Learning with less resources:

minimizing the labeling effort

Negar Rostamzadeh

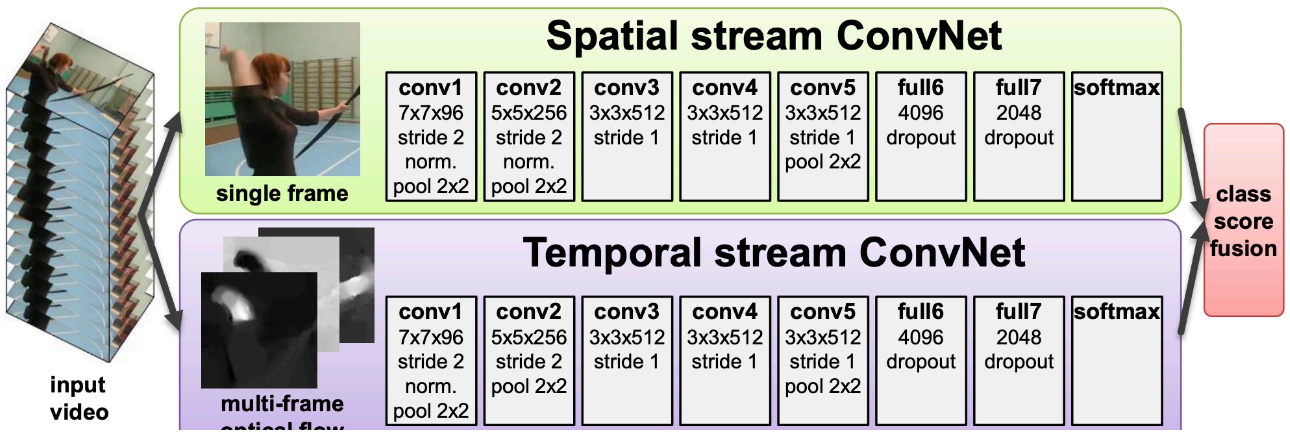
ICLR 2020 workshop on Practical ML for Developing Countries: learning under limited/low resource scenarios

Deep Learning and challenges

ImageNet, Russakovsky et al, 2014

Deep Learning approaches work well with large number of labeled data and good computational power.

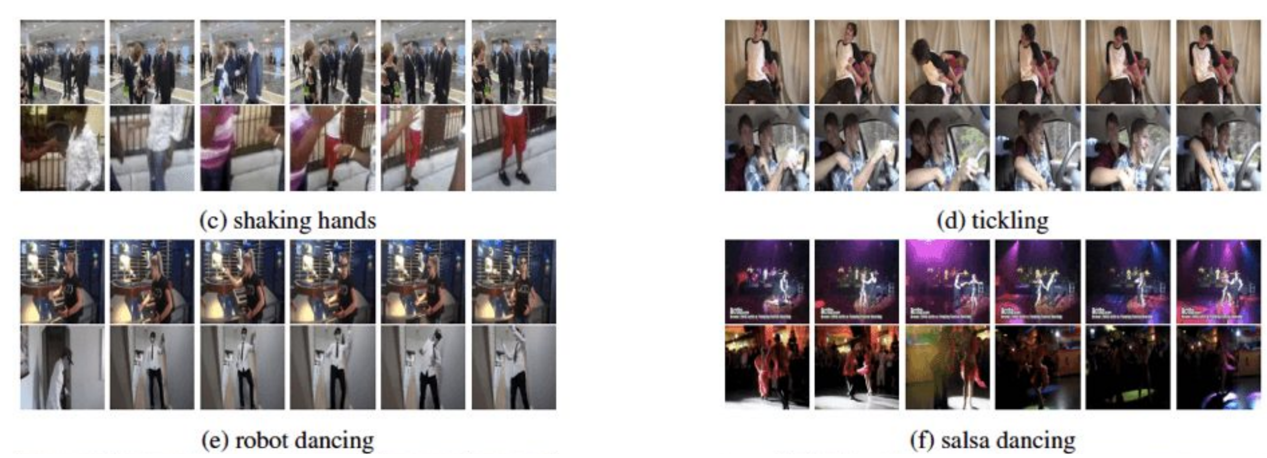
Data annotation challenges- Video understanding



Two-Stream Convolutional Networks in Videos, Simonyan and Zisserman

Learning Spatiotemporal Features with 3D Convolutional Networks, Tran et al.

AVA: A Video Dataset, Gu et al.

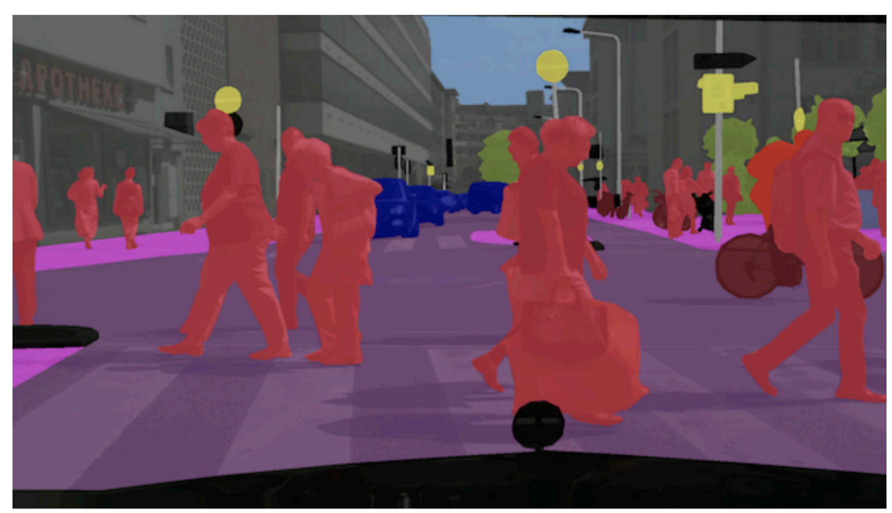


Kinetic dataset, Key et al.

Research question: How can we minimize the labeling effort and still have a good performance?

Data annotation challenges- semantic and instance segmentation

Input Image | Semantic Segmentation | Instance Segmentation



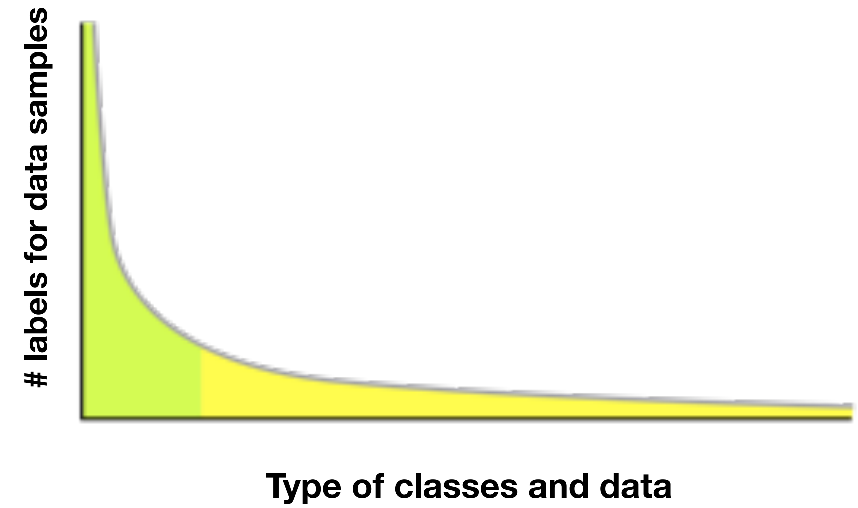
In average 1.5 hour to annotate each image

“The Cityscapes Dataset” M. Cordts et al. CVPR, 2016

Top picture source: https://towardsdatascience.com/review-deepmask-instance-segmentation-30327a072339

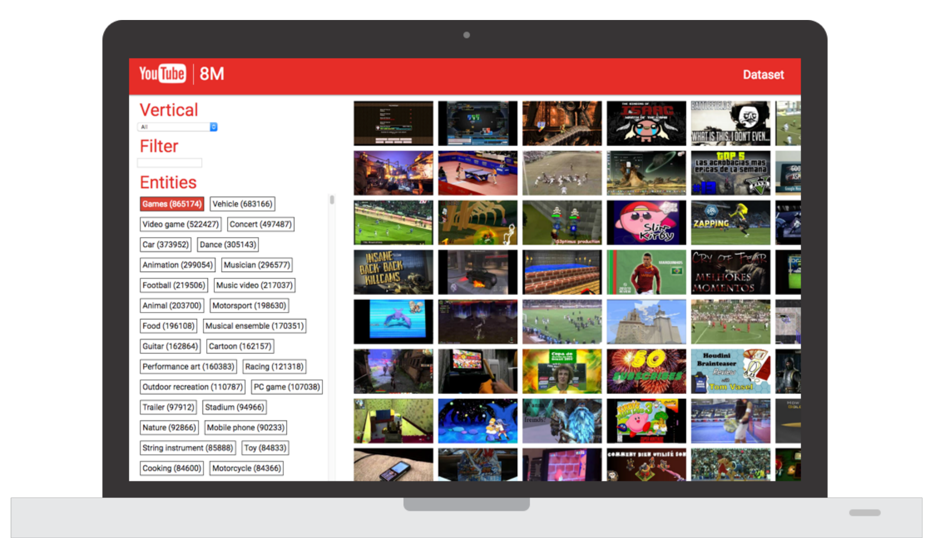
Challenges with scarcity of data

Long tail of data



We have access to multiple sources of data

Text



Videos (Visual/Audio) and text

Visual/Text

Research question: How can we reduce the labeling effort while, maintaining a good performance?

- Q1: Can we have cheaper and easier annotations and still have a competitive performance?
  1. Where are the blobs: Counting by localization with point supervision, Laradji et all, ECCV 2018
  2. Instance Segmentation with Point Supervision, Laradji et al, arXiv:1906.06392

Research question: How can we reduce the labeling effort while, maintaining a good performance?

- Q1: Can we have cheaper and easier annotations and still have a competitive performance?
  1. Where are the blobs: Counting by localization with point supervision, Laradji et all, ECCV 2018
  2. Instance Segmentation with Point Supervision, Laradji et al, arXiv:1906.06392

- Q2: How to exploit the data from a cheaper to annotate domain?

  Domain-Adaptive single-view 3D reconstruction, Pinheiro et al, ICCV 2019

Research question: How can we reduce the labeling effort while, maintaining a good performance?

- Q1: Can we have cheaper and easier annotations and still have a competitive performance?
  1. Where are the blobs: Counting by localization with point supervision, Laradji et all, ECCV 2018
  2. Instance Segmentation with Point Supervision, Laradji et al, arXiv:1906.06392

- Q2: How to exploit the data from a cheaper to annotate domain?

  Domain-Adaptive single-view 3D reconstruction, Pinheiro et al, ICCV 2019

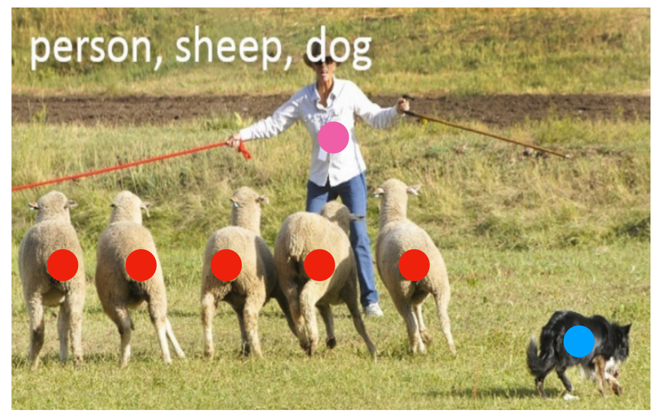
- Q3: How to exploit multiple source of data to solve a problem?
  1. Adaptive cross-modal few-shot learning, Xing et al, NeurIPS 2019
  2. Neural Multisensory Scene Inference, Lim et al, NeurIPS 2019

Can we have cheaper and easier annotations?

Point-level annotation

Where are the blobs: Counting by localization with point supervision,
Issam Laradji, Negar Rostamzadeh, Pedro Pinheiro, David Vazquez, Mark Schmidth, *ECCV 2018*

Input image



Point-level annotated image

5 ships
1 dog
1 person

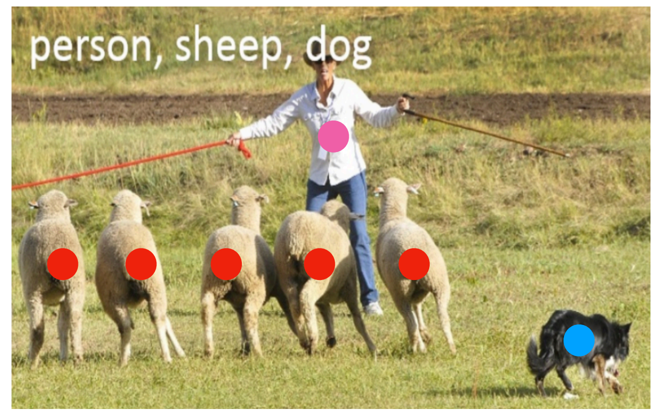
Output: object instance count

Images source: Microsoft COCO: Common Objects in Context, Lin et al.

Point-level annotation

Where are the blobs: Counting by localization with point supervision,
Issam Laradji, Negar Rostamzadeh, Pedro Pinheiro, David Vazquez, Mark Schmidth, *ECCV 2018*

Input image



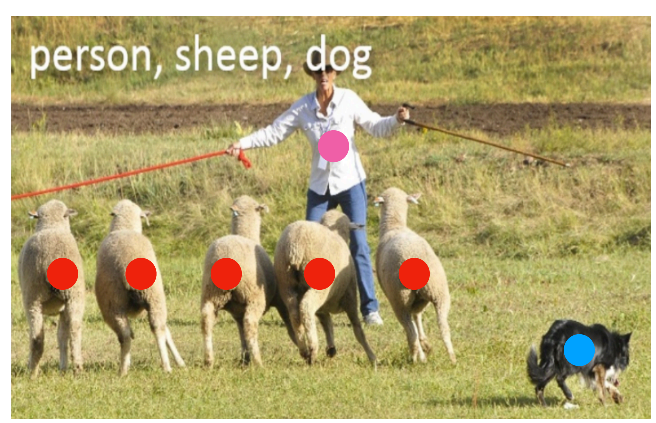
Point-level annotated image

5 ships
1 dog
1 person

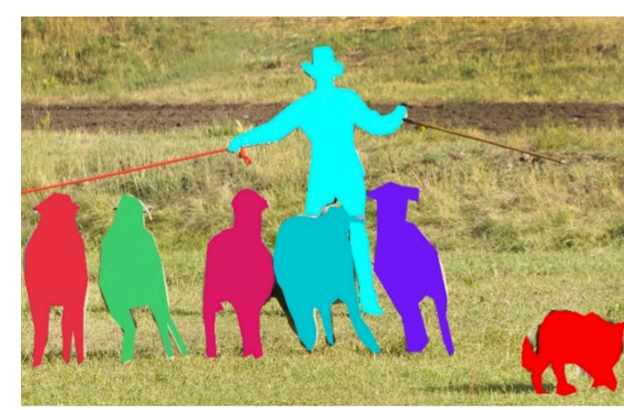
Output: object instance count

Instance Segmentation with Point Supervision,
Issam Laradji, Negar Rostamzadeh, Pedro Pinheiro, David Vazquez, Mark Schmidth, *arXiv: 1906.0639*

