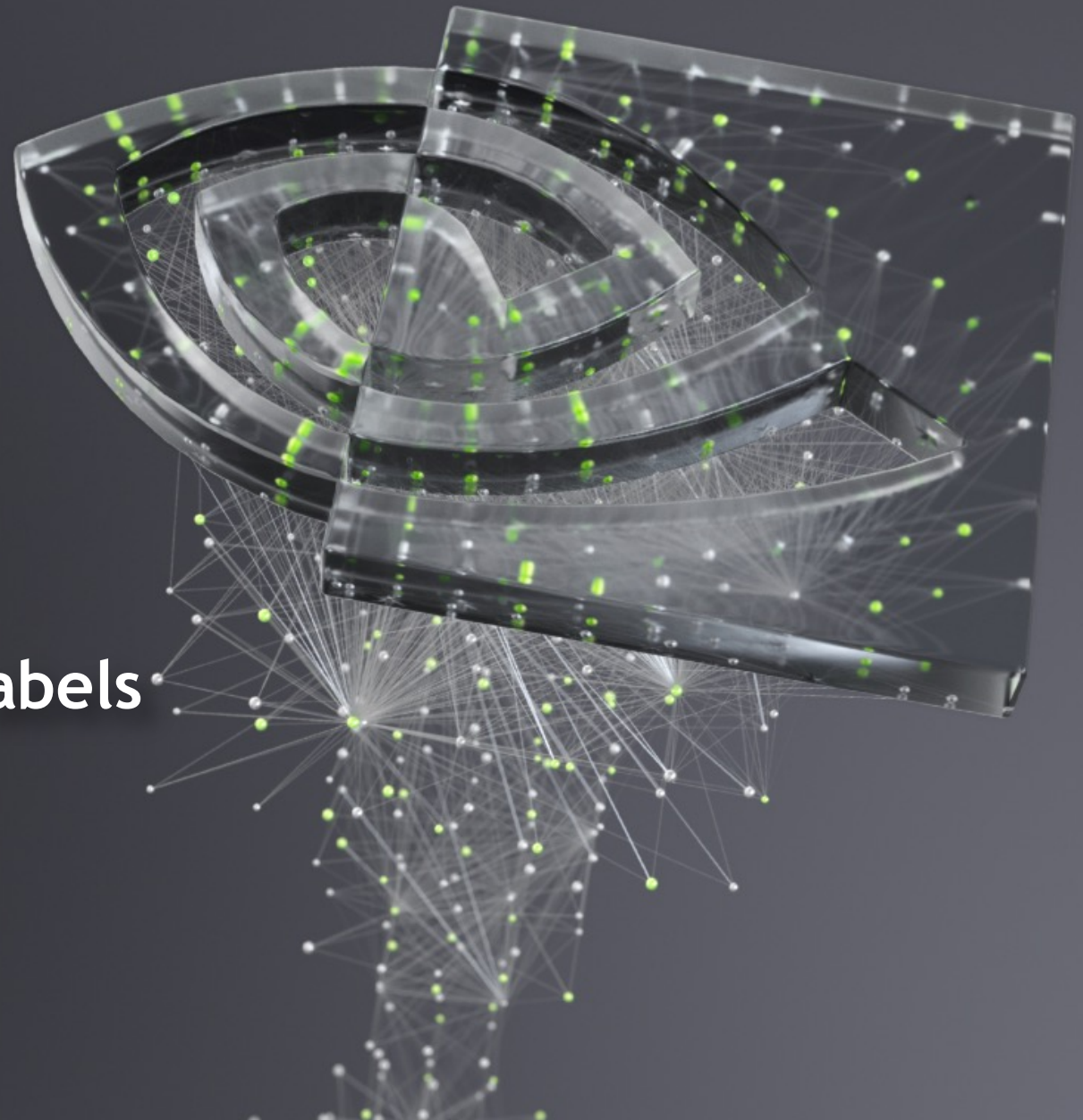
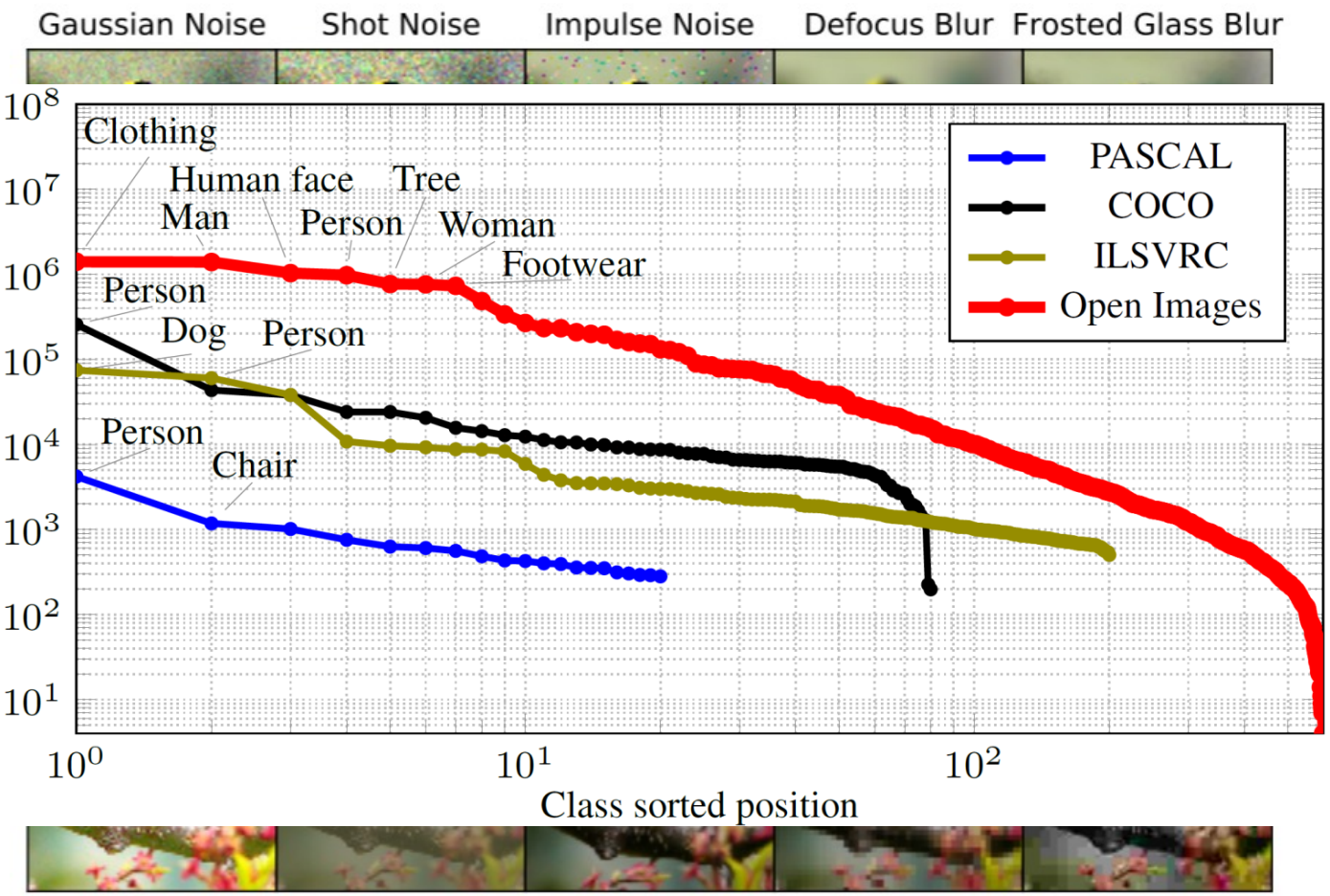


Learning with Imperfect Labels and Visual Data

Anima Anandkumar

Challenge I - Real World Data Are Imperfect

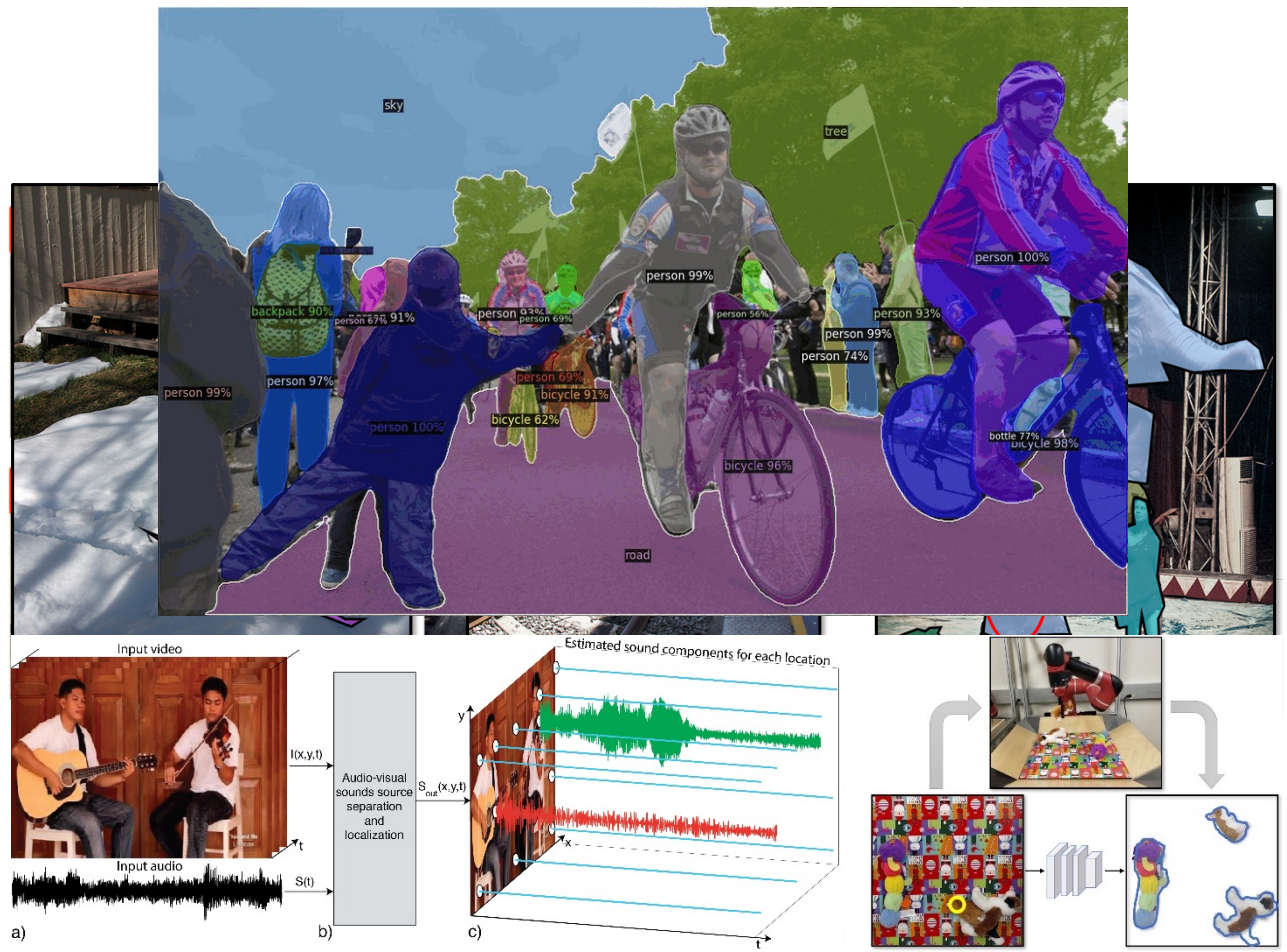
- **Domain gap**
- **Data bias**
- **Data noise**
- Can be *Long tail*
- Can contain *Occlusions*
- Can be *Cluttered*
- Can be *Ambiguous*
- Can be *Multisensory*
- •••



