Towards General Vision Architectures: Attentive Single-Tasking of Multiple Tasks

Neural Architects - ICCV
28 October 2019

Iasonas Kokkinos


Kevis Maninis

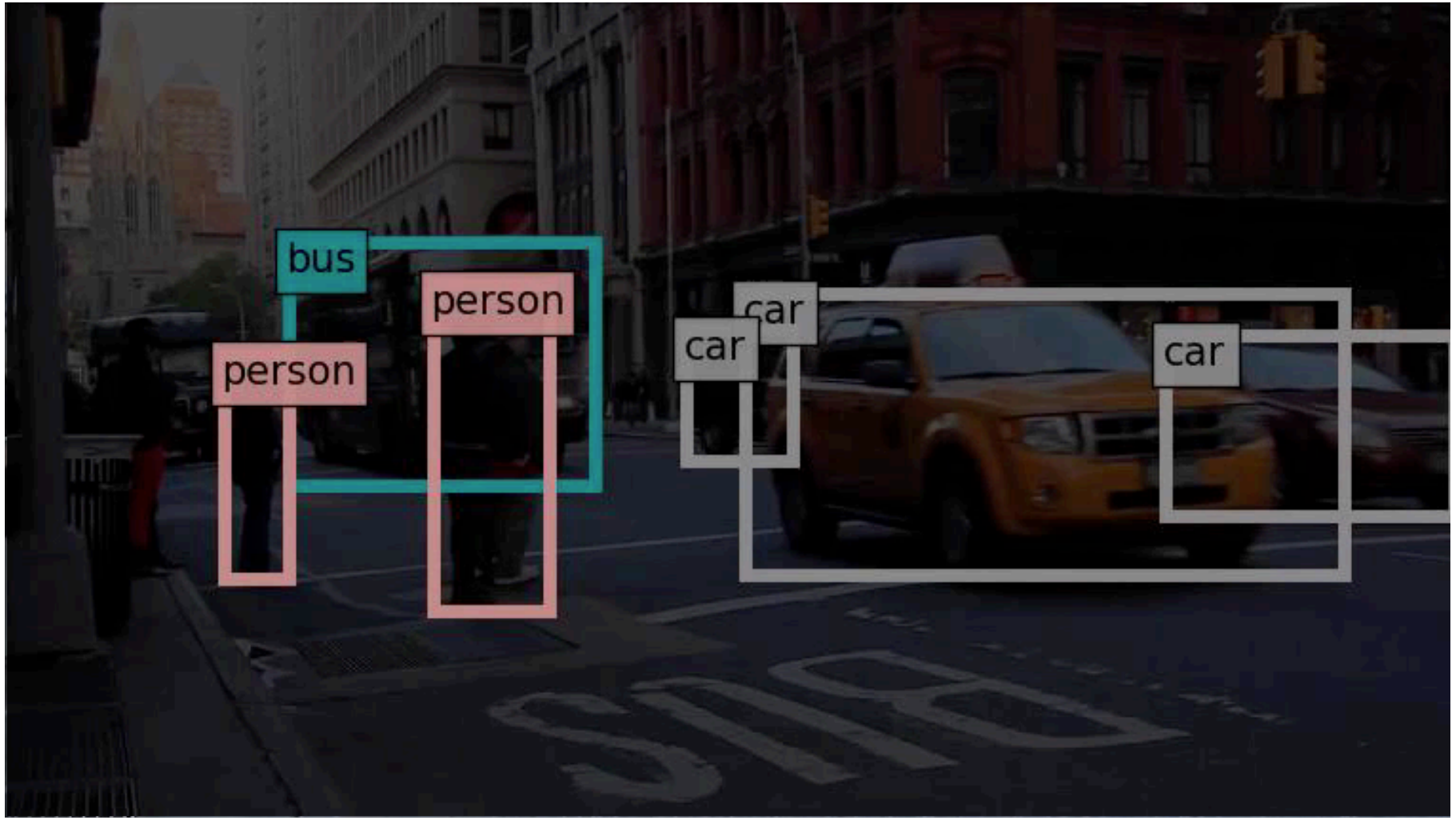

Ilija Radosavovic

Ariel AI

UCL

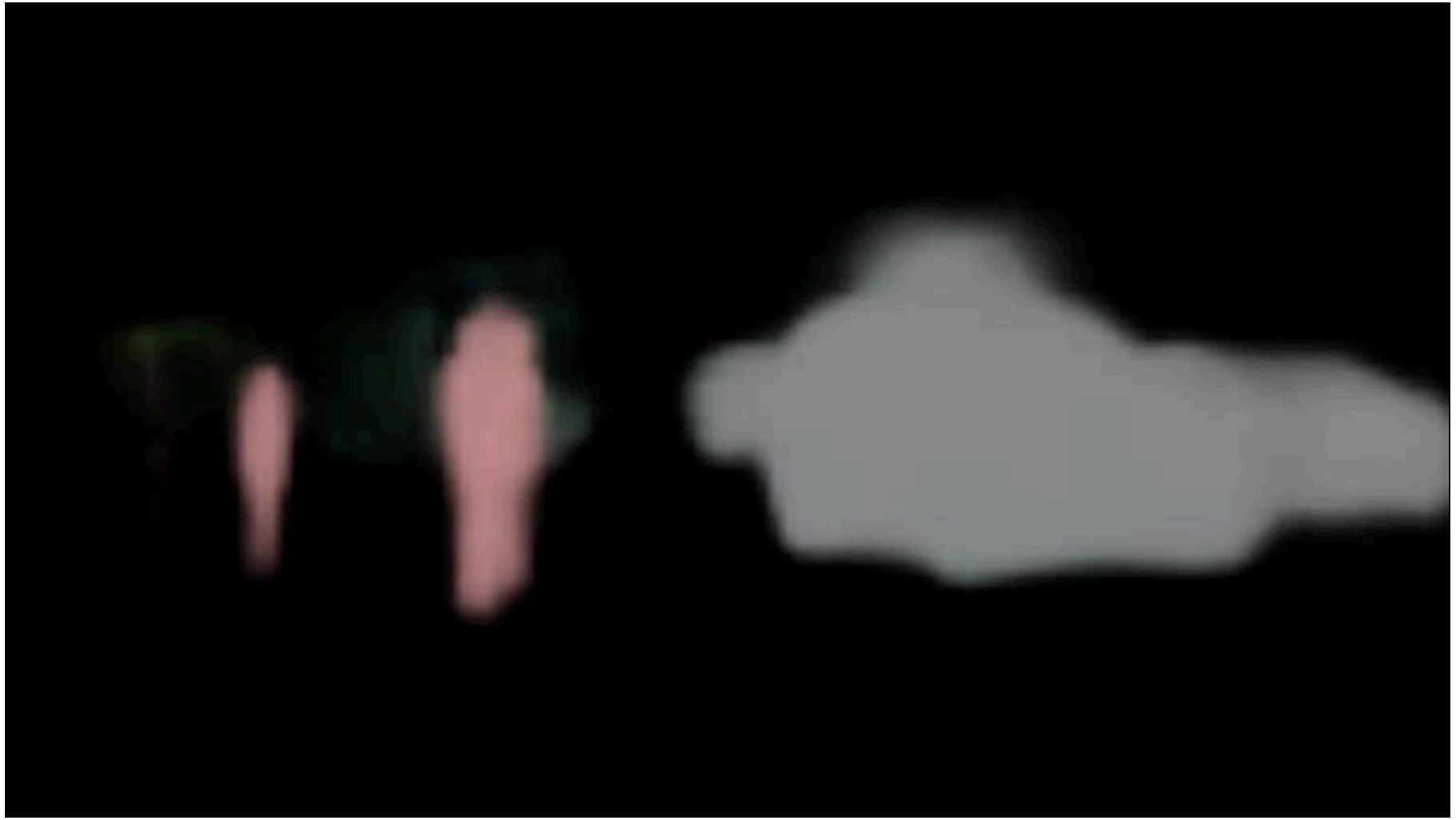
What can we get out of an image?

What can we get out of an image?



Object detection

What can we get out of an image?

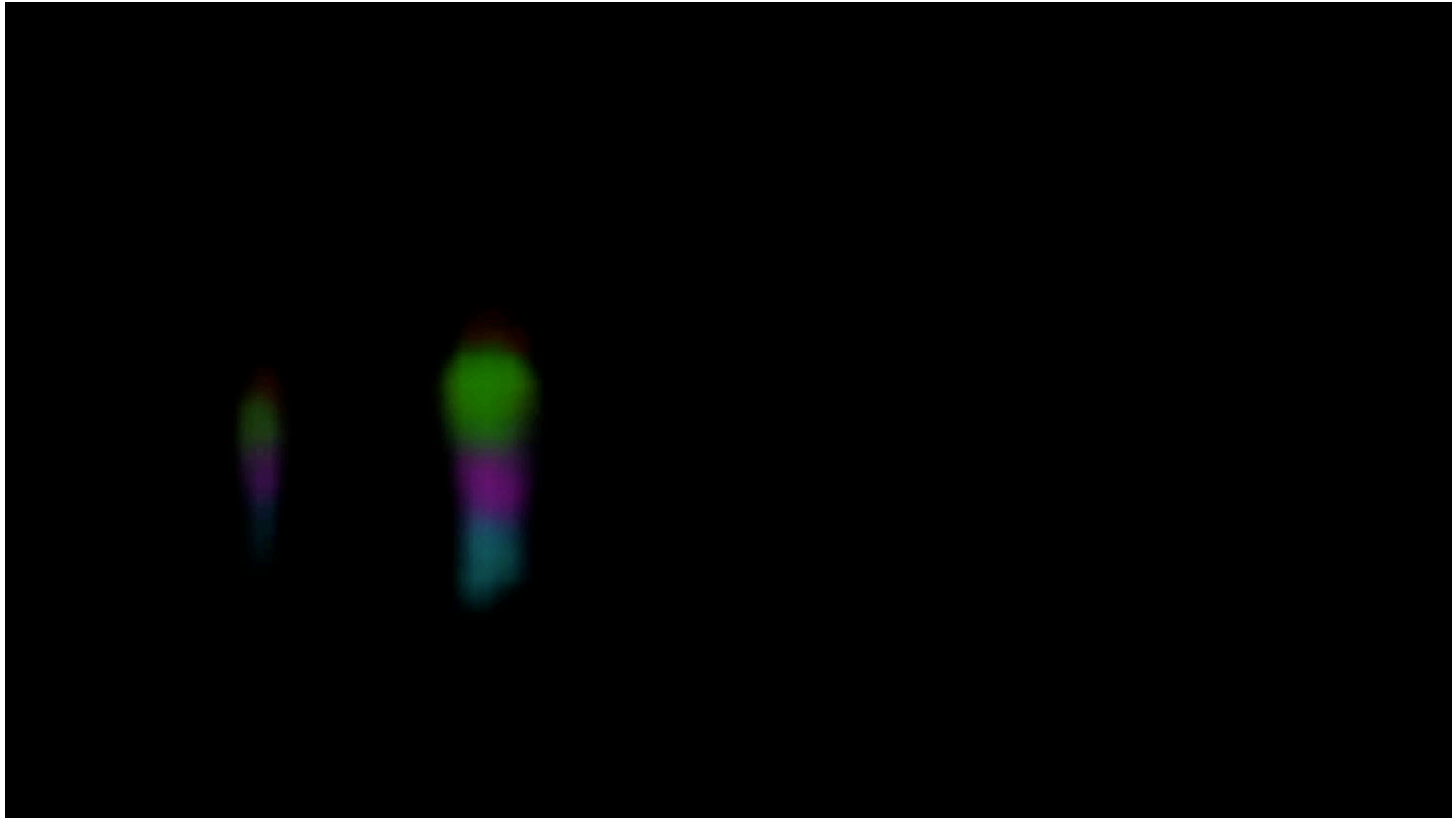


Semantic segmentation

What can we get out of an image?

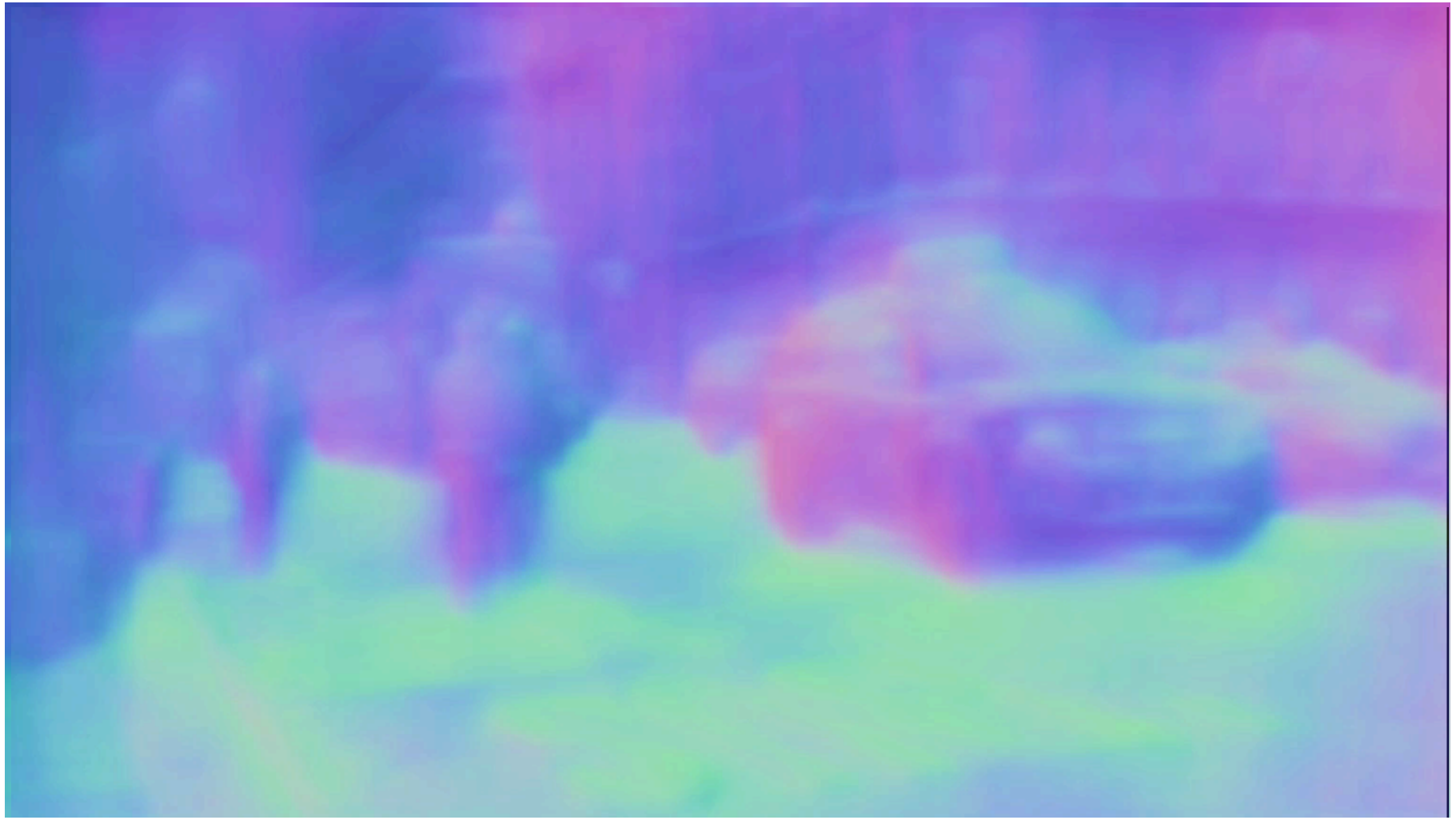
Semantic boundary detection

What can we get out of an image?



Part segmentation

What can we get out of an image?



Surface normal estimation

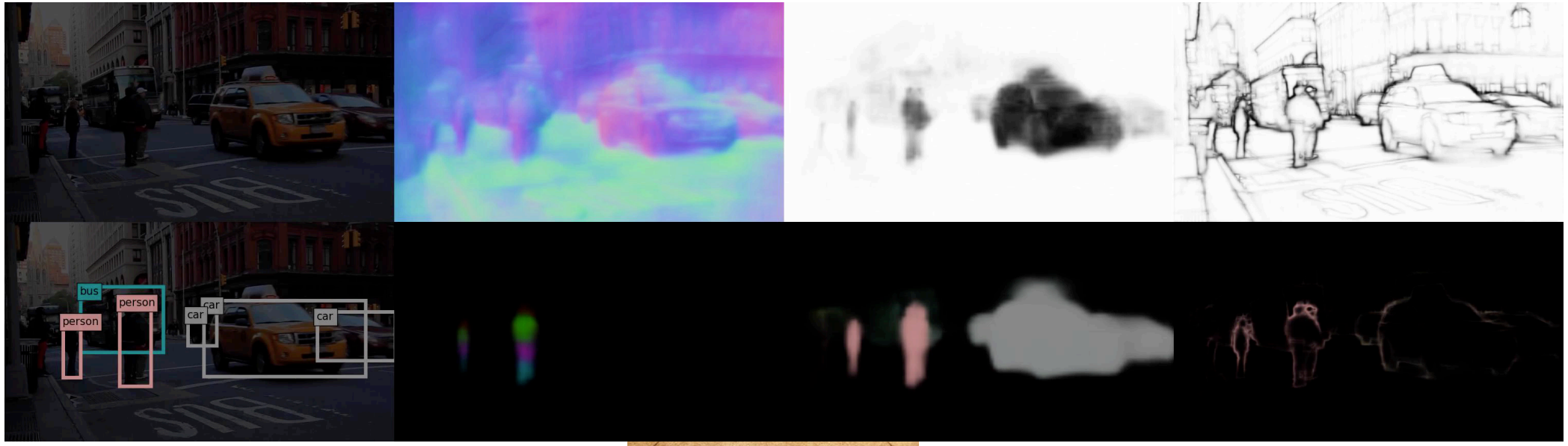
What can we get out of an image?

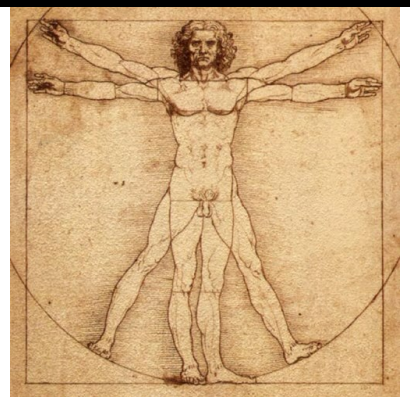
Saliency estimation

What can we get out of an image?

Boundary detection

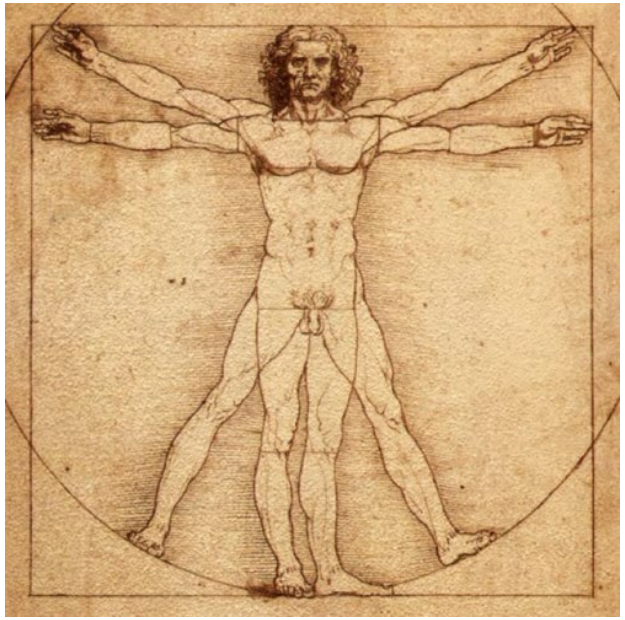
Can we do it all in one network?





I. Kokkinos, UberNet: A Universal Netwok for Low-,Mid-, and High-level Vision, CVPR 2017

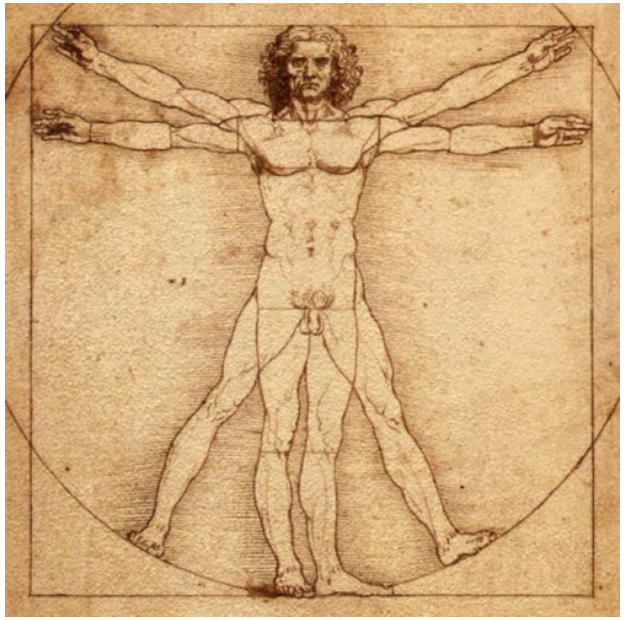
Multi-tasking boosts performance



Detection

Ours, 1-Task	78.7
Ours, Segmentation + Detection	80.1

Multi-tasking boosts performance?



Detection

Ours, 1-Task	78.7
Ours, Segmentation + Detection	80.1
Ours, 7-Task	77.8

Did multi-tasking turn our network to a dilettante?

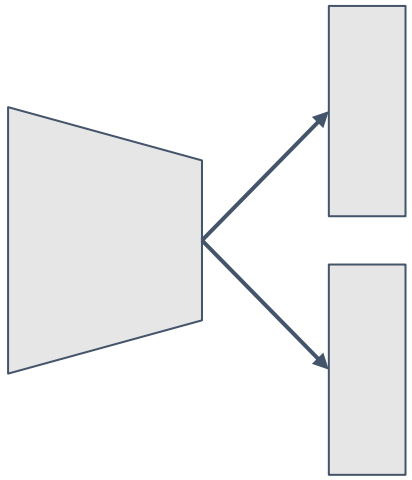
Detection

Ours, 1-Task	78.7
Ours, Segmentation + Detection	80.1
Ours, 7-Task	77.8

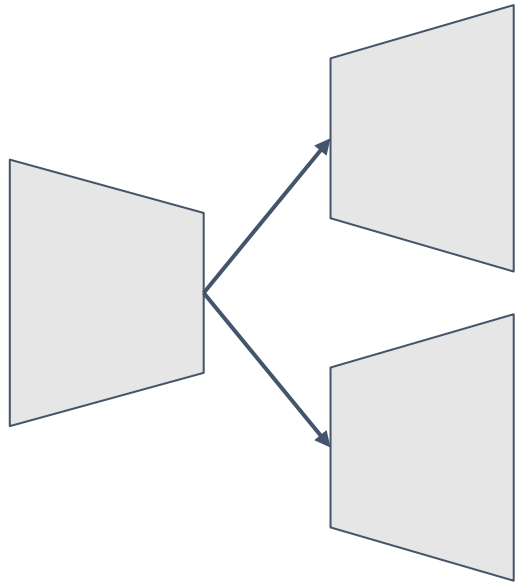
Semantic Segmentation

Ours, 1-Task	72.4
Ours, Segmentation + Detection	72.3
Ours, 7-Task	68.7

Should we just beef-up the task-specific processing?