Input image



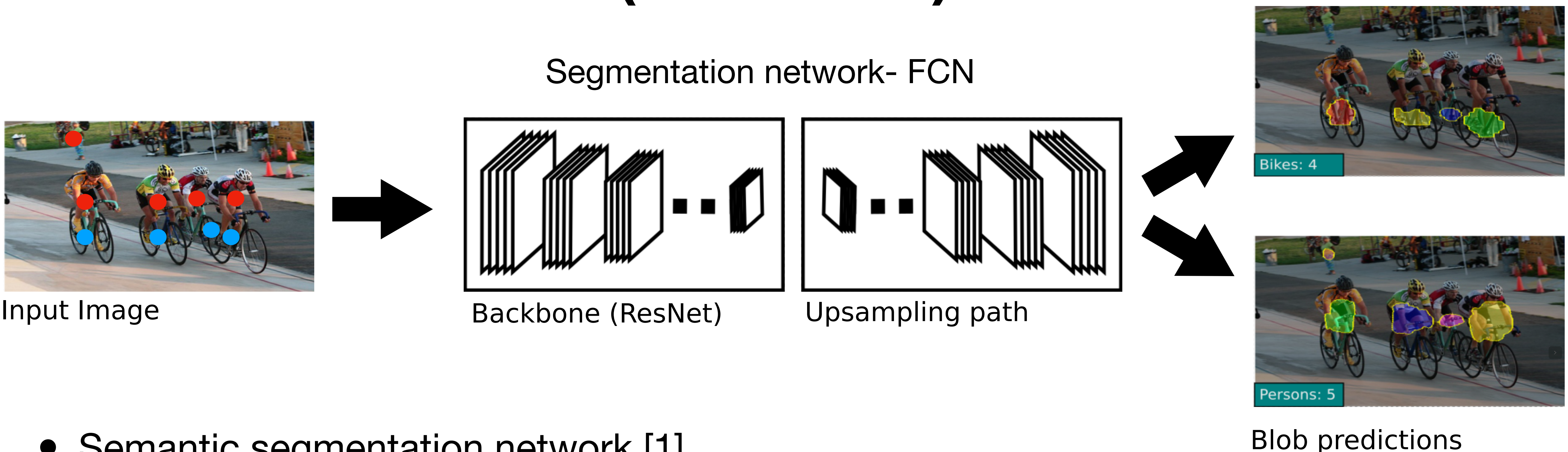
Point-level annotated image



Output: instance segmentation

Images source: Microsoft COCO: Common Objects in Context, Lin et al.

Localization-based Counting FCN (LC-FCN)



- Semantic segmentation network [1]
- The count is the number of predicted blobs
- Trained to output exactly one blob per each object instance

[1] What's the Point: Semantic Segmentation with Point Supervision, Bearman et al, ECCV 2016

Where are the blobs: Counting by localization with point supervision,
Issam Laradji, **<u>Negar Rostamzadeh</u>**, Pedro Pinheiro, David Vazquez, Mark Schmidth, **ECCV 2018**

Localization-based Counting FCN (LC-FCN)

Image-level Loss

Discourage predicting classes not present in the annotations

$$L(S,T) = -\frac{1}{|C_e|}\sum_{c \in C_e} \log(S_{t_c c}) - \frac{1}{|C_{\neg e}|}\sum_{c \in C_{\neg e}} \log(1 - S_{t_c c})$$

S: Output mask the model

T: Ground-truth

Where are the blobs: Counting by localization with point supervision,
Issam Laradji, **Negar Rostamzadeh**, Pedro Pinheiro, David Vazquez, Mark Schmidth, **ECCV 2018**

Localization-based Counting FCN (LC-FCN)

Image-level Loss

Discourage predicting classes not present in the annotations

Point-level Loss

Encourage predicting the classes of the annotated pixels

$$L(S,T) = -\frac{1}{|C_e|}\sum_{c\in C_e}\log(S_{t_c c}) - \frac{1}{|C_{\neg e}|}\sum_{c\in C_{\neg e}}\log(1-S_{t_c c}) - \sum_{i\in \mathcal{I}_s}\log(S_{iT_i})$$

S: Output mask the model

T: Ground-truth

Where are the blobs: Counting by localization with point supervision,
Issam Laradji, **Negar Rostamzadeh**, Pedro Pinheiro, David Vazquez, Mark Schmidth, **ECCV 2018**

Localization-based Counting FCN (LC-FCN)

Point-level segmentation loss [1]

Image-level Loss

Discourage predicting classes not present in the annotations

Point-level Loss

Encourage predicting the classes of the annotated pixels

$$L(S,T) = -\frac{1}{|C_e|}\sum_{c \in C_e} \log(S_{t_c c}) - \frac{1}{|C_{\neg e}|}\sum_{c \in C_{\neg e}} \log(1 - S_{t_c c}) - \sum_{i \in \mathcal{I}_s} \log(S_{iT_i})$$

S: Output mask the model

T: Ground-truth

[1] What's the Point: Semantic Segmentation with Point Supervision, Bearman et al, ECCV 2016

Where are the blobs: Counting by localization with point supervision,
Issam Laradji, **Negar Rostamzadeh**, Pedro Pinheiro, David Vazquez, Mark Schmidth, **ECCV 2018**

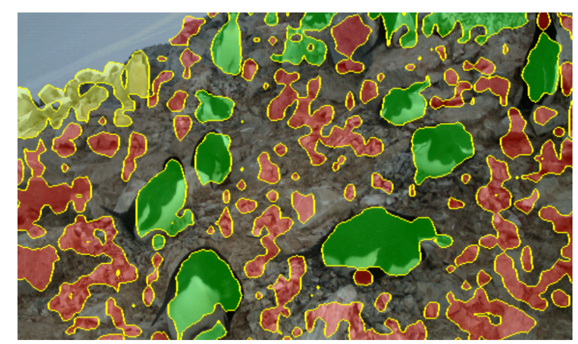
Localization-based Counting FCN (LC-FCN)

Point-level segmentation loss [1]

Image-level Loss	Point-level Loss	Split-level Loss	False positive Loss
Discourage predicting classes not present in the annotations	Encourage predicting the classes of the annotated pixels	discourages the prediction of blobs that have two or more point-annotations	Discourage predicting blobs without point-annotations

$$L(S,T) = -\frac{1}{|C_e|}\sum_{c \in C_e} \log(S_{t_c c}) - \frac{1}{|C_{\neg e}|}\sum_{c \in C_{\neg e}} \log(1 - S_{t_c c}) \;-\sum_{i \in \mathcal{I}_s} \log(S_{iT_i}) \;-\sum_{i \in T_b} \alpha_i \log(S_{i0}) \;-\sum_{i \in B_{fp}} \log(S_{i0})$$






S: Output mask the model

T: Ground-truth

α_i : Number of point-annotations in the blob in which pixel i lies

[1] What's the Point: Semantic Segmentation with Point Supervision, Bearman et al, ECCV 2016

Where are the blobs: Counting by localization with point supervision,
Issam Laradji, **Negar Rostamzadeh**, Pedro Pinheiro, David Vazquez, Mark Schmidth, **ECCV 2018**

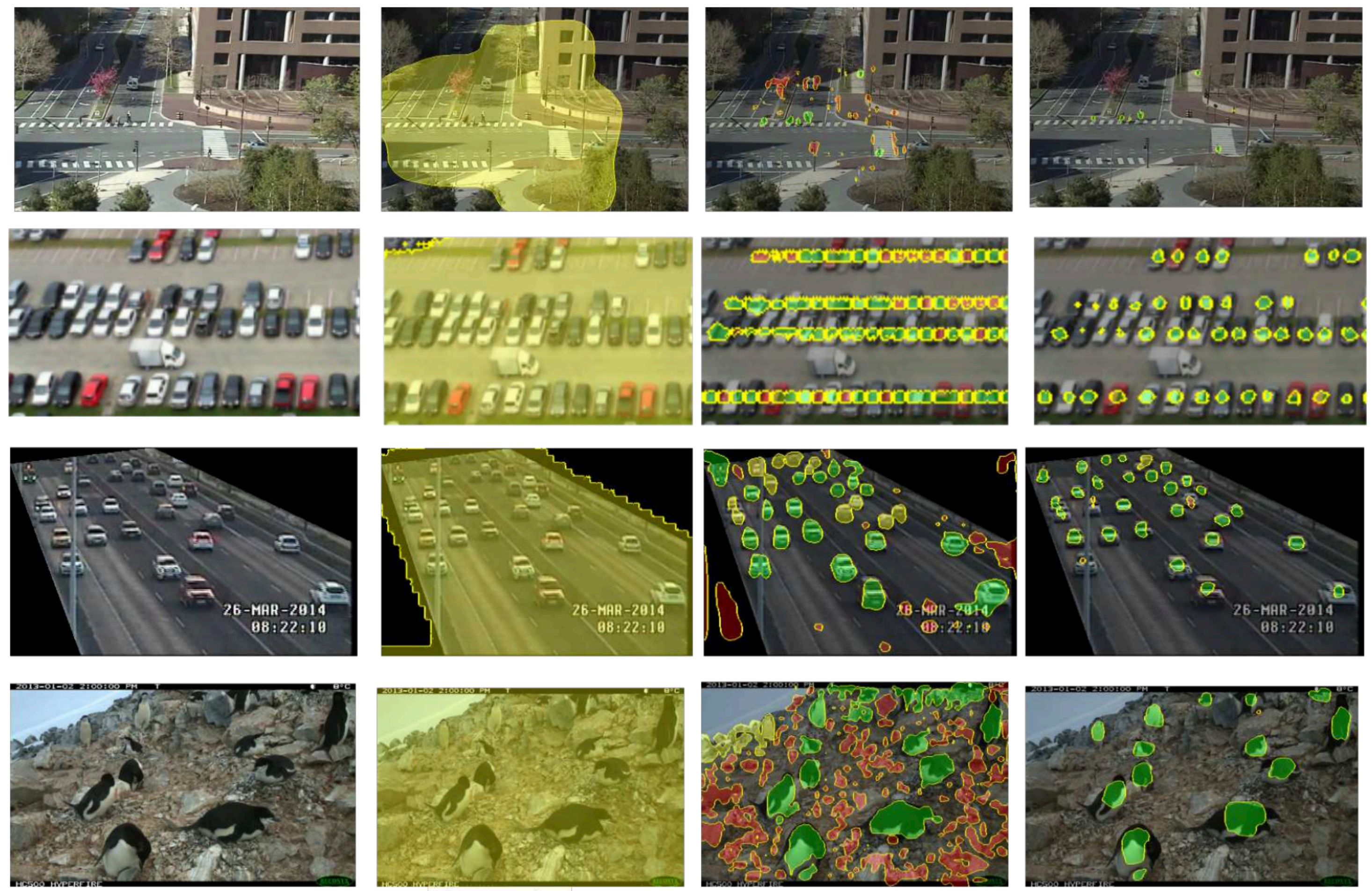
LCFCN: Analyzing loss terms - Qualitative results

	MIT Traffic		PKLot		Trancos		Penguins Separated	
Method	**MAE**	**FS**	**MAE**	**FS**	**MAE**	**FS**	**MAE**	**FS**
Glance	1.57	-	1.92	-	7.01	-	6.09	-
$\mathcal{L}_I + \mathcal{L}_P$	3.11	0.38	39.62	0.04	38.56	0.05	9.81	0.08
$\mathcal{L}_I + \mathcal{L}_P + \mathcal{L}_S$	1.62	0.76	9.06	0.83	6.76	0.56	4.92	0.53
$\mathcal{L}_I + \mathcal{L}_P + \mathcal{L}_F$	1.84	0.69	39.60	0.04	38.26	0.05	7.28	0.04
LC-ResFCN	1.26	**0.81**	10.16	0.84	**3.32**	0.68	3.96	0.63
LC-FCN8	**0.91**	0.69	**0.21**	**0.99**	4.53	**0.54**	**3.74**	**0.61**

$$\mathcal{L}(S,T) = \underbrace{\mathcal{L}_I(S,T)}_{\text{Image-level loss}} + \underbrace{\mathcal{L}_P(S,T)}_{\text{Point-level loss}} + \underbrace{\mathcal{L}_S(S,T)}_{\text{Split-level loss}} + \underbrace{\mathcal{L}_F(S,T)}_{\text{False positive loss}}$$

Where are the blobs: Counting by localization with point supervision,
Issam Laradji, **Negar Rostamzadeh**, Pedro Pinheiro, David Vazquez, Mark Schmidth, **ECCV 2018**

LCFCN: Analyzing loss terms



(a) Original Image (b) $\mathcal{L}_I + \mathcal{L}_P$ (c) $\mathcal{L}_I + \mathcal{L}_P + \mathcal{L}_S$ (d) LC-FCN

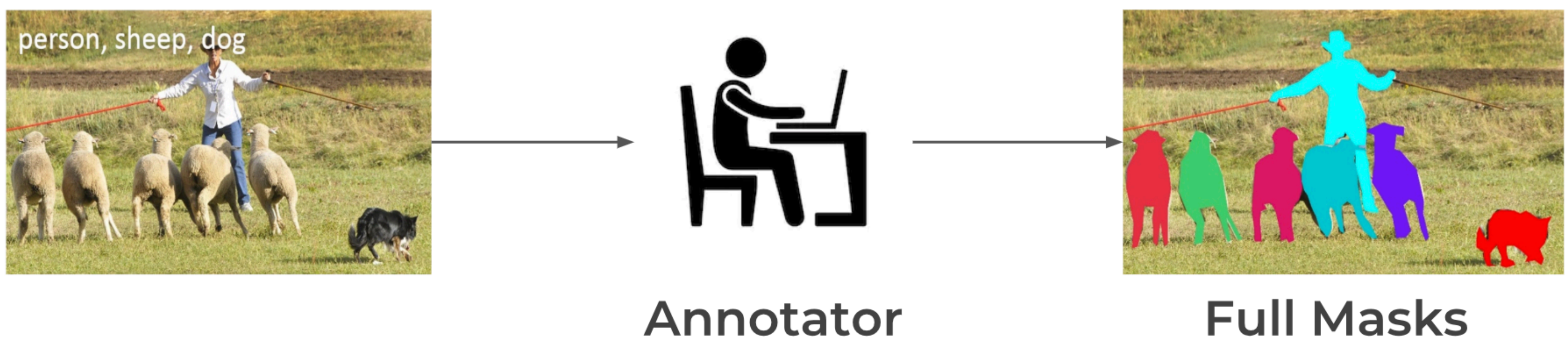
True Positive False Positive More than one point annotation

$$\mathcal{L}(S,T) = \underbrace{\mathcal{L}_I(S,T)}_{\text{Image-level loss}} + \underbrace{\mathcal{L}_P(S,T)}_{\text{Point-level loss}} + \underbrace{\mathcal{L}_S(S,T)}_{\text{Split-level loss}} + \underbrace{\mathcal{L}_F(S,T)}_{\text{False positive loss}}$$

Where are the blobs: Counting by localization with point supervision,
Issam Laradji, **Negar Rostamzadeh**, Pedro Pinheiro, David Vazquez, Mark Schmidth, **ECCV 2018**