Challenge II – Real World Labels Are Imperfect

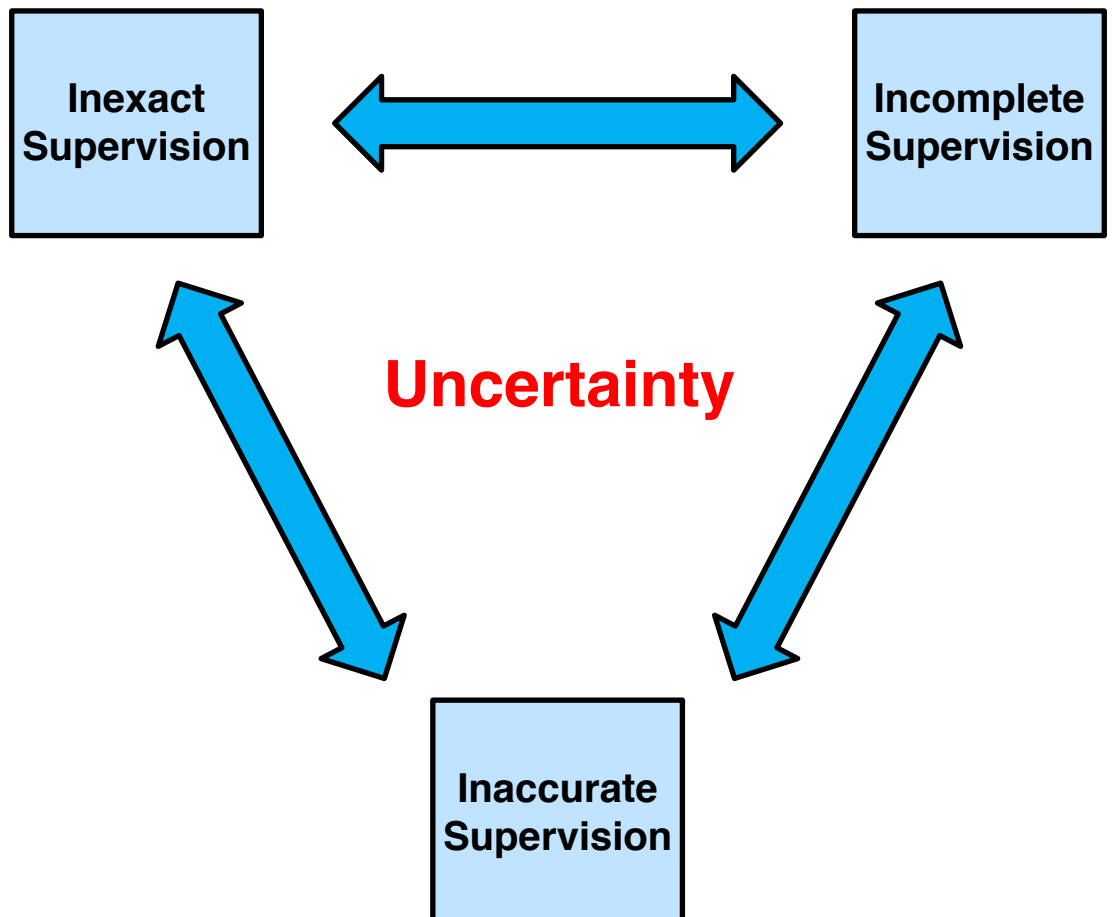
- **Inexact** (indirect) supervision
- **Incomplete** (limited) supervision
- **Inaccurate** (noisy) supervision
- Can be *Versatile*
- Can be *Multimodal*
- Can be *Sequential*
- Can be *Sparse*
- Can be *Interactive*

•••

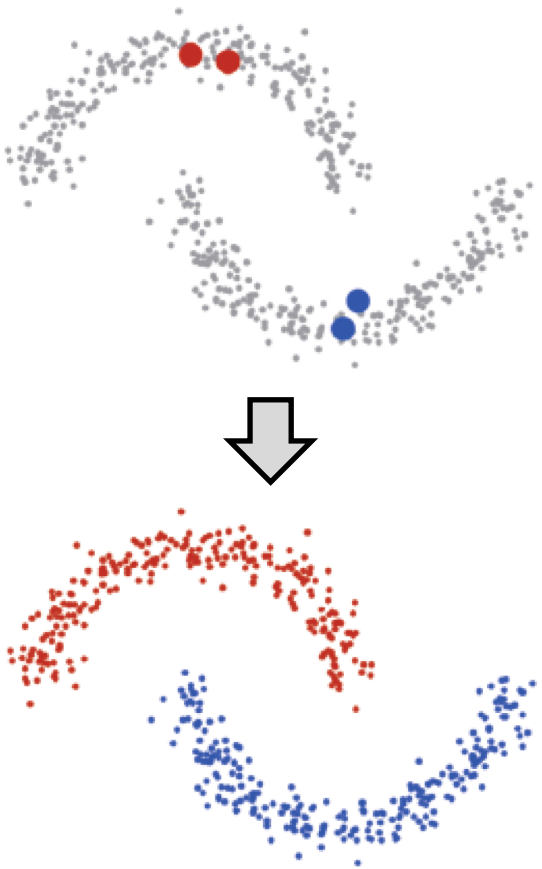


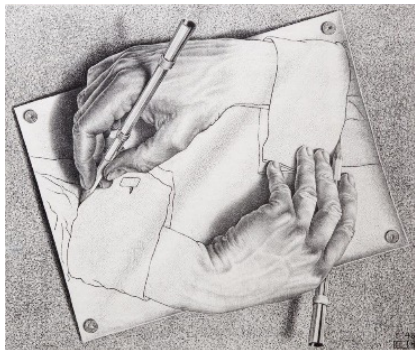
Goal and Challenge

- The ability to improve generalization by learning continuously on new data
- The ability to leverage diverse forms of weak supervision, or self-improving on unlabeled data
- The ability to overcome the **uncertainty** (ill-posed nature) due to the lack of constraints under partial supervision