Ubernet (CVPR 17)

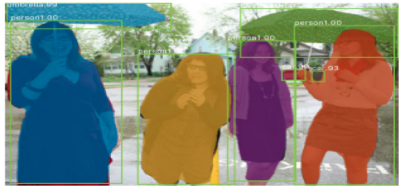


Mask R-CNN (ICCV 17), PAD-Net (CVPR18)

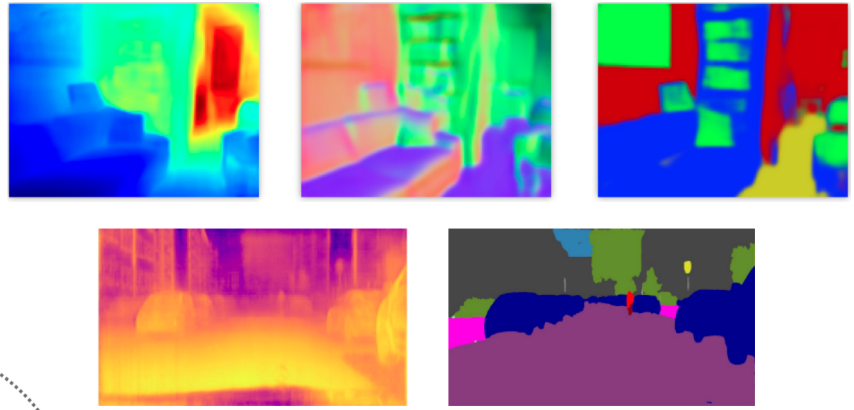
- Memory consumption
- Number of parameters
- Computation
- Effectively no positive transfer across tasks

Multi-tasking can work (sometimes)

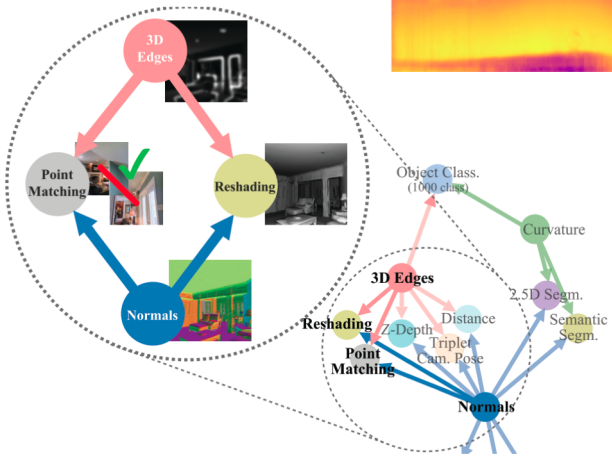
- Mask R-CNN [1]:
  - multi-task: detection + segmentation



- Eigen et al. [2] , PAD-Net [3]
  - multi-task: depth, sem. segmentation



- Taskonomy [4]
  - transfer learning among tasks



[1] He et al., "Mask R-CNN", in ICCV 2017
[2] Eigen and Fergus, "Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture", in ICCV 2015
[3] Xu et al., "PAD-Net: Multi-Tasks Guided Prediction-and-Distillation Network for Simultaneous Depth Estimation and Scene Parsing", in CVPR 2018
[4] Zamir et al., "Taskonomy: Disentangling Task Transfer Learning", in CVPR 2018

Unaligned Tasks

One task's noise is another task's signal

This is not even catastrophic forgetting:
plain task interference

We could even try doing adversarial training on one task to improve performance for the other (force *desired* invariance)

Learning Task Grouping and Overlap in Multi-Task Learning
A. Kumar, H. Daume, ICML 2012
Learning with Whom to Share in Multi-task Feature Learning
Z. Kang, K. Grauman, F. Sha, ICML 2011
Exploiting Unrelated Tasks in Multi-Task Learning,
B. Paredes, A. Argyriou, N. Berthouze, M. Pontil, AISTATS 2012

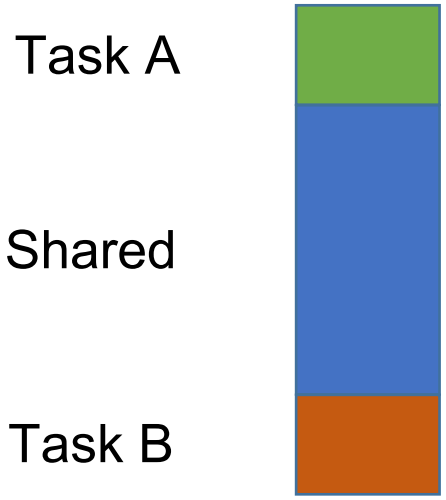
Expression recognition

Identity recognition

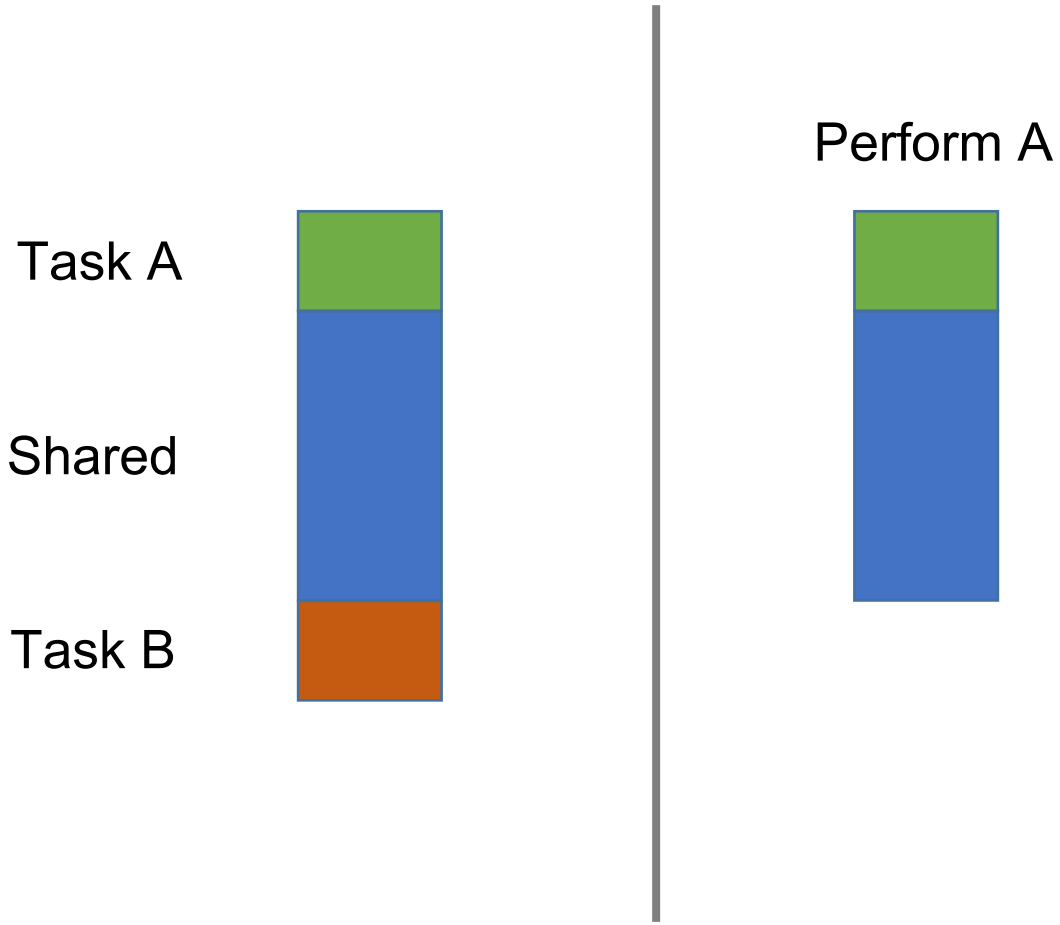
MMI Facial Expression Database

Count the balls!

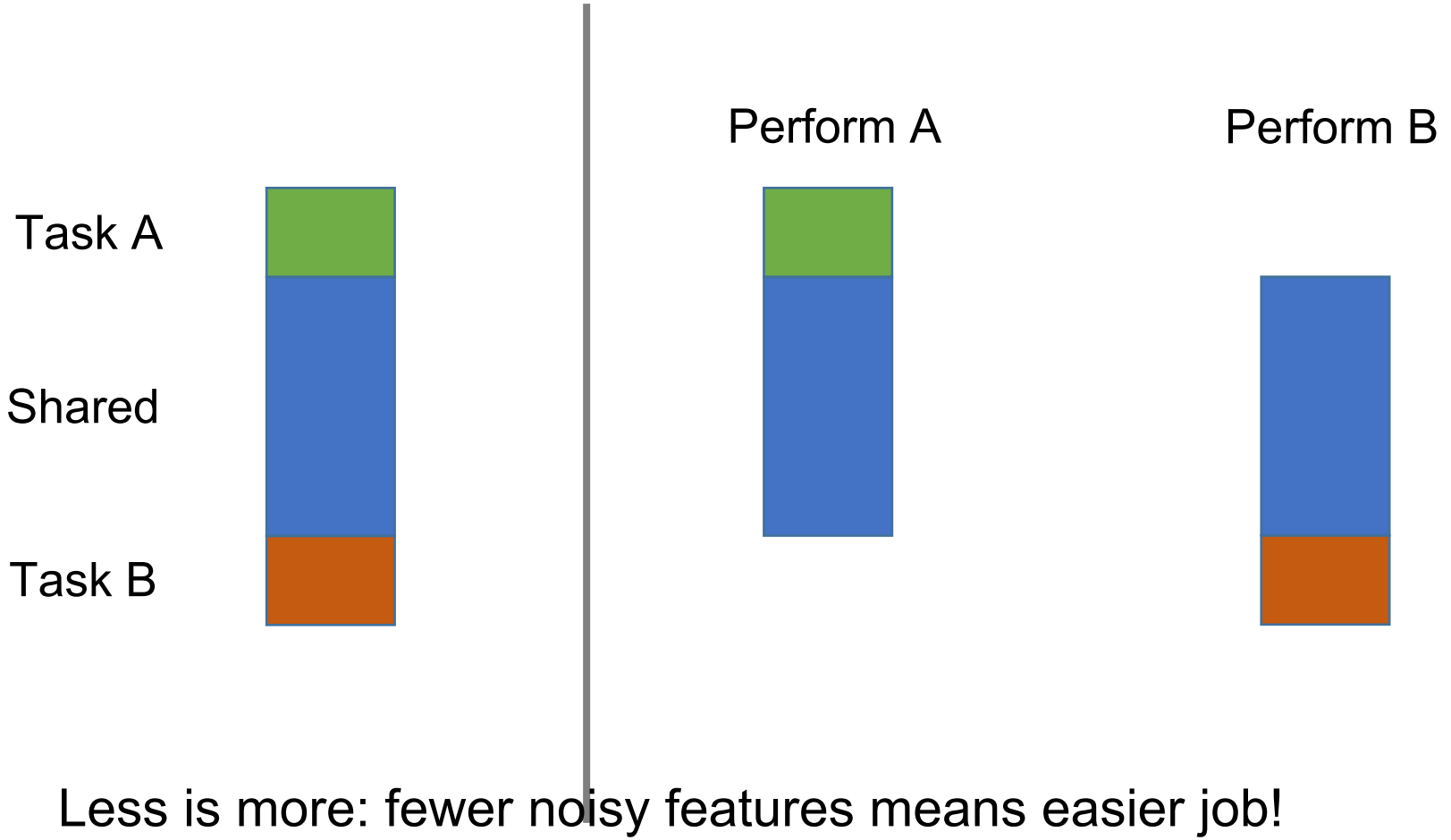
Solution: give each other space



Solution: give each other space



Solution: give each other space

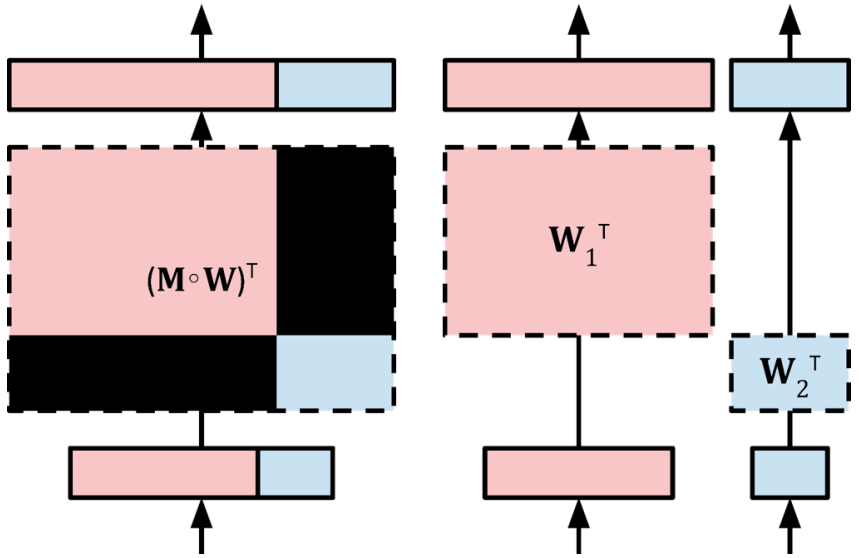


Less is more: fewer noisy features means easier job!

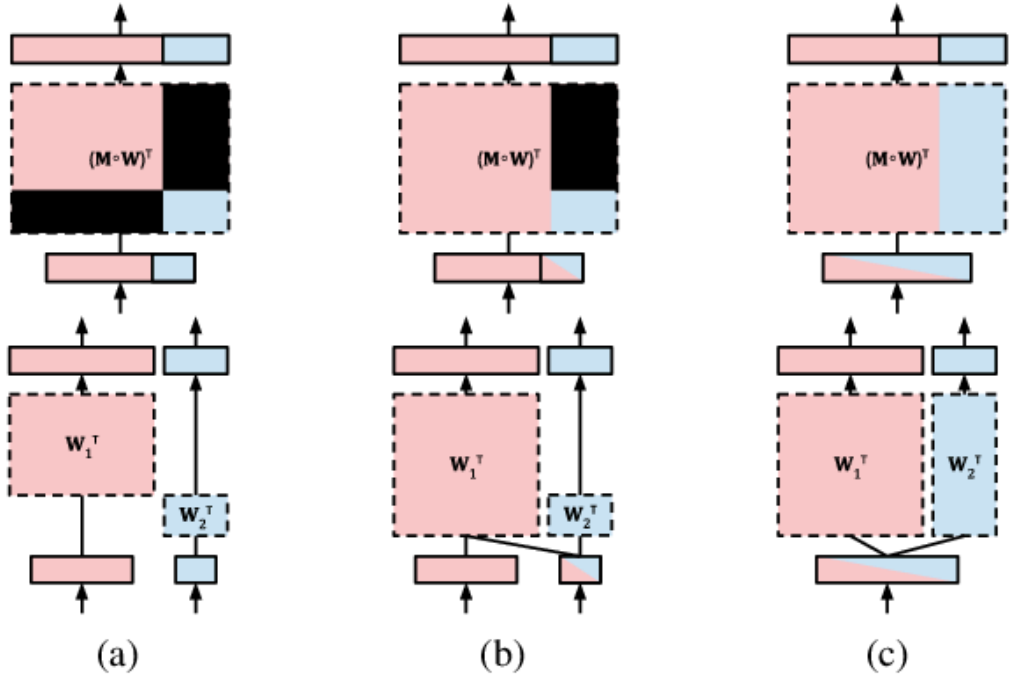
Question: how can we enforce and control the modularity of our representation?

Learning Modular networks by differentiable block sampling

Blockout regularizer

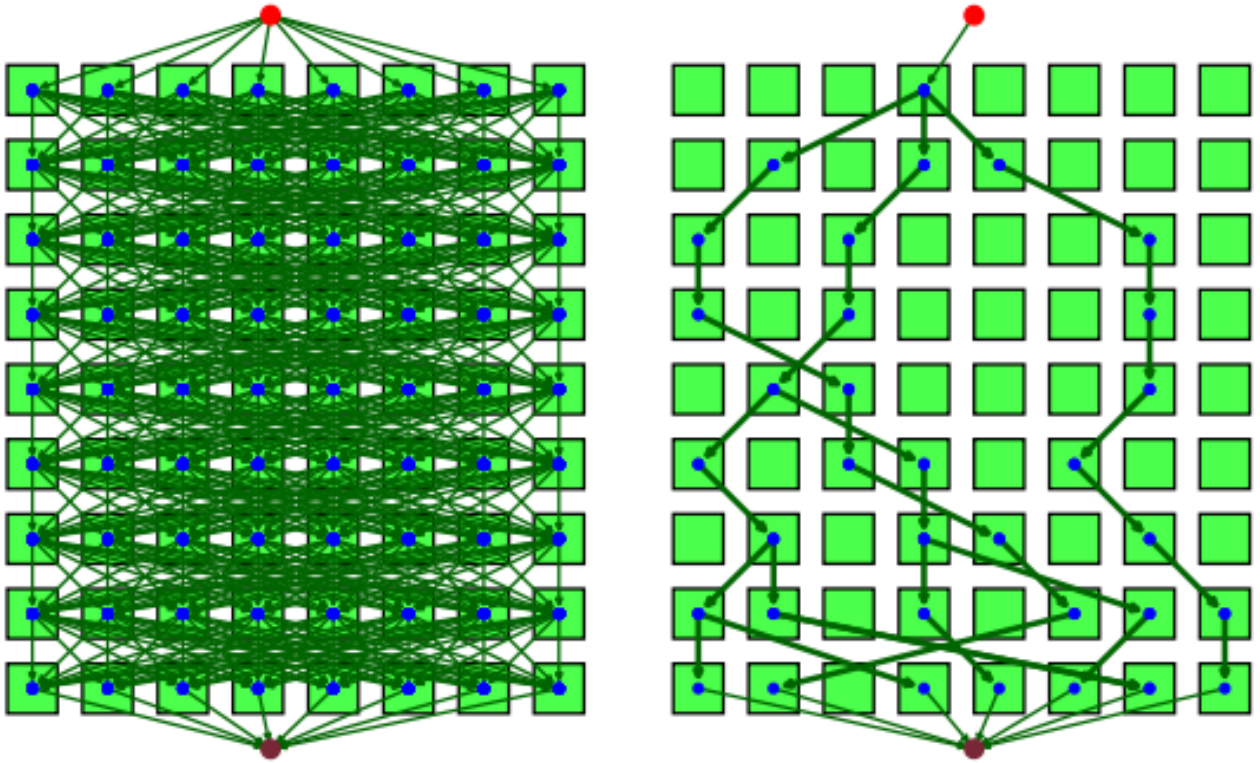


Blocks & induced architectures



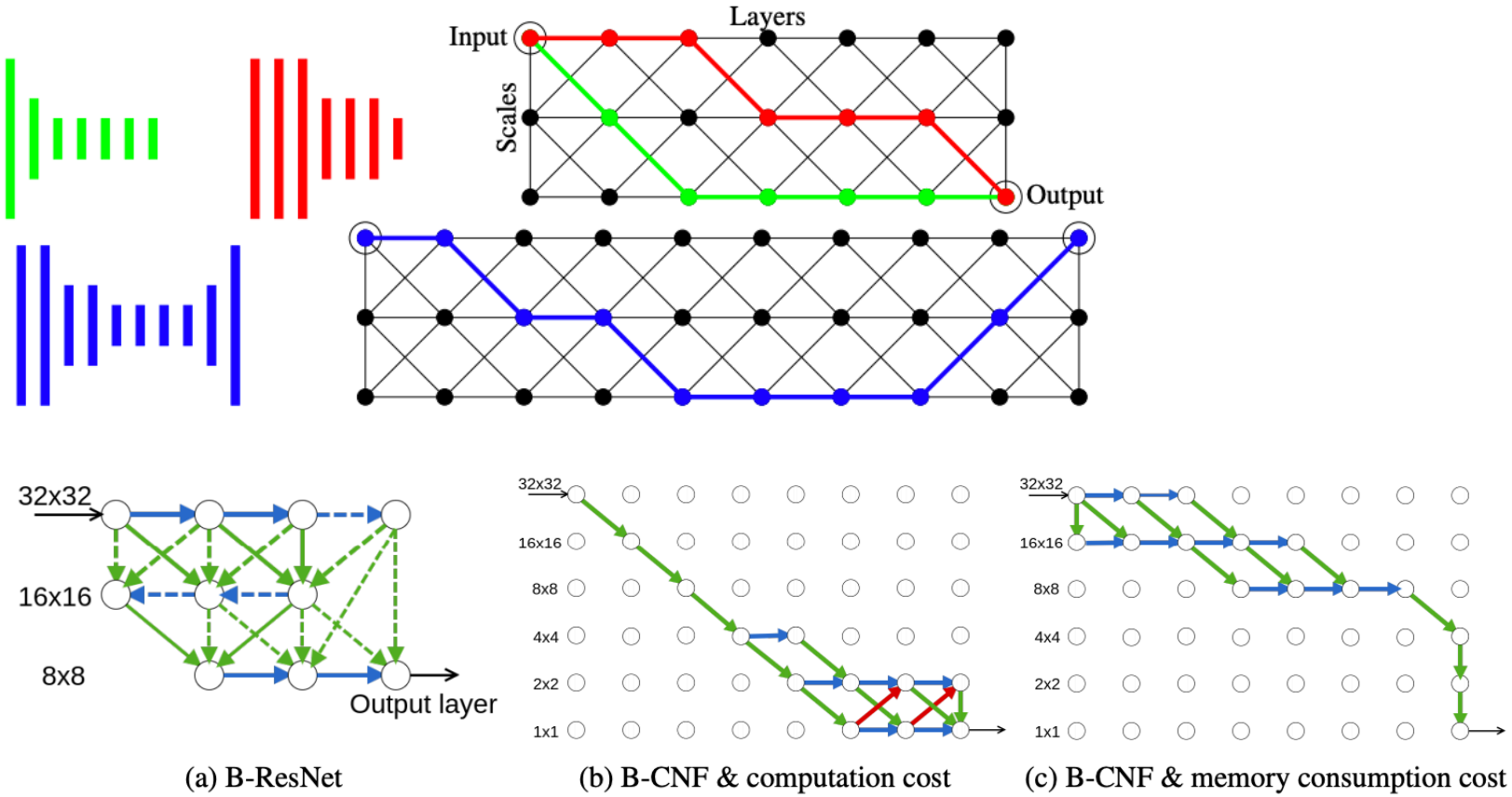
Blockout: Dynamic Model Selection for Hierarchical Deep Networks,
C. Murdock, Z. Li, H. Zhou, T. Duerig, CVPR 2016