Instance segmentation labeling challenge

Traditional annotation: 1.5 hours per image

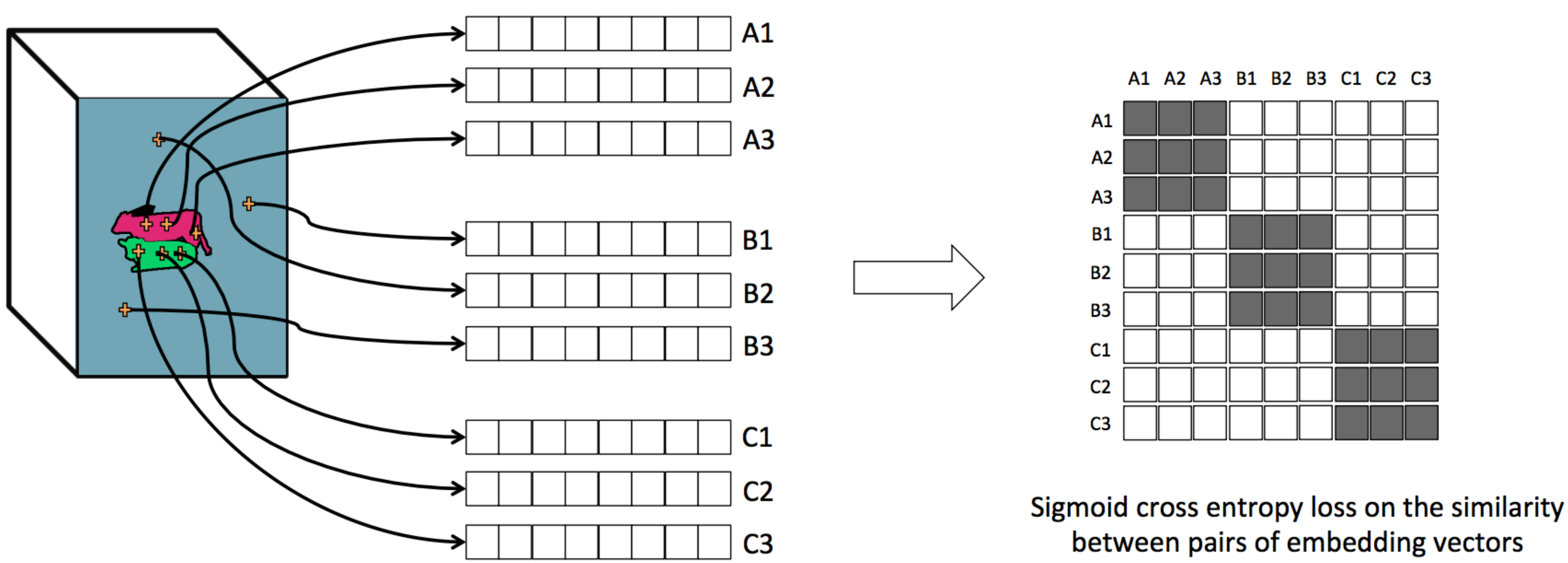


WISE: A few seconds per image

WISE: Instance Segmentation with Point Supervision,
Issam Laradji, **Negar Rostamzadeh**, Pedro Pinheiro, David Vazquez, Mark Schmidth, **ECCV 2018**

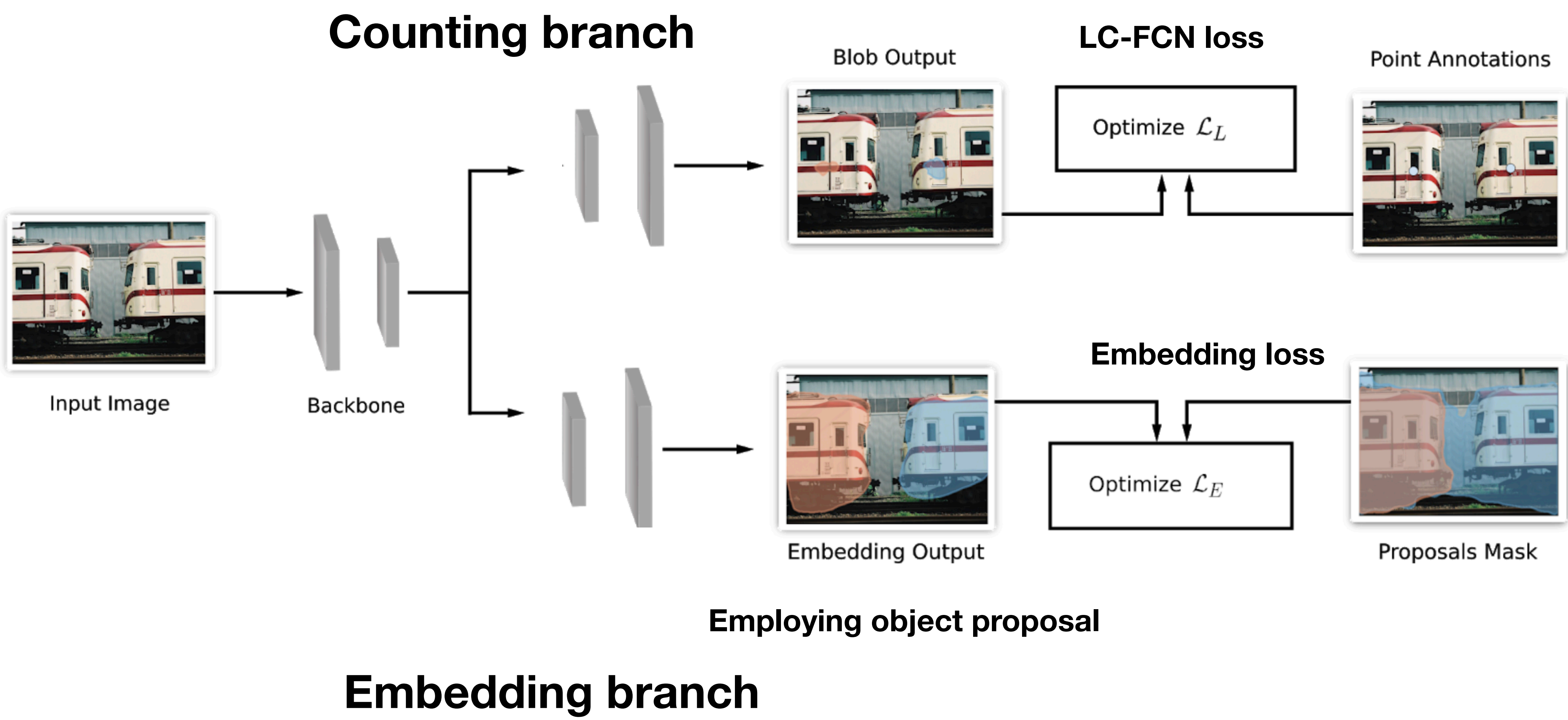
Related work on Instance Segmentation

Metric-based Instance Segmentation



Semantic Instance Segmentation via Deep Metric Learning, Fathi et al, CVPR 2018.

WISE: Weakly-supervised Instance Segmentation



$$\mathcal{L}_W = \lambda \cdot \mathcal{L}_L + (1 - \lambda) \cdot \mathcal{L}_E$$

Instance Segmentation with Point Supervision,
Issam Laradji, **Negar Rostamzadeh**, Pedro Pinheiro, David Vazquez, Mark Schmidth, **arXiv:1906.0639**

Comparison against the SOTA with fixed annotation budget

Method	Annotation	AP_{25}	AP_{50}	AP_{75}
Mask R-CNN (Zhu et al. (2017))	per-pixel	17.1	11.2	03.4
SPN (Zhu et al. (2017))	image-level	26.0	13.0	04.0
PRM (Zhou et al. (2018))	image-level	44.0	27.0	09.0
ILC (Cholakkal et al. (2019))	image-level	**48.5**	30.2	14.4
PRM + E-Net (Ours)	image-level	43.0	32.0	19.0
WISE (Ours)	point-level	47.5	**38.1**	**23.5**

PASCAL VOC- 2012- for 8.13 hours annotation budget

Annotation time per each image, Bearman *et al* [4]):

Per-pixel: 239.7

Point-level: 20.0

Image-level: 22.1

[1] SPN: Soft proposal networks for weakly supervised object localization, Zhou et al, CVPR 2017

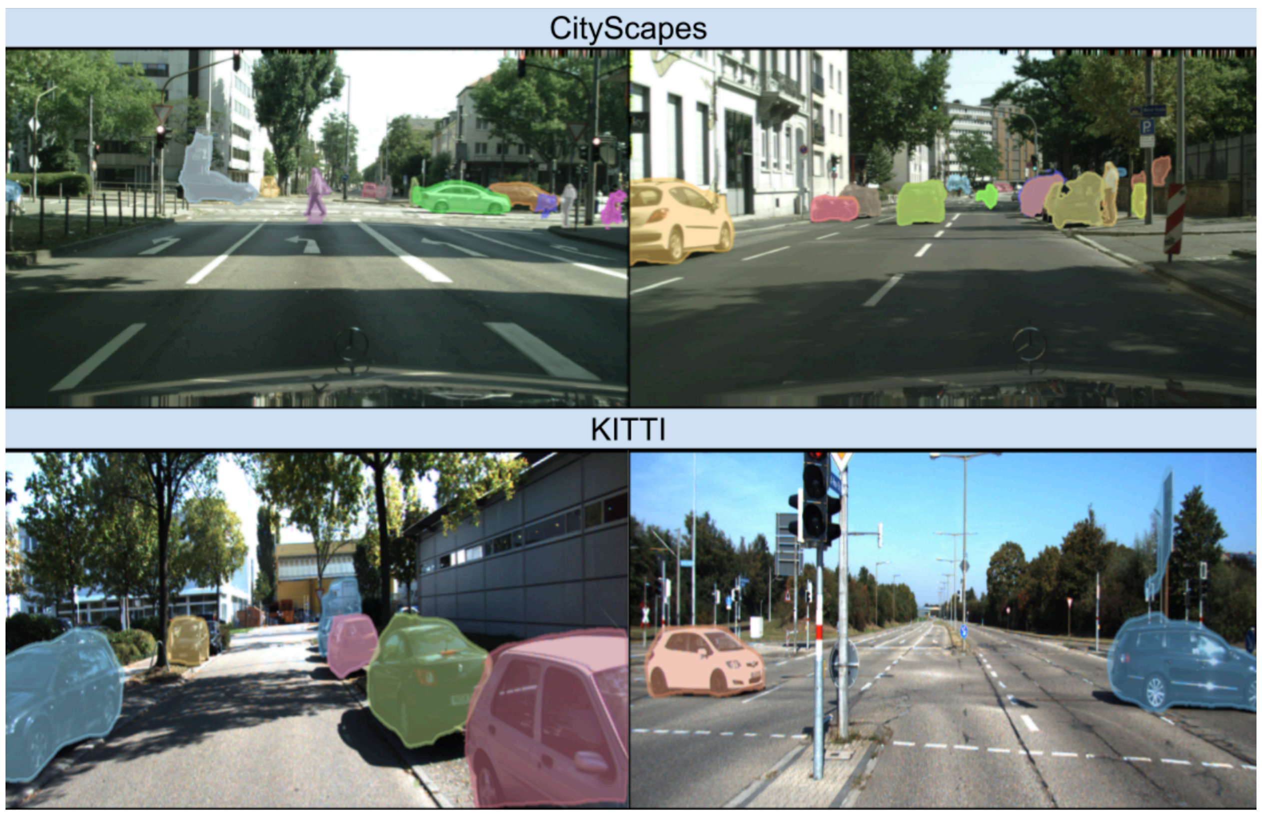
[2] ILC: Object Counting and Instance Segmentation with Image-level Supervision, Cholakkal 2019

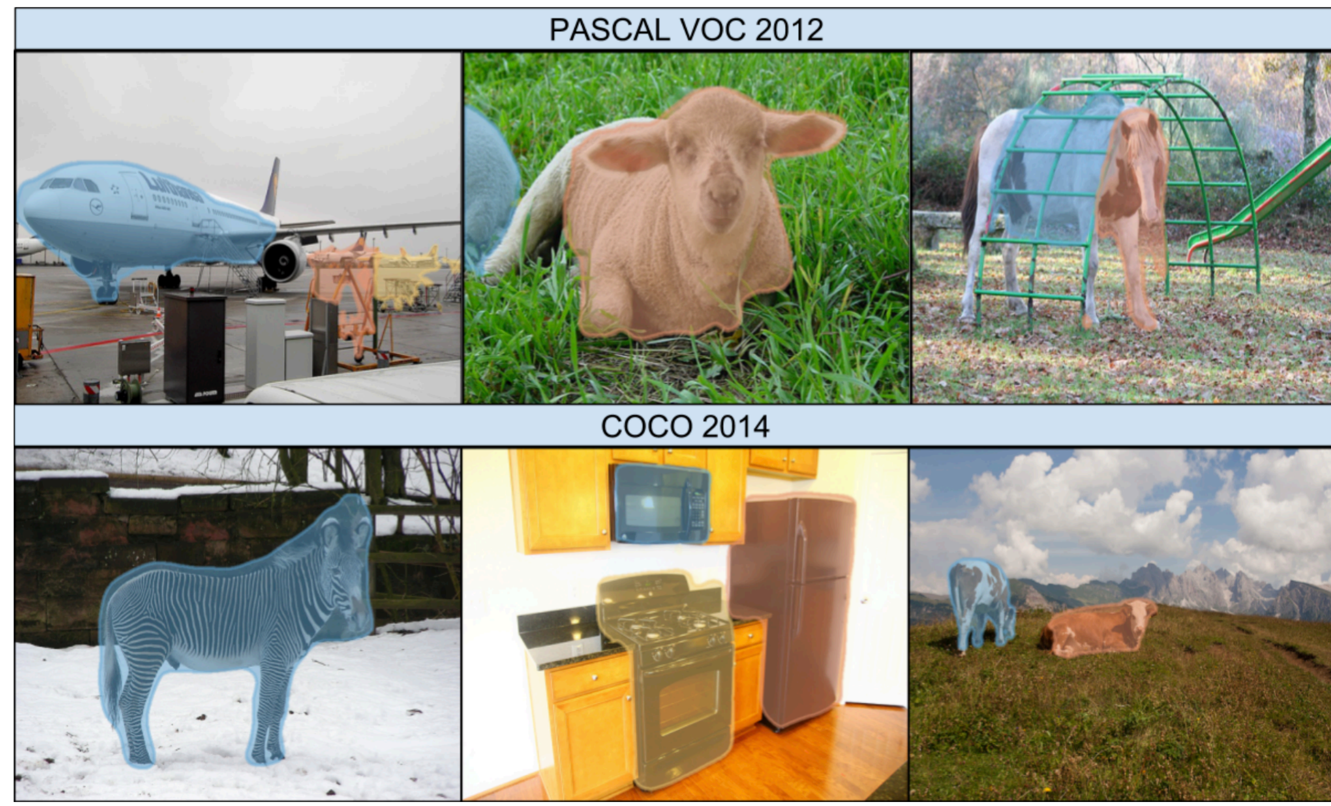
[3] PRM: Weakly supervised instance segmentation using class peak response, Zhou et al, CVPR 2018

[4] Semantic segmentation with point level annotation, Bearman et al, ECCV 2016

Instance Segmentation with Point Supervision,
Issam Laradji, **Negar Rostamzadeh**, Pedro Pinheiro, David Vazquez, Mark Schmidth, **arXiv:1906.0639**

Conclusion on point-level annotation





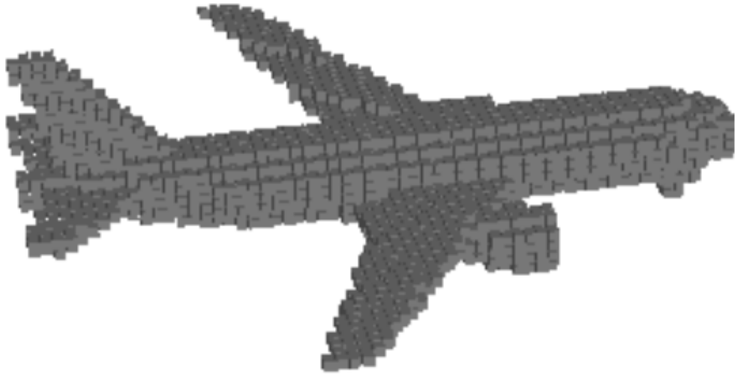
Instance Segmentation with Point Supervision,
Issam Laradji, **Negar Rostamzadeh**, Pedro Pinheiro, David Vazquez, Mark Schmidth, **arXiv:1906.0639**

Can we use labeled data from another domain?

