Cross-Domain Self-supervised Multi-task Feature Learning using Synthetic Imagery

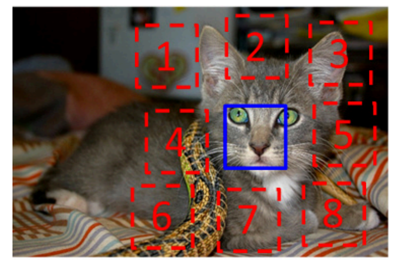
Zhongzheng Ren and Yong Jae Lee

Department of Computer Science, UC Davis

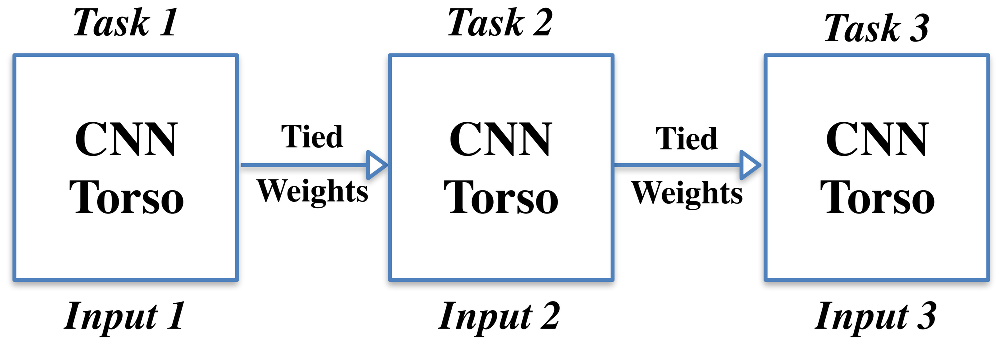
Models, code, and more available at: github.com/jason718/game-feature-learning

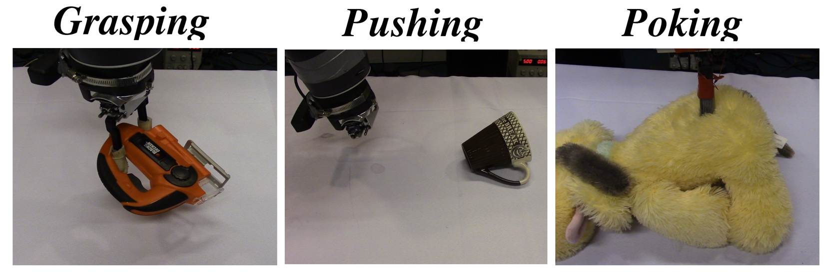
Problem: Learning visual representations without human labels.

Observation I: Existing works mainly leverage a **SINGLE** task.

[1]  [2]  • • •

Observation II: Existing multi-task works are **COMPLEX/LIMITED**.

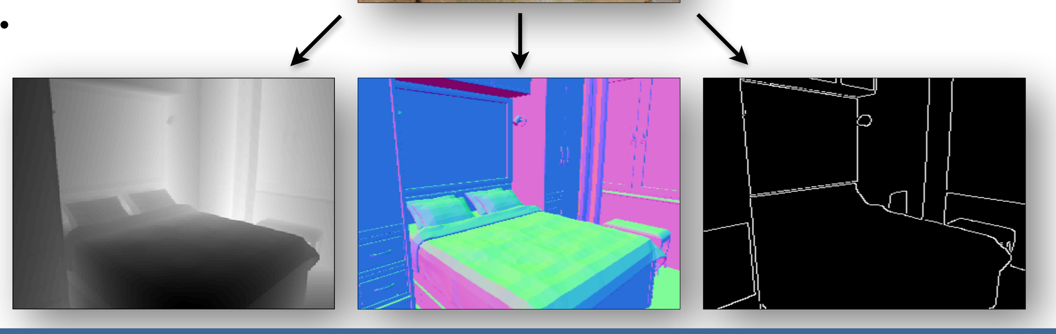
[3] 

[4] 

Learn from Robot Physical Interaction

Our Idea:

- **MULTI-TASK** learning with three outputs for the *same* image input.
- **CG** to render multi-supervision.
- **Domain Adaptation** to better transfer to real-world.