Learning Modular networks by differentiable block sampling



MaskConnect: Connectivity Learning by Gradient Descent,
Karim Ahmed, Lorenzo Torresani, 2017

Learning Modular networks by differentiable block sampling

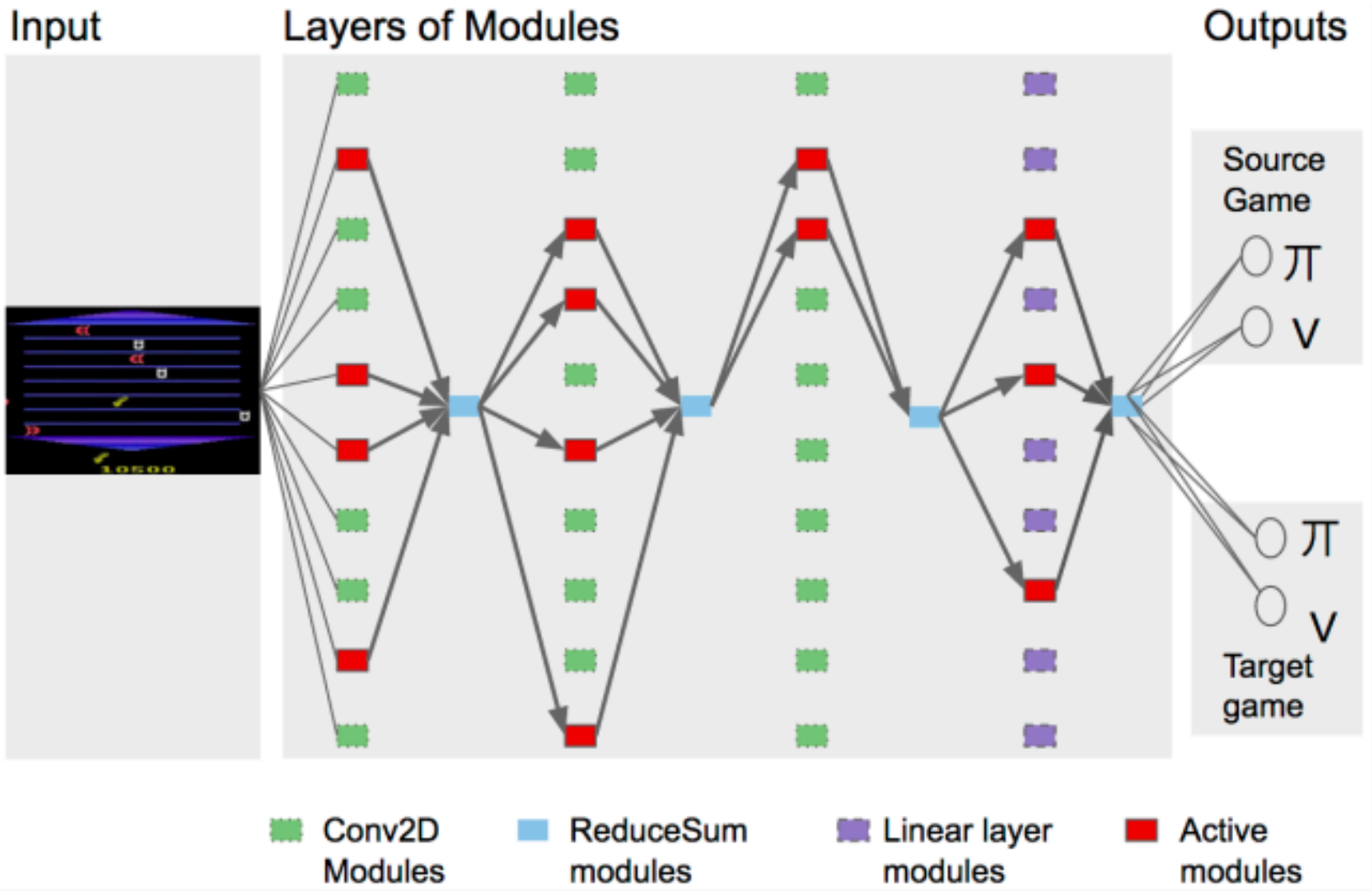


(a) B-ResNet (b) B-CNF & computation cost (c) B-CNF & memory consumption cost

Convolutional Neural Fabrics, S. Saxena and J. Verbeek, NIPS 2016

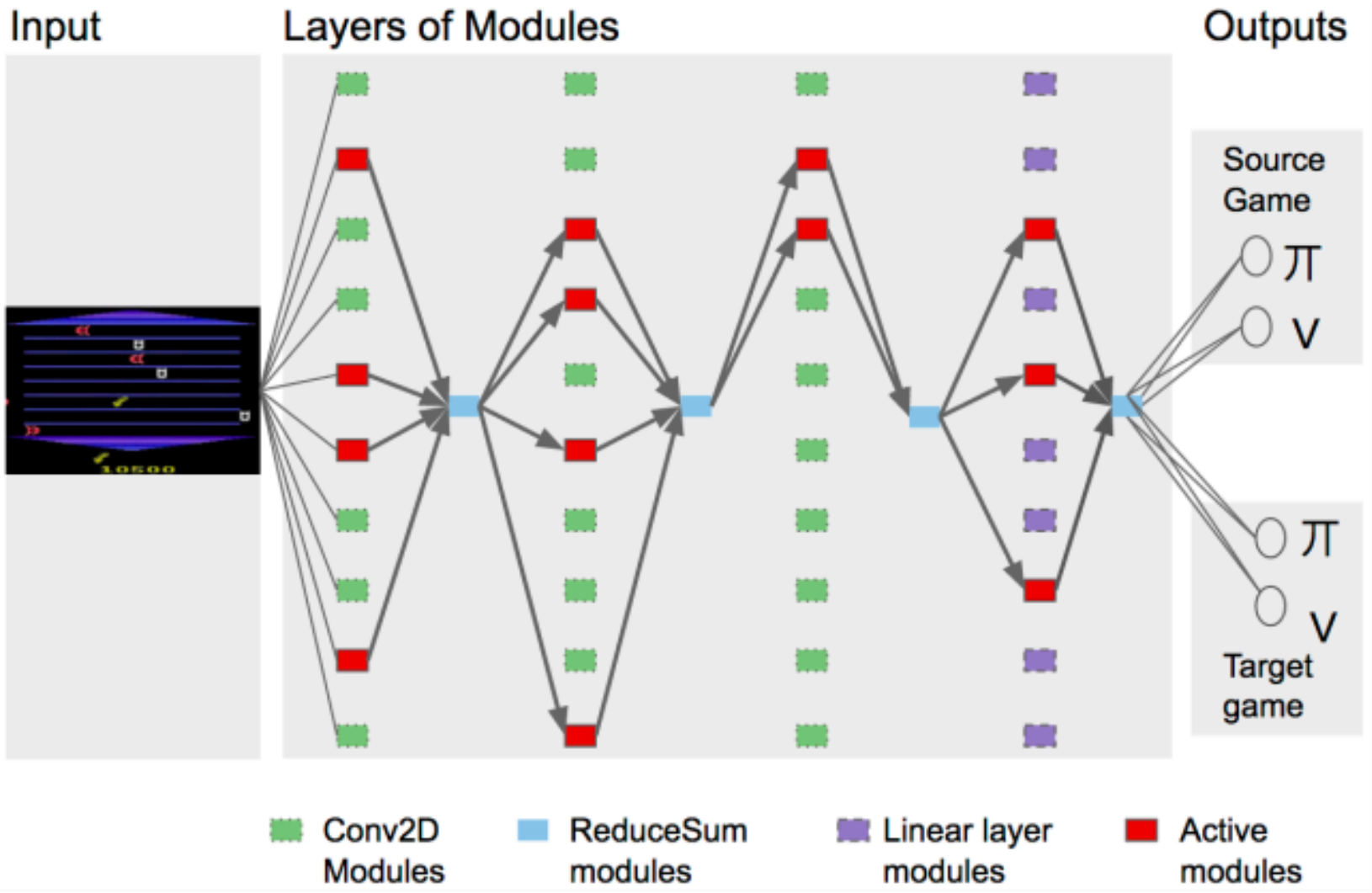
Learning Time/Memory-Efficient Deep Architectures with Budgeted Super Networks, T. Veniat and L. Denoyer, CVPR 2018

Modular networks for multi-tasking



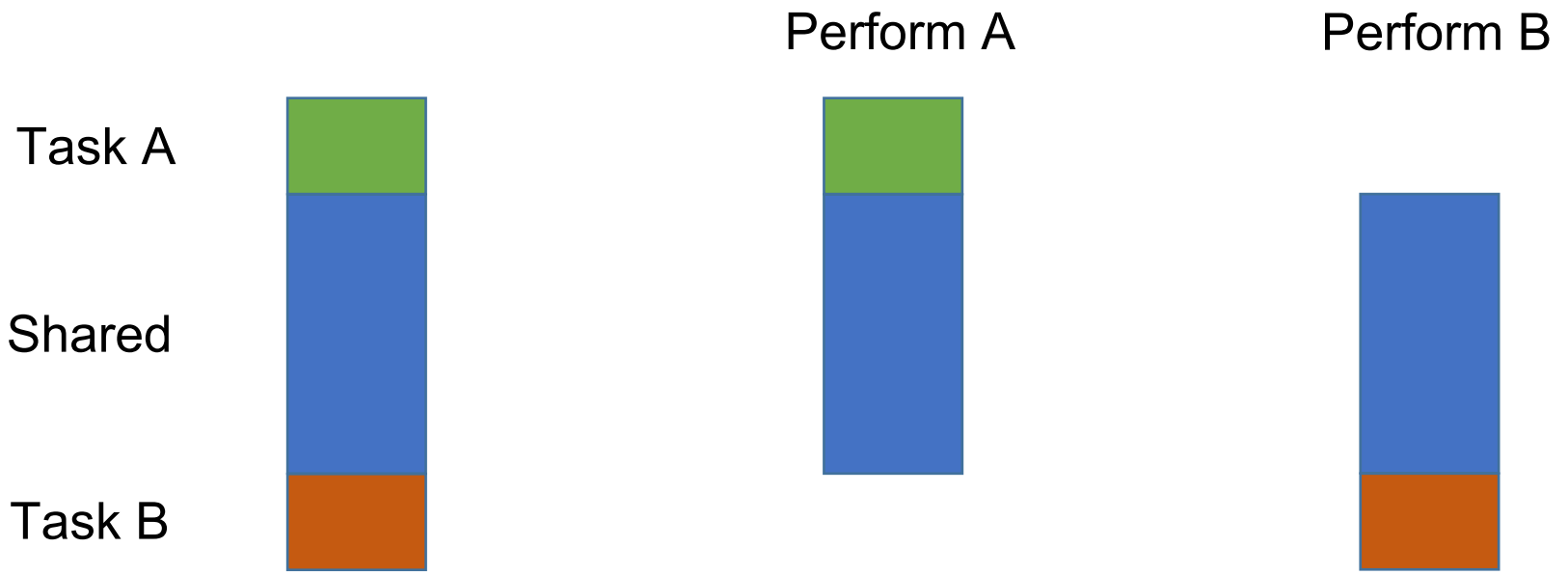
PathNet: Evolution Channels Gradient Descent in Super Neural Networks, Fernando et al., 2017

Modular networks for multi-tasking



PathNet: Evolution Channels Gradient Descent in Super Neural Networks, Fernando et al., 2017

Aim: differentiable & modular multi-task networks



How to avoid combinatorial search over feature-task combinations?

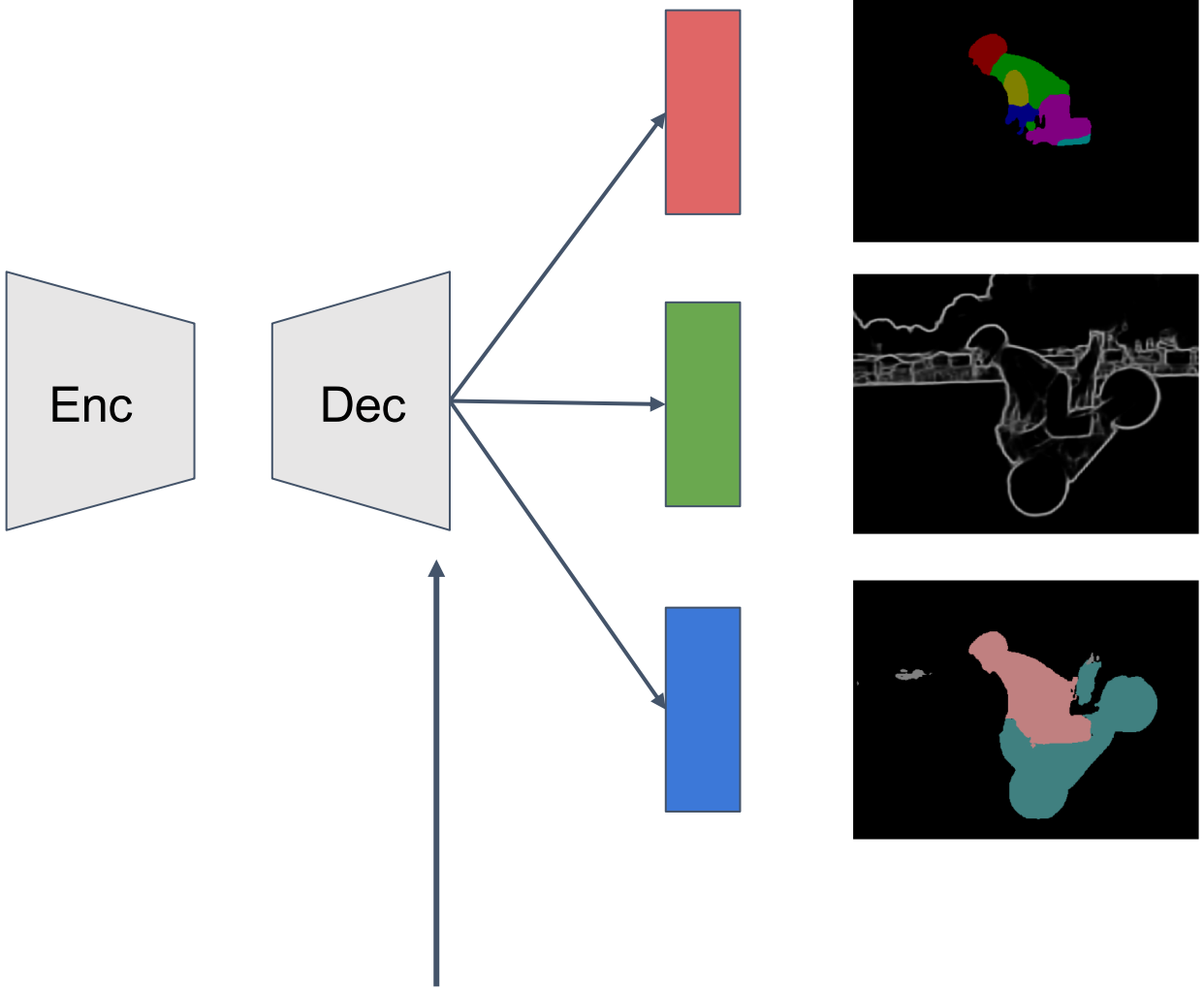
Attentive Single-Tasking of Multiple Tasks

- Approach
  - Network performs one task at a time
  - Accentuate relevant features
  - Suppress irrelevant features

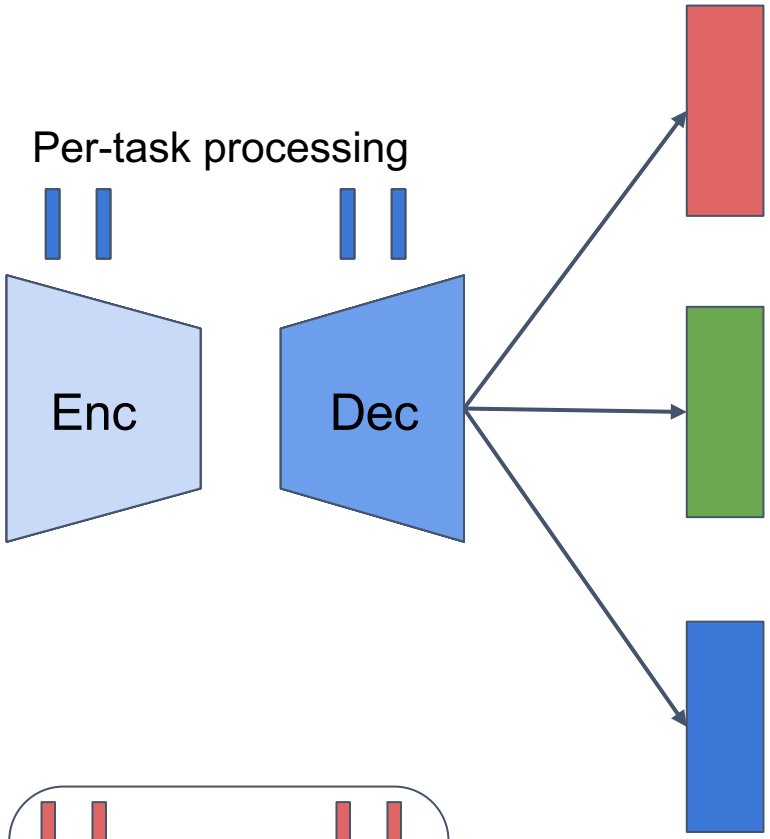
http://www.vision.ee.ethz.ch/~kmaninis/astmt/

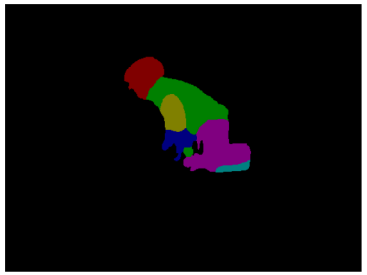
Kevis Maninis, Ilija Radosavovic, I.K. “Attentive single Tasking of Multiple Tasks”, CVPR 2019

Multi-Tasking Baseline



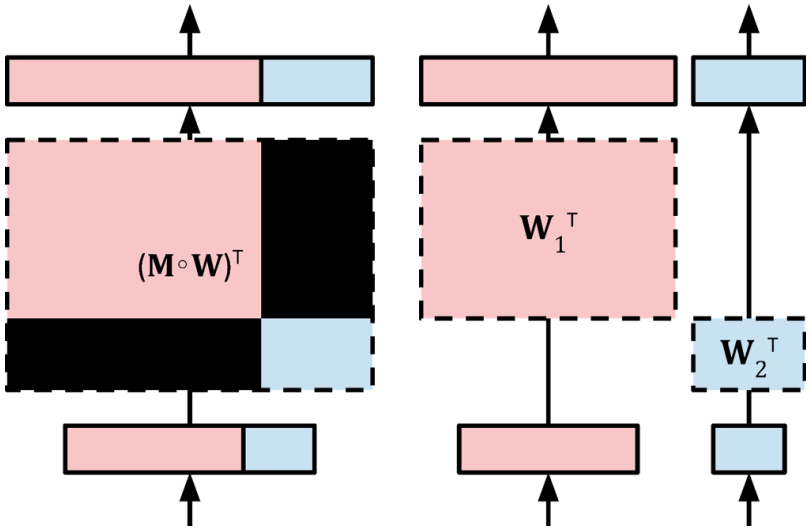
Attention to Task - Ours



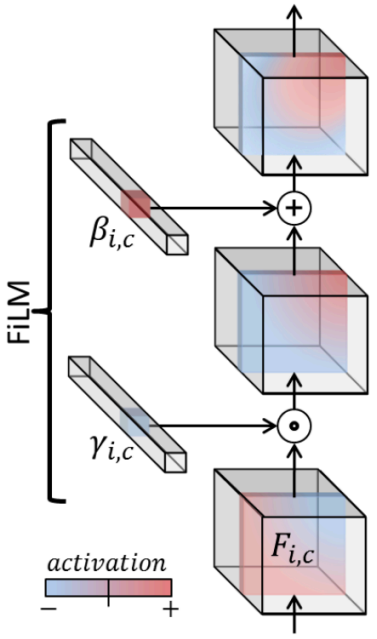


- Attention to task:
  Focus on one task at a time
- Accentuate relevant features
- Suppress irrelevant features