Single-View 3D reconstruction

Single view 3D shape reconstruction

Natural image



3D voxel occupancy grid

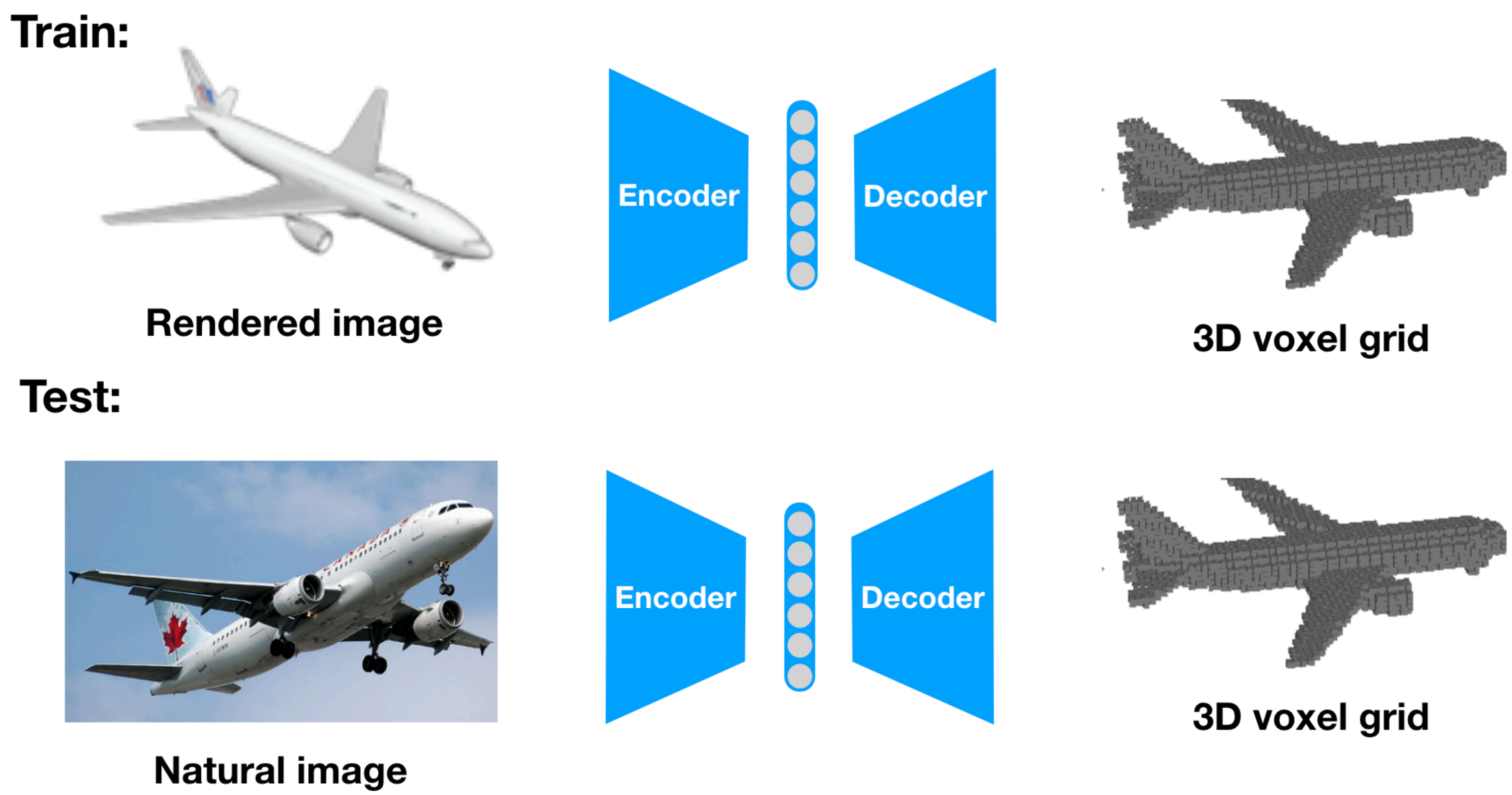
Challenges:

- Acquiring large number of views from natural images is impractical
- 3D annotation of natural images is a very **label heavy** task.
- This is an **ill-posed** problem.

Domain-Adaptive single-view 3D reconstruction,
Pedro Pinheiro, **Negar Rostamzadeh**, Sungjin Ahn, **ICCV 2019**

Recent work on 3D reconstruction

- Use **easy to access 3D CAD** repositories as **synthetic source** of data (pairs of rendered images and voxels)

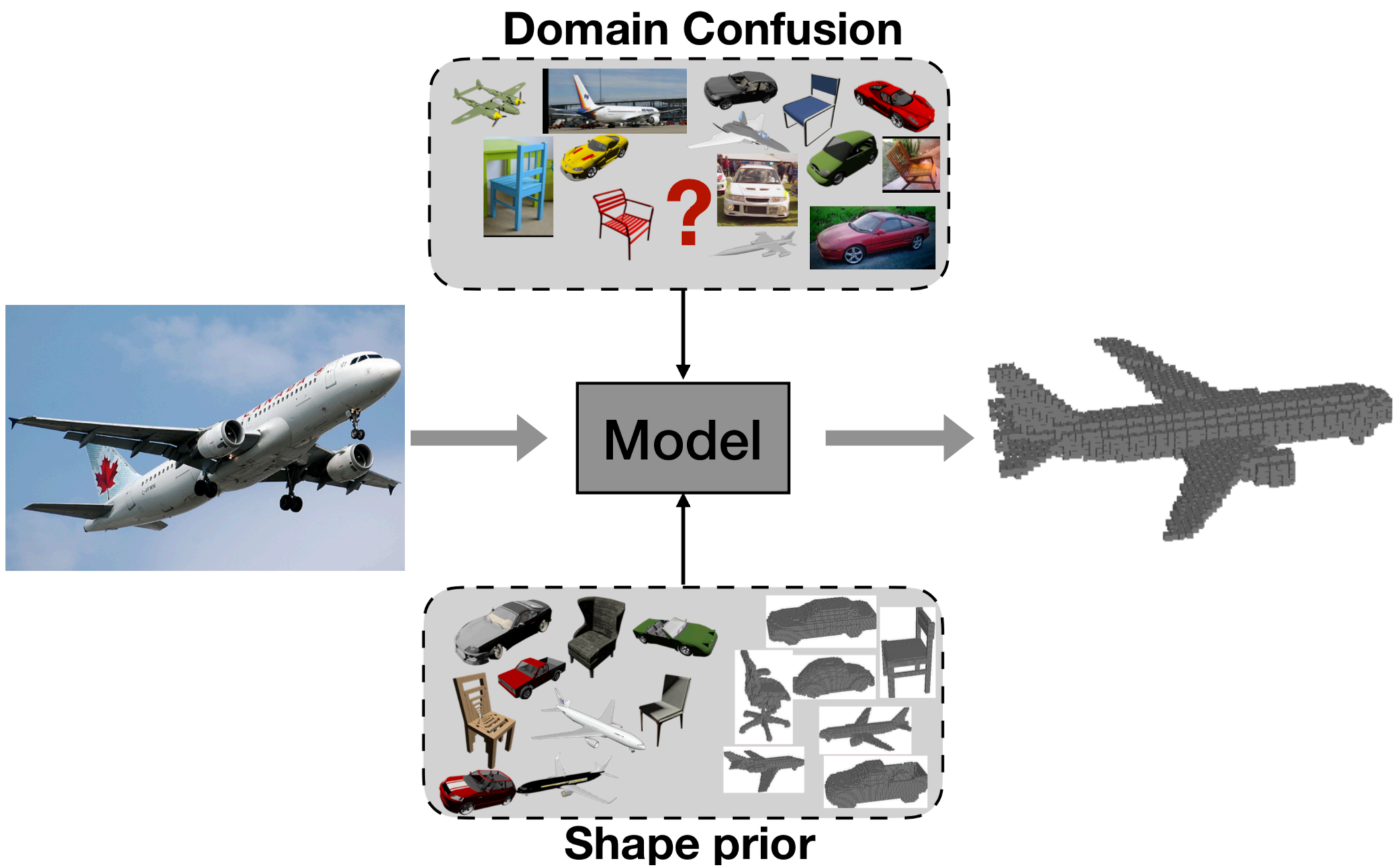


Challenges:

- ***Domain shift*** between rendered images and natural images.
- Unrealistic reconstructed shape.

Domain-Adaptive single-view 3D reconstruction,
Pedro Pinheiro, **Negar Rostamzadeh**, Sungjin Ahn, **ICCV 2019**

DAREC: Domain-Adaptive RE-Construction

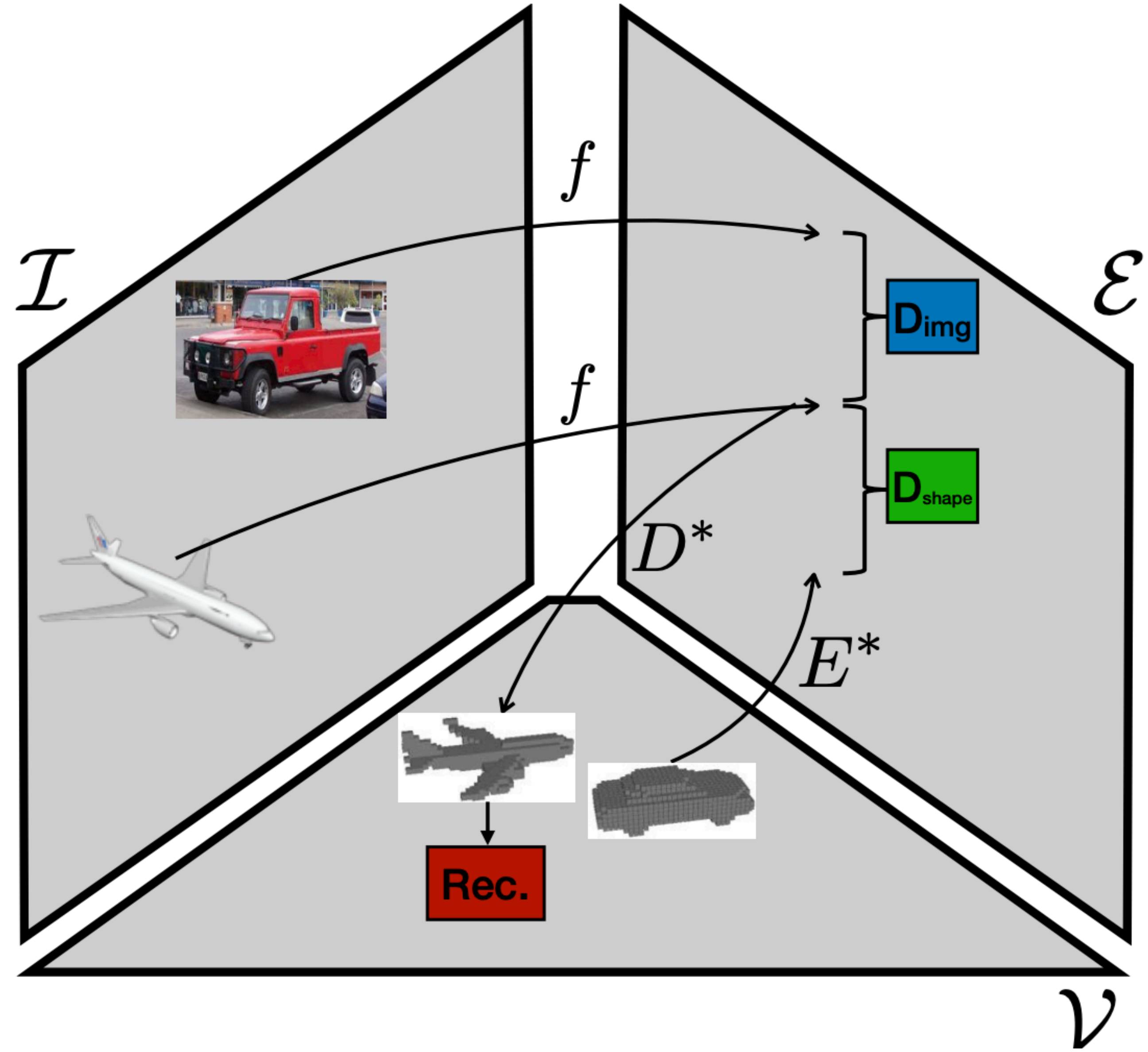


Domain-Adaptive single-view 3D reconstruction,
Pedro Pinheiro, **Negar Rostamzadeh**, Sungjin Ahn, **ICCV 2019**

DAREC: Domain-Adaptive RE-Construction

2 steps training:

1. **Shape autoencoder**

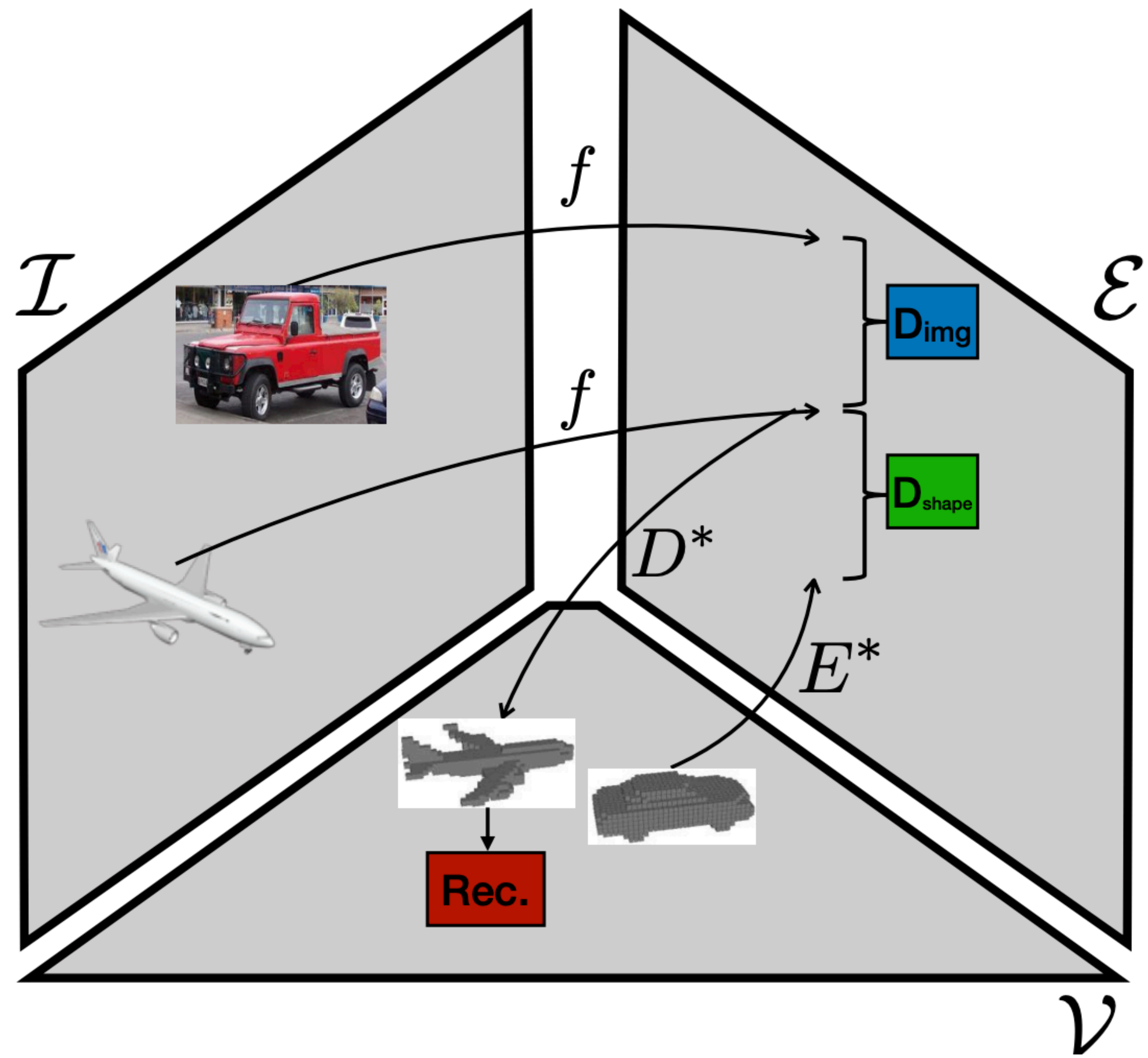


Domain-Adaptive single-view 3D reconstruction,
Pedro Pinheiro, **Negar Rostamzadeh**, Sungjin Ahn, **ICCV 2019**