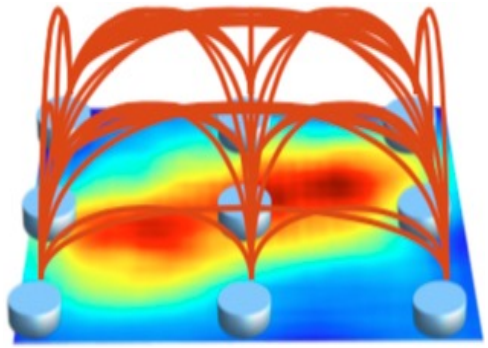


Overcoming Uncertainties from Imperfect Labels

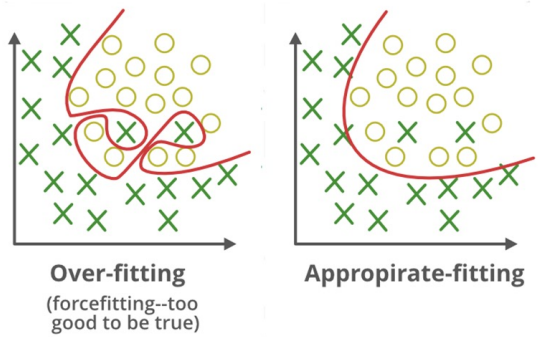




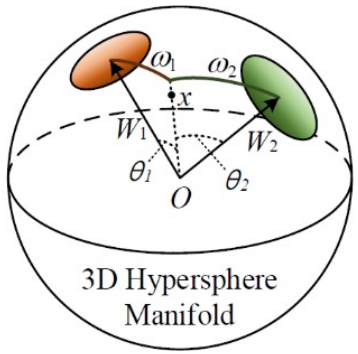
Self-supervision



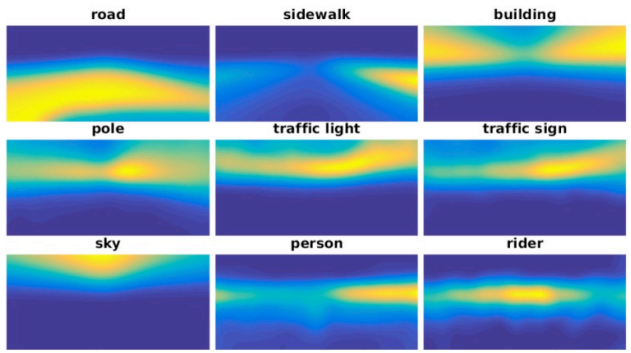
Structuredness



Regularization



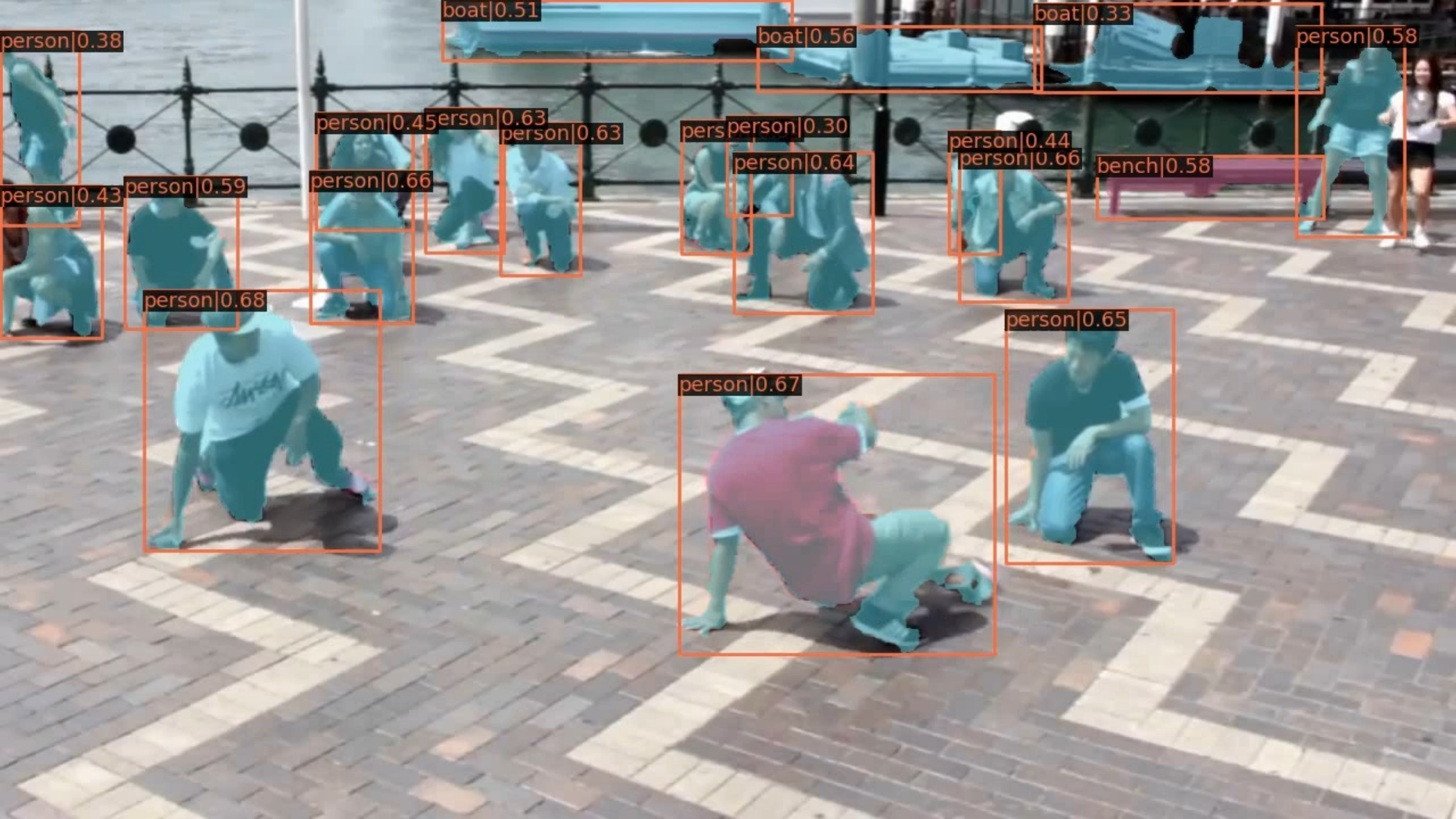
Inductive Bias



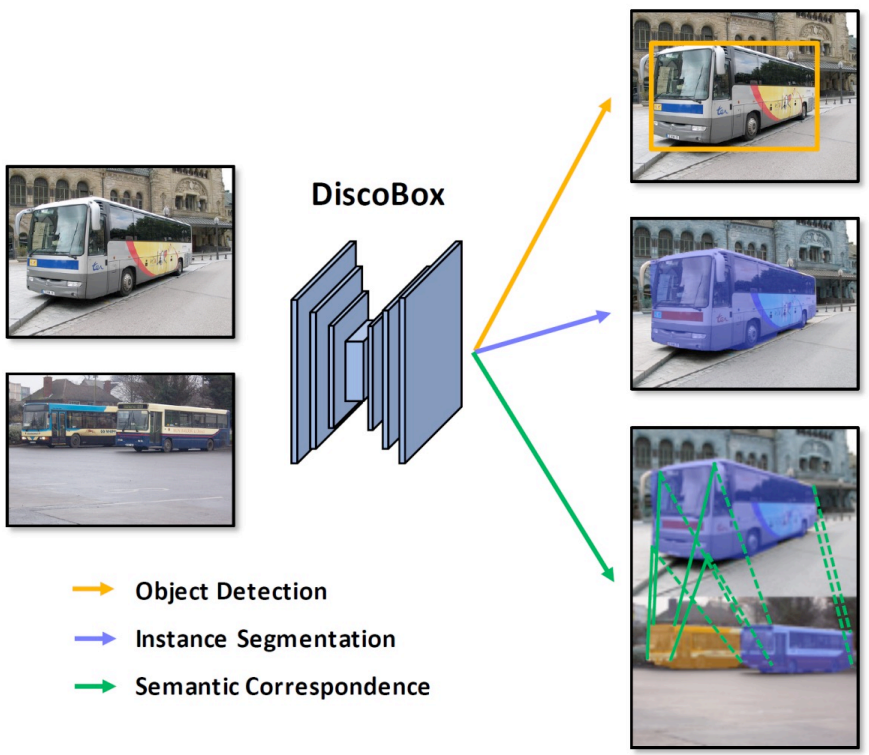
Domain Prior

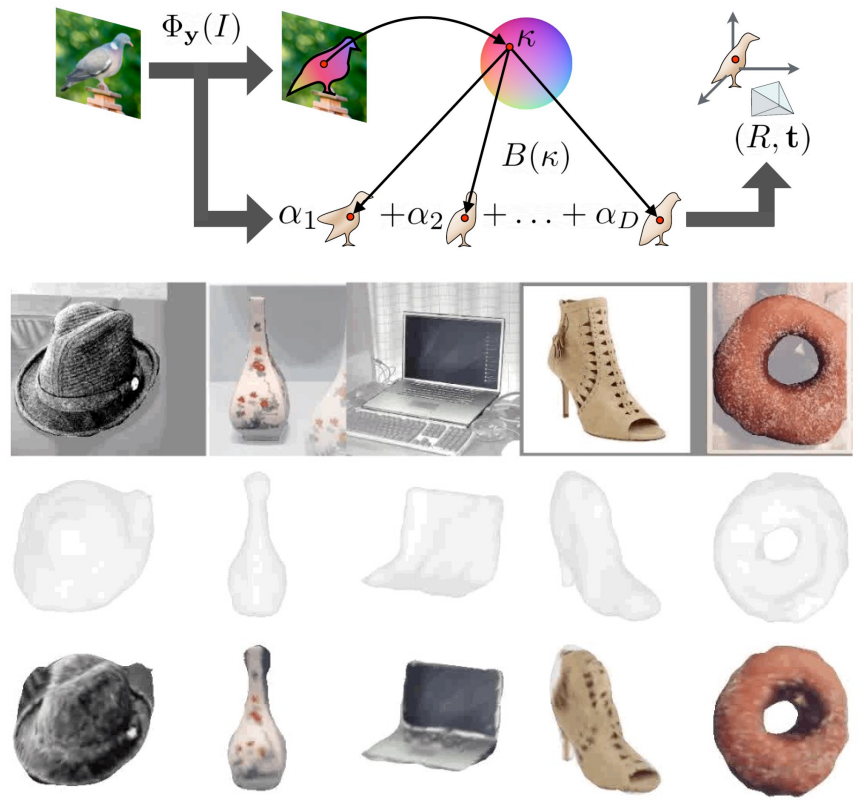
Synthetic Data

DiscoBox


boat|0.51
boat|0.56
boat|0.33
person|0.58
person|0.38
person|0.45
erson|0.63
person|0.63
pers
person|0.30
person|0.64
person|0.44
person|0.66
bench|0.58
person|0.43
person|0.59
person|0.66
person|0.68
person|0.67
person|0.65

