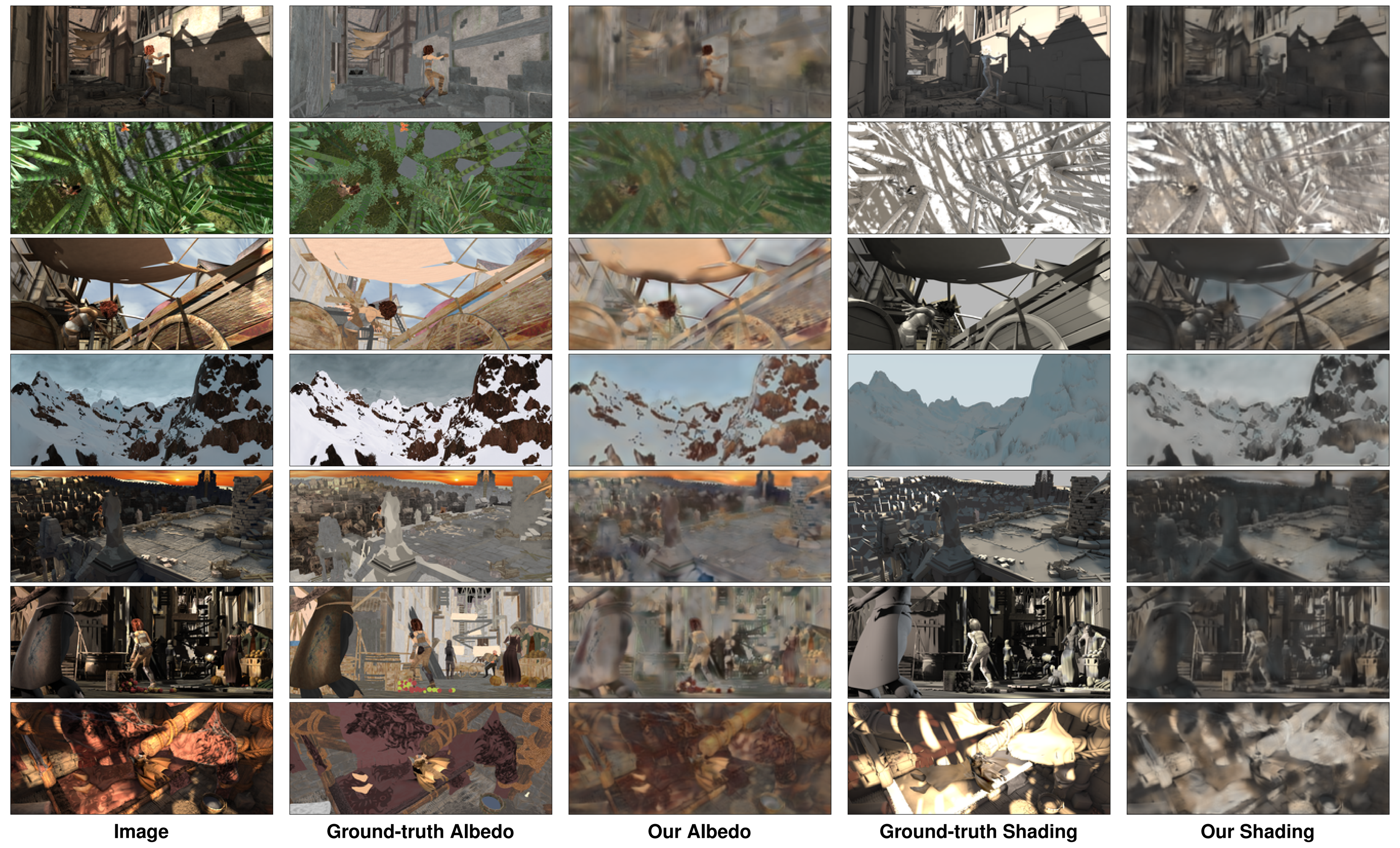
Competitive Performance on MIT

MIT Training & Testing: Our Split	MSE			LMSE		
	Albedo	Shading	Avg	Albedo	Shading	Total [5]
*Ours: MSCR+dropout+deconv+DA+GenMIT	0.0105	**0.0083**	0.0094	0.0296	**0.0163**	0.0234
*Ours without deconv	0.0123	0.0135	0.0129	0.0304	0.0164	0.0249
Ours without DA	0.0107	0.0086	0.0097	0.0300	0.0167	0.0239
Ours without GenMIT	0.0106	0.0097	0.0102	0.0302	0.0184	0.0252
Ours + Sintel	0.0110	0.0103	0.0107	0.0293	0.0182	0.0243
*Ours + ResynthSintel	**0.0096**	0.0085	**0.0091**	**0.0267**	0.0172	**0.0224**