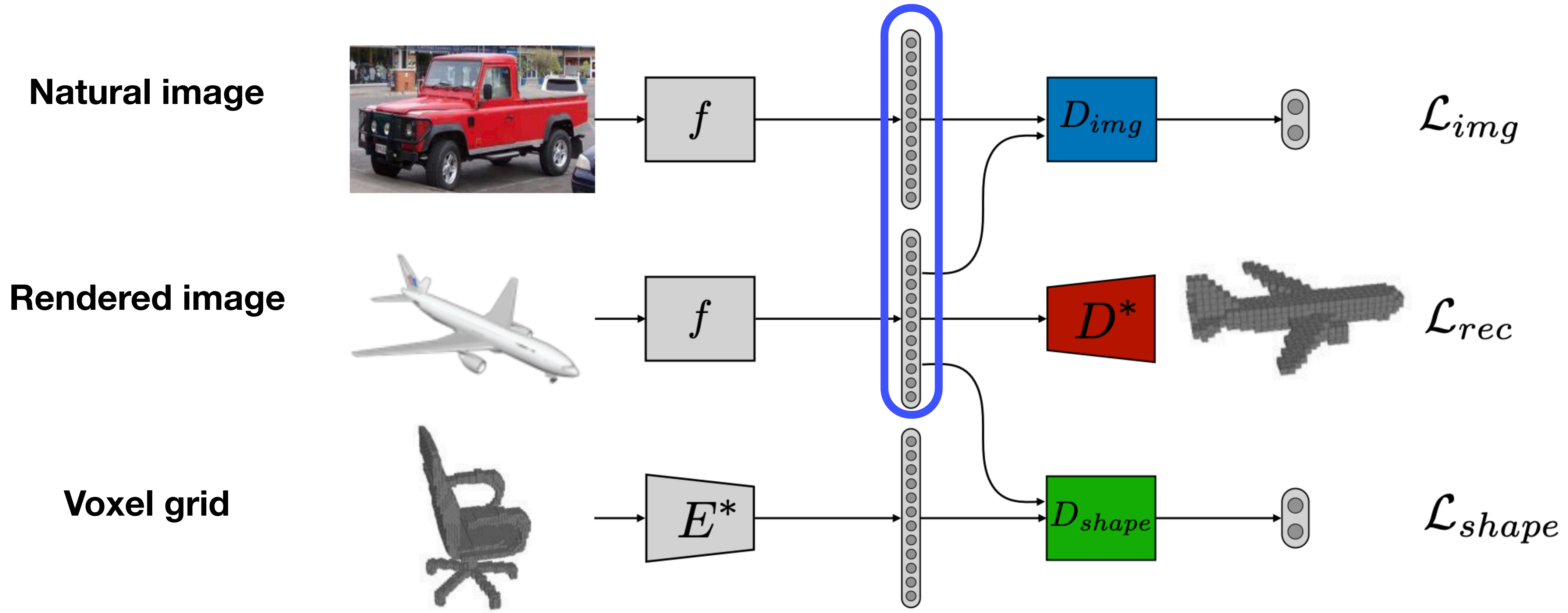
DAREC: Domain-Adaptive RE-Construction

2 steps training:

1. **Shape autoencoder**
2. **3D reconstruction network**



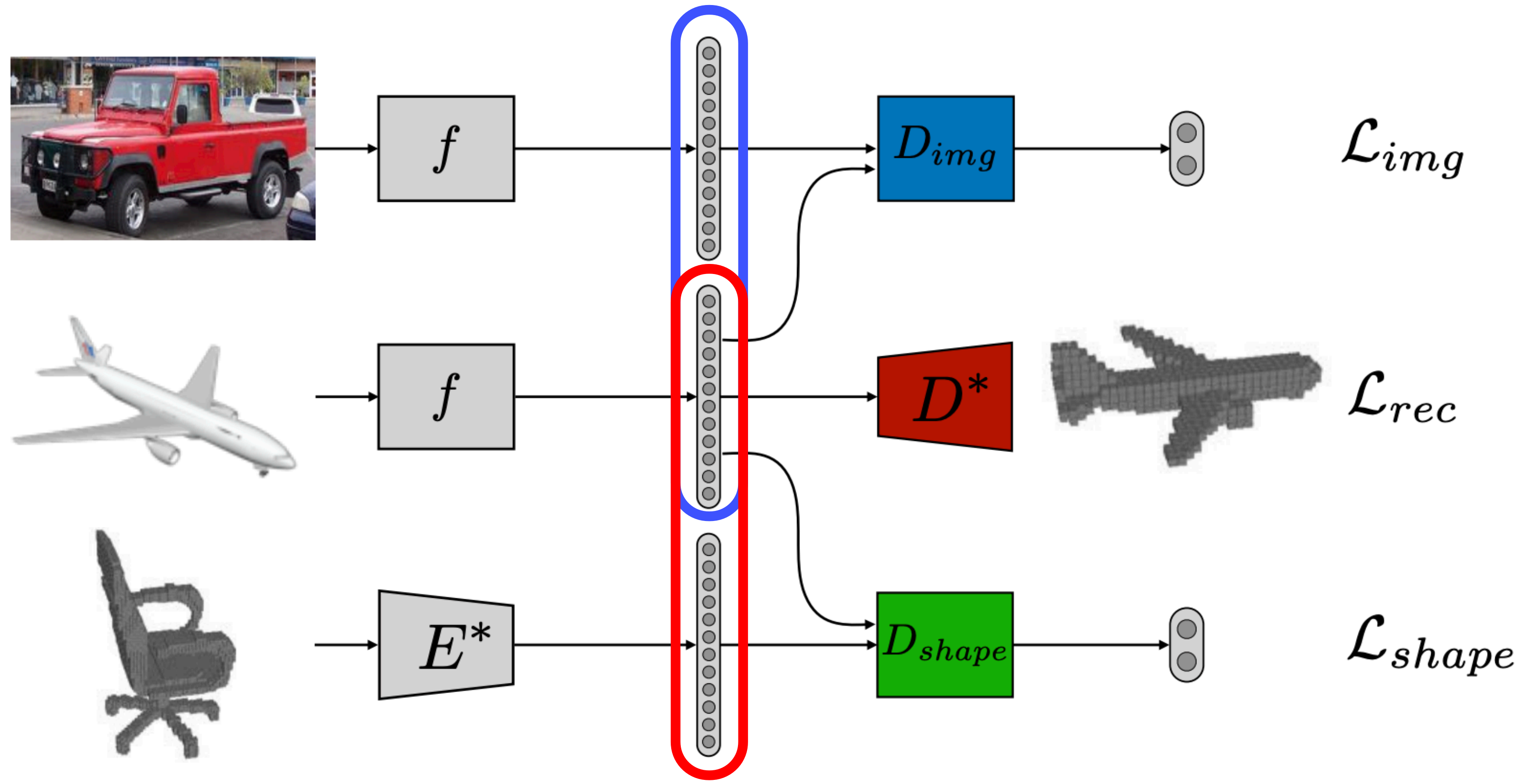
Domain-Adaptive single-view 3D reconstruction,
Pedro Pinheiro, **Negar Rostamzadeh**, Sungjin Ahn, **ICCV 2019**



$$\mathcal{L}_{img}(\theta_f, \theta_{img}) = -\mathop{\mathbb{E}}_{x^r \sim p_r} \log D_{img}(f(x^r)) + \\ -\mathop{\mathbb{E}}_{x^n \sim p_n} \log \left(1 - D_{img}(f(x^n))\right)$$

DANN: Domain-adversarial training of neural networks. Ganin et al, JMLR, 2016

Domain-Adaptive single-view 3D reconstruction,
Pedro Pinheiro, **Negar Rostamzadeh**, Sungjin Ahn, **ICCV 2019**

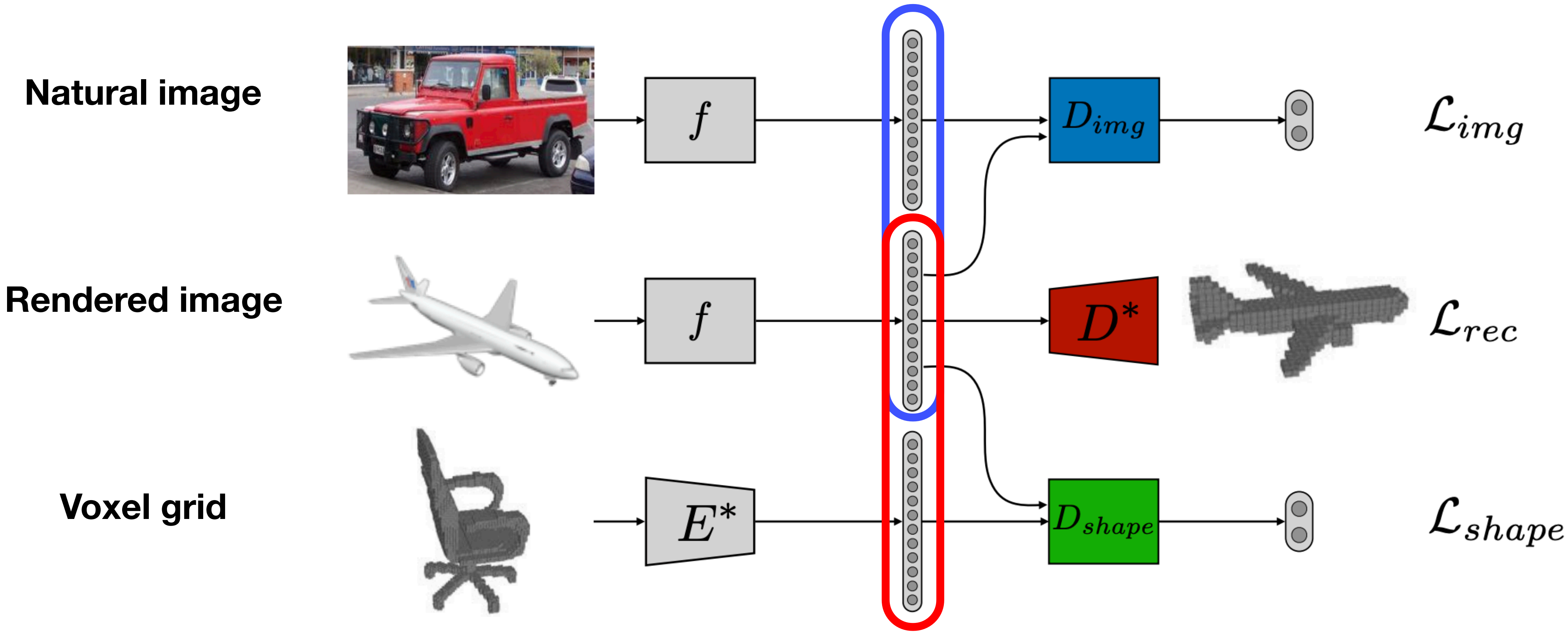


$$\mathcal{L}_{img}(\theta_f, \theta_{img}) = -\mathop{\mathbb{E}}_{x^r \sim p_r} \log D_{img}(f(x^r)) + - \mathop{\mathbb{E}}_{x^n \sim p_n} \log\left(1 - D_{img}(f(x^n))\right)$$

$$\mathcal{L}_{shape}(\theta_f, \theta_{shape}) = -\mathop{\mathbb{E}}_{x^r \sim p_r} \log D_{shape}(f(x^r)) + - \mathop{\mathbb{E}}_{v \sim p_r} \log\left(1 - D_{shape}(E^*(v^r)\right)$$

DANN: Domain-adversarial training of neural networks. Ganin et al, JMLR, 2016

Domain-Adaptive single-view 3D reconstruction,
Pedro Pinheiro, **Negar Rostamzadeh**, Sungjin Ahn, **ICCV 2019**



$$\mathcal{L}_{img}(\theta_f, \theta_{img}) = -\mathop{\mathbb{E}}_{x^r \sim p_r} \log D_{img}(f(x^r)) + - \mathop{\mathbb{E}}_{x^n \sim p_n} \log\,(1 - D_{img}(f(x^n)))$$

$$\mathcal{L}_{shape}(\theta_f, \theta_{shape}) = -\mathop{\mathbb{E}}_{x^r \sim p_r} \log D_{shape}(f(x^r)) + - \mathop{\mathbb{E}}_{v \sim p_r} \log\,(1 - D_{shape}(E^*(v^r))$$

$$\min_{\theta_f} \max_{\theta_{img}, \theta_{shape}} L_{rec}(\theta_f) - \lambda_i L_{img}(\theta_f, \theta_{img}) - \lambda_s L_{shape}(\theta_f, \theta_{shape})$$

DANN: Domain-adversarial training of neural networks. Ganin et al, JMLR, 2016

Domain-Adaptive single-view 3D reconstruction,
Pedro Pinheiro, **Negar Rostamzadeh**, Sungjin Ahn, **ICCV 2019**

DAREC—Analyzing loss terms

			Pix3D	
$\mathcal{L}_{rec}$	$\mathcal{L}_{img}$	$\mathcal{L}_{shape}$	voxel	point cloud
✓			.220	.148
✓		✓	.196	.140
✓	✓		.156	.129
✓	✓	✓	.140	.112

Results measured by Chamfer Distance- CD (lower is better)

$$CD(P1,P2) = \frac{1}{|P_1|} \sum_{x \in P_1} min_{y \in P_2} \| x - y \| + \frac{1}{|P_2|} \sum_{x \in P_2} min_{y \in P_1} \| x - y \|$$

Domain-Adaptive single-view 3D reconstruction,
Pedro Pinheiro, **Negar Rostamzadeh**, Sungjin Ahn, **ICCV 2019**