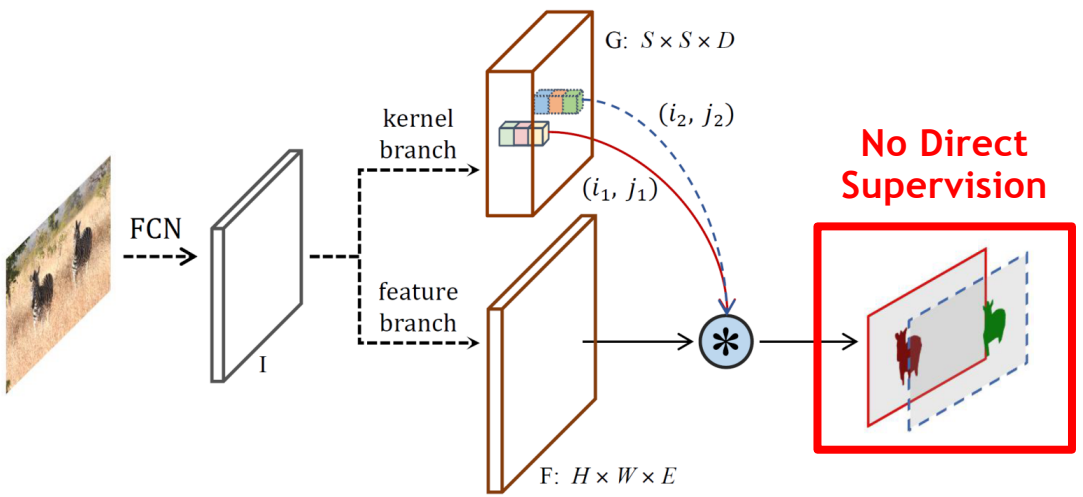
Method Overview



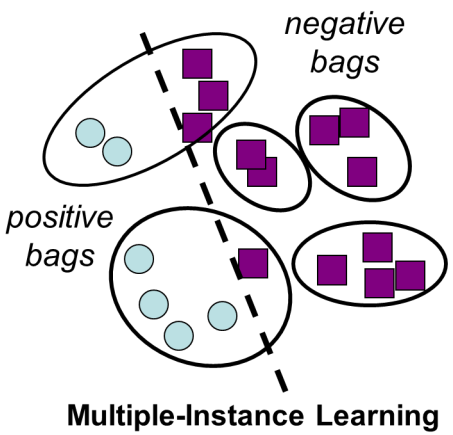


Task Network Design

SOLOv2

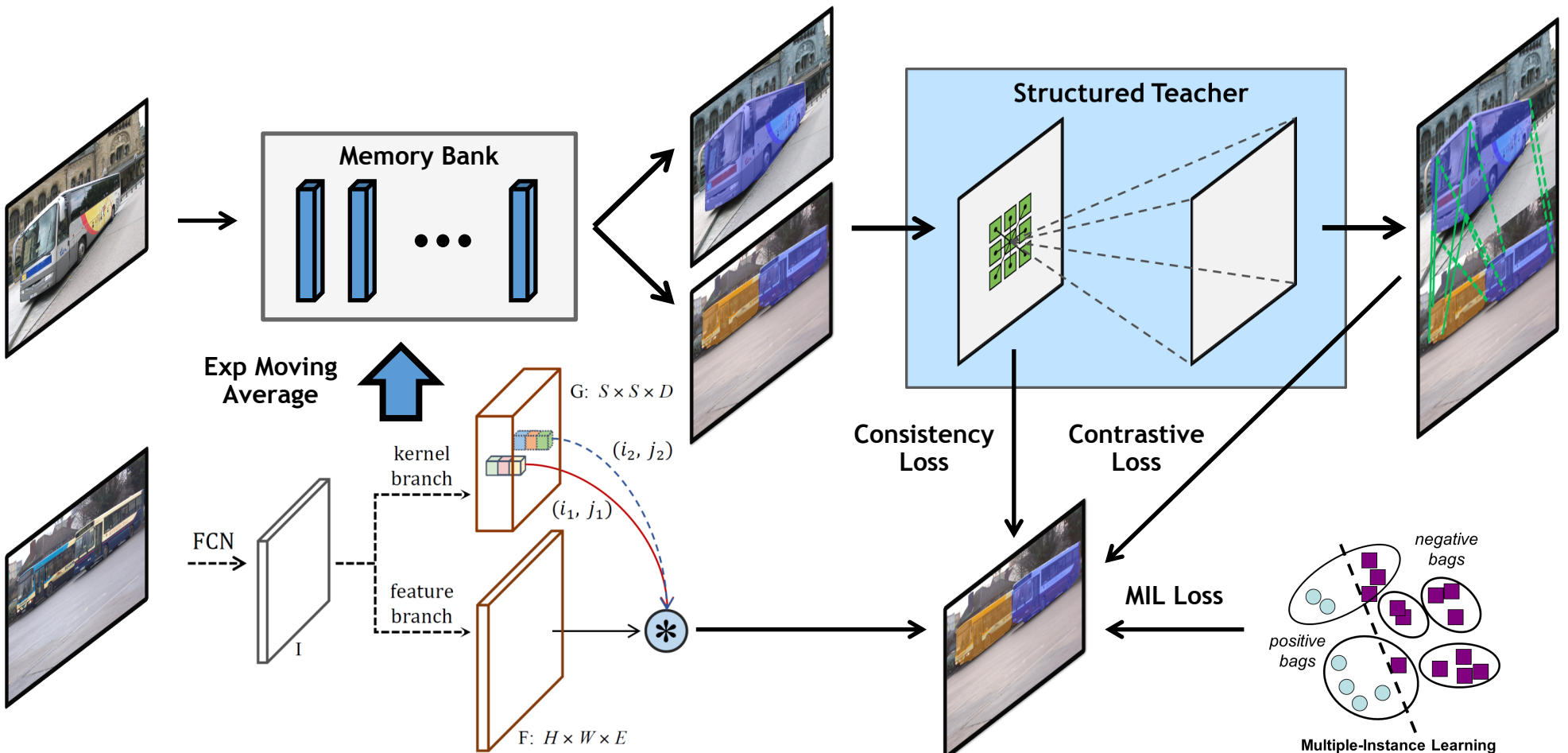


BBTP



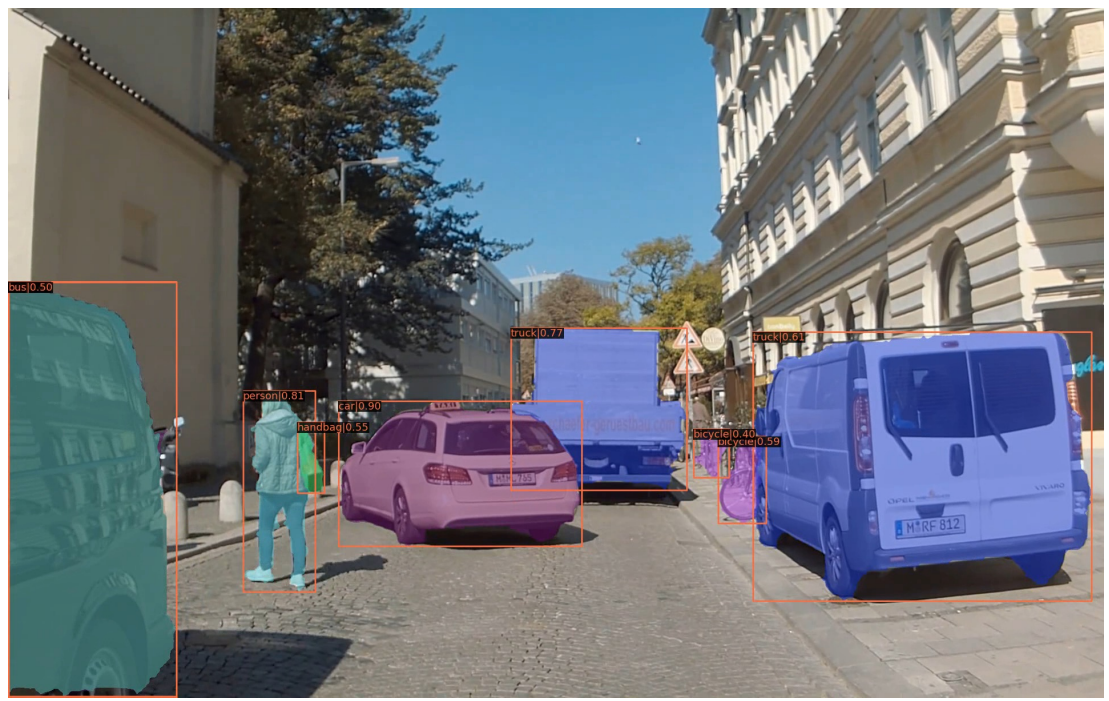
Wang et al., **SOLOv2: Dynamic and Fast Instance Segmentation**, NeurIPS20
Hsu et al., **Weakly supervised instance segmentation using the bounding box tightness prior**, NeurIPS19