Adversarial Training Process

Input: *Synthetic images X, real images Y*

Output: *Domain adapted base network B*

0 *while in training loop:*

1 *Sample real images $\{x \in X\}$ and synthetic $\{y \in Y\}$*

2 *Extract features $z_x = B(x)$, $z_y = B(y)$*

3 *Keep D frozen, update B via three tasks w/ input y*

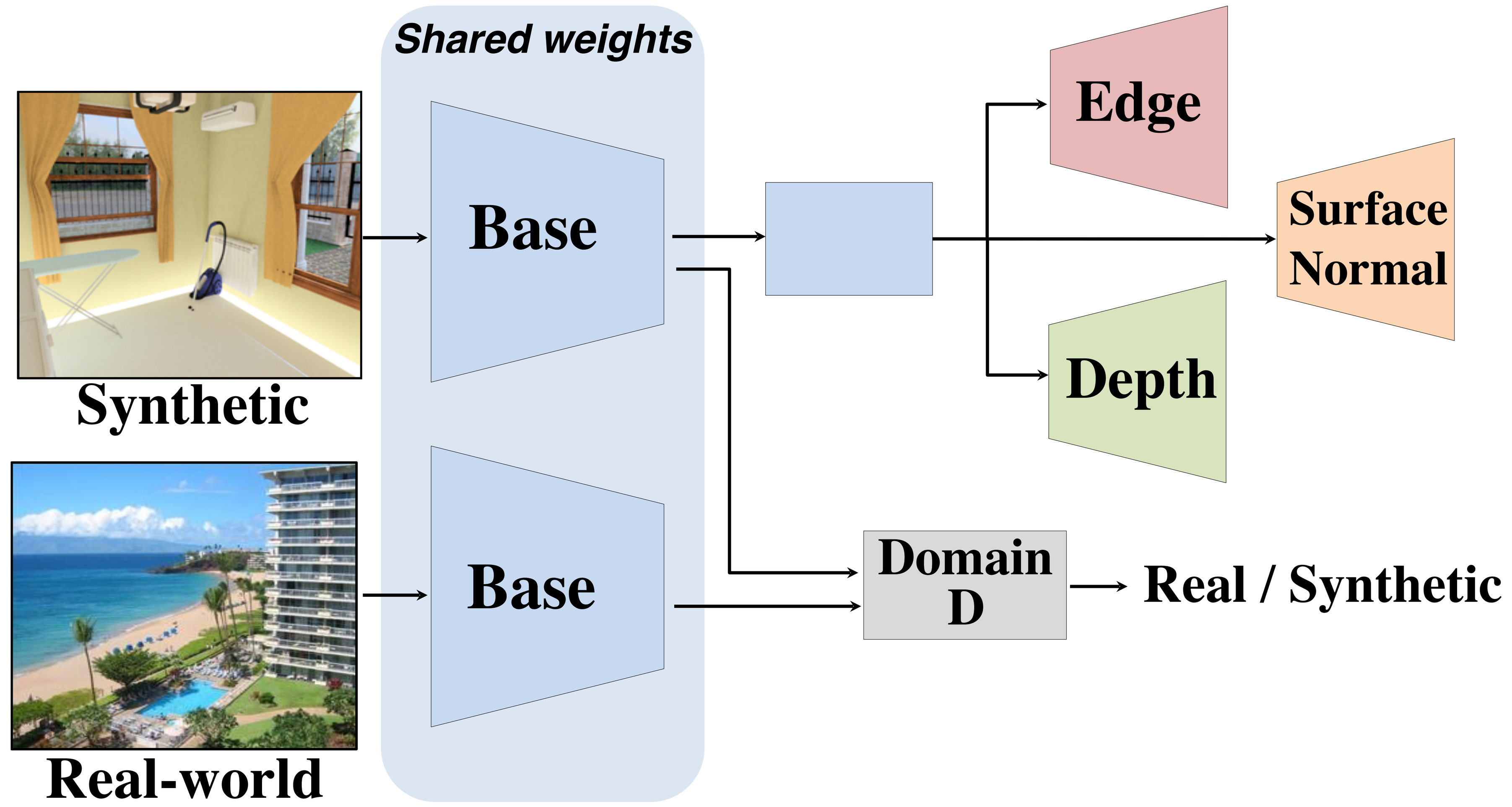
$$L(\phi_B, \phi_{tasks} \mid z_x) = -\sum \log(1 - D(z_x)) + L_{edge} + L_{depth} + L_{normal}$$

4 *Keep B frozen, update D via adversarial loss w/ input (x, y)*

$$L(\phi_D \mid z_x, z_y) = -\sum \log(D(z_x)) - \sum \log(1 - D(z_y))$$

1. C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. ICCV 2015.
2. R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. ECCV 2016.
3. C. Doersch and A. Zisserman. Multi-task Self-Supervised Visual Learning. ICCV 2017.
4. L. Pinto, et al. The Curious Robot: Learning Visual Representations via Physical Interactions. ECCV 2016.