Continuous search over blocks with attention

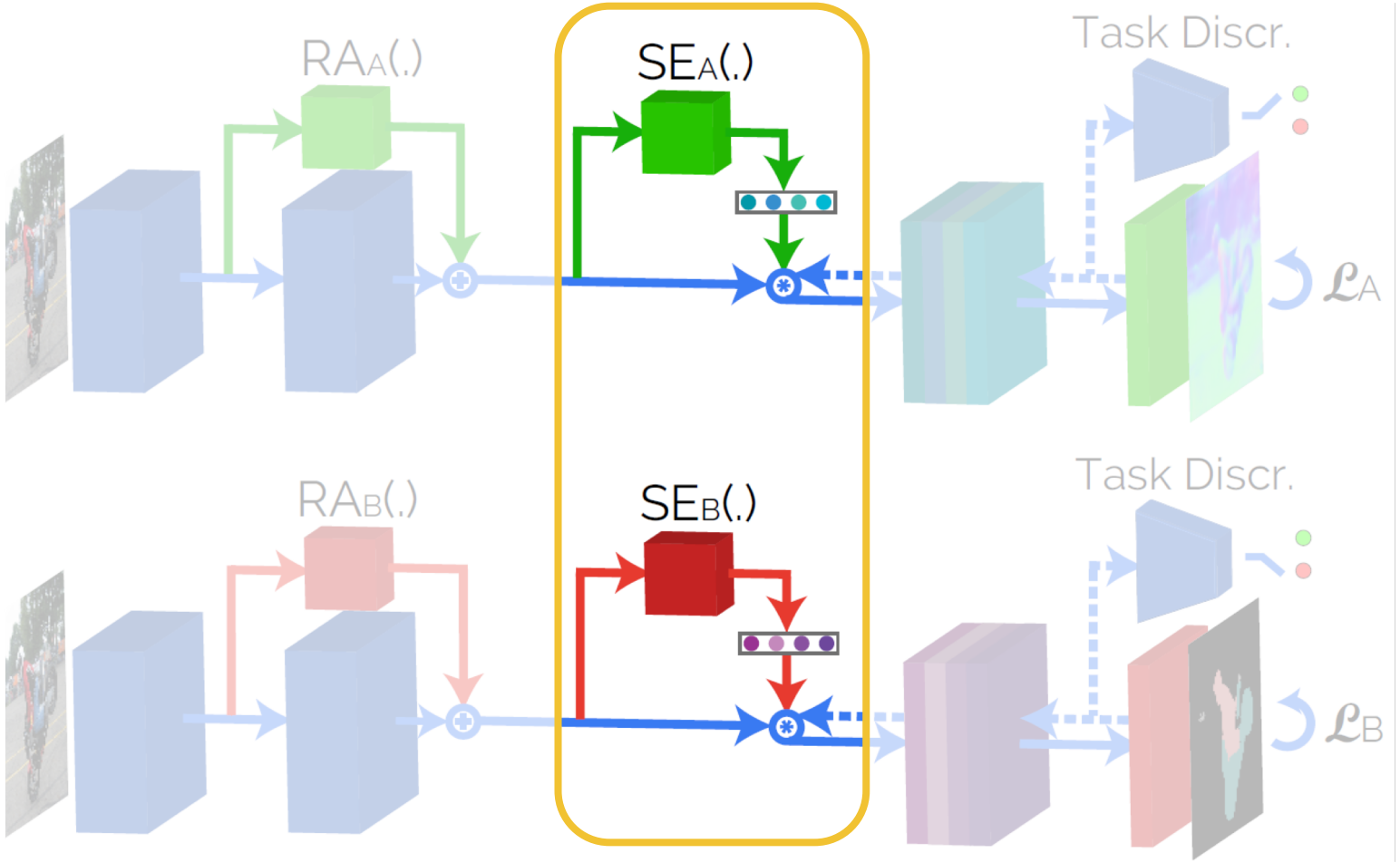


Modularity through Modulation:
we can recover any task-specific block
by shunning the remaining neurons



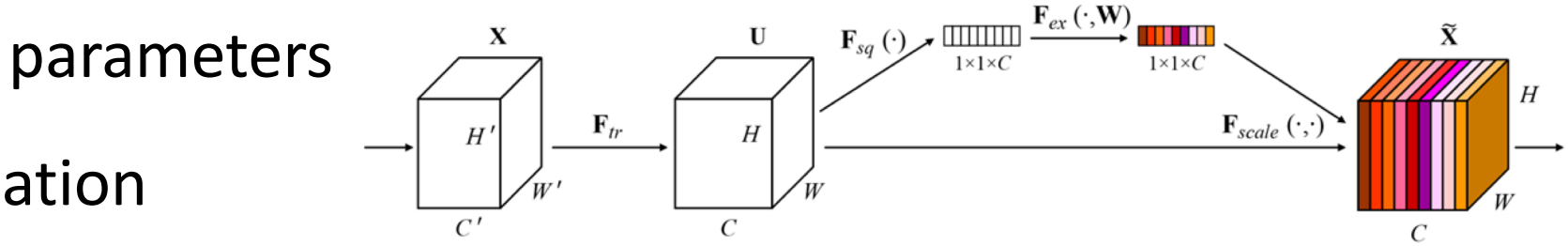
A Learned Representation For Artistic Style., V. Dumoulin, J. Shlens, and M. Kudlur. ICLR, 2017.
FiLM: Visual Reasoning with a General Conditioning Layer, E. Perez, Florian Strub, H. Vries, V. Dumoulin, A. Courville, AAAI 2018
Learning Visual Reasoning Without Strong Priors, E. Perez, H. Vries, F. Strub, V. Dumoulin, A. Courville, 2017
Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization, Xun Huang, Serge Belongie, 2018
A Style-Based Generator Architecture for Generative Adversarial Networks, T. Karras, S. Laine, T. Aila, CVPR 2019

Modulation: Squeeze and Excitation



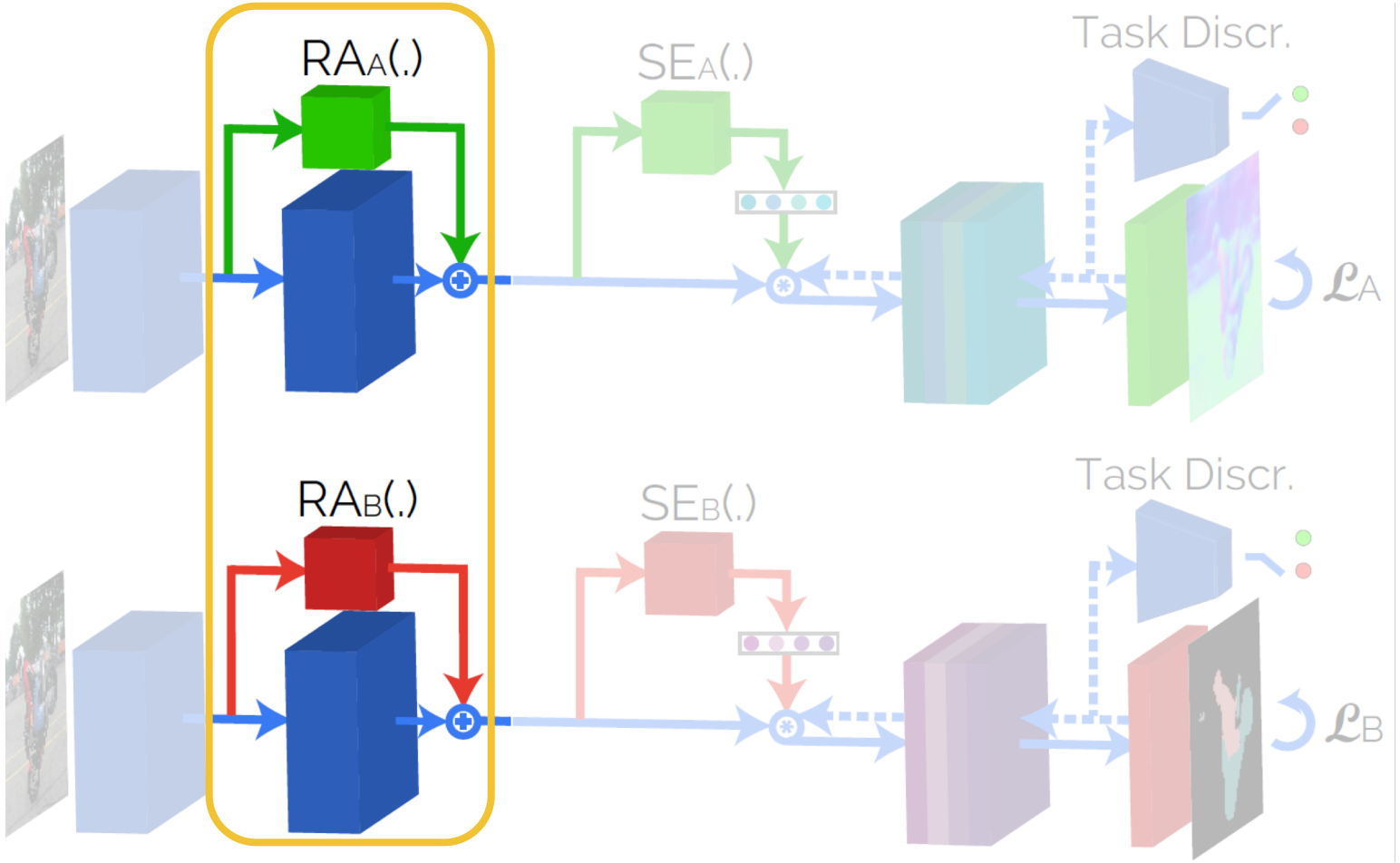
Squeeze and Excitation (SE)

- Negligible amount of parameters
- Global feature modulation



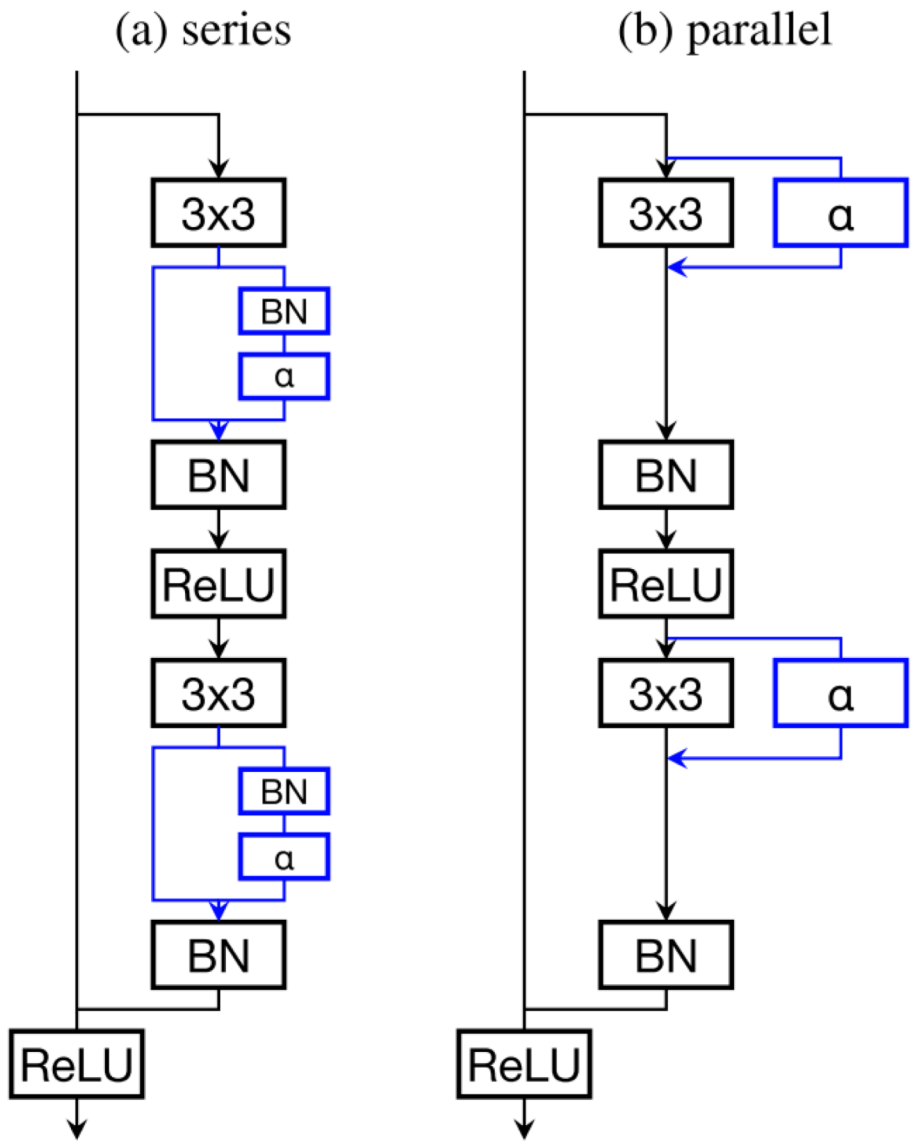
Hu et al., "Squeeze and Excitation Networks", in CVPR 2018

Feature Augmentation: Residual Adapters



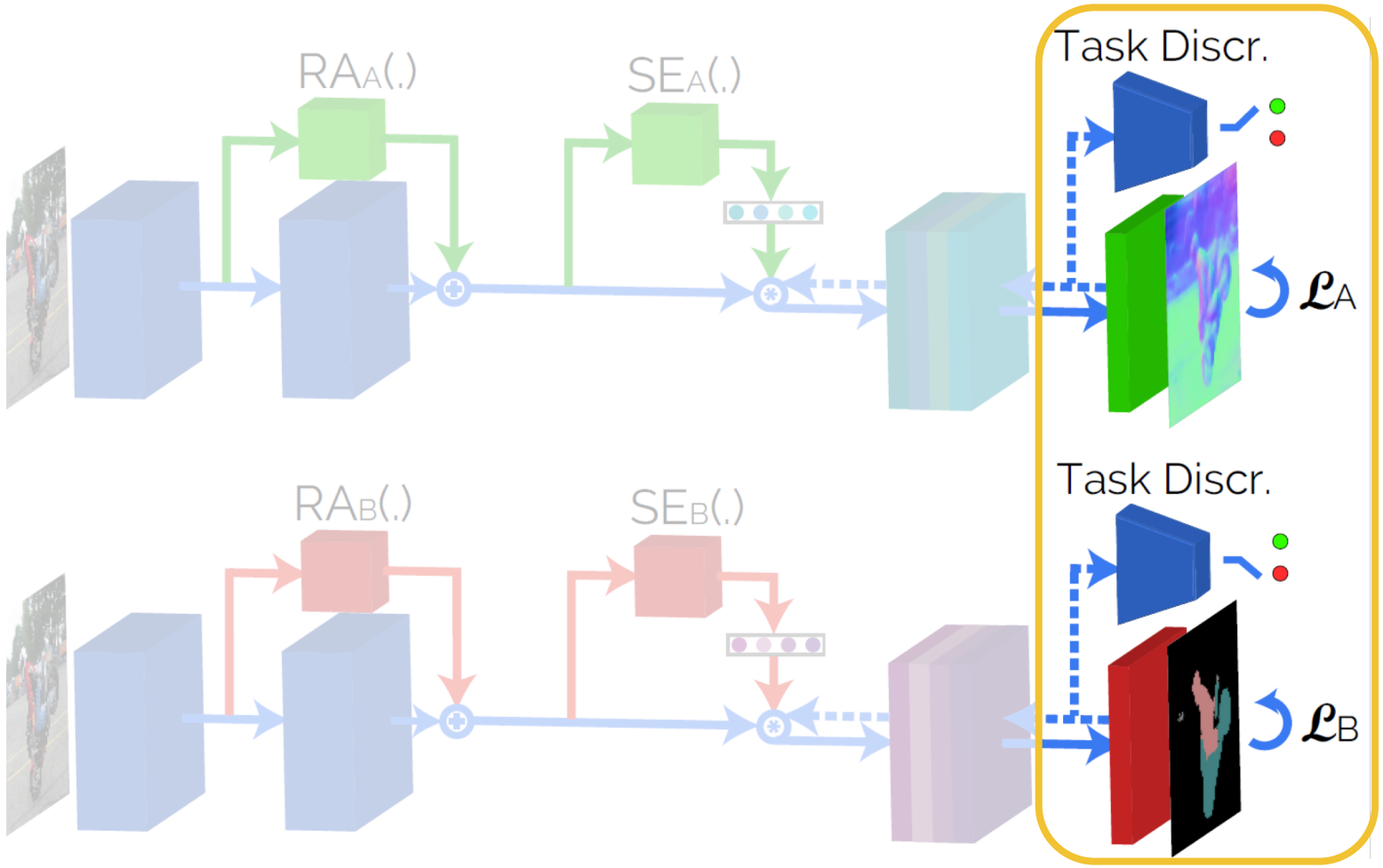
Residual Adapters (RA)

- Original used for Domain adaptation
- Negligible amount of parameters
- In this work: parallel residual adapters

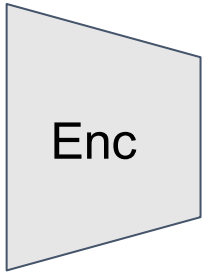


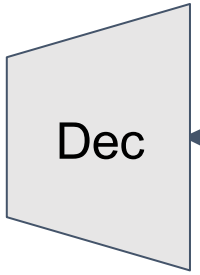
Rebuffi et al., "Learning multiple visual domains with residual adapters", in NIPS 2017
Rebuffi et al., "Efficient parametrization of multi-domain deep neural networks", in CVPR 2018