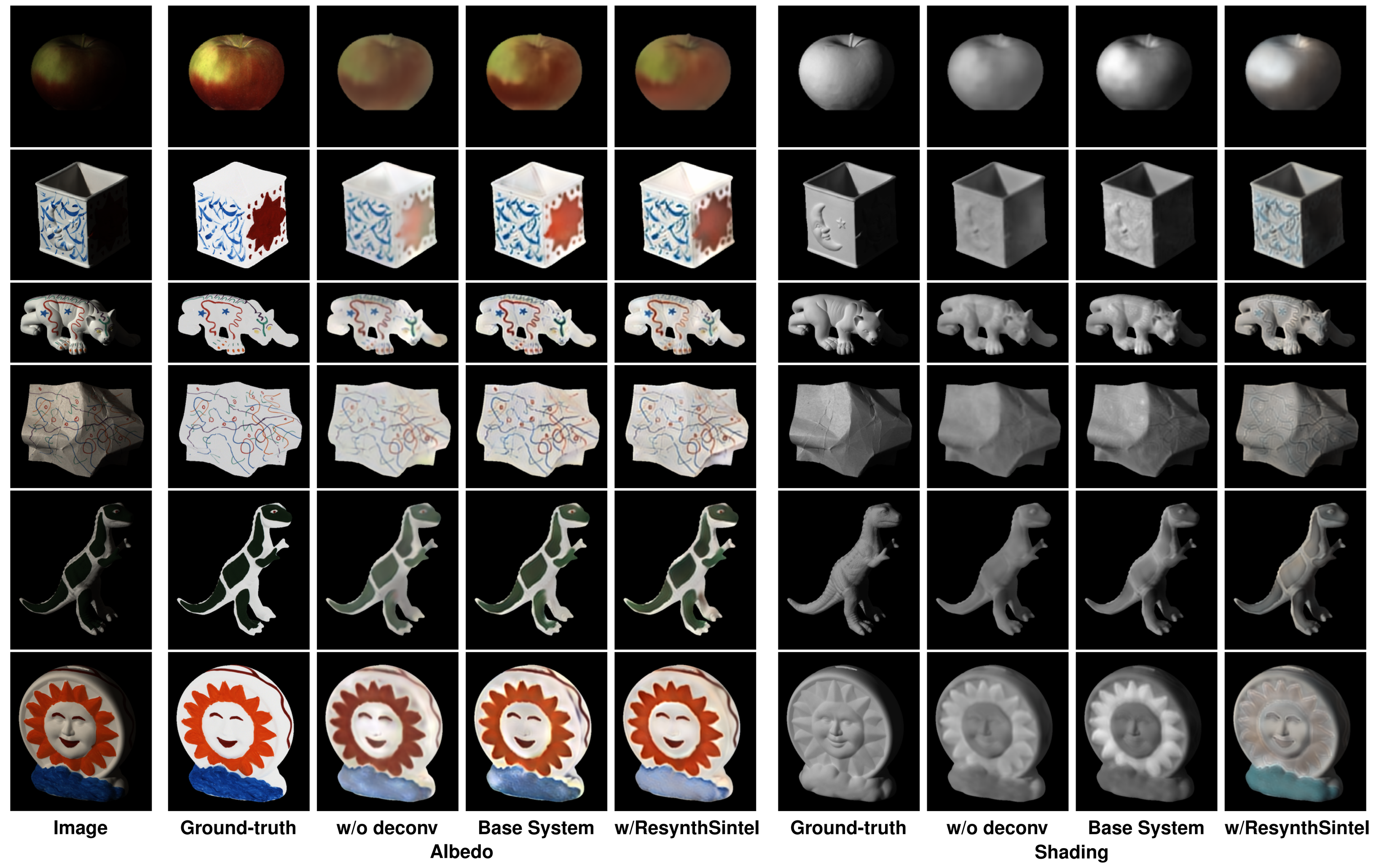
MIT Training & Testing: Barron *et al.*'s Split	MSE			LMSE		
	Albedo	Shading	Avg	Albedo	Shading	Total [5]
Naive Baseline (from [1], uniform shading)	0.0577	0.0455	0.0516	–	–	0.0354
Barron *et al.* [1]	**0.0064**	0.0098	**0.0081**	**0.0162**	**0.0075**	**0.0125**
Ours + ResynthSintel	0.0096	**0.0080**	0.0088	0.0275	0.0152	0.0218

Key: GL = gradient loss HC = hypercolumns DA = data augmentation (scaling, rotation)
GenMIT = add MIT w/generated shading to training
Sintel = add Sintel data to training
ResynthSintel = add resynthesized Sintel data to training

Consistent Result Quality across Sintel Images



Deconvolution and Resynthesis Improve Results on MIT



[1] J. T. Barron and J. Malik. Shape, Illumination, and Reflectance from Shading. *PAMI*, 2015.

[2] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A Naturalistic Open Source Movie for Optical Flow Evaluation. *ECCV*, 2012.

[3] Q. Chen and V. Koltun. A Simple Model for Intrinsic Image Decomposition with Depth Cues. *ICCV*, 2013.

[4] D. Eigen and R. Fergus. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture. *CVPR*, 2015.

[5] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman. Ground Truth Dataset and Baseline Evaluations for Intrinsic Image Algorithms. *ICCV*, 2009.

[6] K. J. Lee, Q. Zhao, X. Tong, M. Gong, S. Izadi, S. U. Lee, P. Tan, and S. Lin. Estimation of Intrinsic Image Sequences from Image+Depth Video. *ECCV*, 2012.

[7] T. Narihira, M. Maire, and S. X. Yu. Learning Lightness from Human Judgement on Relative Reflectance. *CVPR*, 2015.