DAREC—Comparison against the SOTA

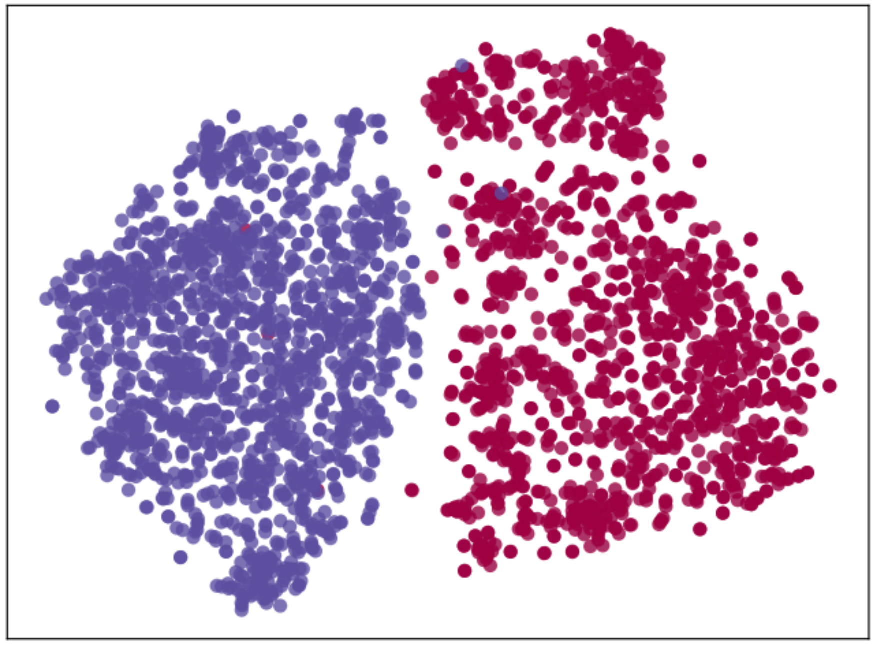
Pix3D dataset	IoU	CD
3D-R2N2 (Choy et al. (2016))	0.136	0.239
3D-VAE-GAN (Wu et al. (2016))	0.171	0.182
PSGN (Fan et al. (2017))	-	0.199
MarrNet (Wu et al. (2017))	0.231	0.144
DRC (Tulsiani et al. (2017))	0.265	0.160
AtlasNet (Groueix et al. (2018))	-	0.126
ShapeHD (Wu et al. (2018))	0.284	0.123
DAREC(ours)	0.237	0.136

IoU (higher is better), CD (lower is better)

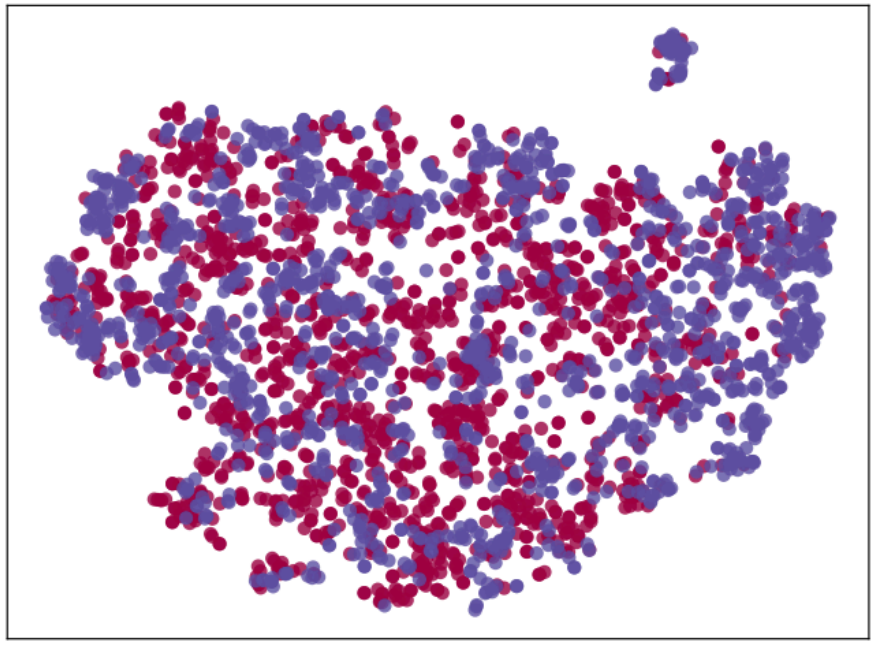
Domain-Adaptive single-view 3D reconstruction,
Pedro Pinheiro, **Negar Rostamzadeh**, Sungjin Ahn, **ICCV 2019**

DAREC — Feature visualization

Before



After

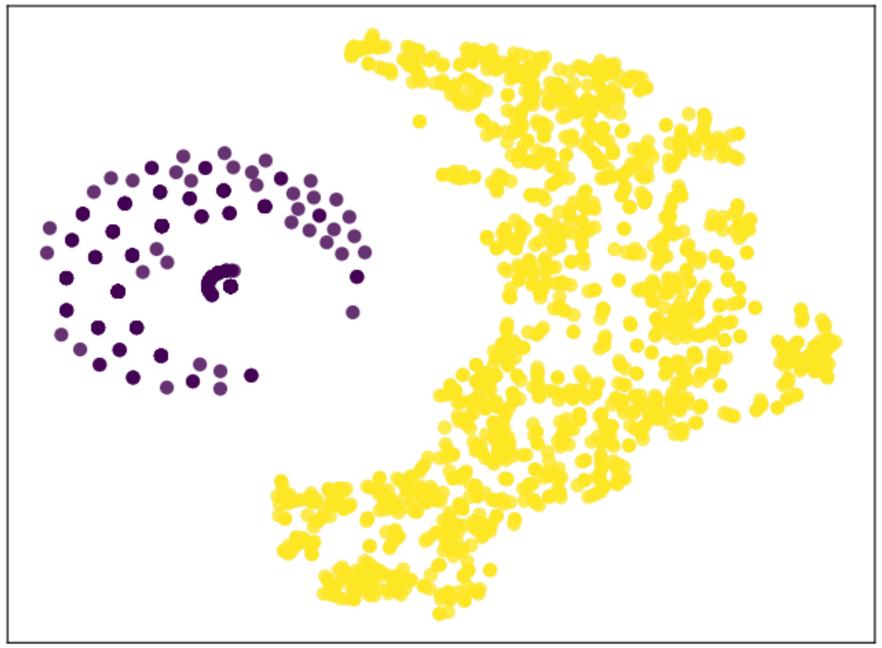


t-SNE visualization of Rendered and Natural images, before and domain confusion

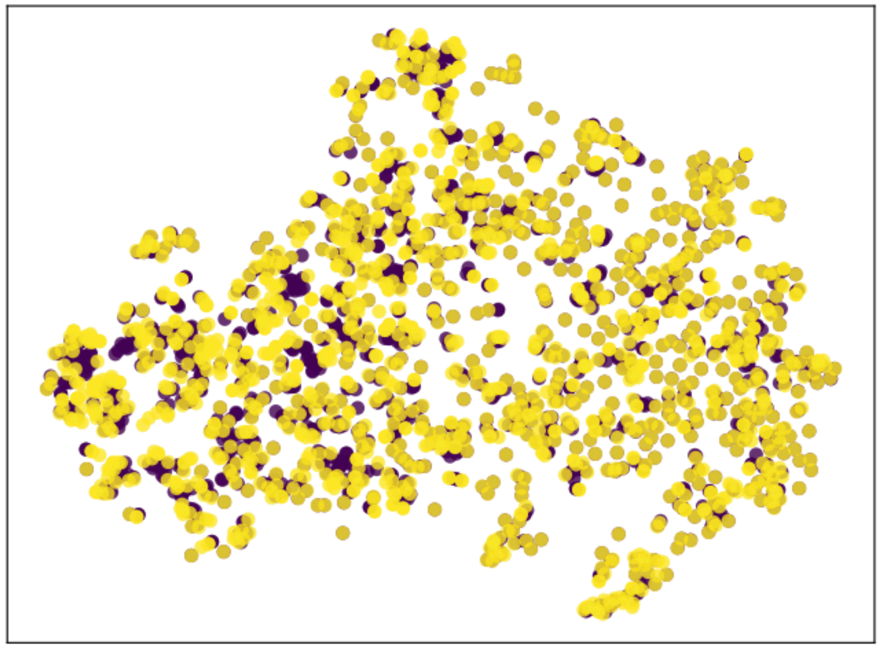
Domain-Adaptive single-view 3D reconstruction,
Pedro Pinheiro, **Negar Rostamzadeh**, Sungjin Ahn, **ICCV 2019**

DAREC—Feature visualization

Before



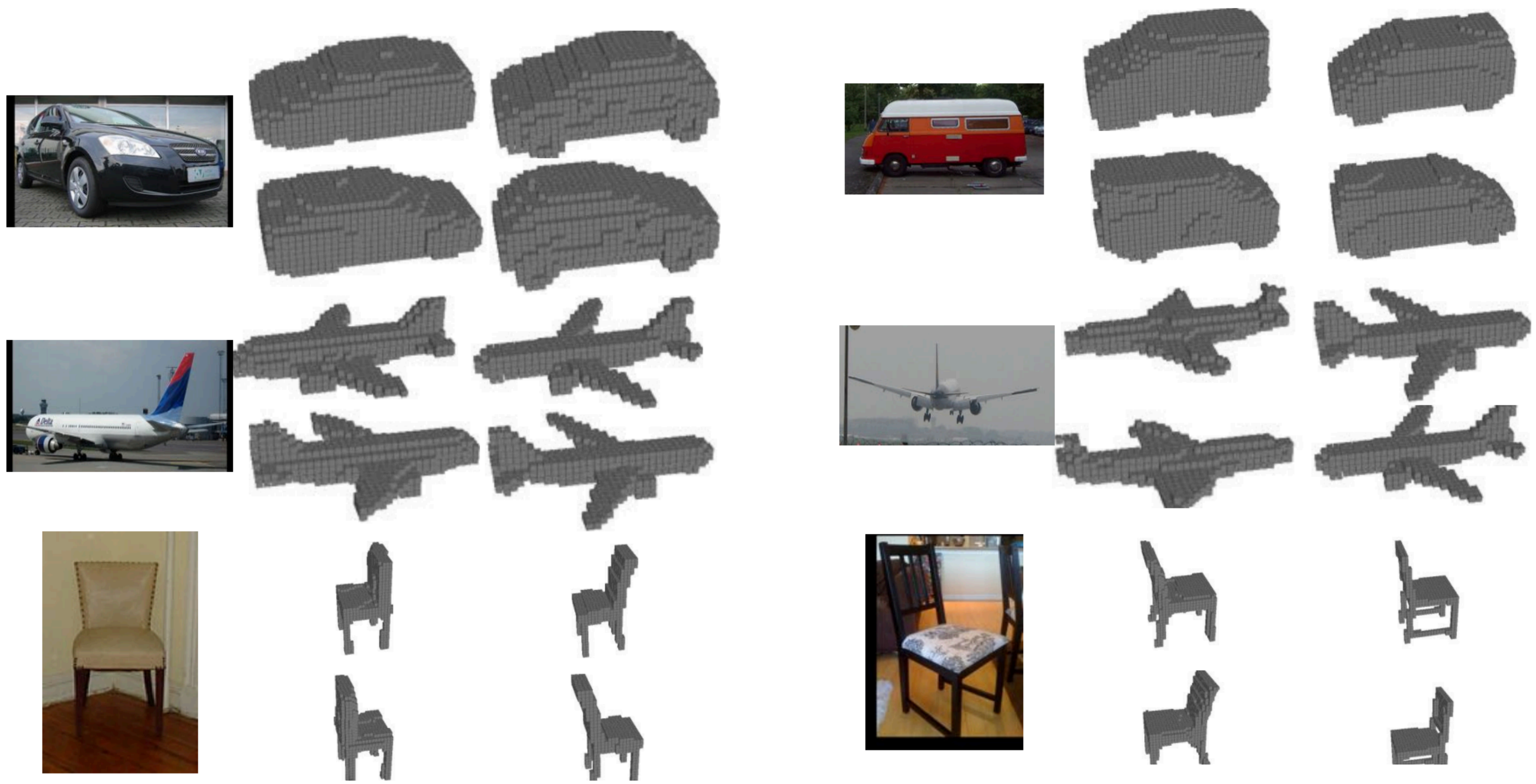
After



t-SNE visualization of 2D rendered embedding and points from shape manifold before and after training

Domain-Adaptive single-view 3D reconstruction,
Pedro Pinheiro, **Negar Rostamzadeh**, Sungjin Ahn, **ICCV 2019**

Conclusion on single-view 3D reconstruction



Domain-Adaptive single-view 3D reconstruction,
Pedro Pinheiro, **Negar Rostamzadeh**, Sungjin Ahn, **ICCV 2019**

Multimodal learning

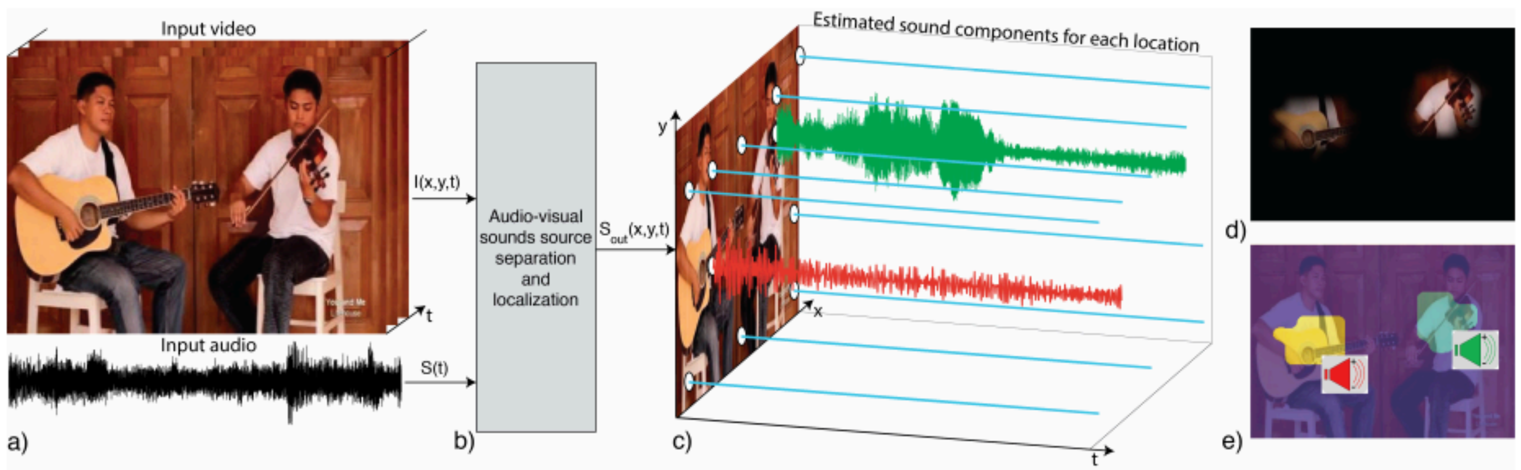
Motivation: human Neuro-psychological studies

- **Degeneracy in neural structure:** Any single function can be carried out by more than one configuration of neural signals and different neural clusters participate in a number of different functions.

- **Edelman's idea of re-entrance:** Even in explicitly unimodal tasks, multiple modalities contribute.