Architecture



Dataset:

- Real world: **Places-365** (Zhou et al. PAMI'17)
- Synthetic: **SUNCG** (Song et al. CVPR'17) / **SceneNet** (McCormac et al. ICCV'17)

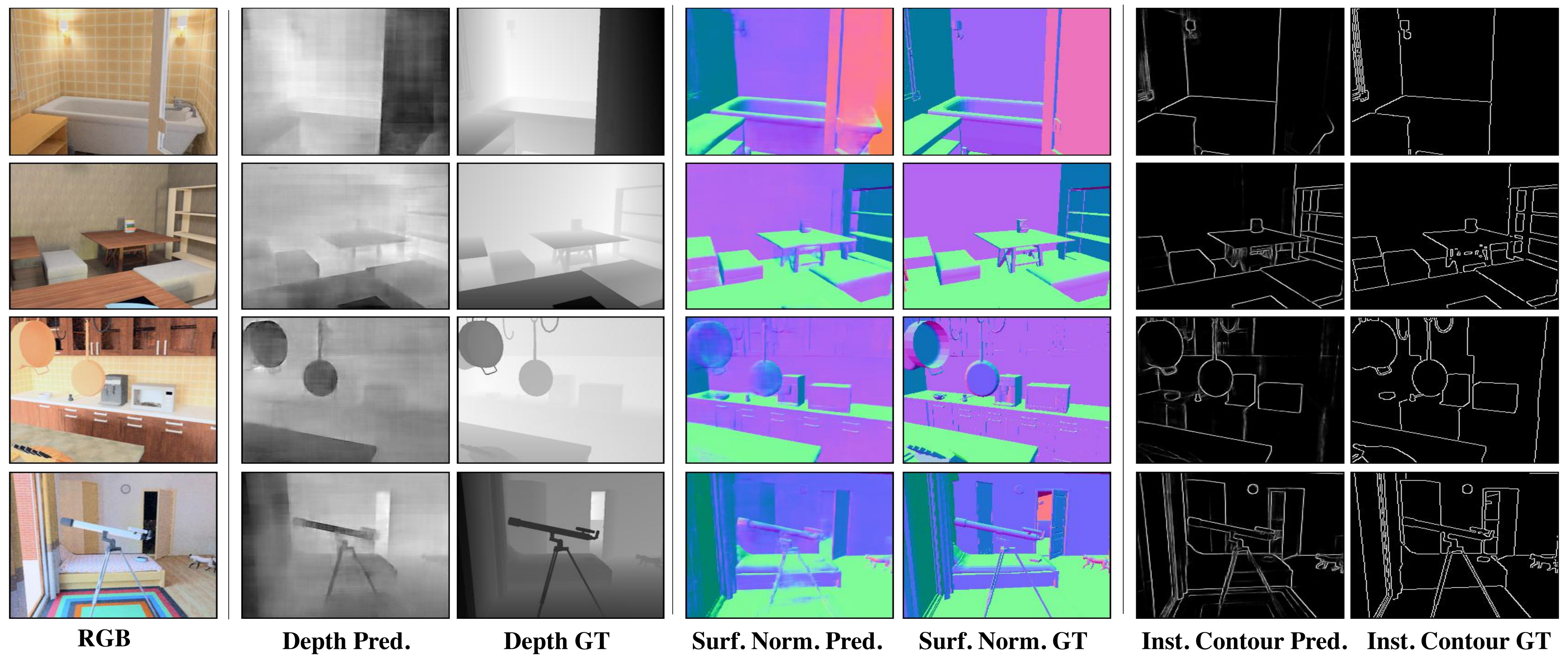
Transfer Learning Results

Method	*conv1*	*conv2*	*conv3*	*conv4*	*conv5*	*07 Cls.*	*07 Det.*	*12 Det.*
ImageNet	19.3	36.3	44.2	48.3	50.5	79.9	56.8	56.5
Gaussian	11.6	17.1	16.9	16.3	14.1	53.4	41.3	-
Krahenbuhl et al.	17.5	23.0	24.5	23.2	20.6	56.6	45.6	42.8
Context	16.2	23.3	30.2	31.7	29.6	65.3	51.1	49.9
BiGAN	17.7	24.5	31.0	29.9	28.0	58.6	46.2	44.9
Context-Encoder	14.1	20.7	21.0	19.8	15.5	56.5	44.5	-
Colorization	12.5	24.5	30.4	31.5	30.3	65.9	46.9	44.5
Jigsaw	**18.2**	28.8	34.0	33.9	27.1	67.6	**53.2**	-
Split-Brain	17.7	29.3	**35.4**	**35.2**	**32.8**	67.1	46.7	43.8
Counting	18.0	**30.6**	34.3	32.5	25.7	67.7	51.4	-
Ours	16.5	27.0	30.5	30.1	26.5	**68.0**	52.6	**50.0**

ImageNet Classification w/o finetuning
Comparable to methods learned on ImageNet

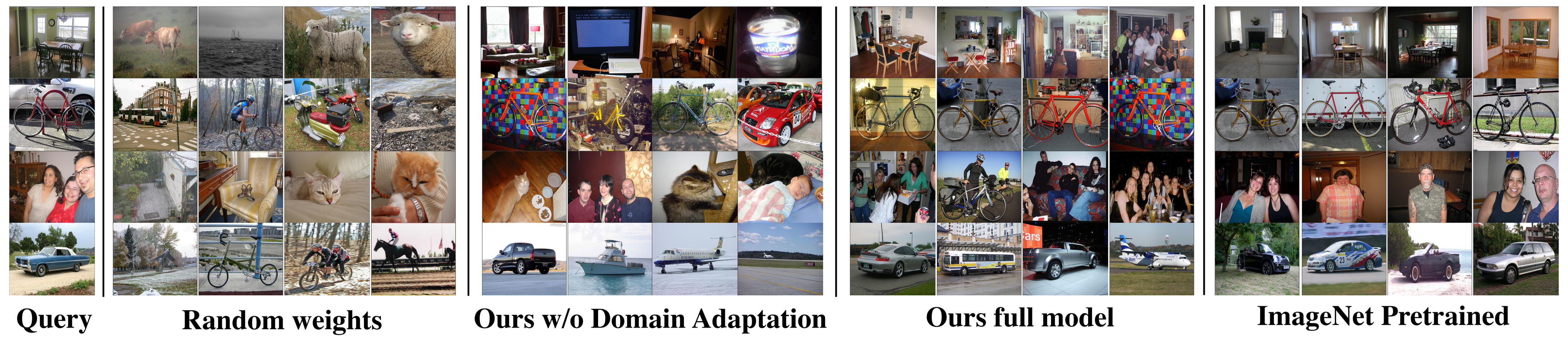
VOC w/ finetuning
SOTA results

Qualitative Results



The better our model performs on these tasks, the better transferable features it's likely to get.