Self-Ensembling with Structured Teacher



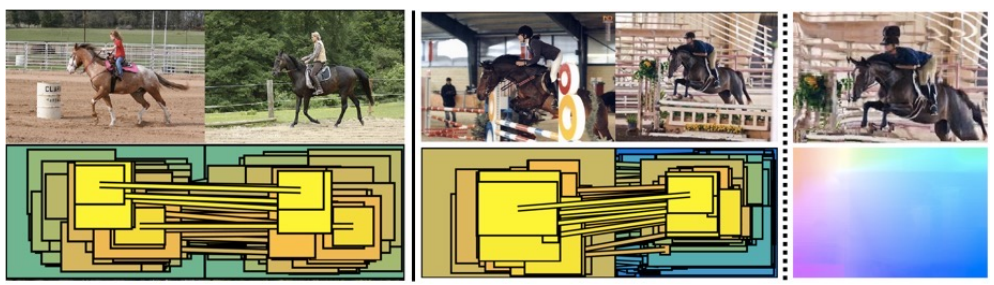
Instance Segmentation

NVIDIA DriveAV Data

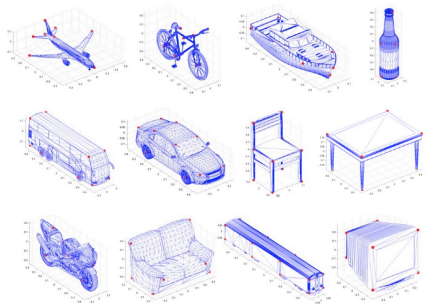


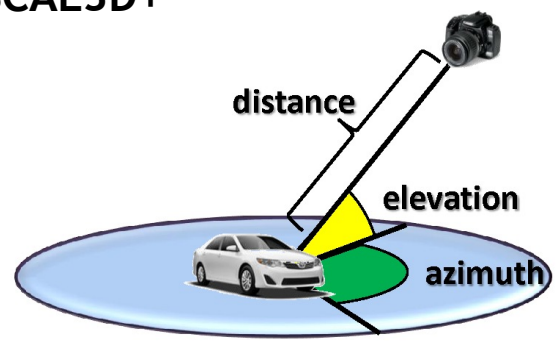
Semantic Correspondence

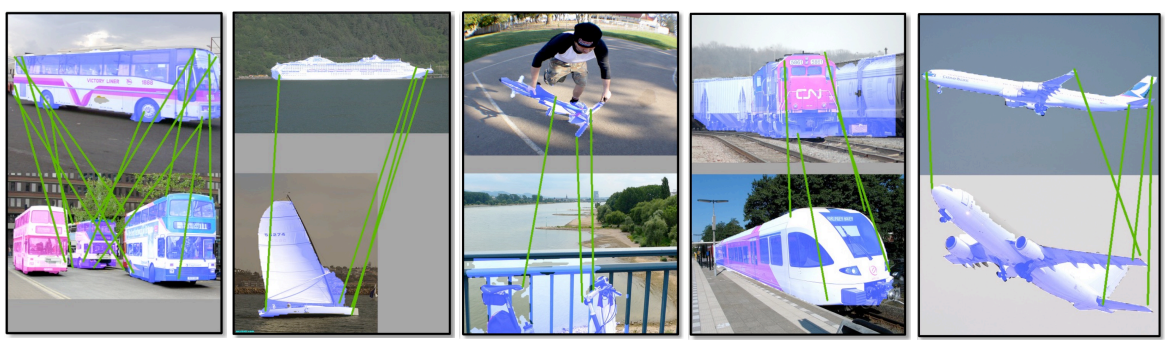
PF-PASCAL



PASCAL3D+



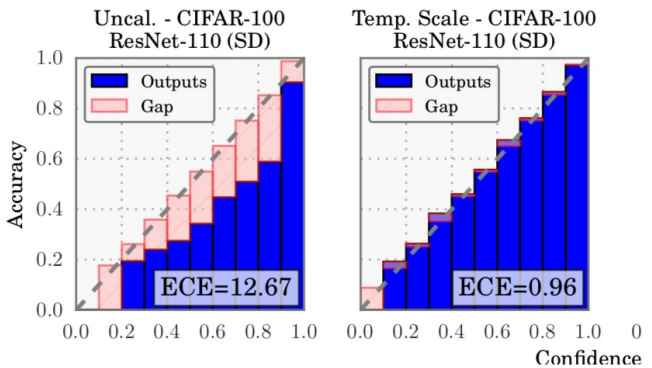




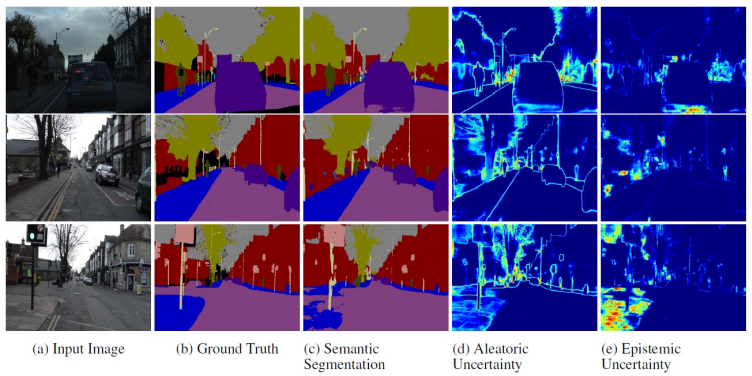
Ham et al., **Proposal flow,** CVPR16
Xiang et al., **Beyond PASCAL: A benchmark for 3d object detection in the wild,** WACV14

Syn-to-real Adaptation

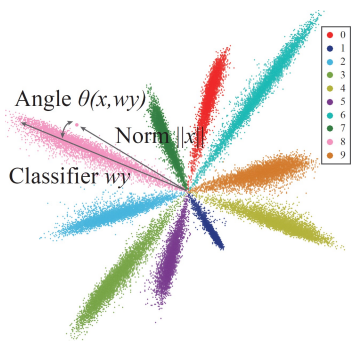
Reliable Uncertainty Estimation Under Domain Shift



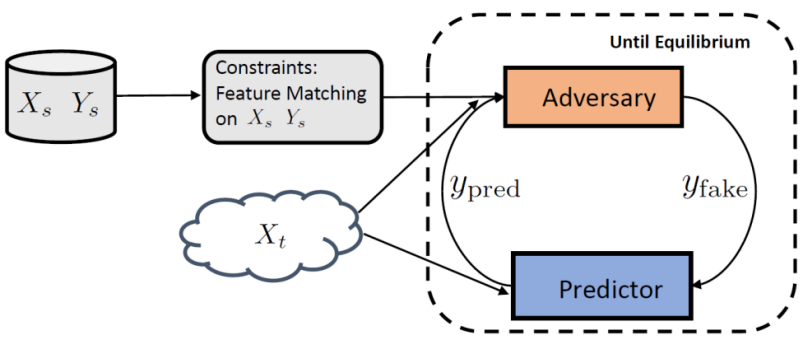
Temperature Scaling



Bayesian Deep Learning



Angular Distance



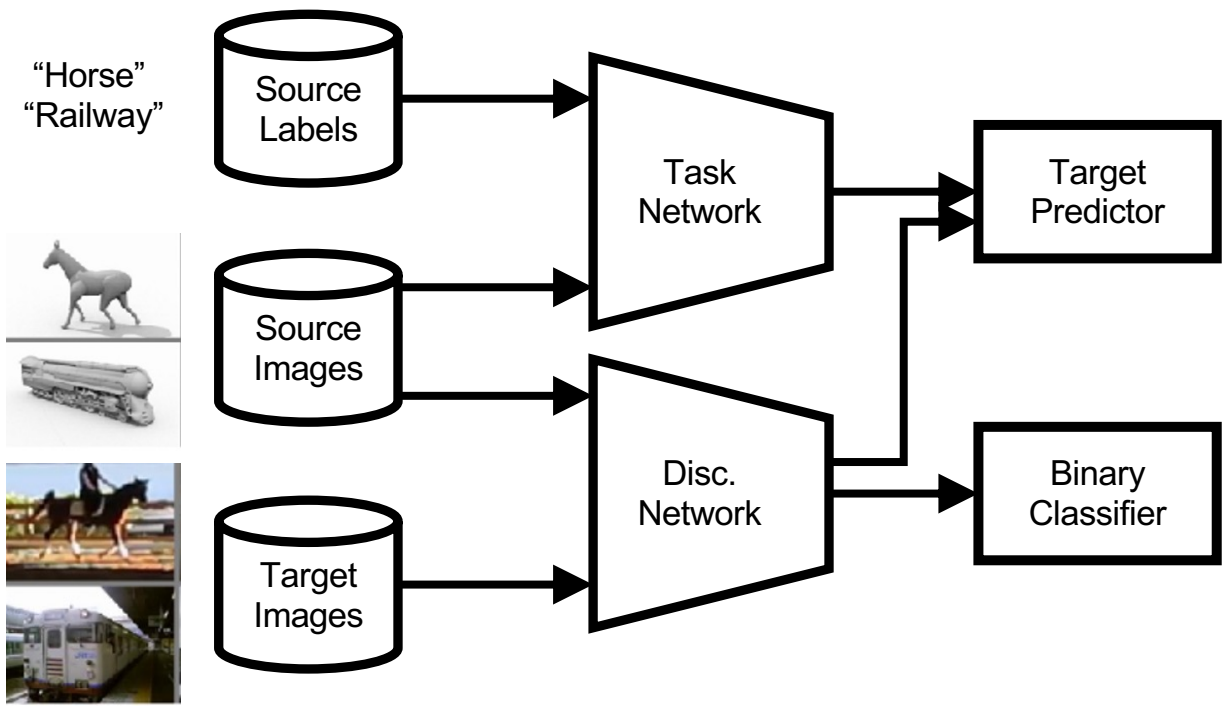
Distributionally Robust Learning

Credit: Li et al., Improving Confidence Estimates for Unfamiliar Examples, CVPR20

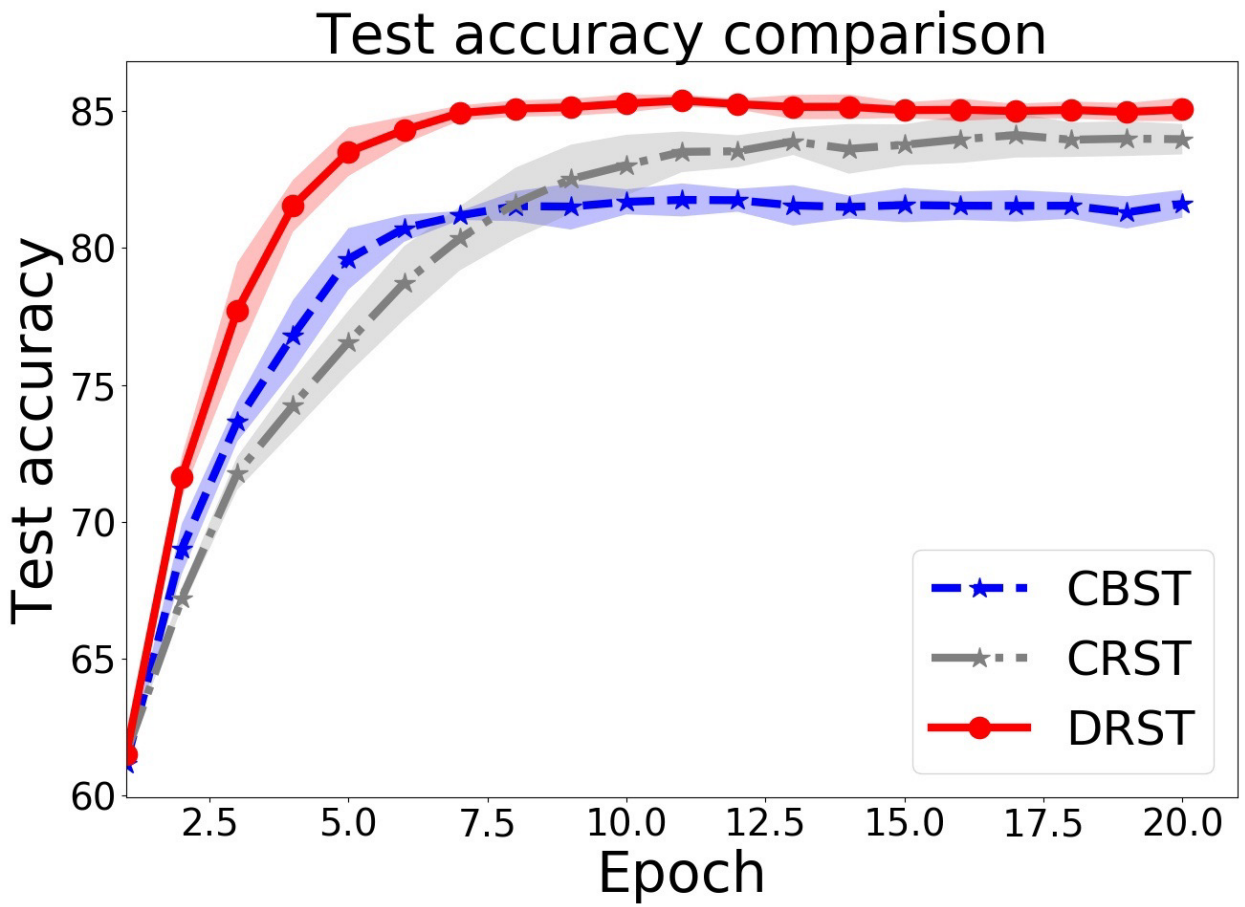
Want et al., **Deep distributionally robust learning for calibrated uncertainties under domain shift**, under submission at NeurIPS21

DENSITY RATIO ESTIMATION

$$\hat{P}(y|x) \propto \exp\left(\frac{P_{\text{src}}(x)}{P_{\text{trg}}(x)}\theta \cdot \phi(x, y)\right)$$



ACCURACY OF DOMAIN ADAPTION (VISDA2017)



CBST: Class-Balanced
CRST: Confidence-Regularized
DRST: Ours

We achieve the SOTA results in self-training domain adaptation.

