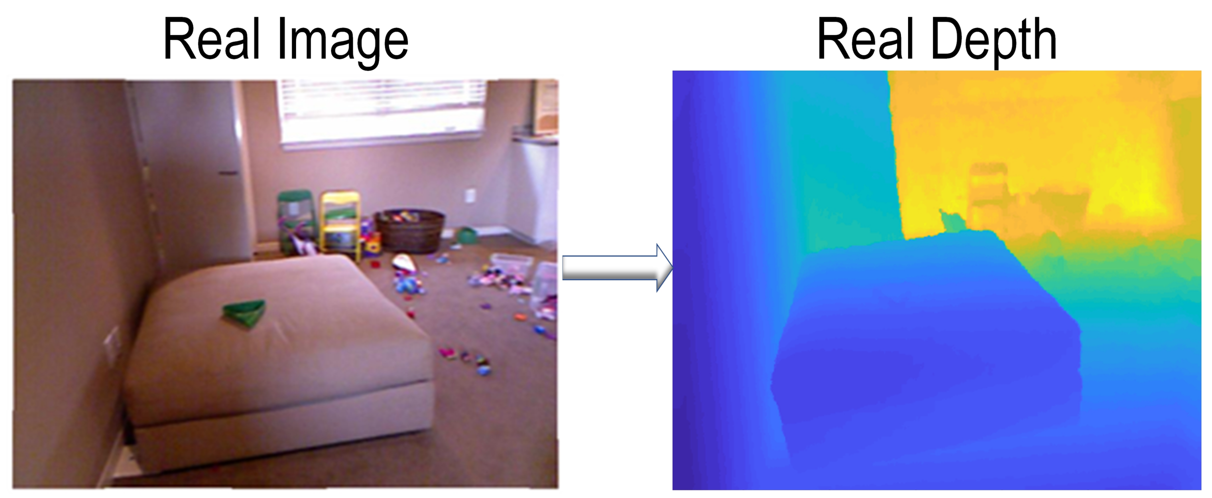
T[2]Net: Synthetic-to-Realistic Translation for Solving Single-Image Depth Estimation Tasks

Chuanxia Zheng, Tat-Jen Cham, Jianfei Cai
School of Computer Science and Engineering
Nanyang Technological University Singapore

Motivation

Goal: Single-Image Depth Estimation

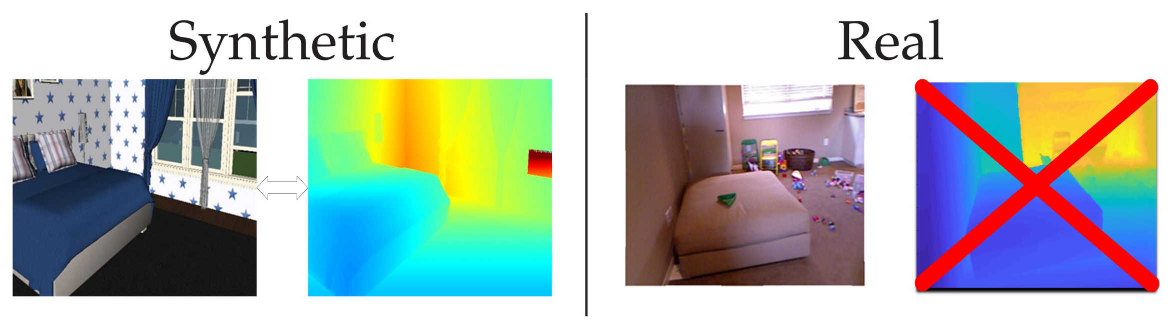


Problem:

1. Real image-depth paired datasets not widely available
2. Real depth sensory data are sparse/noisy

Approach: Train only on synthetic paired data and unpaired real images

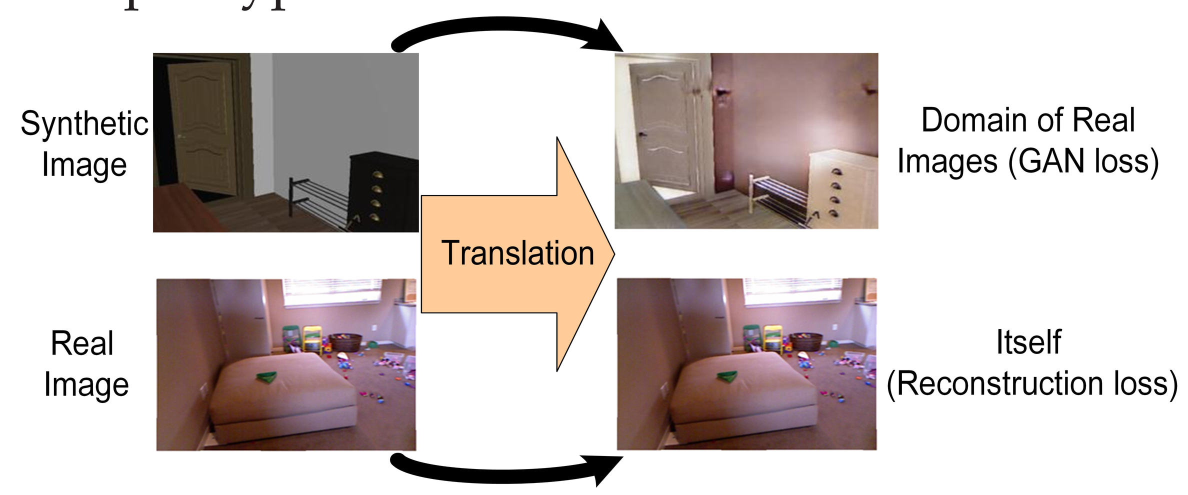
Training:



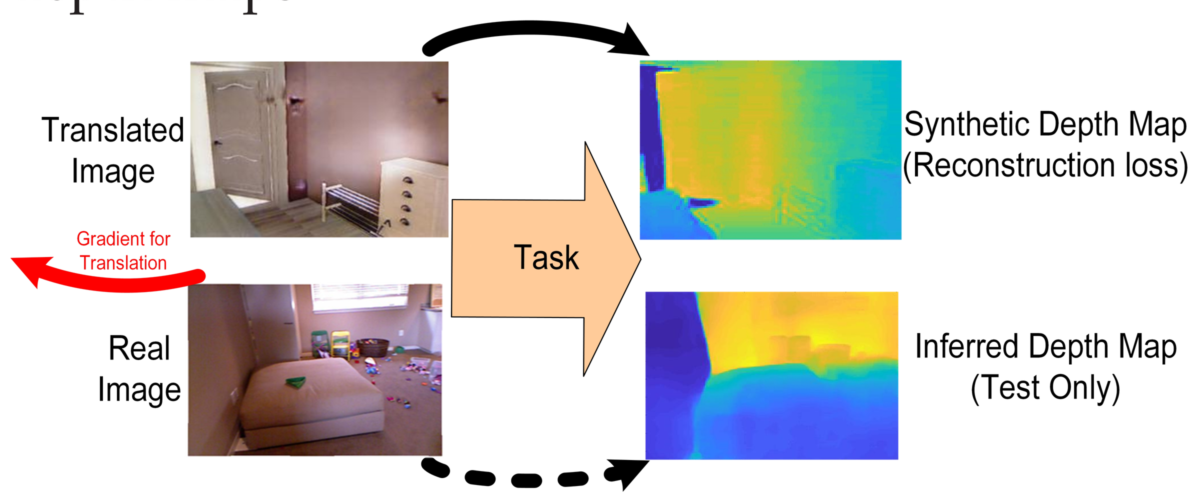
Challenge: Large gap between synthetic images and real images

Key Insights

1. Propose **Wide-Spectrum GAN** for training domain translation
   —**Switch** between loss functions, depending on input type