Nearest Neighbor Results



An initial evidence that our model can capture high-level semantics on real-world data.

Ablation Study

Task	Adaptation	#Train	07-Cls.	07-Det.
Edge	-	*0.5M*	63.9	44.8
Dep	-	*0.5M*	61.9	45.8
Surf.	-	*0.5M*	65.3	45.4
3 tasks	-	*0.5M*	**65.6**	**47.2**
3 tasks	*conv1*	*0.5M*	61.9	46.0
3 tasks	*conv2*	*0.5M*	63.4	46.3
3 tasks	***conv5***	*0.5M*	**67.4**	**49.2**
3 tasks	*conv6*	*0.5M*	66.9	48.2
3 tasks	*conv5*	***1.5M***	**68.0**	**50.0**

Multi-task! Multi-data!
Domain Adaptation helps!

NYUD Results

		Lower the better		*Higher the better*		
GT	*Methods*	*Mean*	*Median*	*11.25°*	*22.5°*	*30°*
1	*Zhang et al. CVPR'17*	22.1	14.8	**39.6**	65.6	75.3
1	*Ours*	**21.9**	**14.6**	39.5	**66.7**	**76.5**
2	*Wang et al. ICCV'17*	26.0	18.0	33.9	57.6	67.5
2	*Ours*	**23.8**	**16.2**	**36.6**	**62.0**	**72.9**

The learned features also transfer well on original three tasks.

Acknowledgment

This work was supported in part by the National Science Foundation (NSF) under Grant No. 1748387. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.