Adversarial Task Discriminator



Handling Conflicting Gradients: Adversarial Training

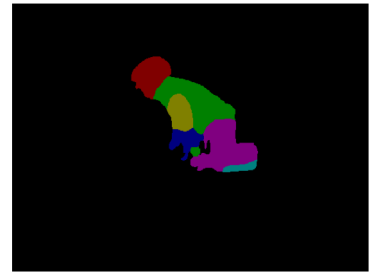




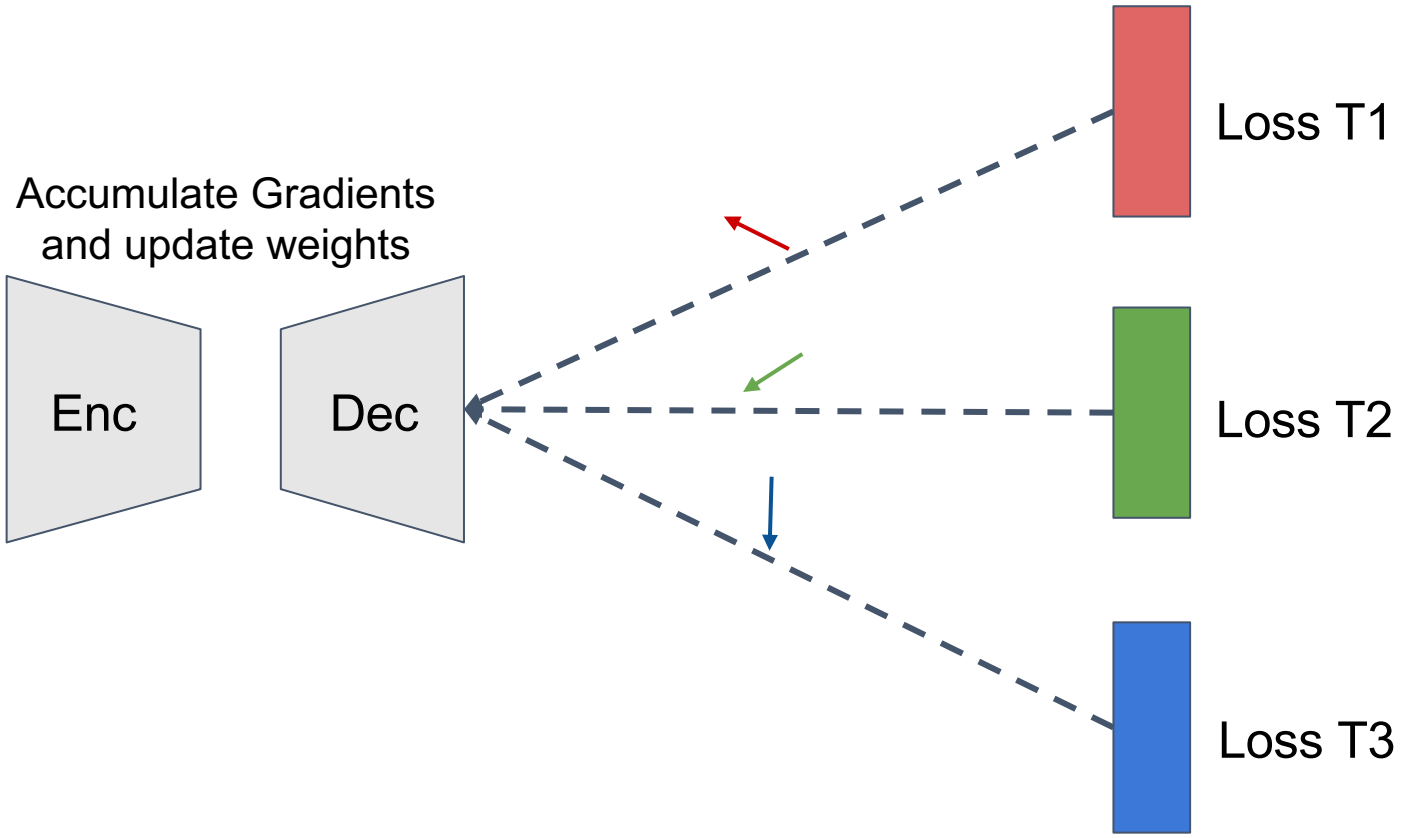
Loss T1

Loss T2

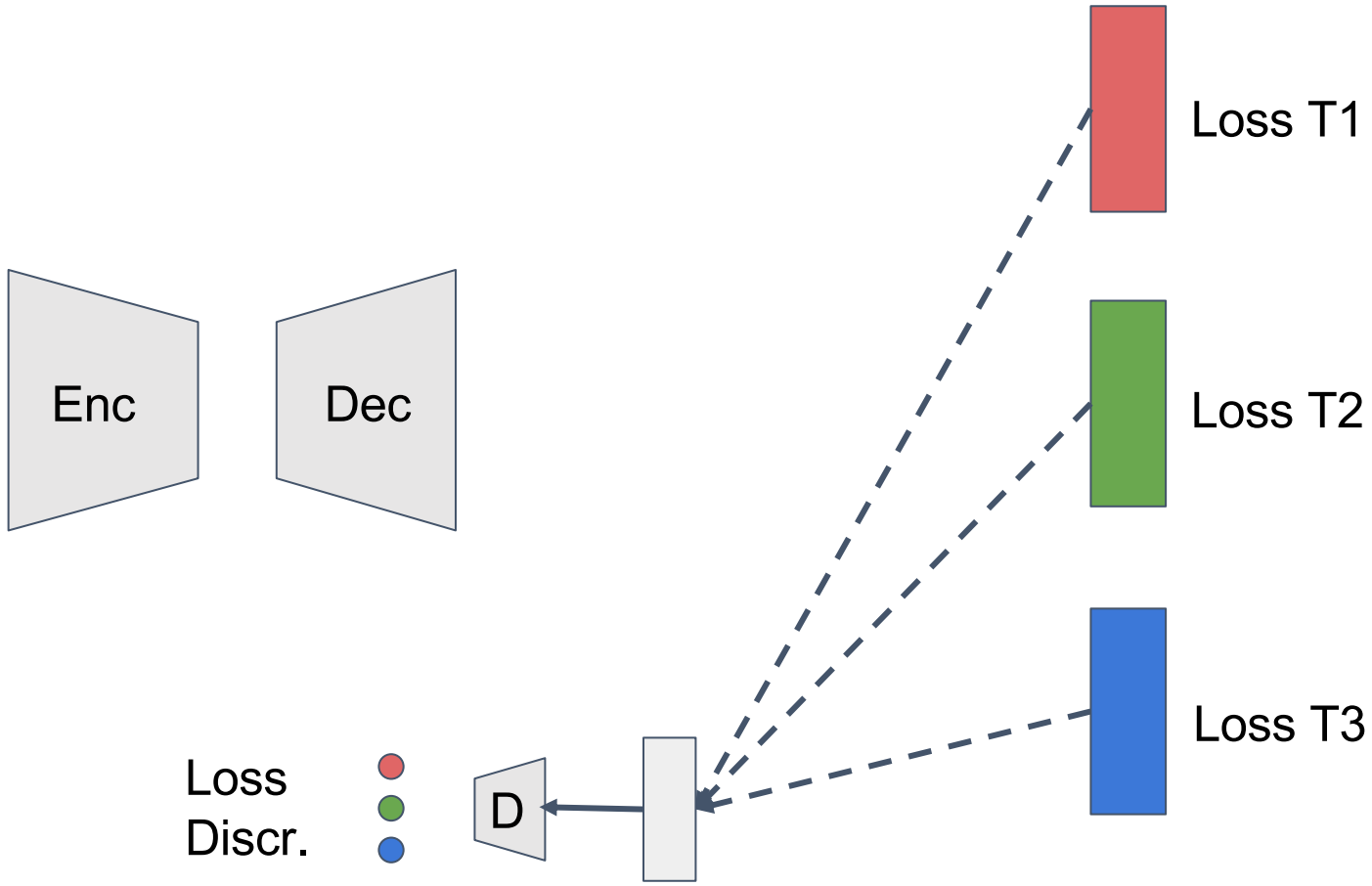
Loss T3



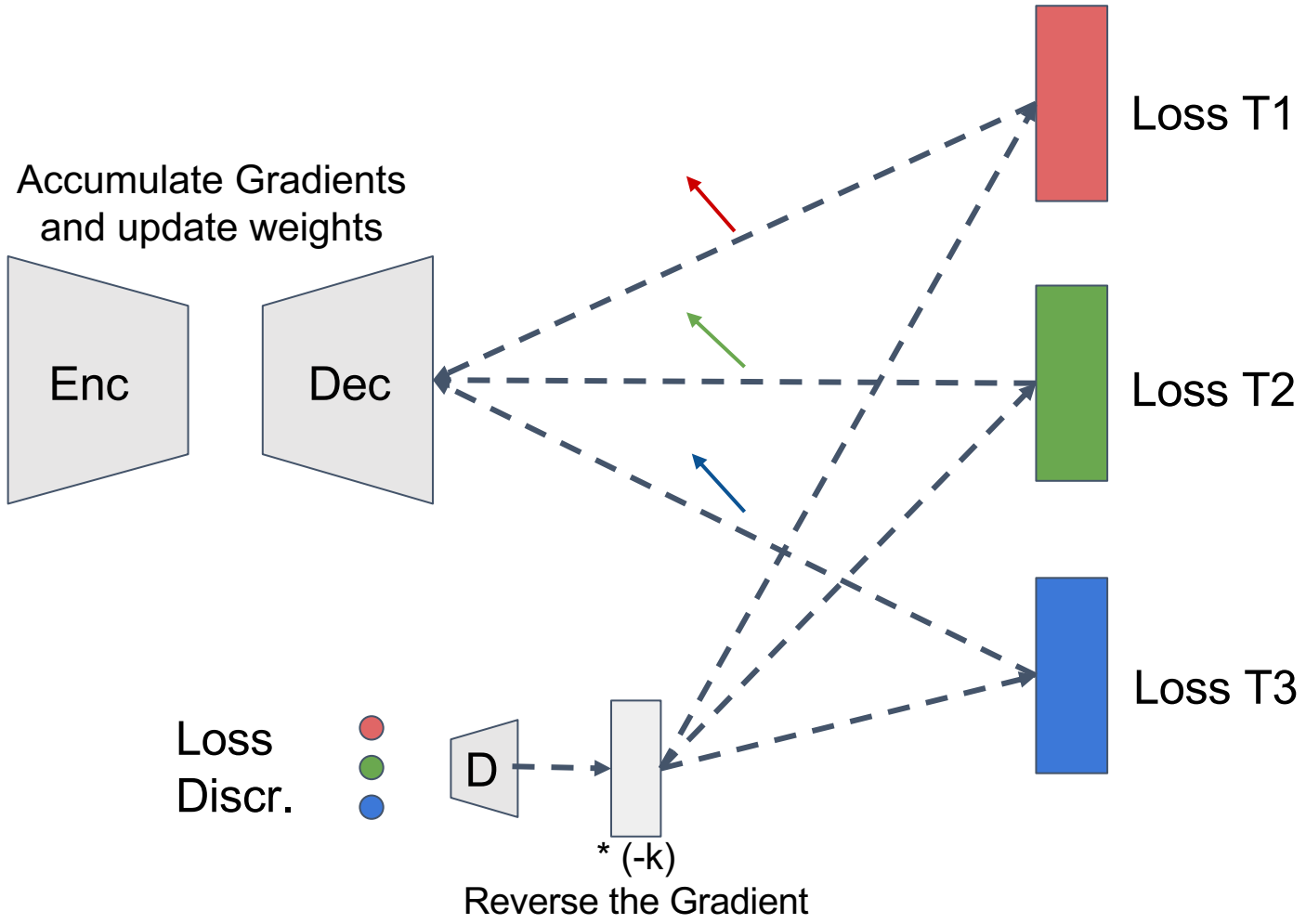
Handling Conflicting Gradients: Adversarial Training



Handling Conflicting Gradients: Adversarial Training



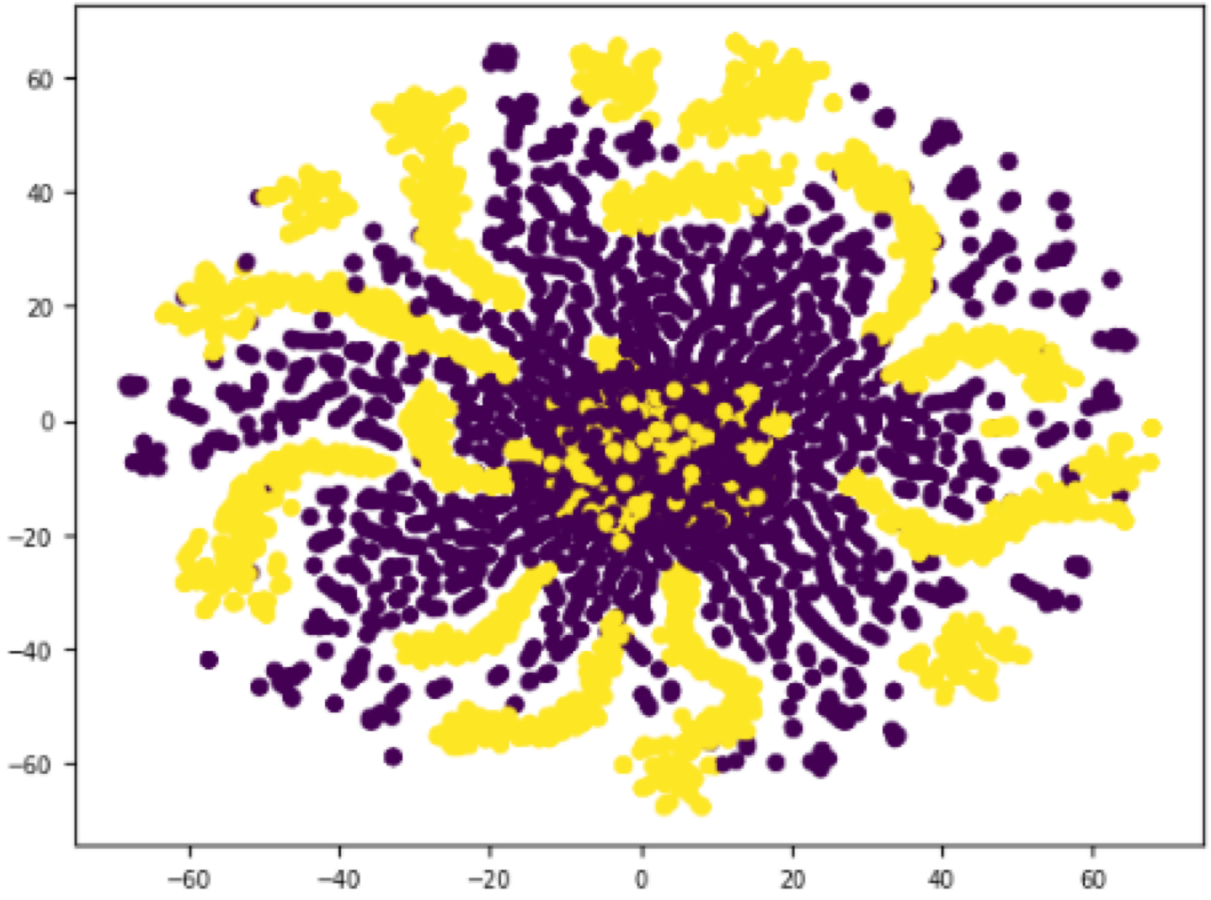
Handling Conflicting Gradients: Adversarial Training



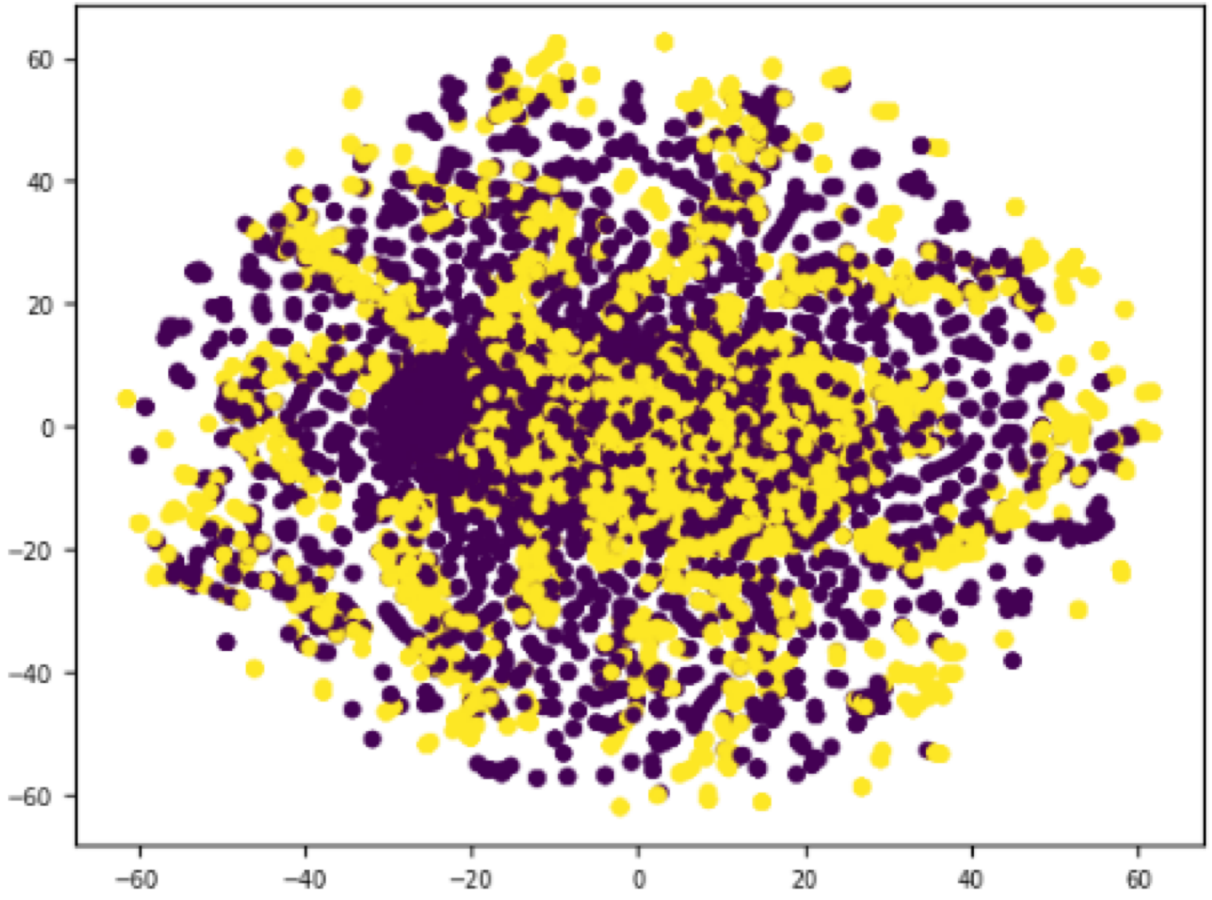
Ganin and Lempitsky, "Unsupervised Domain Adaptation by Backpropagation", in ICML 15

Effect of adversarial training on gradients

t-SNE visualizations of gradients for 2 tasks, without and with adversarial training



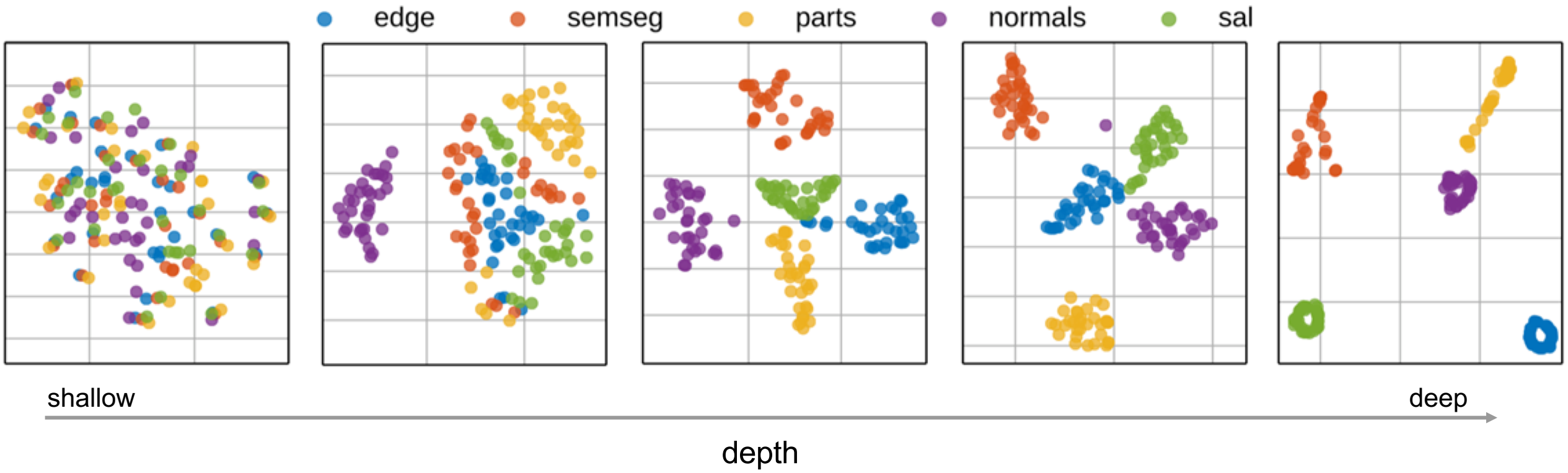
w/o adversarial training



w/ adversarial training

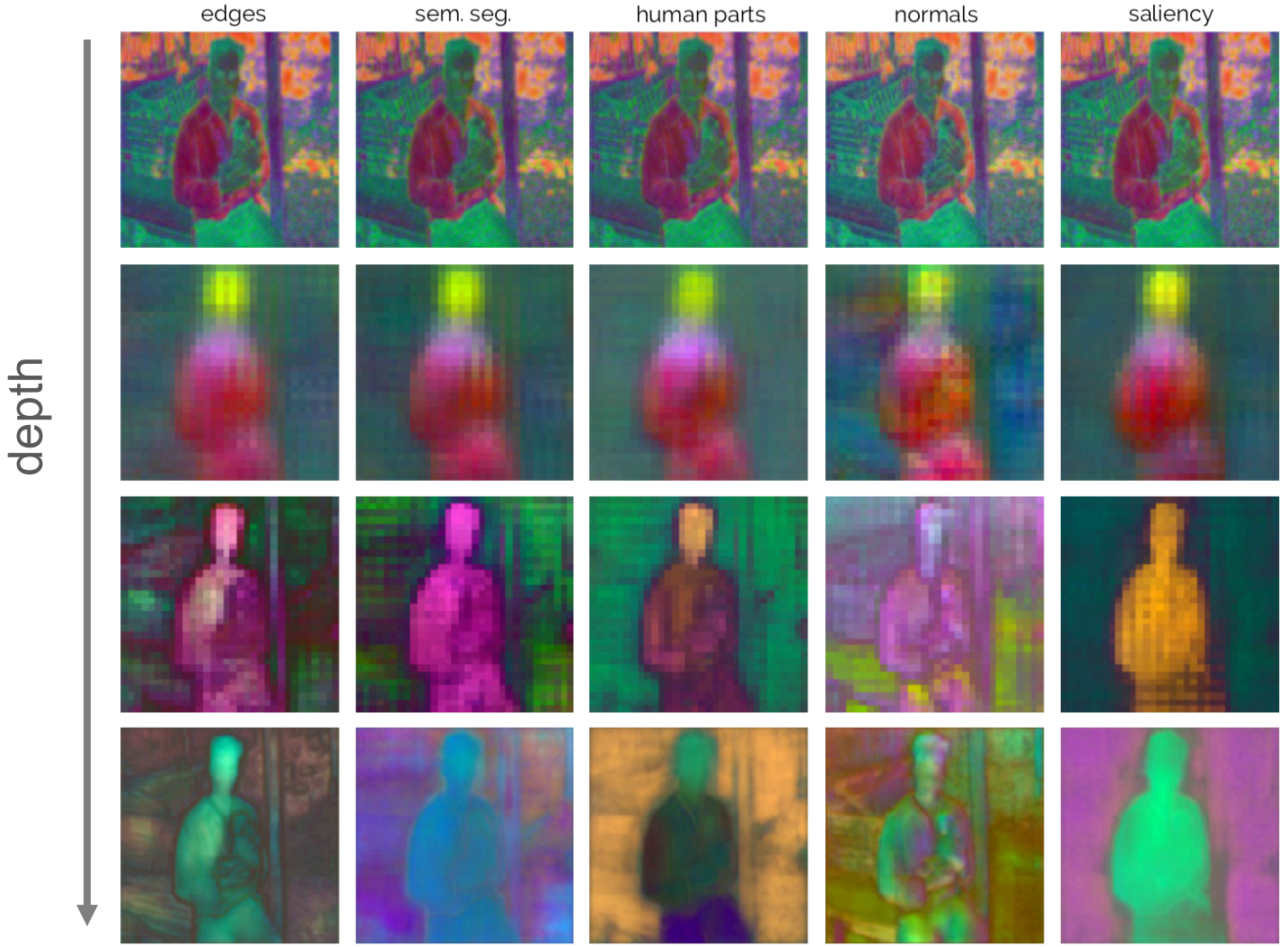
Learned task-specific representation

t-SNE visualizations of SE modulations for the first 32 val images in various depths of the network