INTERPRETABLE DENSITY RATIOS

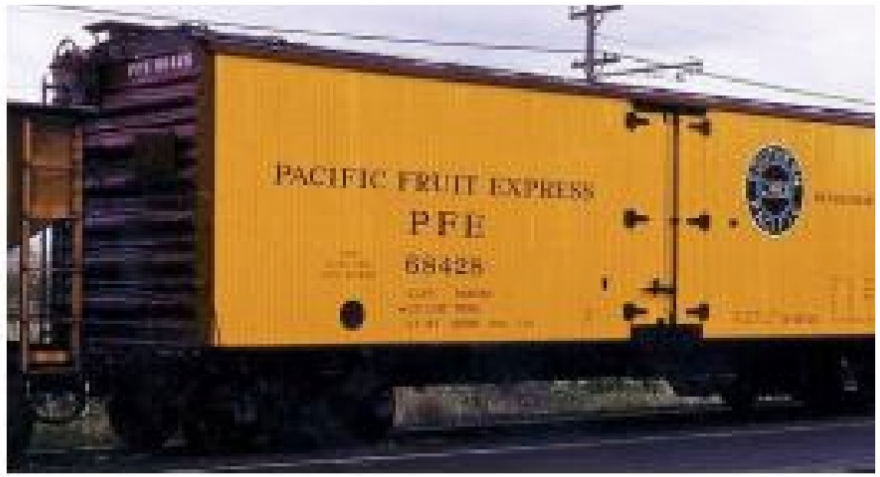
Noisy target data has lower source support

Source

A less typical “railway”

Density Ratio: 1.004

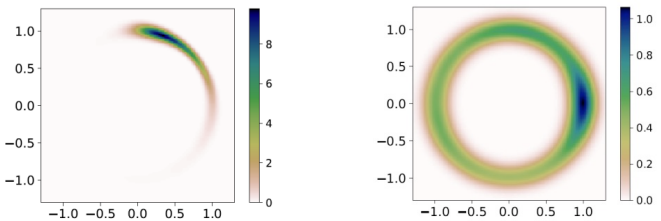
A more typical “railway”

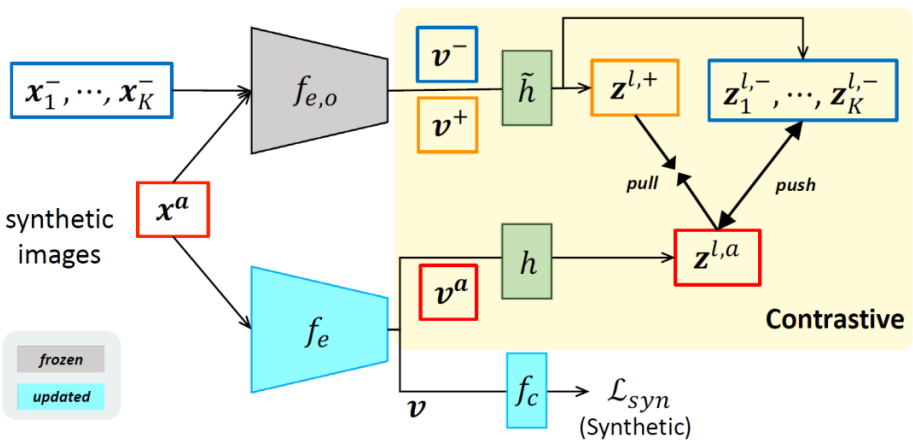


Density Ratio: 2.223

Automated Syn-to-Real Generalization

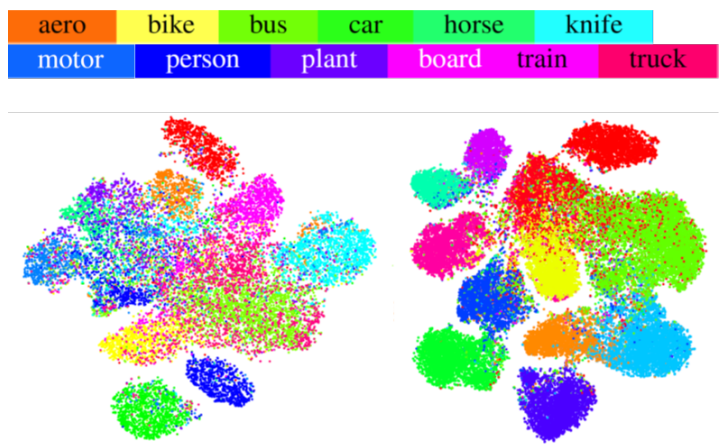
Synthetic training leads to collapsed representations.





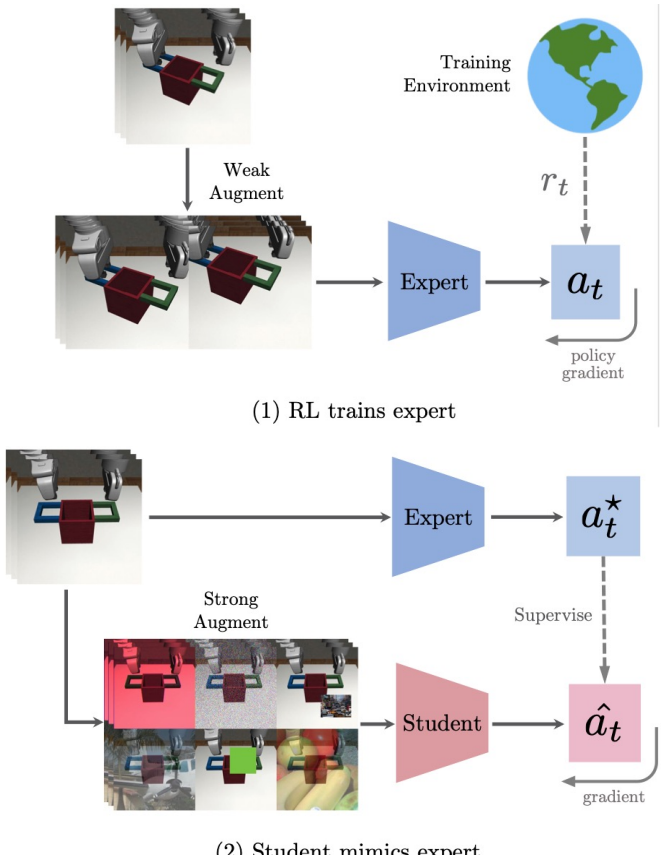
Contrastive knowledge distillation

Improved representation and accuracy

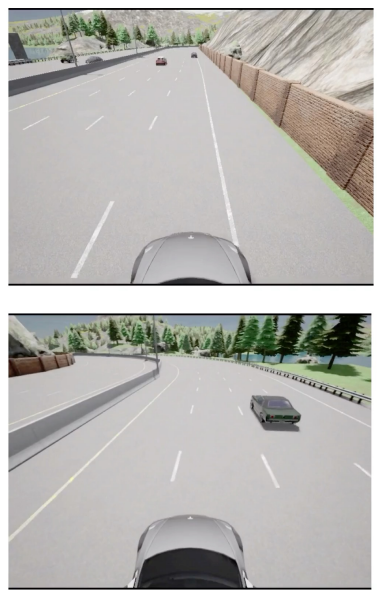


Chen et al., **Automated Synthetic-to-Real Generalization**, ICML20.
Chen et al., **Contrastive Syn-to-Real Generalization**, ICLR21.

Sequential Decision Making



CARLA: Driving in Diverse Weathers



Fan et al., SECANT: Self-Expert Cloning for Zero-Shot Generalization of Visual Policies, ICML 2021

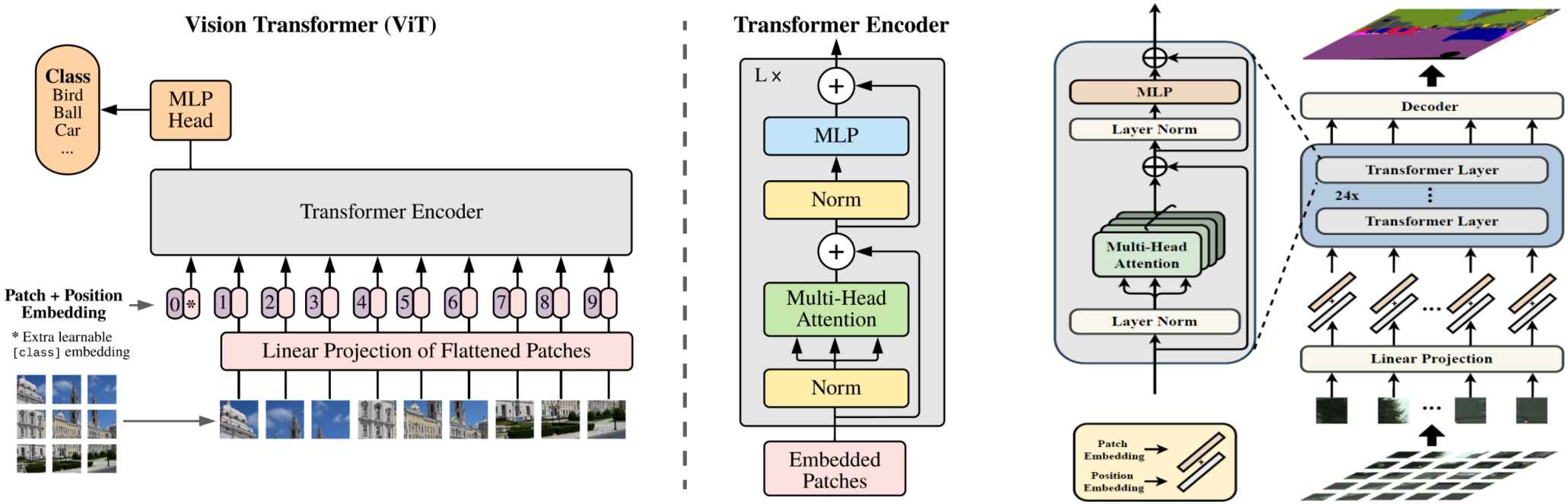
SegFormer: Semantic Segmentation with Vision Transformers

SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, Ping Luo

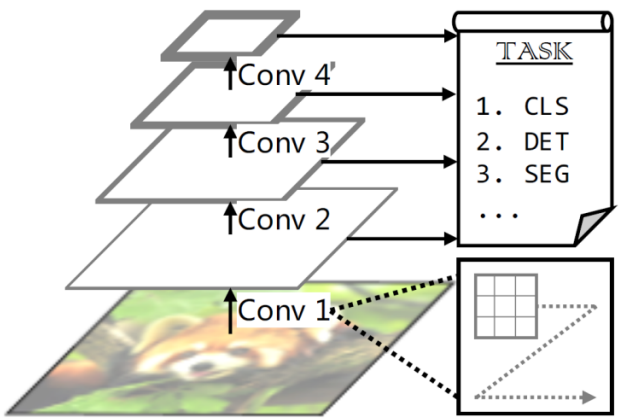
The University of Hong Kong Nanjing University NVIDIA Caltech

Vision Transformer (ViT)

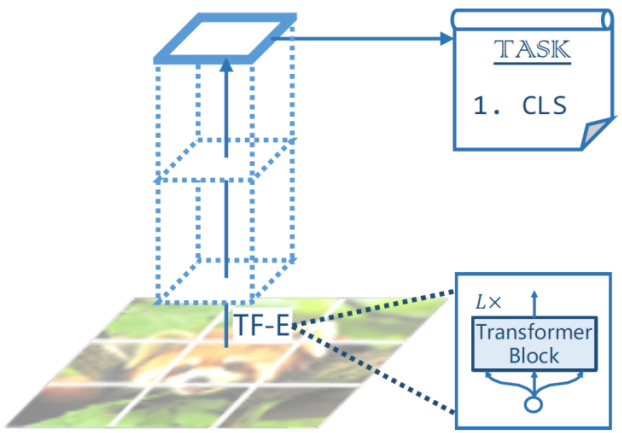


Dosovitskiy et al., **An image is worth 16x16 words: Transformers for image recognition at scale**, ICLR21.
Zheng et al., **Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers**, CVPR21.

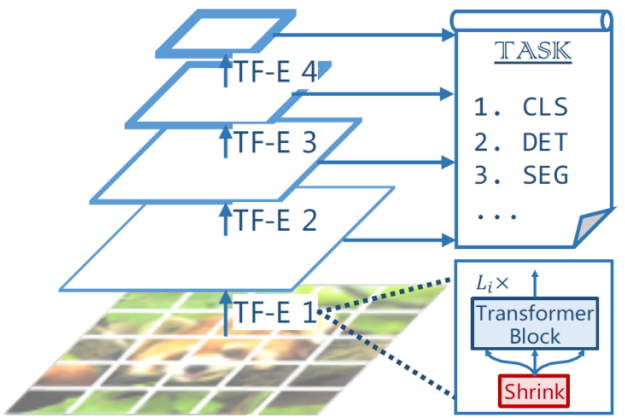
Pyramid Vision Transformer (PVT)



(a) CNNs: VGG [41], ResNet [15], *etc.*



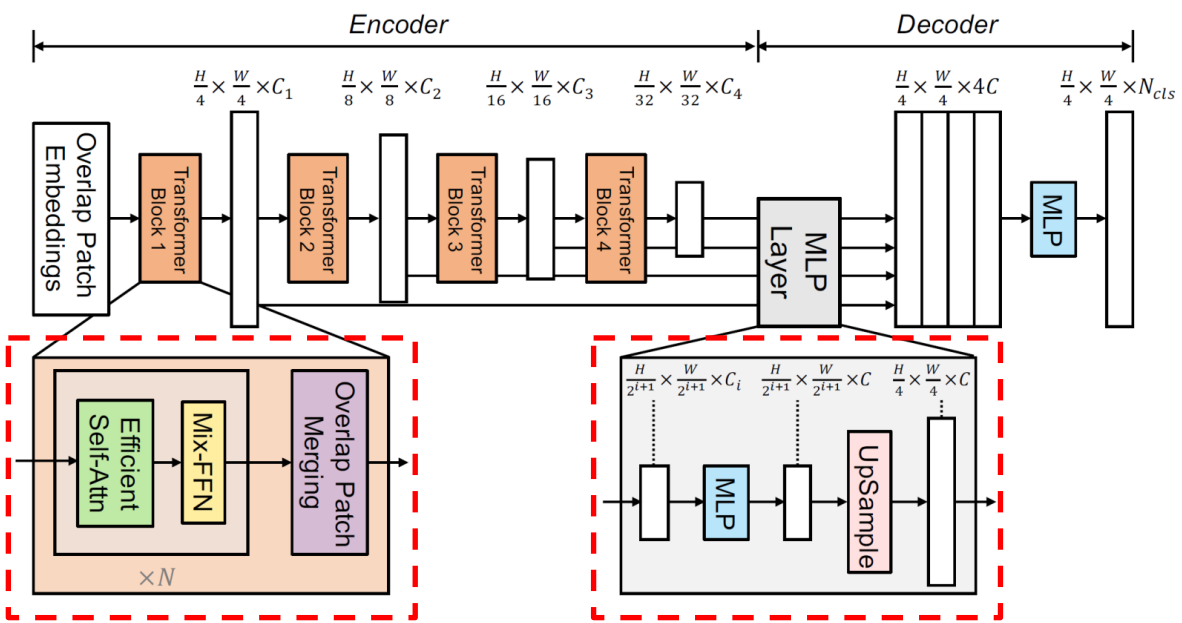
(b) Vision Transformer [10]



(c) Pyramid Vision Transformer (ours)

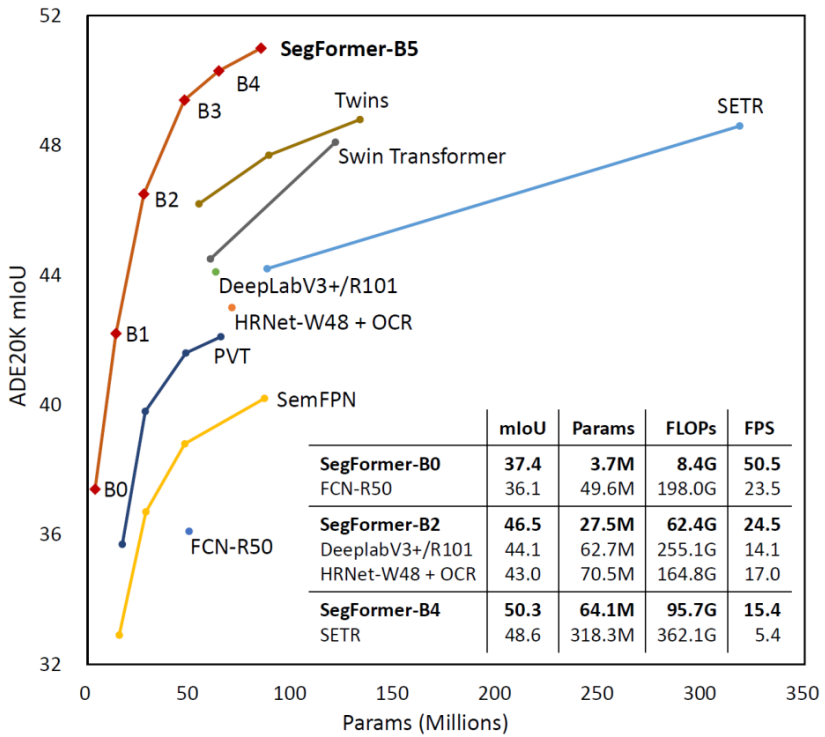
Wang et al., **Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions**, arXiv21.