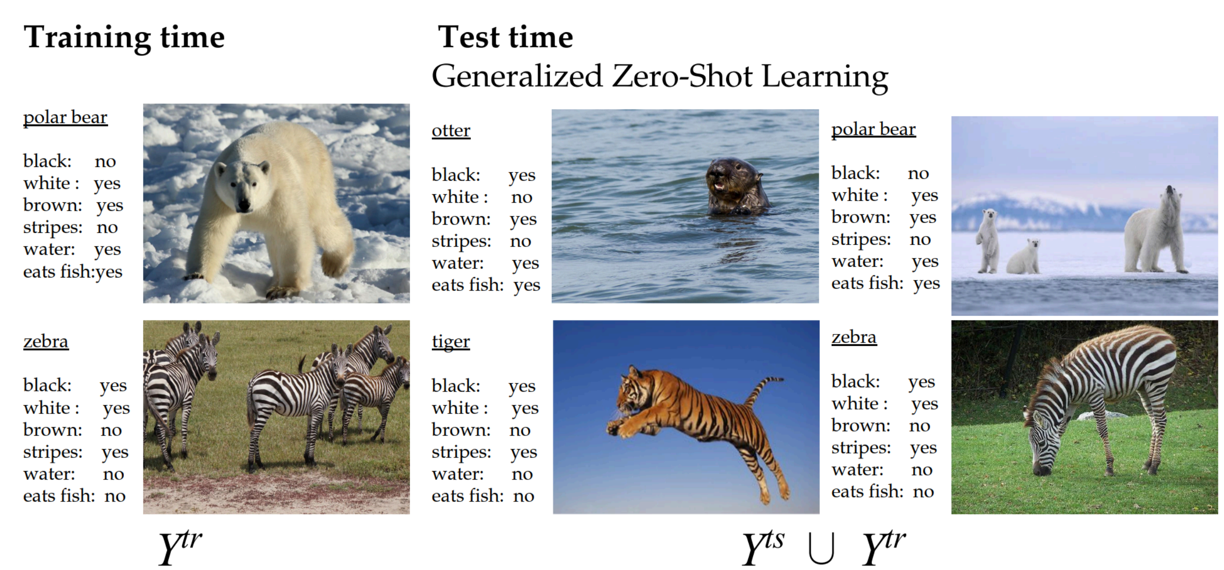
The Development of Embodied Cognition: Six Lessons from Babies, Smith et al

Motivation: available large scale multimodal data



Sounds of the Pixels, Zhao et al

Fashion-Gen dataset and challenge, Rostamzadeh et al.



Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly, Xian et al

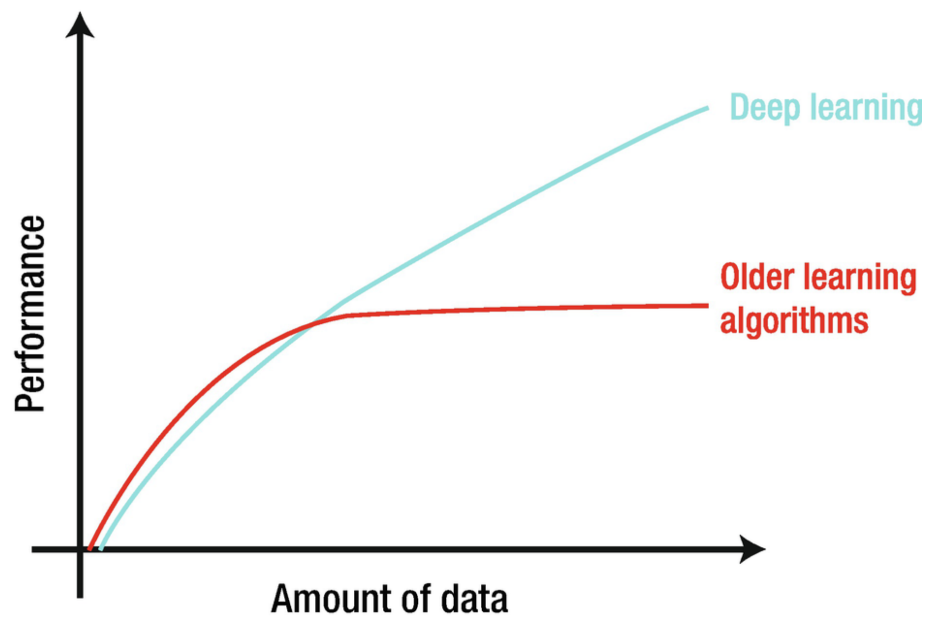
AM3:
Adaptive Cross-Modal Few-Shot Learning

Chen Xing, **Negar Rostamzdeh**, Boris N. Oreshkin, Pedro O. Pinheiro,

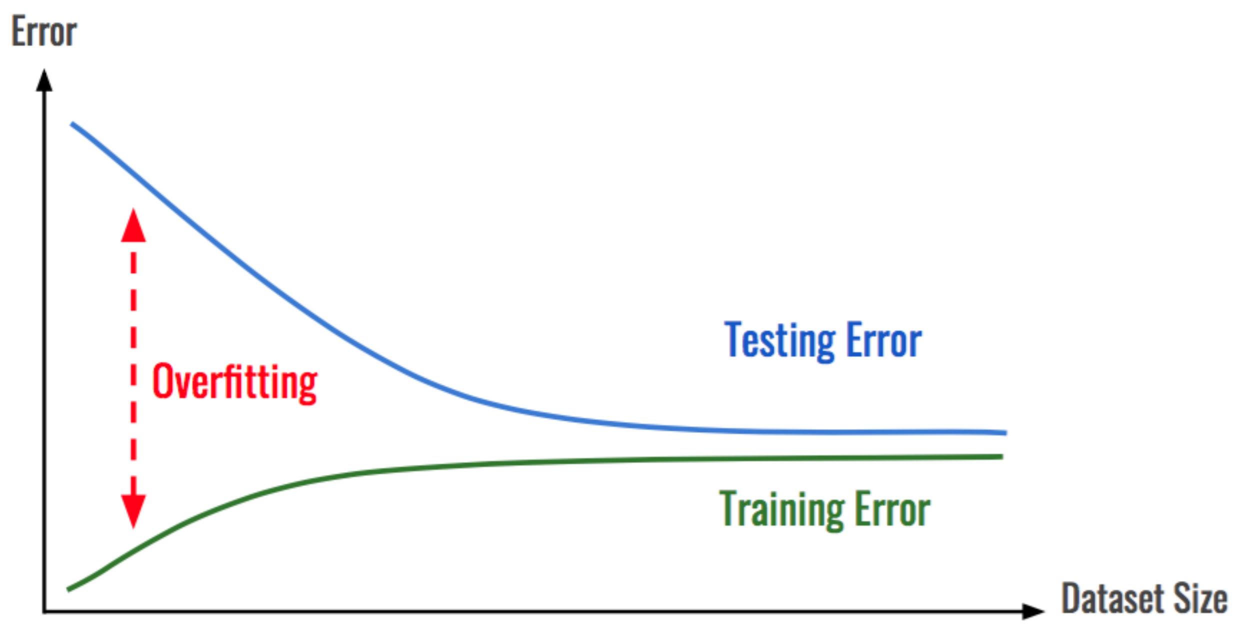
NeurIPS 2019

Deep learning and dataset size

- Deep learning models are data hungry



- Overfitting risk in small data size



Introduction to Natural Language Processing and Deep Learning, Goyal et al.

Image source: https://hackernoon.com/memorizing-is-not-learning-6-tricks-to-prevent-overfitting-in-machine-learning-820b091dc42

Humans are faster learners!

Prior → **Supporting shots** → **Learning**

Dogs

Cat?

A seen dog?

Yorkie
(an unseen breed)

Adaptive Cross-Modal Few-Shot Learning,
Chen Xing, **Negar Rostamzadeh**, Boris N. Oreshkin, Pedro O. Pinheiro, **NeurIPS 2019**

Few-shot classification definition

- **Learning new classes with the help of few samples (shots) per class.**

- **Train and Test sets are disjoint.**

$$\mathcal{C}_{\text{train}} \cap \mathcal{C}_{\text{test}} = \emptyset$$

- **During test, K supporting shots are given for every new class to help classification.**

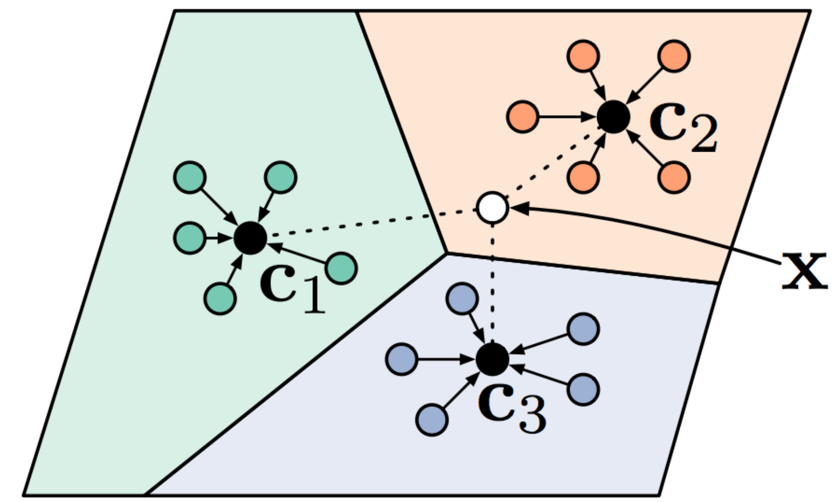
- **Episodic training**

Adaptive Cross-Modal Few-Shot Learning,
Chen Xing, **Negar Rostamzadeh**, Boris N. Oreshkin, Pedro O. Pinheiro, **NeurIPS 2019**

Related work on few shot learning

Metric-based Meta-learning

- Prototypical network (Snell et al)
- TADAM (Oreshkin et al)
- …

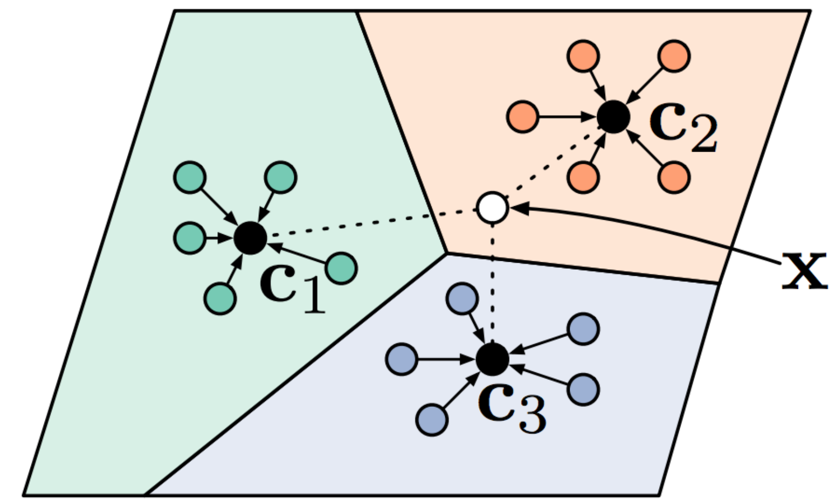


Adaptive Cross-Modal Few-Shot Learning,
Chen Xing, **Negar Rostamzadeh**, Boris N. Oreshkin, Pedro O. Pinheiro, **NeurIPS 2019**

Related work on few shot learning

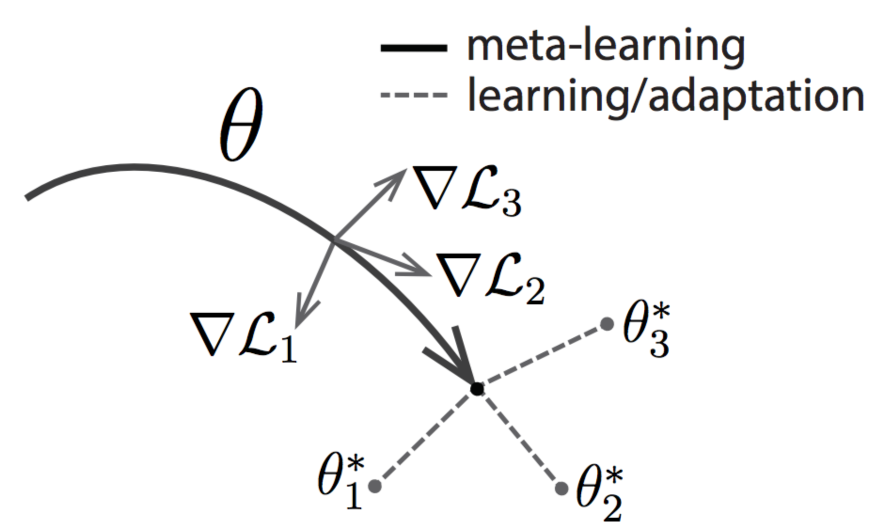
Metric-based Meta-learning

- Prototypical network (Snell et al)
- TADAM (Oreshkin et al)
- …



Gradient-based Meta-learning

- MAML (Finn et al)
- CAML (Zintgraf et al)
- SNAIL (Mishra et al)
- LEO (Rusu et al)
- ….

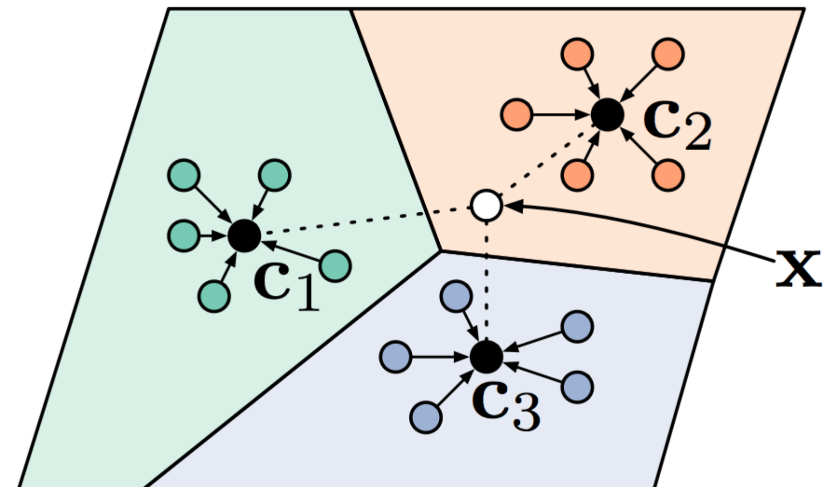


Adaptive Cross-Modal Few-Shot Learning,
Chen Xing, **Negar Rostamzadeh**, Boris N. Oreshkin, Pedro O. Pinheiro, **NeurIPS 2019**

Related work on few shot learning

Metric-based Meta-learning

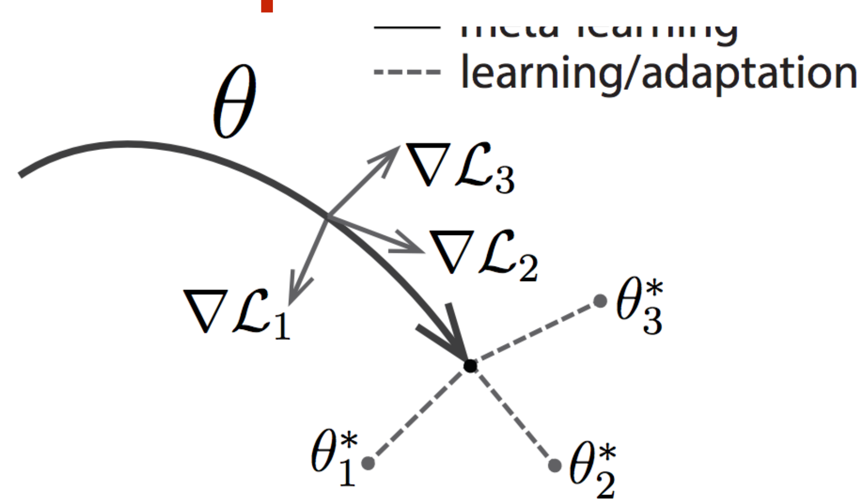
- Prototypical network (Snell et al)
- TADAM (Oreshkin et al)
- …



Exploiting language semantic structure in few-shot image classification is not explored.