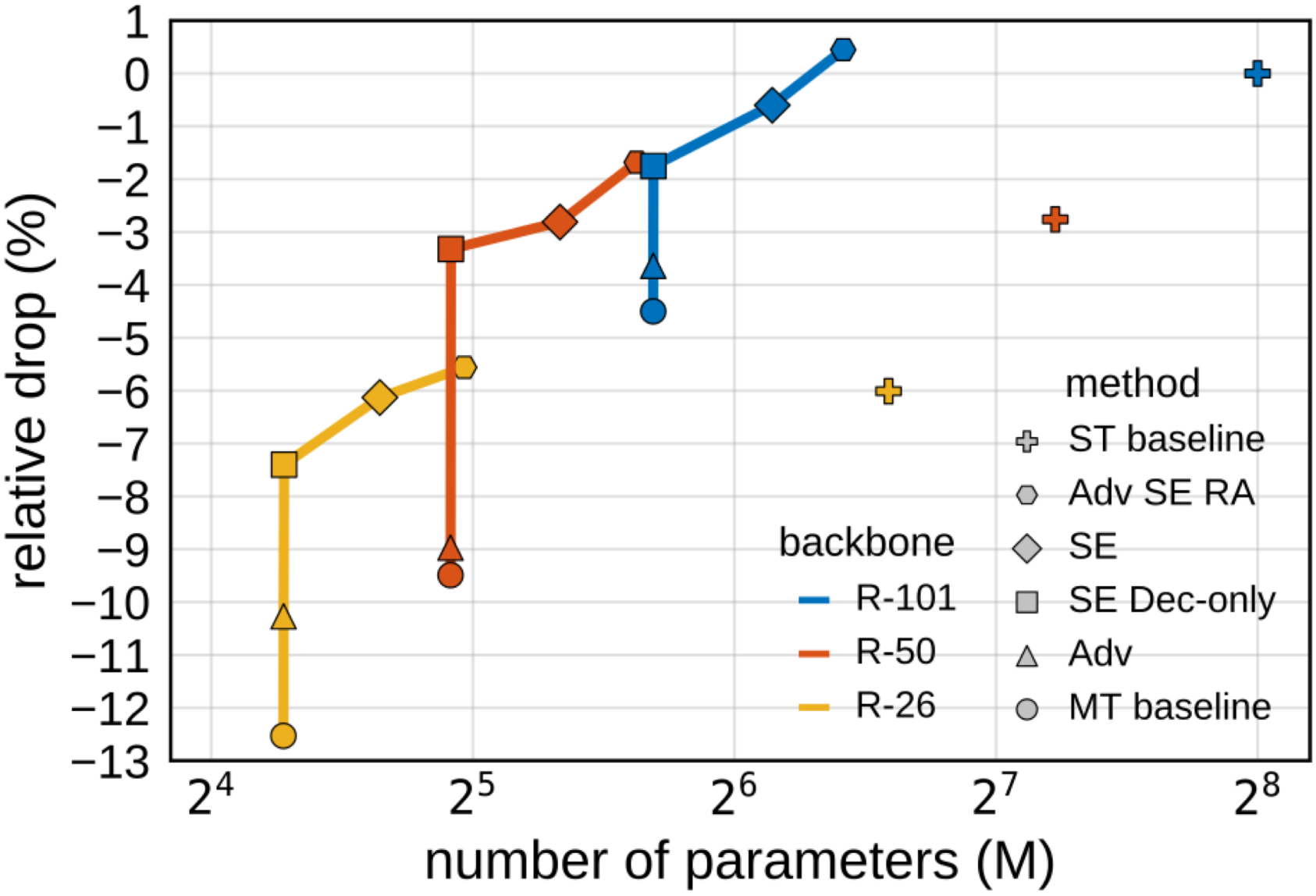


Learned task-specific representation

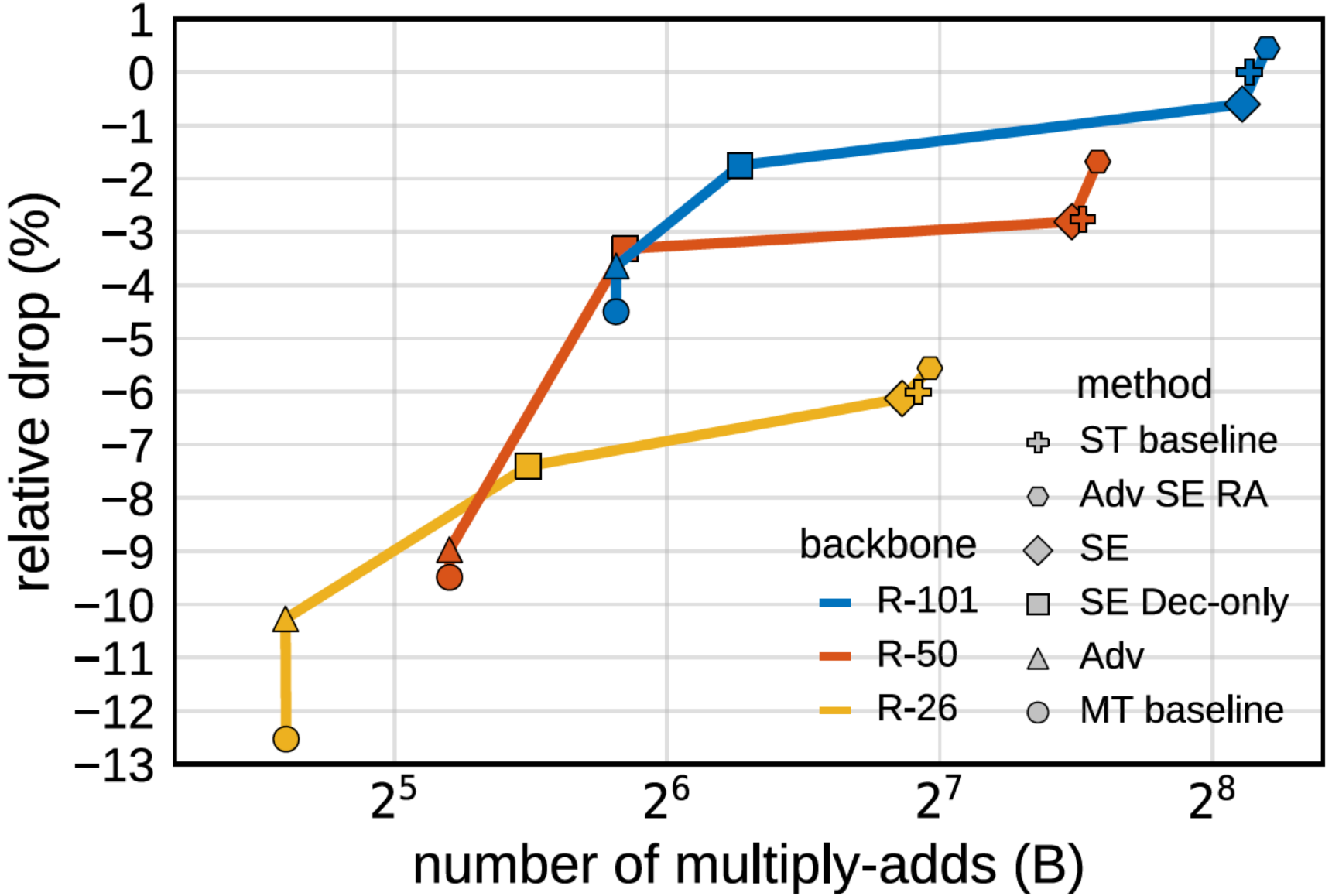
PCA projections
into "RGB" space



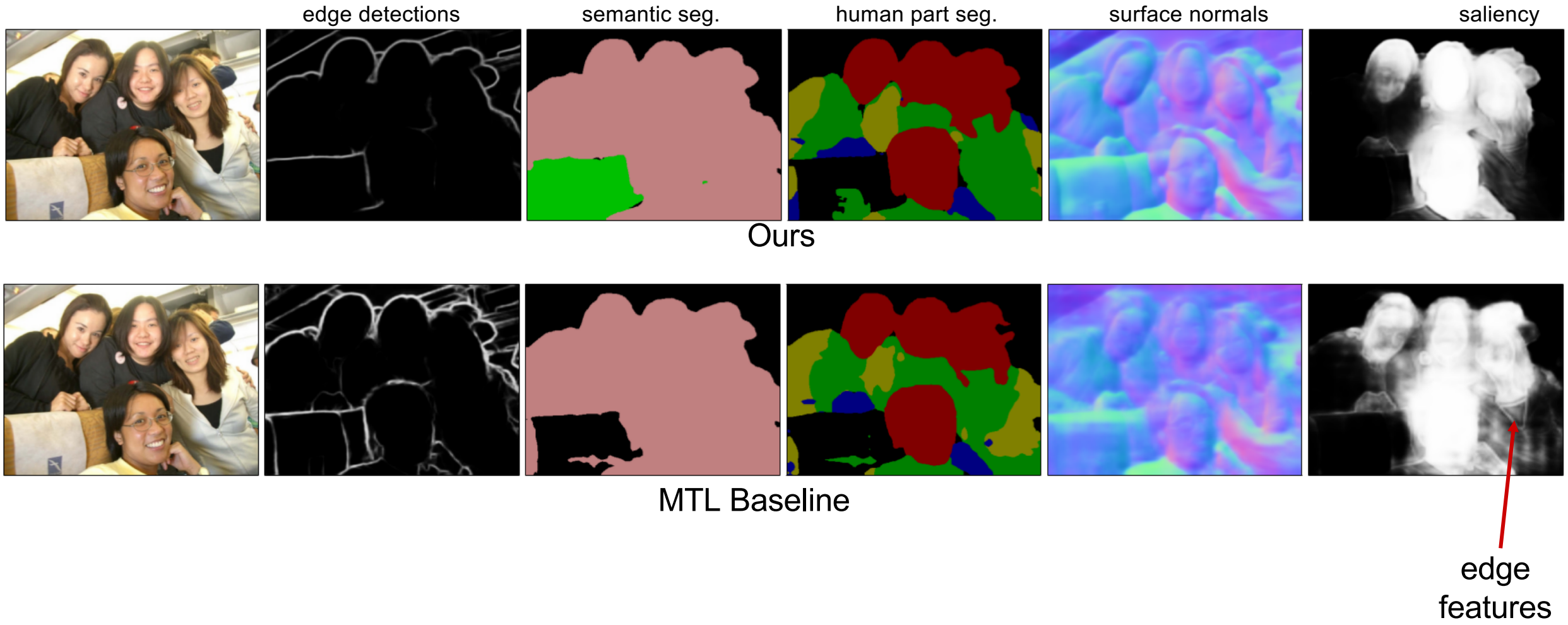
Relative average drop vs. # Parameters



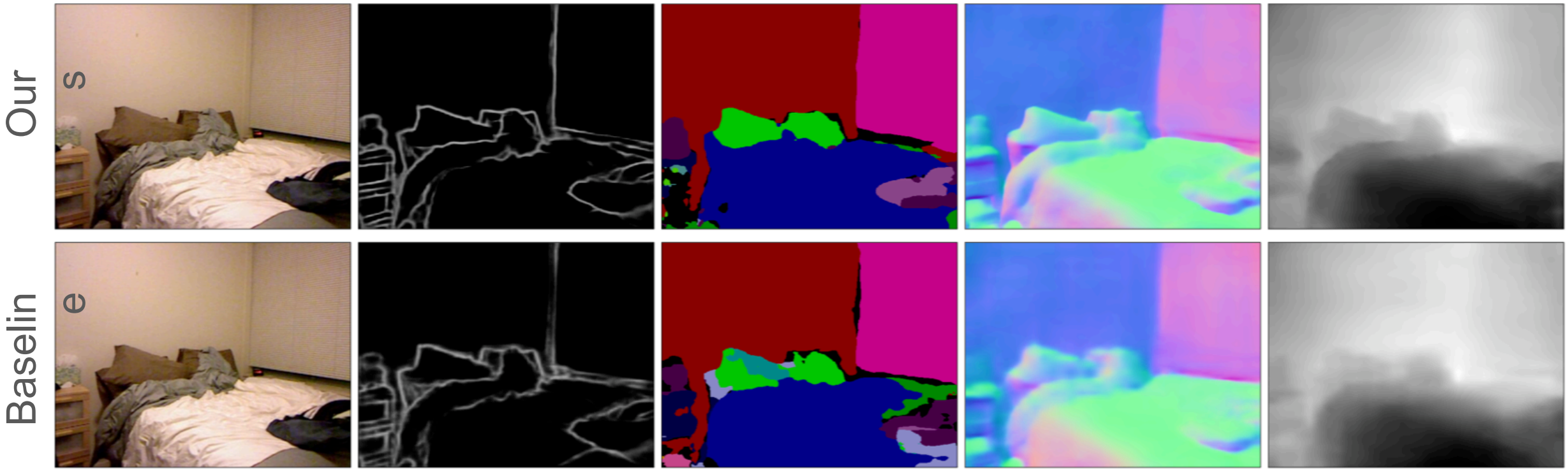
Relative average drop vs. FLOPS



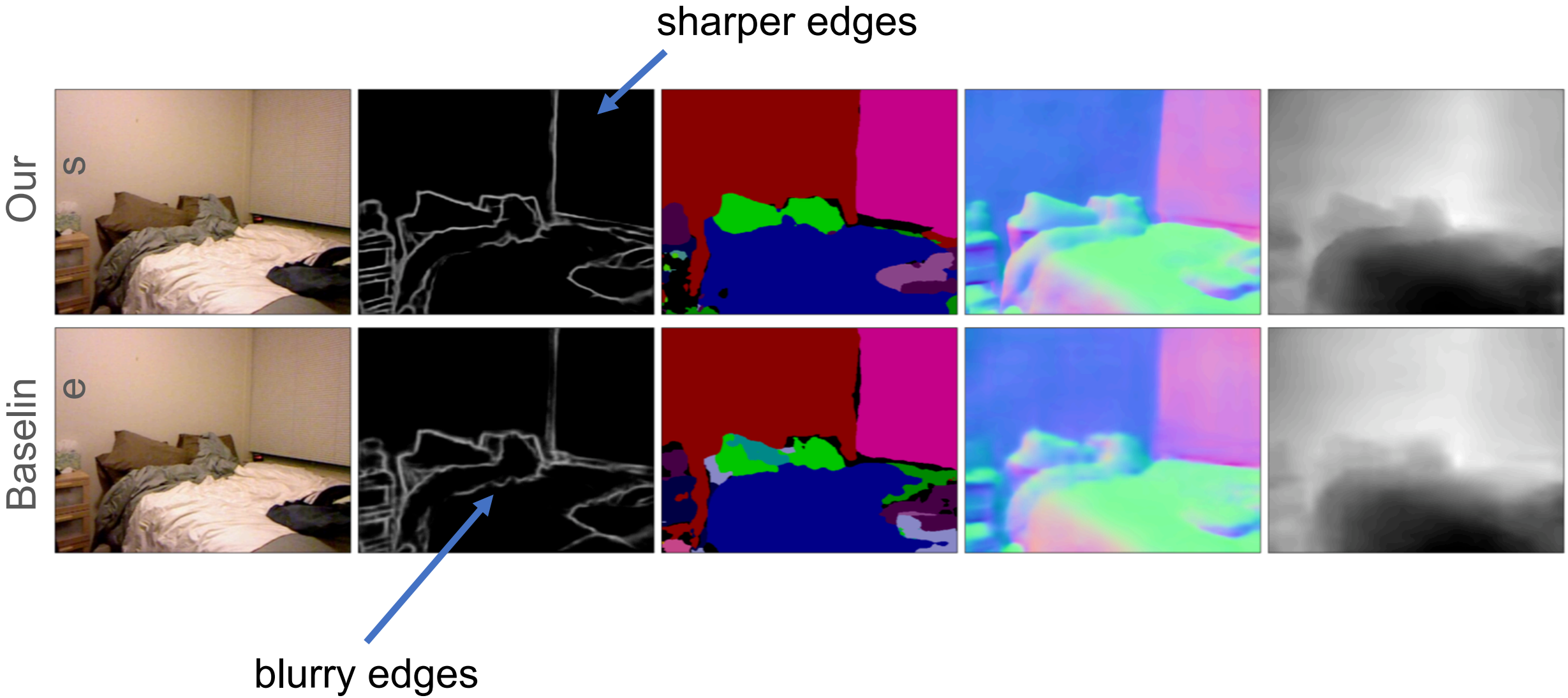
Qualitative Results: PASCAL



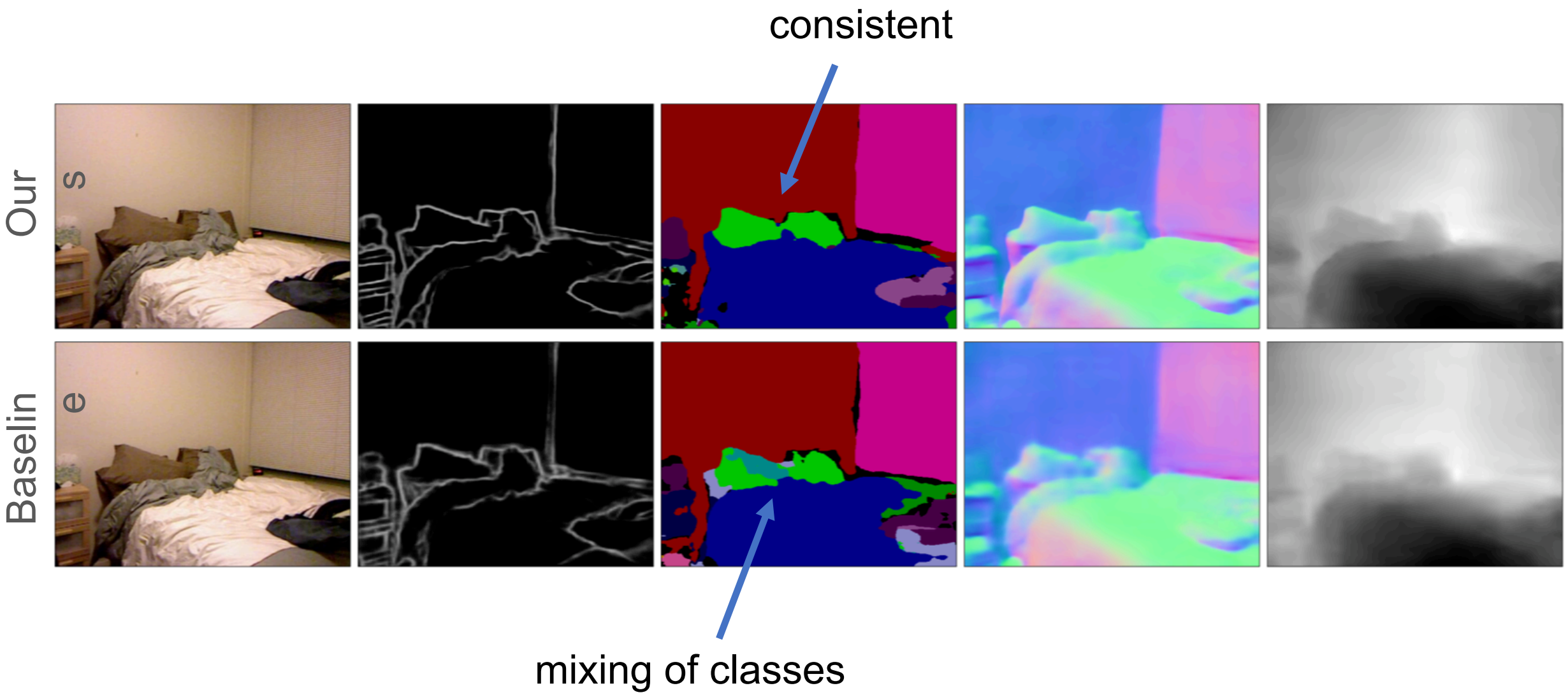
Qualitative Results



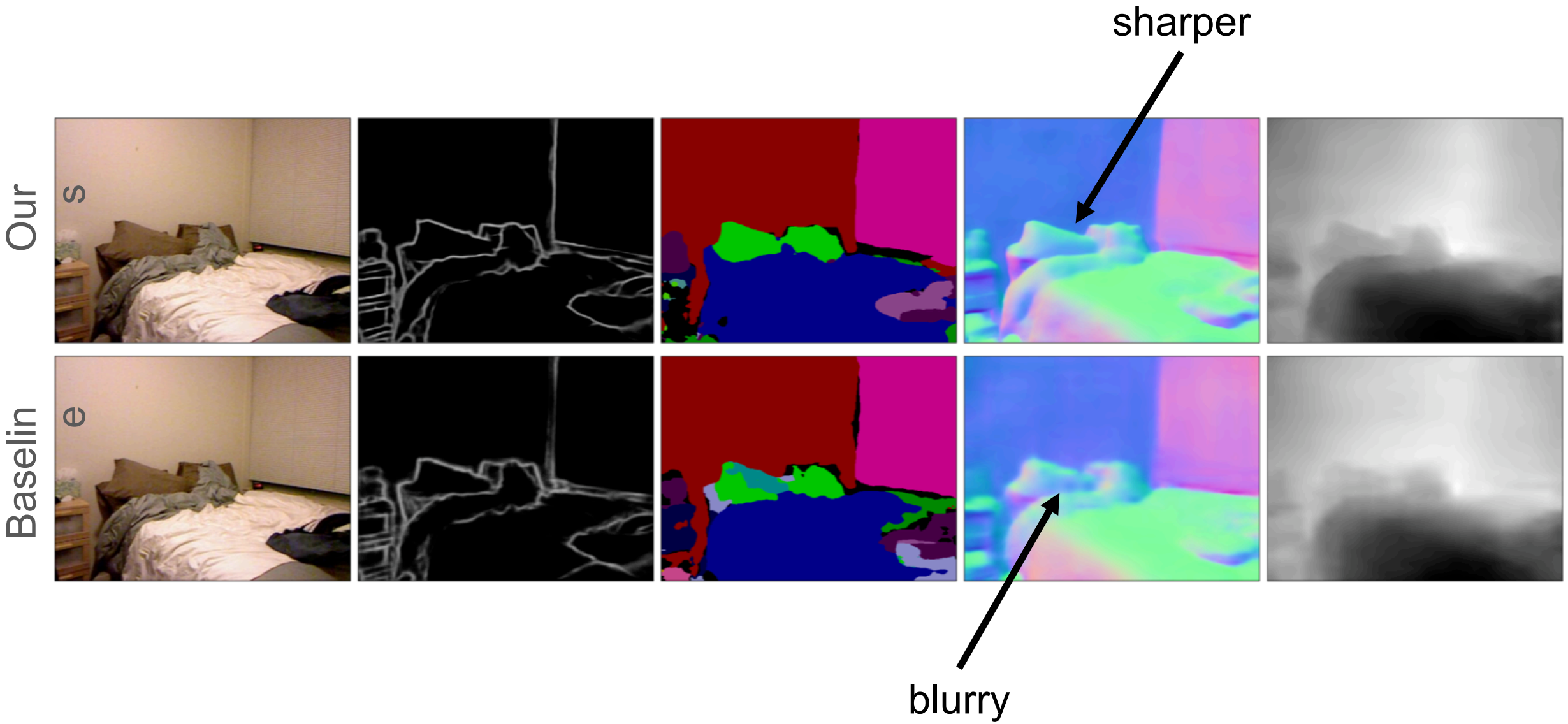
Qualitative Results



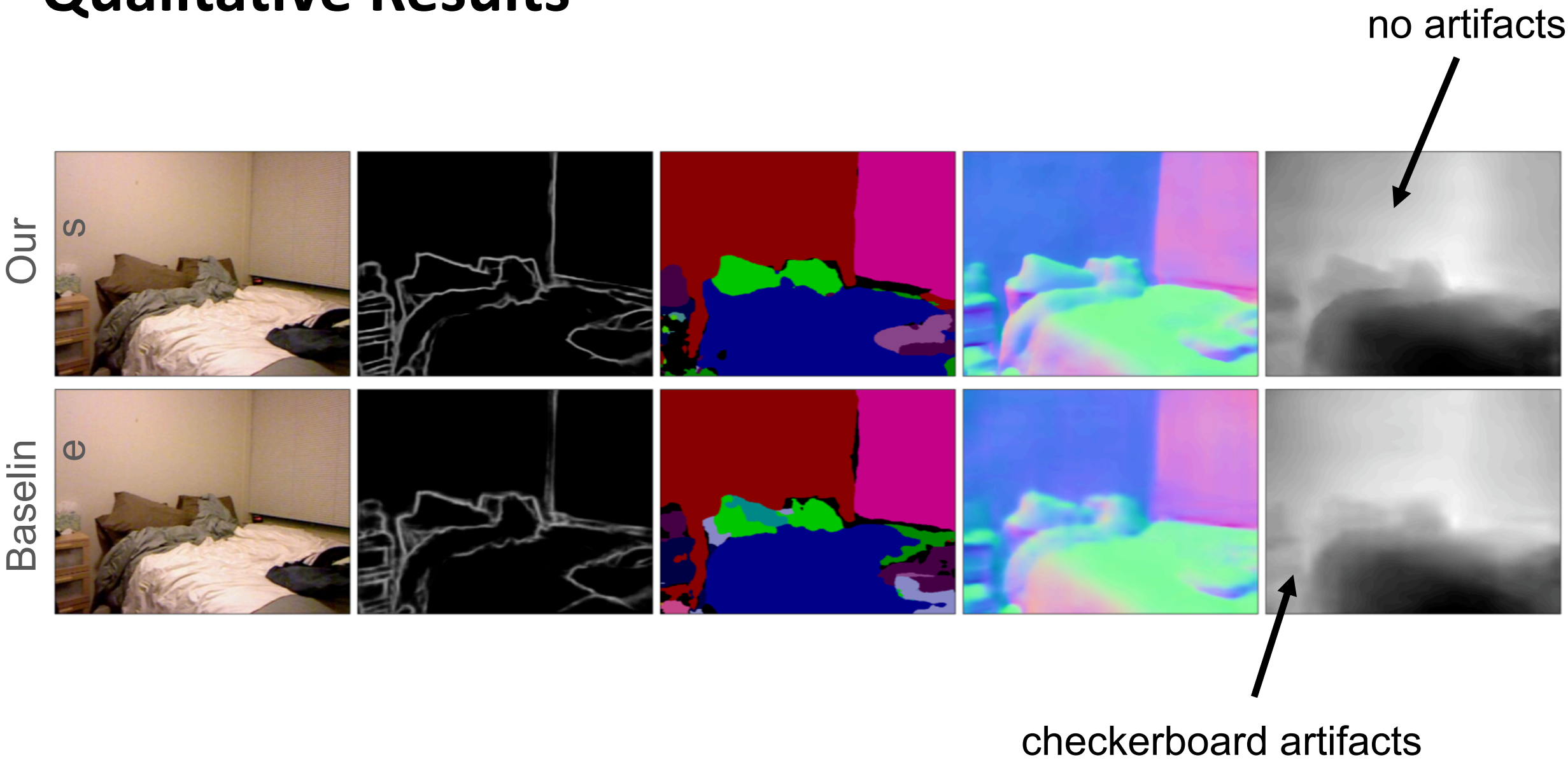
Qualitative Results



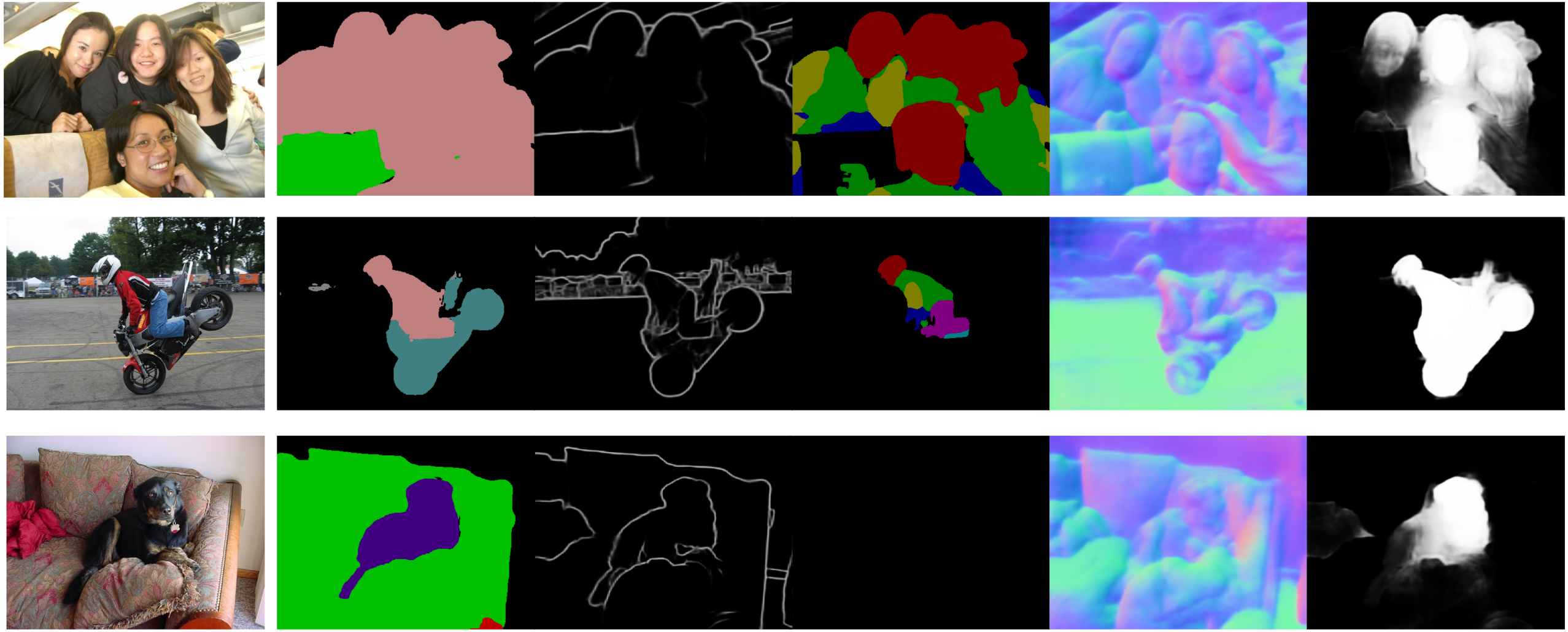
Qualitative Results



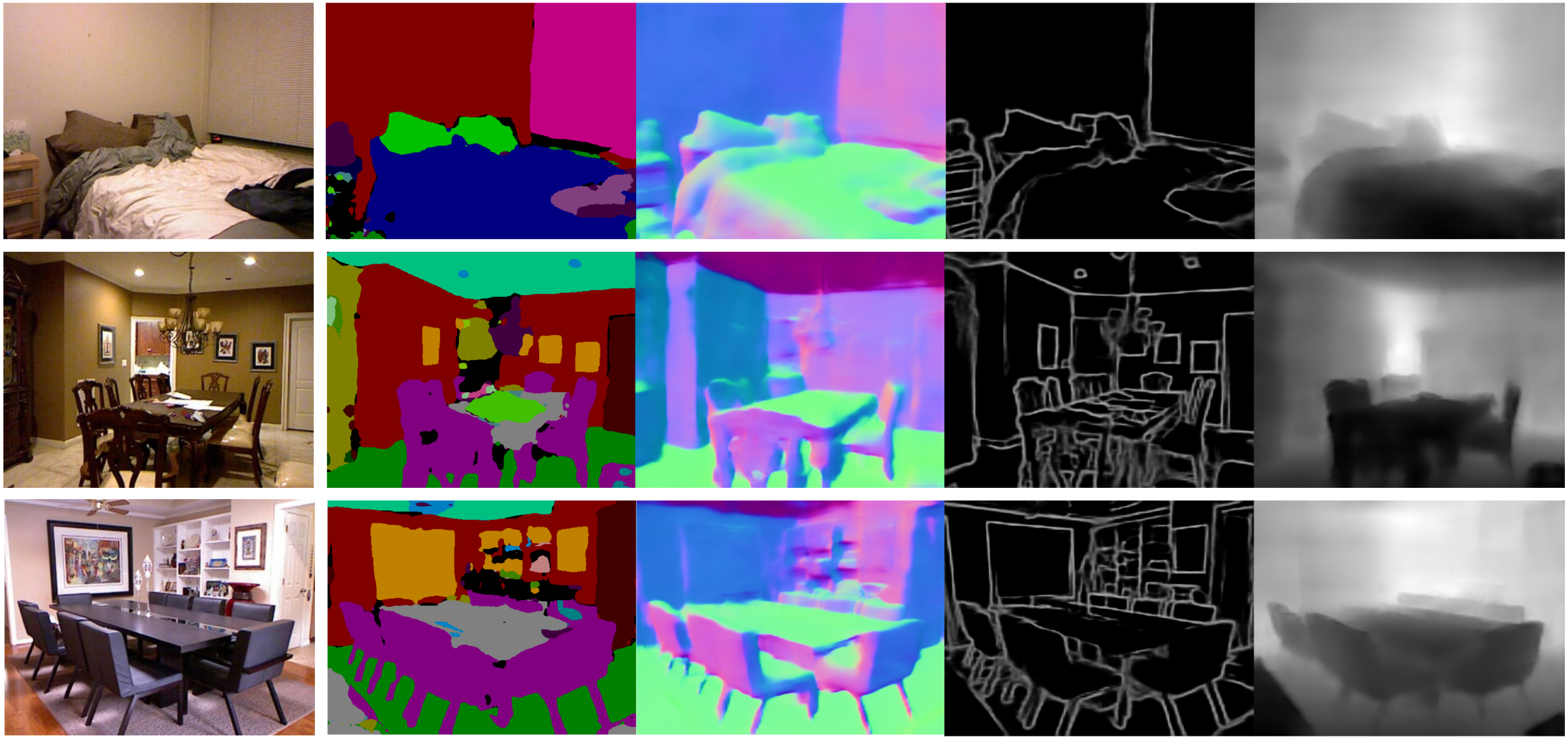
Qualitative Results



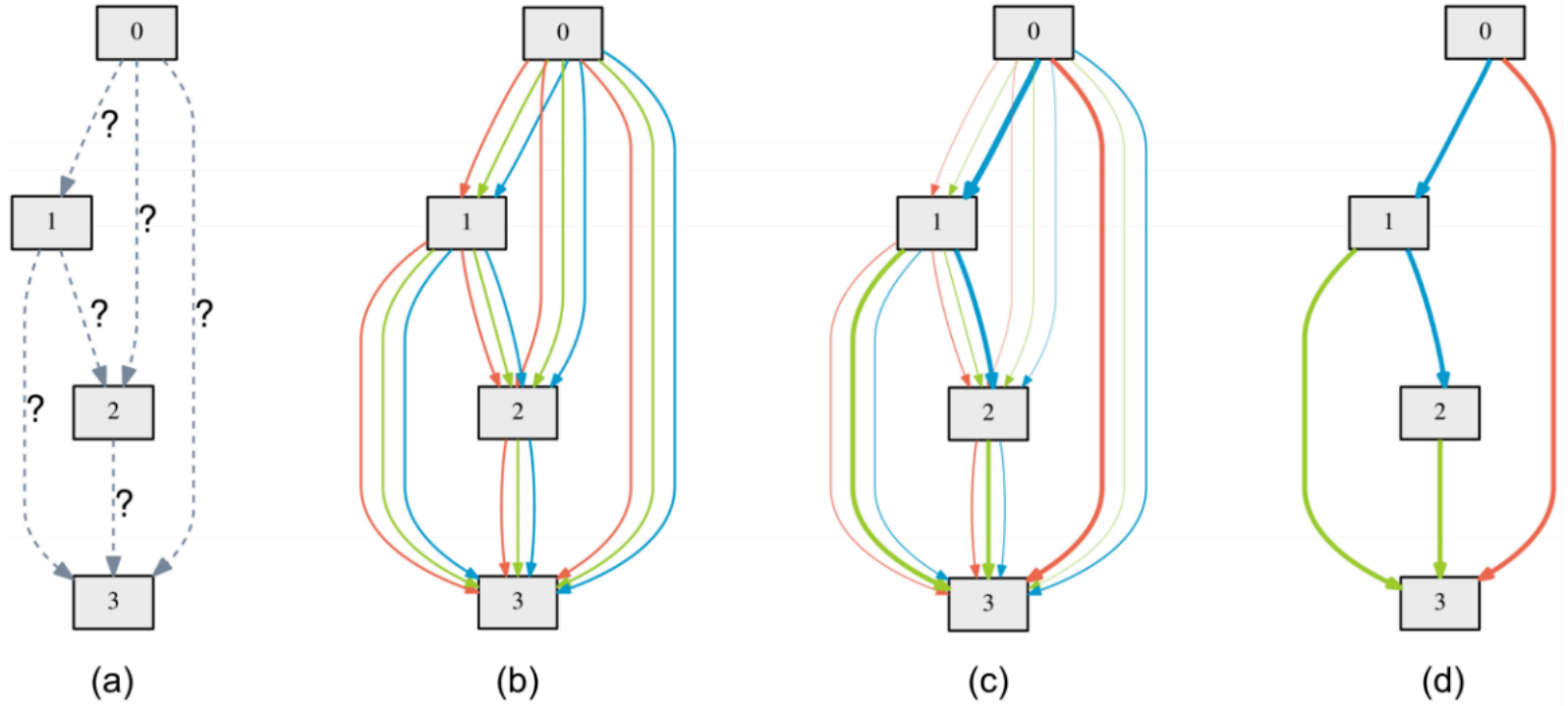
More qualitative Results