SegFormer: Architecture Overview



Positional encoding free, pyramid structured encoder

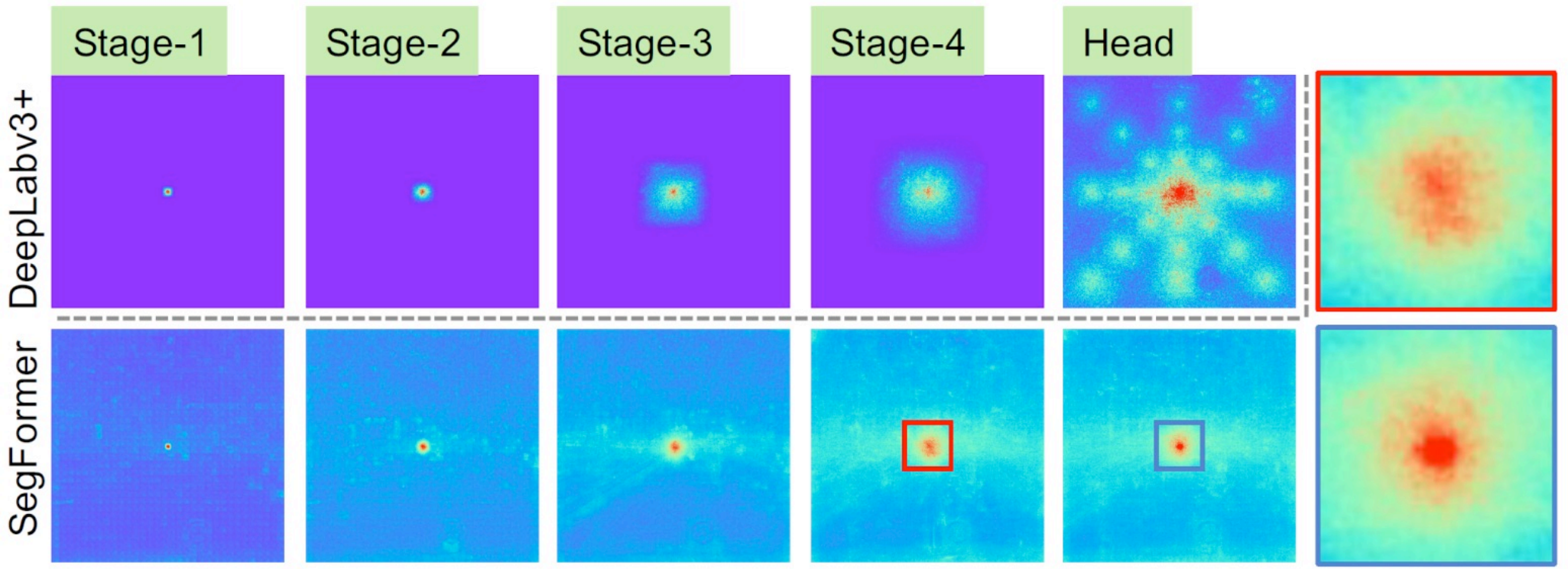
Lightweight all-MLP decoder



	mIoU	Params	FLOPs	FPS
SegFormer-B0	**37.4**	**3.7M**	**8.4G**	**50.5**
FCN-R50	36.1	49.6M	198.0G	23.5
SegFormer-B2	**46.5**	**27.5M**	**62.4G**	**24.5**
DeeplabV3+/R101	44.1	62.7M	255.1G	14.1
HRNet-W48 + OCR	43.0	70.5M	164.8G	17.0
SegFormer-B4	**50.3**	**64.1M**	**95.7G**	**15.4**
SETR	48.6	318.3M	362.1G	5.4

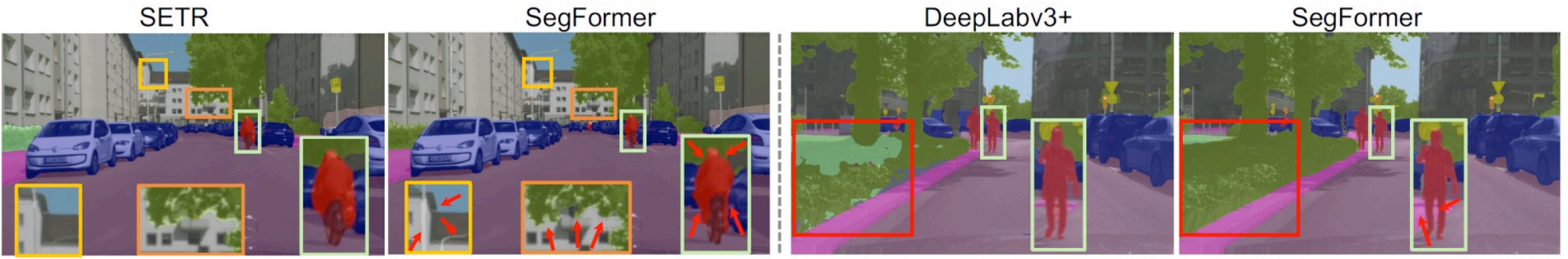
Xie et al., **SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers**, arXiv21.

SegFormer: Decoder Design

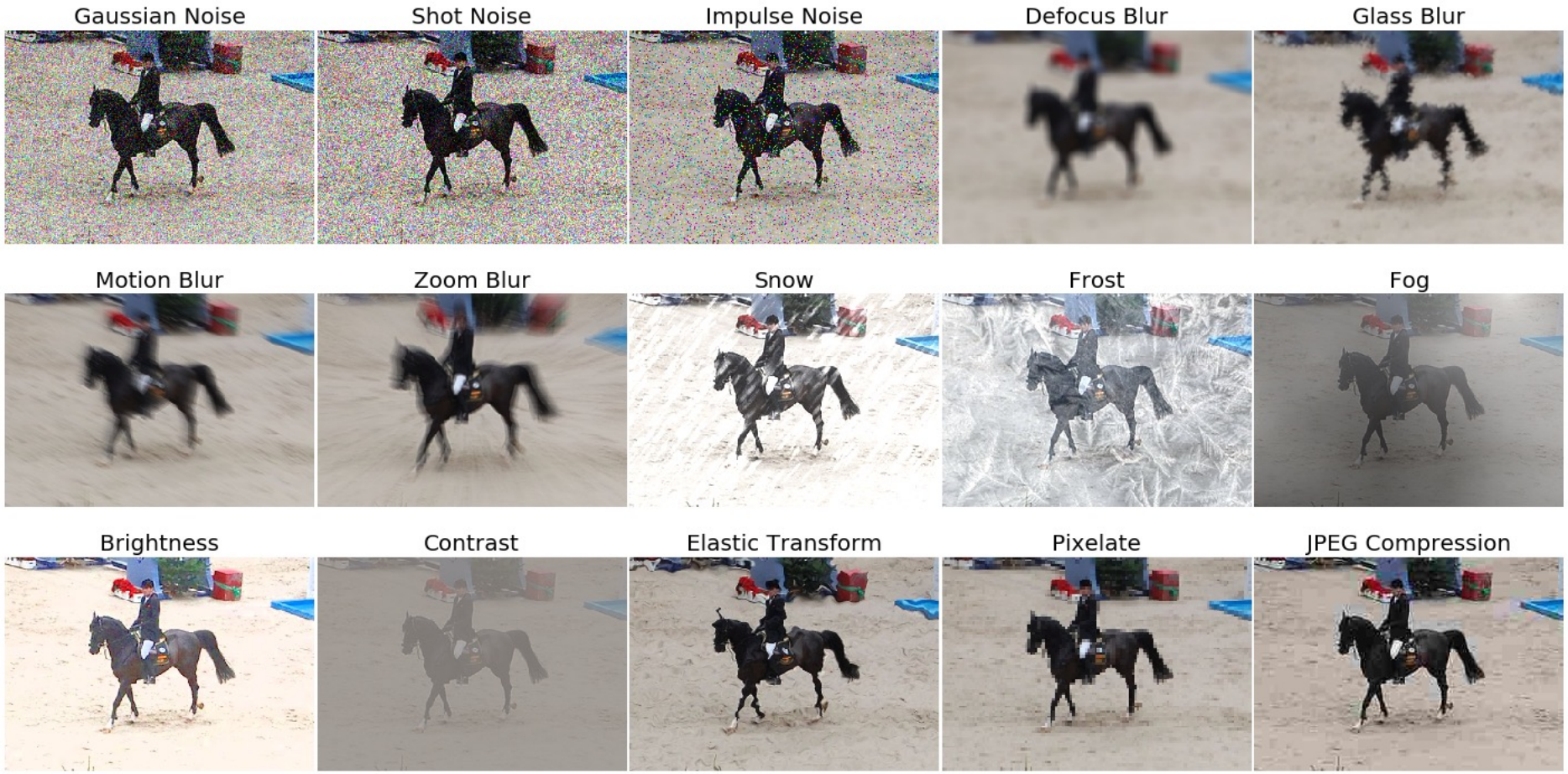


Luo et al., **Understanding the Effective Receptive Field in Deep Convolutional Neural Networks**, NeurIPS16.

Experiment: Qualitative Results

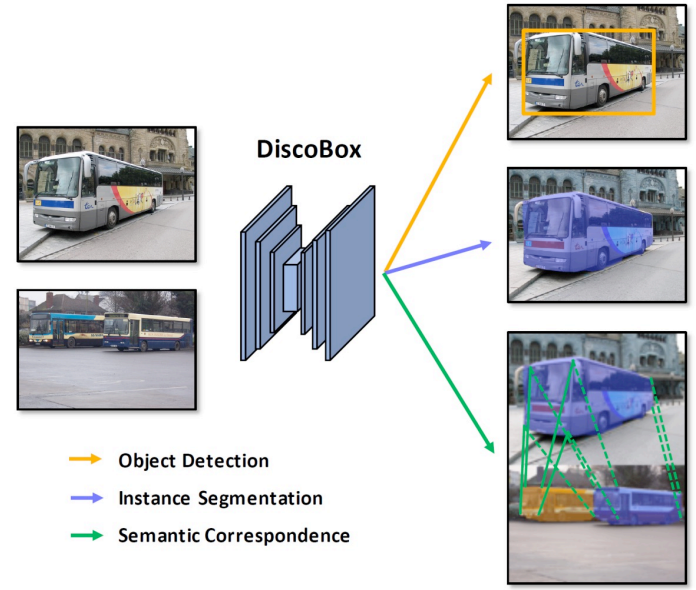


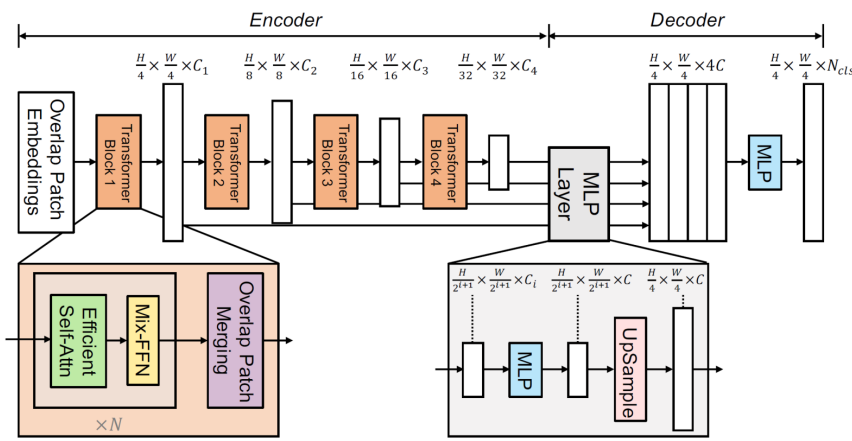
Experiment: Robustness to Corruptions



Conclusion

- Mask information probably can be totally removed in future for instance segmentation problems.
- Auto-labeling is promising for many dense prediction tasks.
- The emergence of Transformers will probably add even more to the above directions.
- Synthetic data can be seamlessly adapted to real-world tasks.





Thank You!