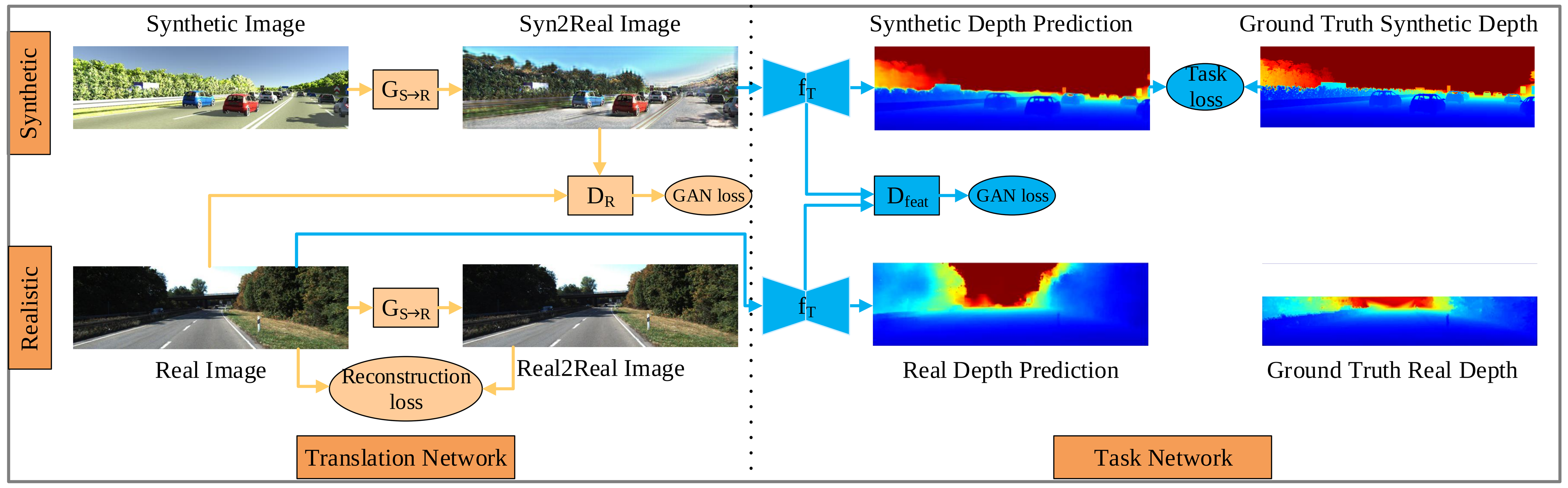


2. Leverage *easily-generated* and *precise* synthetic depth maps
   —Minimize need to depend on real sensor depth maps



Notes:
— No paired real data needed
— Framework can be trained end-to-end

The Proposed T[2]Net Framework



Translation Network

Adversarial loss (for *synthetic* images):

$$\mathcal{L}_{\mathrm{GAN}}(G_{S\to R}, D_R) = \mathbb{E}_{x_r\sim X_R}[\log D_R(x_r)] + \mathbb{E}_{x_s\sim X_S}[\log(1 - D_R(G_{S\to R}(x_s)))]$$

Target reconstruction loss (for *real* images):

$$\mathcal{L}_r(G_{S\to R}) = ||G_{S\to R}(x_r) - x_r||_1$$

Task Network

Task loss (for *synthetic* depth):

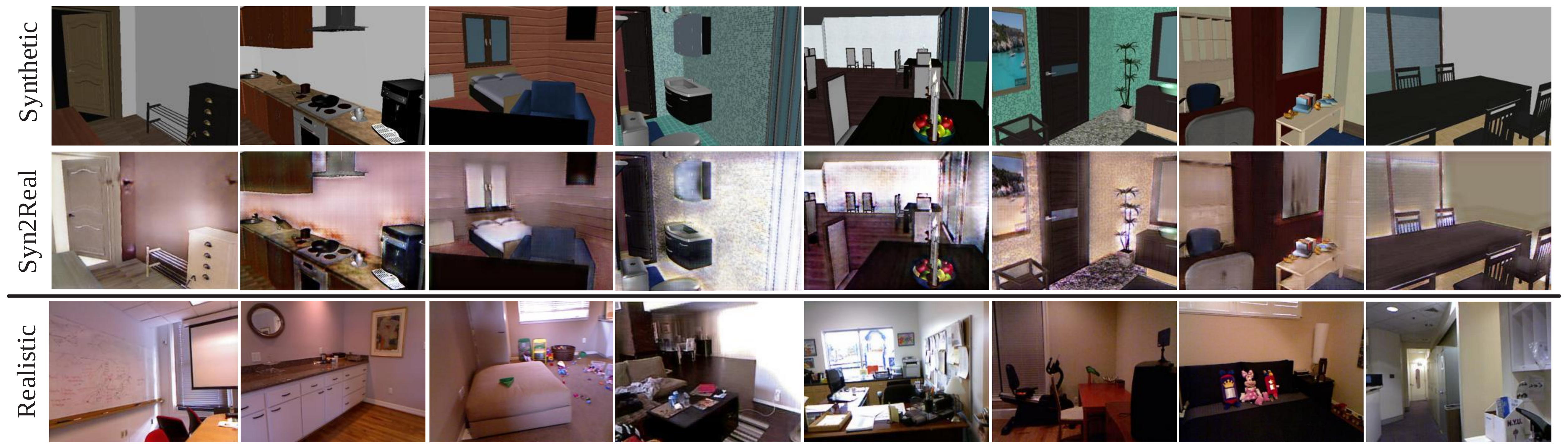
$$\mathcal{L}_t(f_T) = ||f_T(\hat{x}_s) - y_s||_1$$

Smoothness loss (for *real* depth):

$$\mathcal{L}_s(f_T) = |\partial_x f_T(x_r)|e^{-|\partial_x x_r|} + |\partial_y f_T(x_r)|e^{-|\partial_y x_r|}$$

Image Translation

Unpaired indoor image translation results:



Quantitative Results

Quantitative results on KITTI:

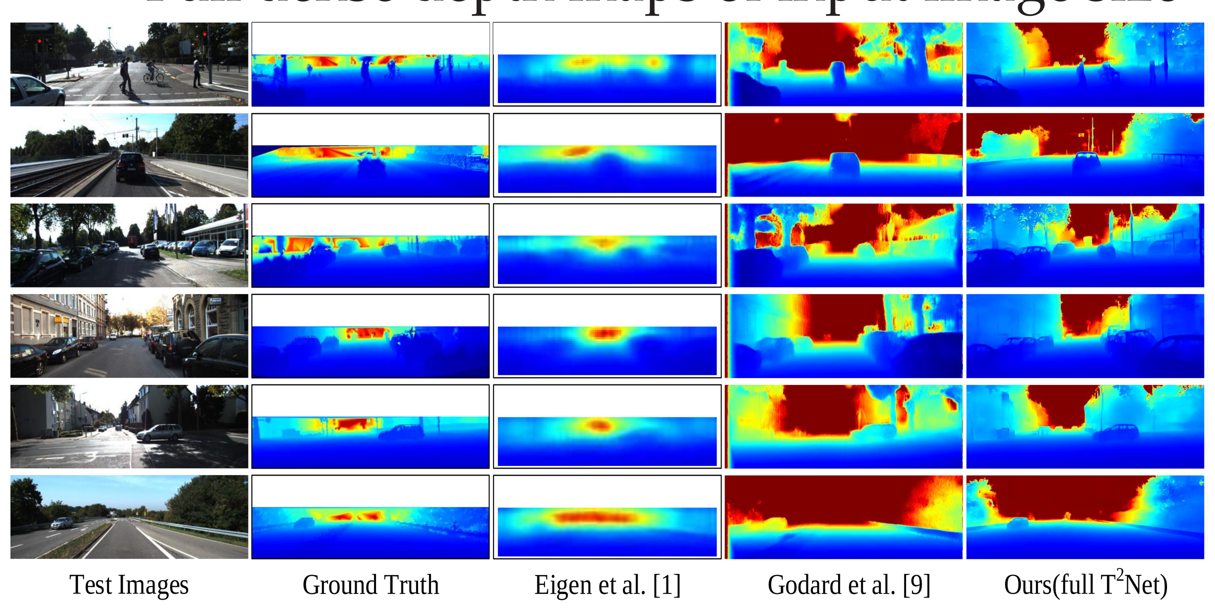
Method	Dataset	cap	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
			lower is better				higher is better		
Eigen et al. [4] Fine	K(I+D)	0-80m	0.190	1.515	7.156	0.270	0.692	0.899	0.967
Garg et al. [7] L12 Aug.8x	K(L+R)	1-50m	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Godard et al. [10]	CS+K(L+R)	1-50m	0.117	0.762	3.972	0.206	0.860	0.948	0.976
Kuznietsov et al. [20]	K(D+L+R)	1-50m	0.108*	0.595*	3.518*	0.179	0.875*	0.964*	0.988*
Baseline, train set mean	vK(I+D)	1-50m	0.521	11.024	10.598	0.473	0.638	0.755	0.835
Our f_T, all-real	K(I+D)	1-50m	0.114	0.627	3.549	0.178*	0.867	0.960	0.986
Our f_T, all-synthetic	vK(I+D)	1-50m	0.278	3.216	6.268	0.322	0.681	0.854	0.929
Our T[2]Net, D_{feat} only	vK(I+D) + K(I)	1-50m	0.233	2.902	6.285	0.300	0.743	0.880	0.938
Our T[2]Net, D_{image} only	vK(I+D) + K(I)	1-50m	**0.168**	**1.199**	**4.674**	**0.243**	**0.772**	**0.912**	**0.966**
Our full T[2]Net	vK(I+D) + K(I)	1-50m	0.169	1.230	4.717	0.245	0.769	**0.912**	0.965

Quantitative results of ablation study:

Method	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
	lower is better				higher is better		
baseline, synthetic only	0.278	3.216	6.268	0.322	0.681	0.854	0.929
vanilla task network, synthetic only	0.295	3.793	8.403	0.363	0.600	0.817	0.912
vanilla task network, full approach	0.259	2.891	6.380	0.324	0.694	0.853	0.927
separated training	0.234	2.706	6.068	0.293	0.747	0.882	0.942
separated training with CycleGAN	0.212	1.973	5.340	0.269	0.750	0.895	0.952
self-domain reconstruction	0.199	1.517	5.349	0.298	0.695	0.866	0.9420
No reconstruction loss(epoch 3)	0.201	1.941	5.619	0.286	0.741	0.882	0.945
No feature loss	**0.168**	**1.199**	**4.674**	**0.243**	**0.772**	**0.912**	**0.966**
No image GAN loss	0.233	2.902	6.285	0.300	0.743	0.880	0.938
our full approach	0.169	1.230	4.717	0.245	0.769	0.912	0.965

Depth Estimation

Real world depth estimation results:
—Full dense depth maps of input image size



Source Code

The source code and video are available at
`https://github.com/lyndonzheng/Synthetic2Realistic`

Analysis