More qualitative Results



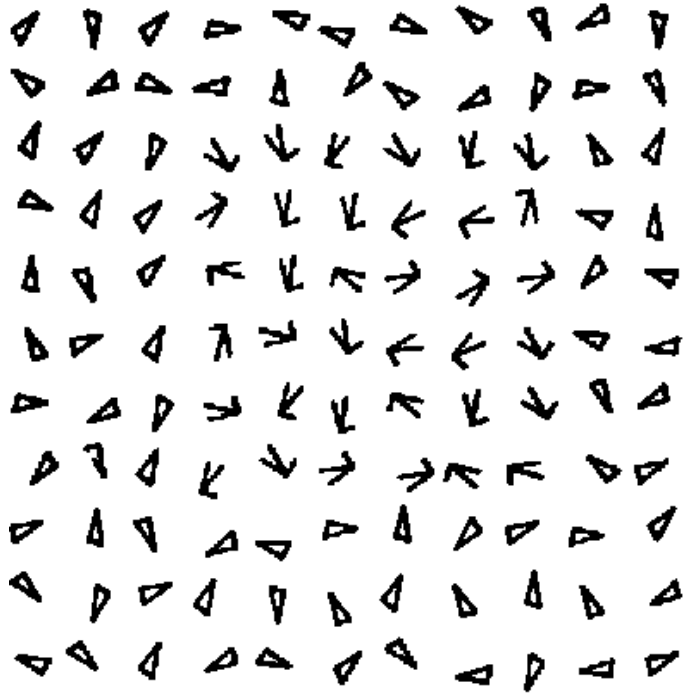
Big picture: continuous optimization vs search

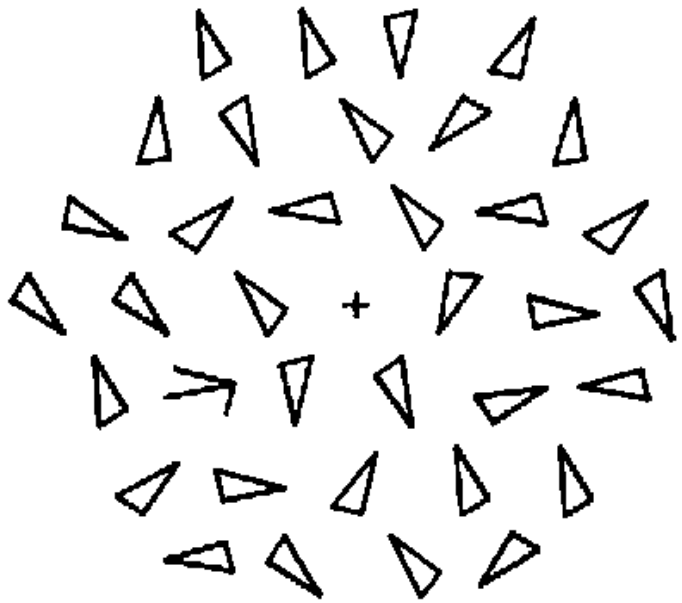


DARTS: Differentiable Architecture Search, H. Liu, K. Simonyan, Y. Yang

Pre-attentive vs. attentive vision

(a)

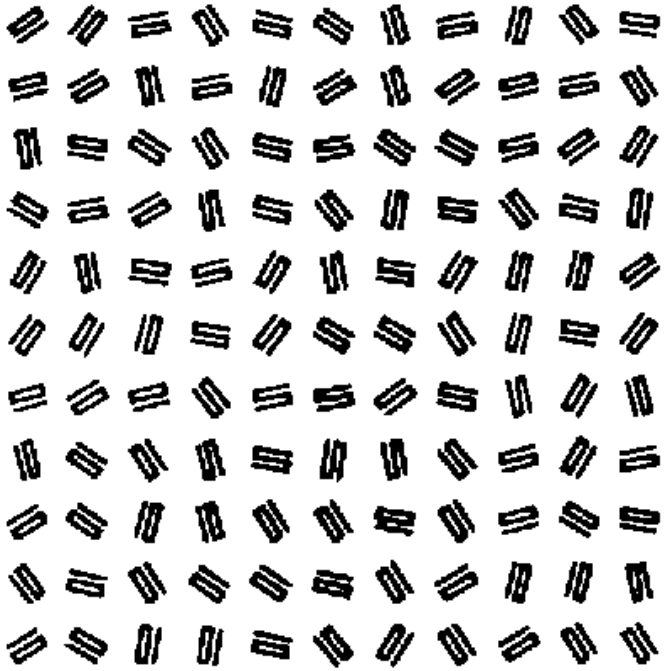


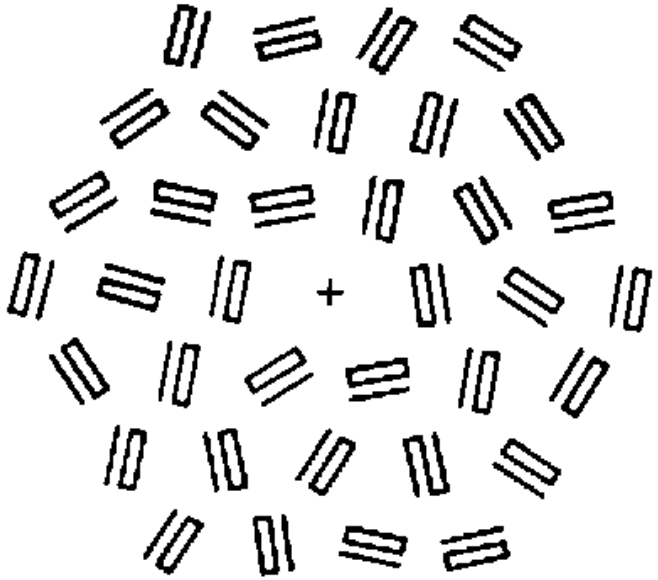


Human factors and behavioral science: Textons, the fundamental elements in preattentive vision and perception of textures, Bela Julesz, James R. Bergen, 1983

Pre-attentive vs. attentive vision

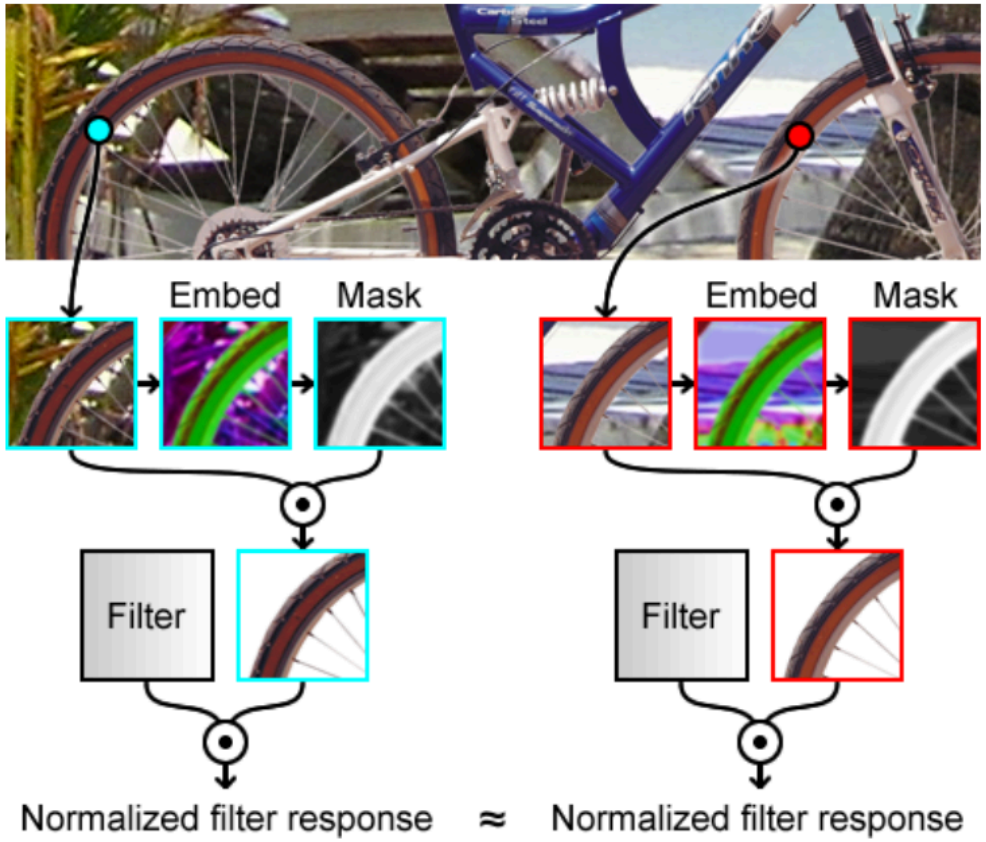
(a)

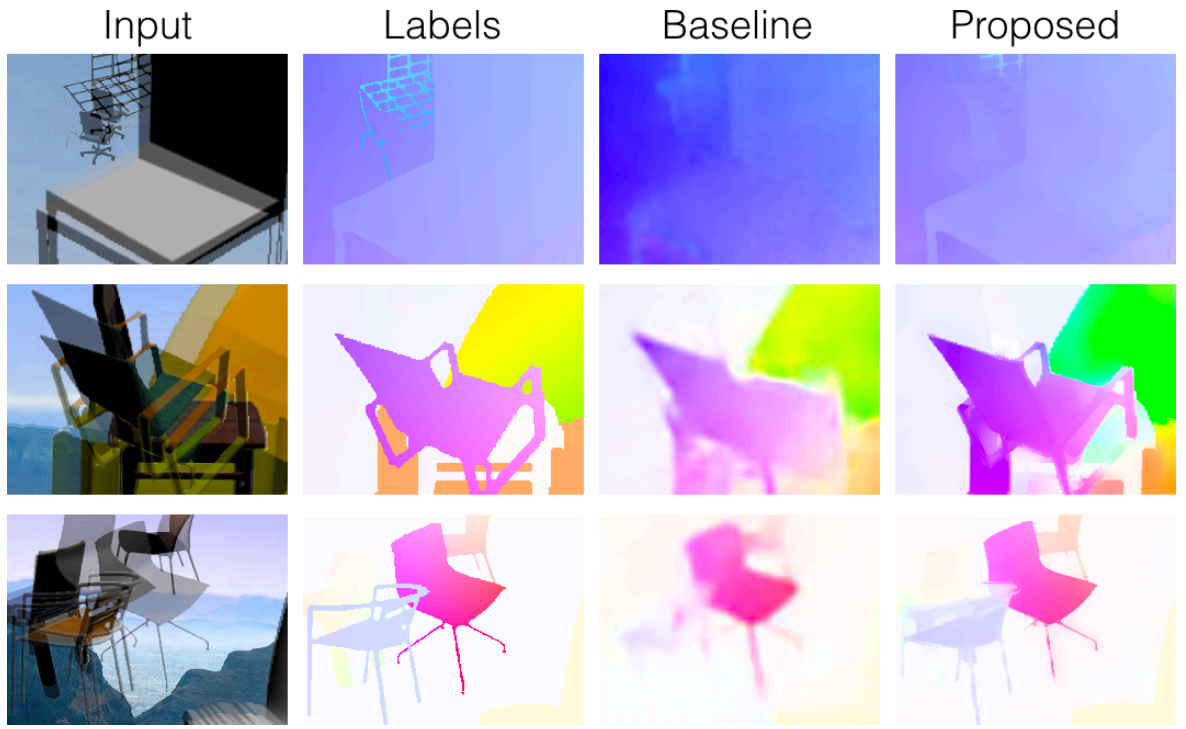




Human factors and behavioral science: Textons, the fundamental elements in preattentive vision and perception of textures, Bela Julesz, James R. Bergen, 1983

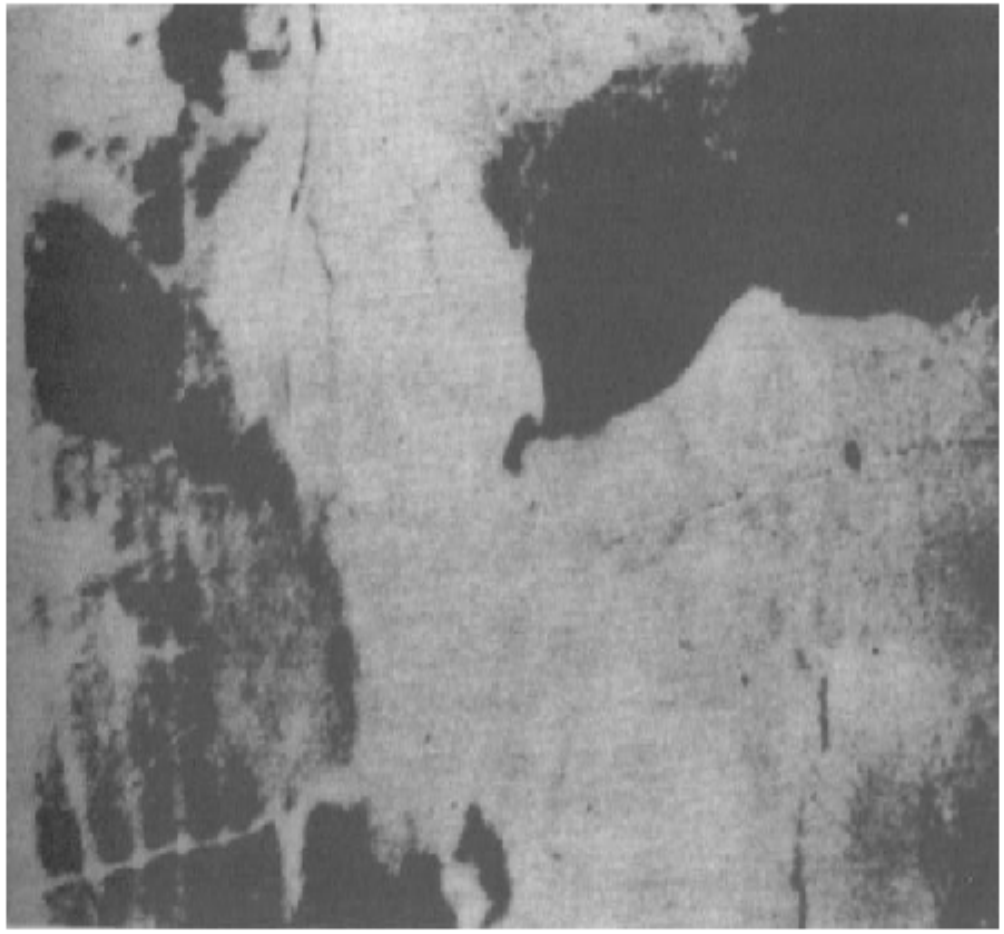
Local attention: Harley et al, ICCV 2017



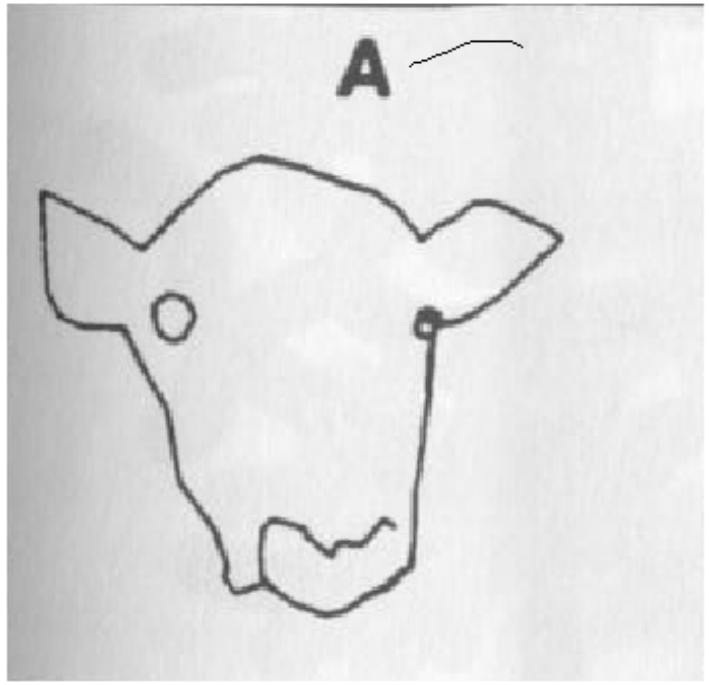


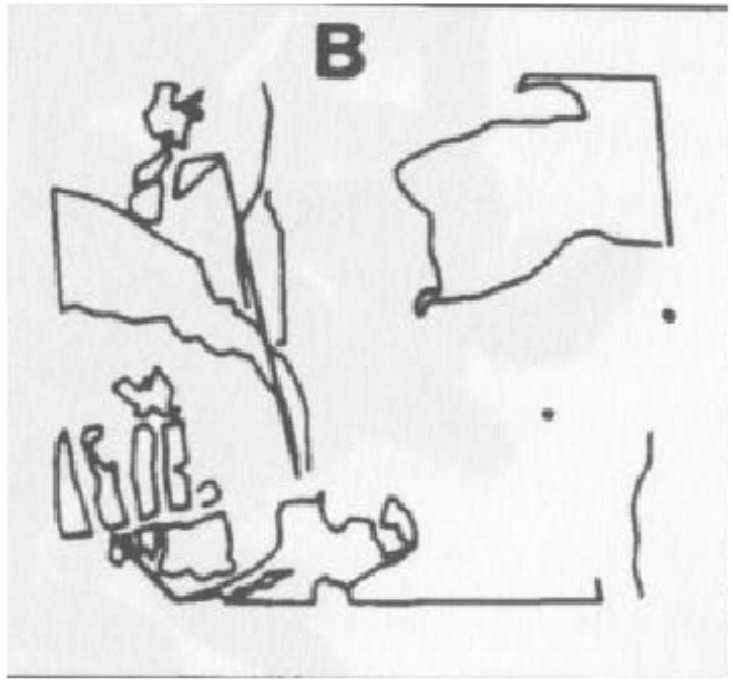
Segmentation-Aware Networks using Local Attention Masks,
A. Harley, K. Derpanis, I. Kokkinos, ICCV 2017

Object-level priming

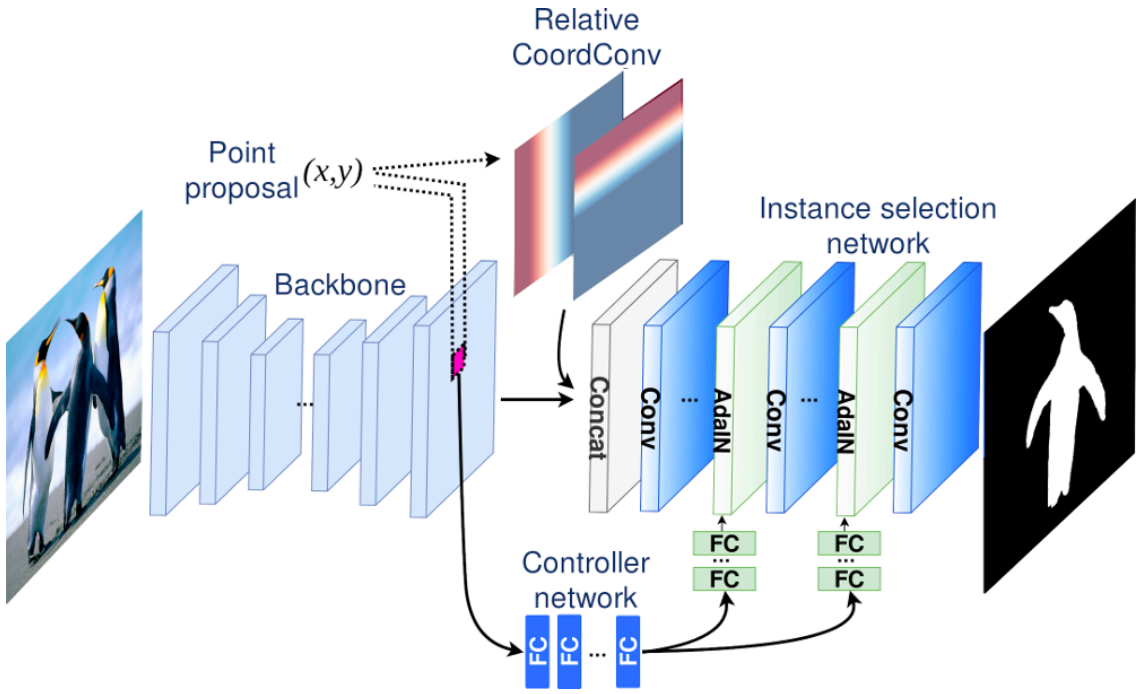


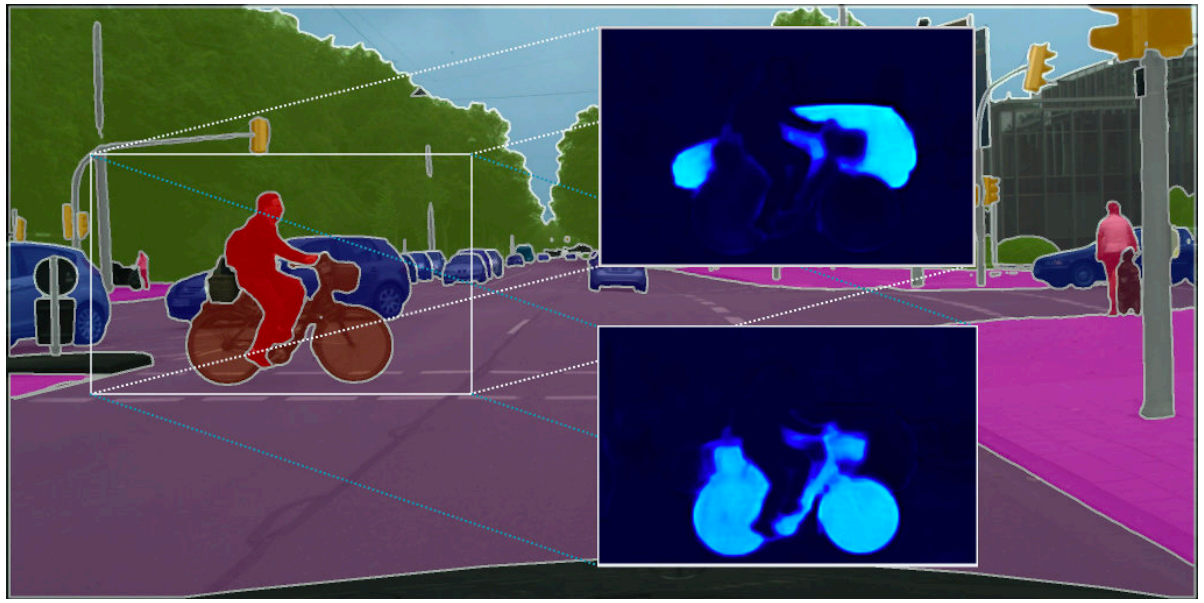
a.k.a. top-down image segmentation





Object & position-level priming





AdaptIS: Adaptive Instance Selection Network, Konstantin Sofiiuk, Olga Barinova, Anton Konushin, ICCV 2019
Priming Neural Networks Amir Rosenfeld , Mahdi Biparva , and John K.Tsotsos, CVPR 2018

Task-level priming: count the balls!

Attentive Single-Tasking of Multiple Tasks

- Approach
  - Network performs one task at a time
  - Accentuate relevant features
  - Suppress irrelevant features

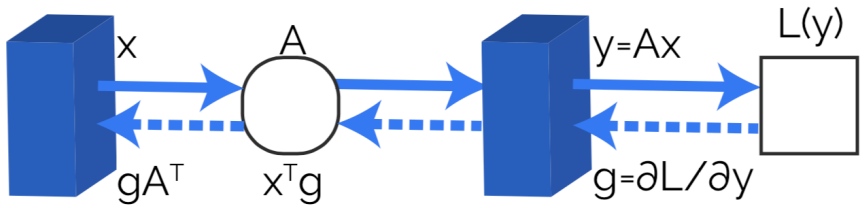
http://www.vision.ee.ethz.ch/~kmaninis/astmt/

Kevis Maninis, Ilija Radosavovic, I.K. “Attentive single Tasking of Multiple Tasks”, CVPR 2019

Thank you for your **attention**.

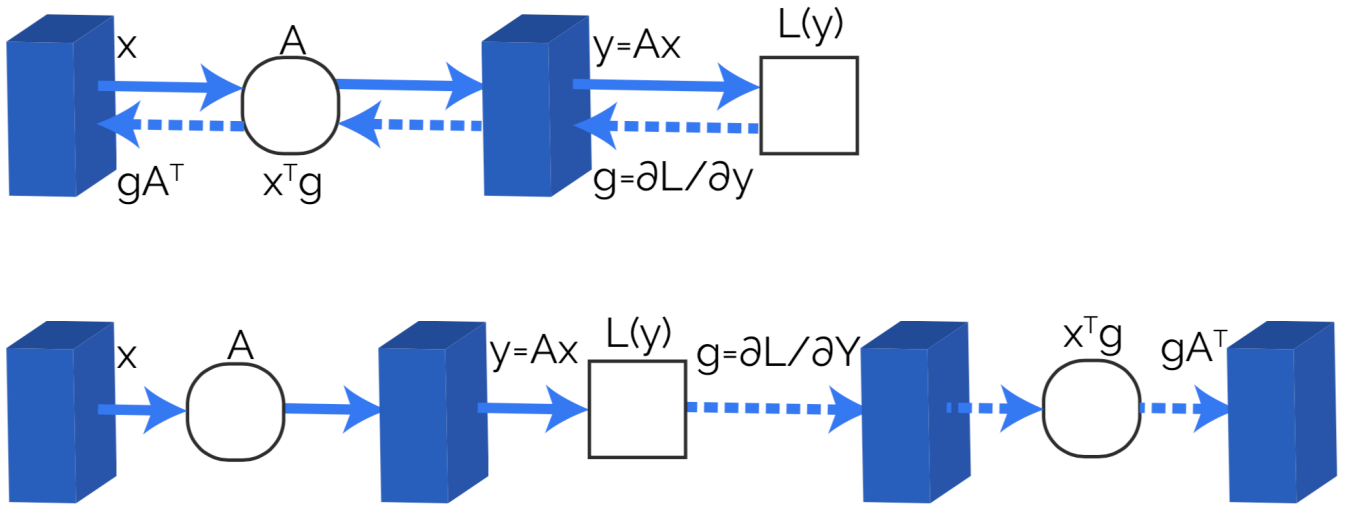
http://www.vision.ee.ethz.ch/~kmaninis/astmt/

Double back-propagation



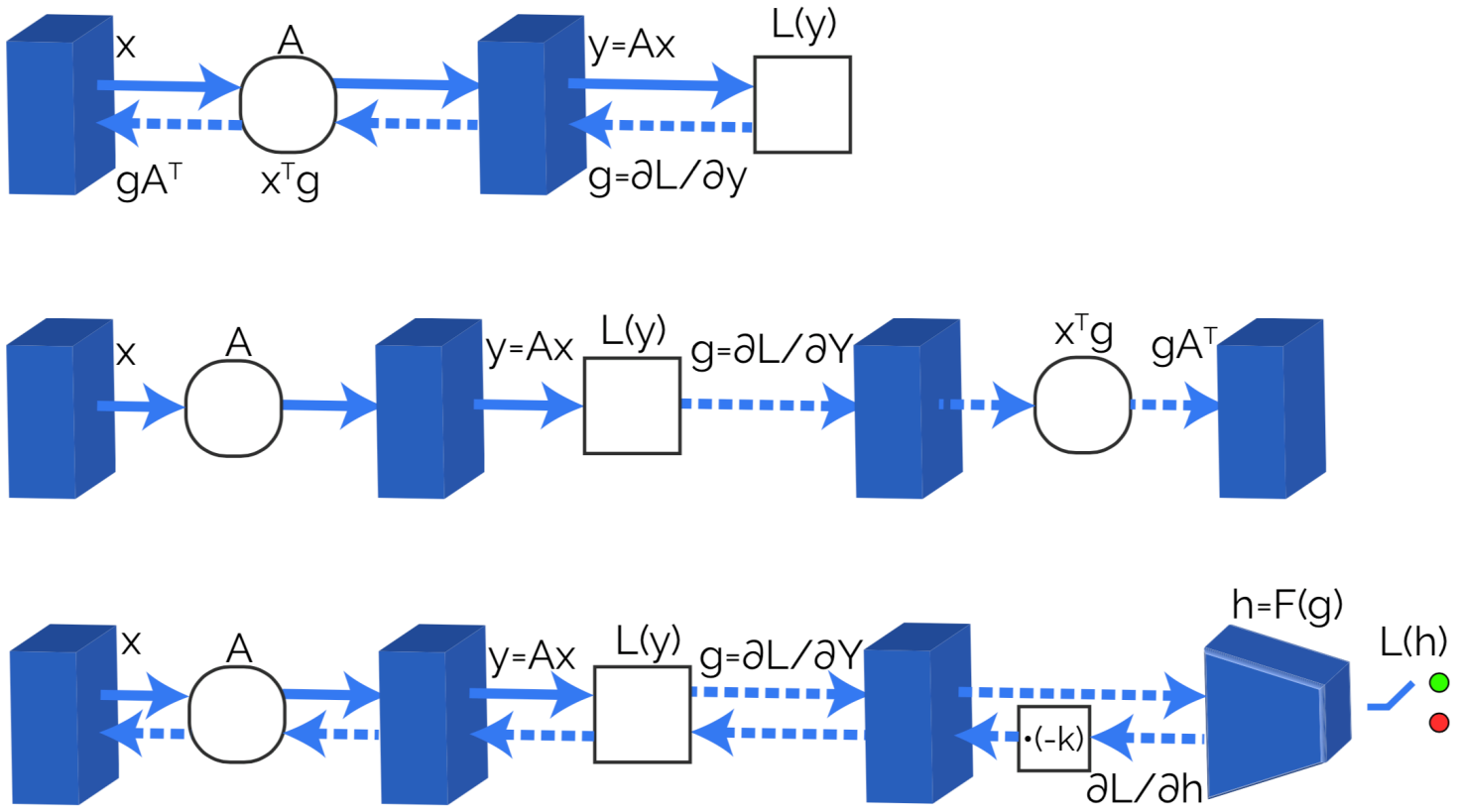
Harris Drucker, Yann LeCun, “Double Backpropagation Increasing Generalization Performance”, IJCNN 1991

Double back-propagation



Harris Drucker, Yann LeCun, “Double Backpropagation Increasing Generalization Performance”, IJCNN 1991

Adversarial Training using Double Back-Propagation



Deeplab v3+: Sanity Check

Benchmark our re-implementation on popular benchmarks for different (single) tasks:

low-, mid-, and high-level tasks

Task	Dataset	Metric	R-101	strong baseline
Edge	BSDS500	odsF ↑	82.5	81.3 [28]
S.Seg	VOC	mIoU ↑	78.9	79.4 [6]
H. Parts	P. Context	mIoU ↑	64.3	64.9* [5]
Normals	NYUD	mErr ↓	20.1	19.0 [3]
Saliency	PASCAL-S	maxF ↑	84.0	83.5 [29]
Depth	NYUD	RMSE ↓	0.56	0.58 [61]

* COCO pre-training

Chen et al., "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation", in ECCV 2018

Ablation on PASCAL: Modulation

SE	RA	#T	Edge ↑	Seg ↑	Parts ↑	Norm ↓	Sal ↑	drop ↓
		1	71.3	64.9	57.1	14.9	64.2	
		5	69.2	60.2	54.1	17.0	62.1	6.6
	✓	5	70.5	62.8	56.4	15.3	64.8	**1.4**
✓		5	71.1	64.0	56.8	15.1	64.4	**0.6**

Type of modulation

enc	dec	#T	Edge ↑	Seg ↑	Parts ↑	Norm ↓	Sal ↑	drop ↓
		1	71.3	64.9	57.1	14.9	64.2	
		5	69.2	60.2	54.1	17.0	62.1	6.6
	✓	5	70.6	63.3	56.7	15.1	63.2	**1.4**
✓	✓	5	71.1	64.0	56.8	15.1	64.4	**0.6**

Location of modulation

Attention-to-task almost reaches single-tasking performance

Ablation on PASCAL: Adversarial training

mod	A	#T	Edge ↑	Seg ↑	Parts ↑	Norm ↓	Sal ↑	drop ↓
		1	71.3	64.9	57.1	14.9	64.2	
		5	69.2	60.2	54.1	17.0	62.1	6.6
	✓	5	69.7	62.2	55.0	16.2	62.2	**4.4**
✓		5	71.1	64.0	56.8	15.1	64.4	0.6
✓	✓	5	71.0	64.6	57.3	15.0	64.7	**0.1**

Adversarial training helps! Gains smaller but free of additional computation

Experiments on NYUD and FSV

SEA	#T	Edge ↑	Seg ↑	Norm ↓	Depth ↓	drop ↓
	1	74.4	32.8	23.3	0.6	
	4	73.2	30.9	23.3	0.7	5.4
✓	4	74.5	32.2	23.2	0.6	**-1.2**

(a) Results on **NYUD-v2**.

SEA	#T	Seg ↑	Albedo ↓	Disp ↓	drop ↓
	1	71.2	0.086	0.063	
	3	66.9	0.093	0.078	7.04
✓	3	70.7	0.085	0.063	**-0.02**

(b) Results on **FSV**.

Results equal or better to the single-tasking baselines

Ablation: Different backbones

backbone	SEA	#T	Edge ↑	Seg ↑	Parts ↑	Norm ↓	Sal ↑	drop ↓
R-26		1	71.3	64.9	57.1	14.9	64.2	
R-26		5	69.2	60.2	54.1	17.0	62.1	6.6
R-26	✓	5	71.0	64.6	57.3	15.0	64.7	**0.1**
R-50		1	72.7	68.3	60.7	14.6	65.4	
R-50		5	69.2	63.2	55.1	16.0	63.6	6.8
R-50	✓	5	72.4	68.0	61.1	14.8	65.7	**0.04**
R-101		1	73.5	69.8	63.5	14.2	67.4	
R-101		5	70.5	66.4	61.5	15.4	66.4	4.5
R-101	✓	5	73.5	68.5	63.4	14.4	67.7	**0.6**

Results consistent across backbones