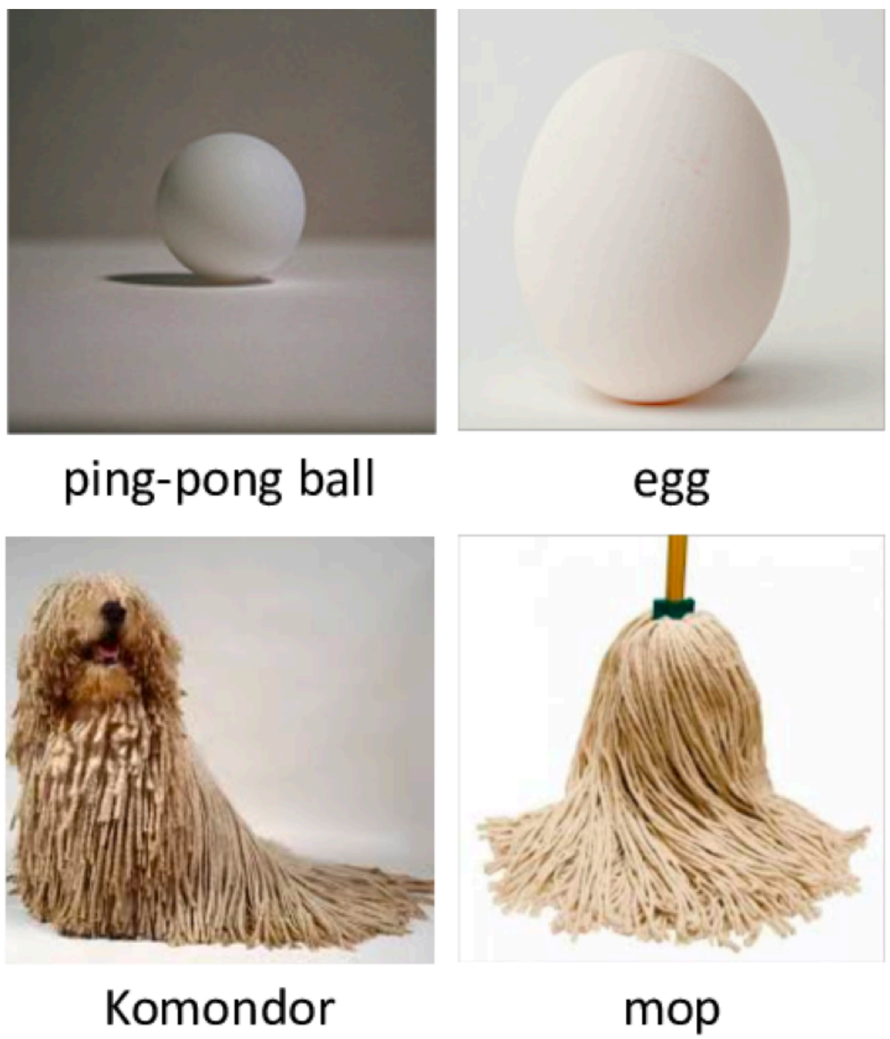
Gradient-based Meta-learning

- MAML (Finn et al)
- CAML (Zintgraf et al)
- SNAIL (Mishra et al)
- LEO (Rusu et al)
- ….



Adaptive Cross-Modal Few-Shot Learning,
Chen Xing, **Negar Rostamzadeh**, Boris N. Oreshkin, Pedro O. Pinheiro, **NeurIPS 2019**

Language semantics information can be orthogonal to visual information



Visually close, semantically different

Visually different, semantically close

Adaptive Cross-Modal Few-Shot Learning,
Chen Xing, **Negar Rostamzadeh**, Boris N. Oreshkin, Pedro O. Pinheiro, **NeurIPS 2019**

AM3—Preliminaries: Episodic Training

- **Episodic training mimics the test scenario.**
- **Models are trained on K-shot, N-way episodes.**
- **For a random sampled episode e:**

  - *Support* Set $\mathcal{S}_e = \{(s_i, y_i)\}_{i=1}^{N \times K}$ contains K samples of N categories.

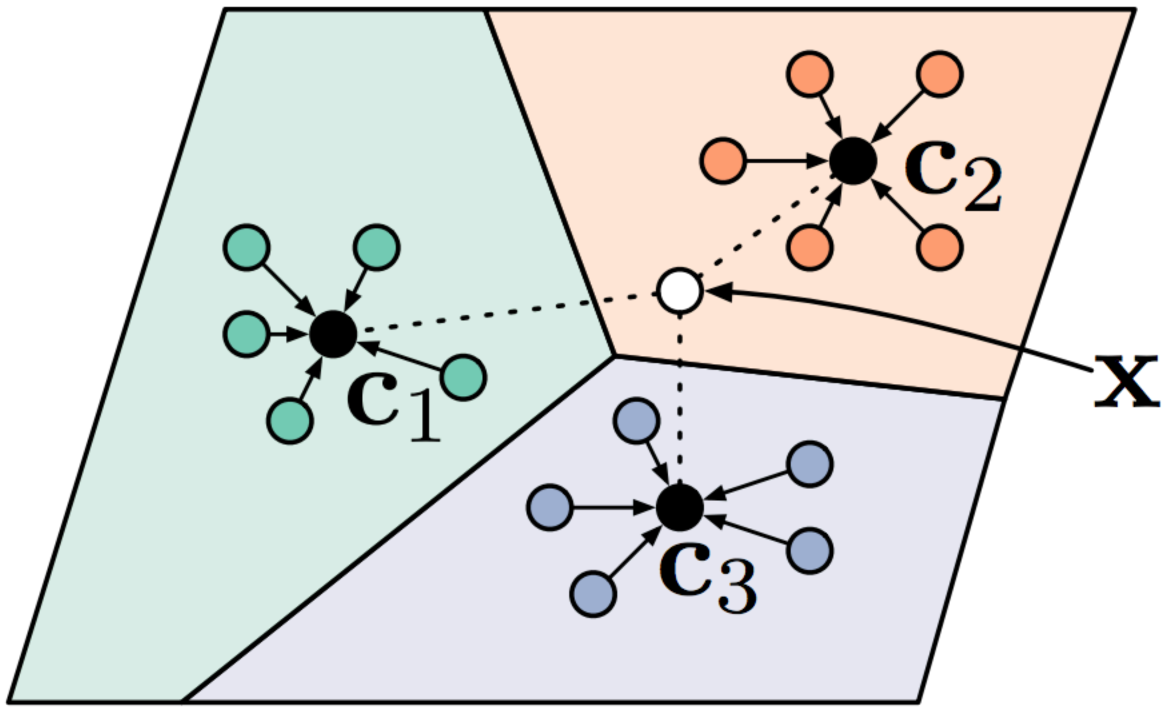
  - *Query* Set $\mathcal{Q}_e = \{(q_j, y_j)\}_{j=1}^{Q}$ contains samples from N categories.

- **Episode Loss:**

$$\mathcal{L}(\theta) = \mathop{\mathbb{E}}_{(\mathcal{S}_e, \mathcal{Q}_e)} - \sum_{t=1}^{Q} \log p_\theta(y_t | q_t, \mathcal{S}_e)$$

Adaptive Cross-Modal Few-Shot Learning,
Chen Xing, **Negar Rostamzadeh**, Boris N. Oreshkin, Pedro O. Pinheiro, **NeurIPS 2019**

AM3—Preliminaries: Prototypical Nets



- **For each category c in episode e, support set -> centroid (prototype)**

$$\mathbf{p}_c = \frac{1}{|S_e^c|} \sum_{(s_i, y_i) \in \mathcal{S}_e^c} f(s_i)$$

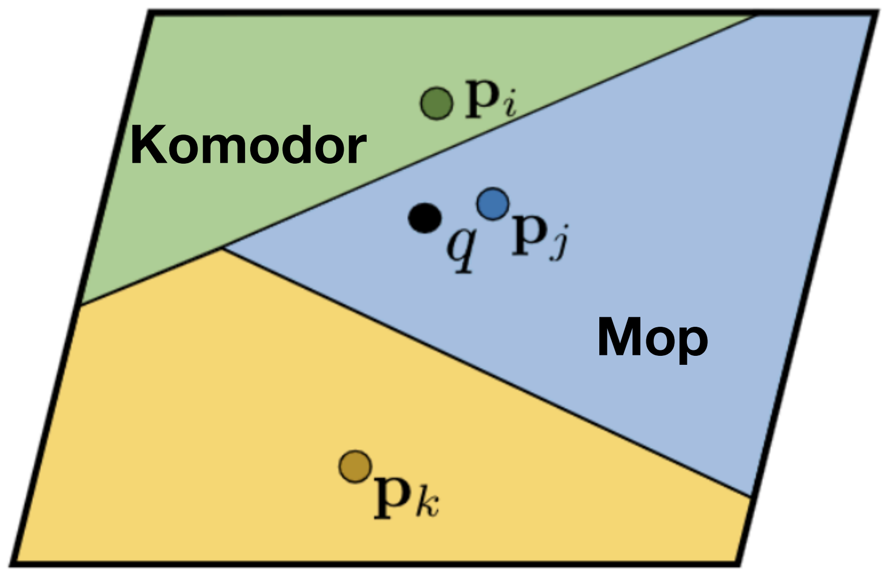
- **Embedded query points are classified via a softmax over negative distances to class prototypes**

$$p(y = c | q_t, S_e, \theta) = \frac{\exp(-d(f(q_t), \mathbf{p}_c))}{\sum_k \exp(-d(f(q_t), \mathbf{p}_k))}$$

Adaptive Cross-Modal Few-Shot Learning,
Chen Xing, **Negar Rostamzadeh**, Boris N. Oreshkin, Pedro O. Pinheiro, **NeurIPS 2019**

AM3: komodor or a mop?

Visual info only

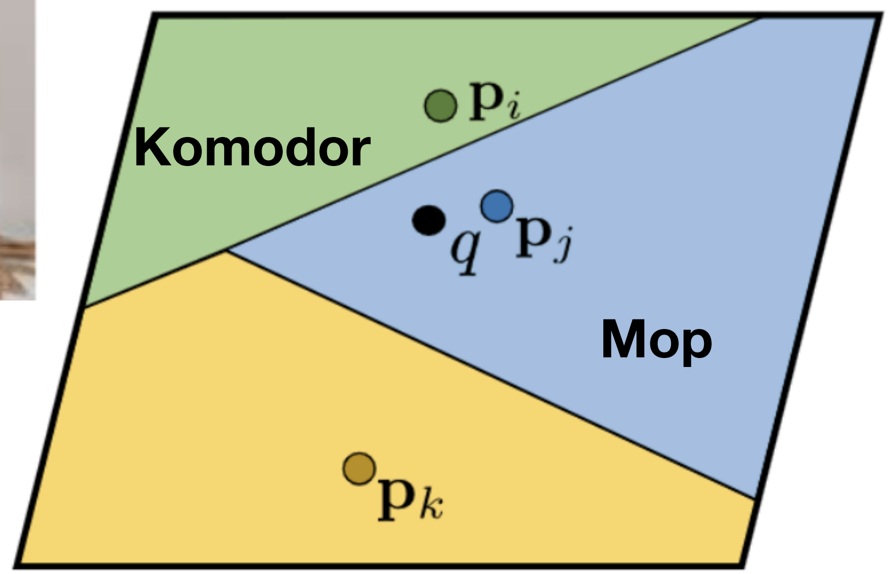


Adaptive Cross-Modal Few-Shot Learning,
Chen Xing, **Negar Rostamzadeh**, Boris N. Oreshkin, Pedro O. Pinheiro, **NeurIPS 2019**

AM3: komodor or a mop?

This should be a mop!

Visual info only

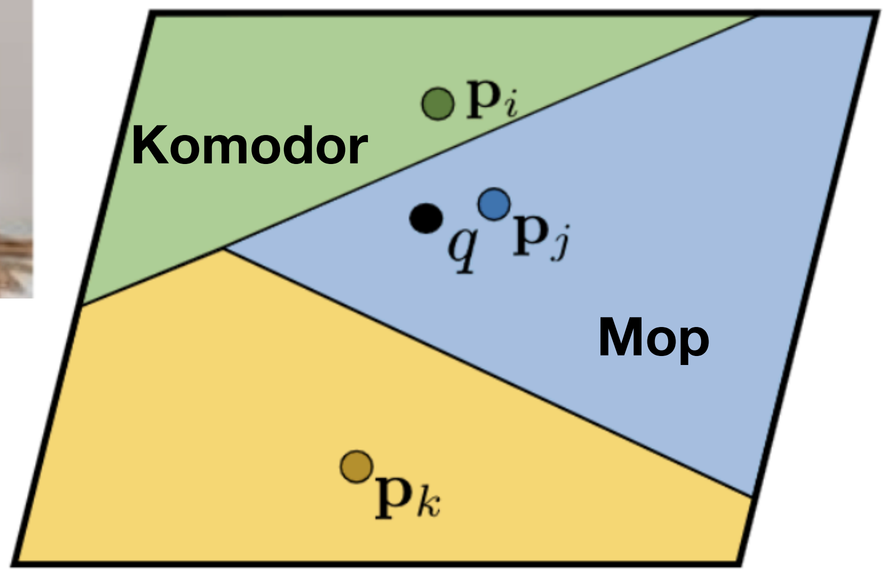


Adaptive Cross-Modal Few-Shot Learning,
Chen Xing, **Negar Rostamzadeh**, Boris N. Oreshkin, Pedro O. Pinheiro, **NeurIPS 2019**

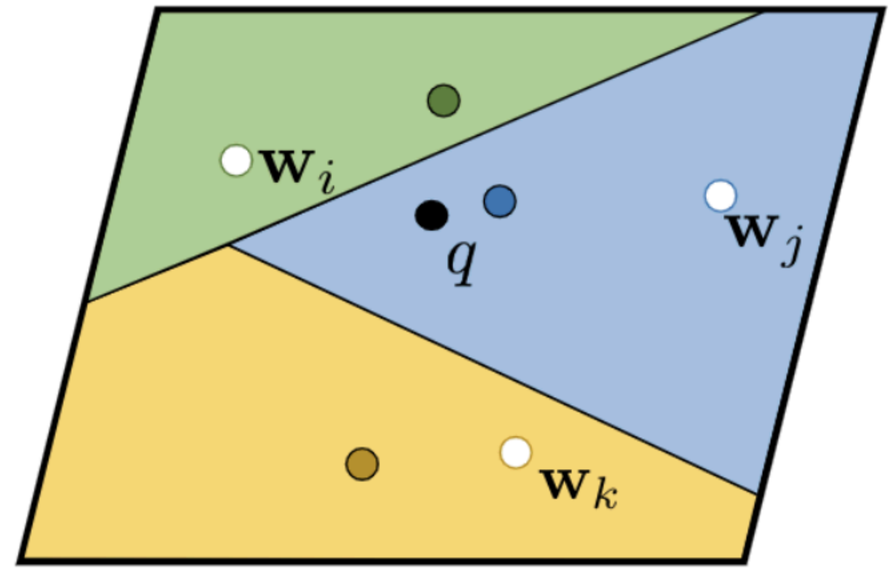
AM3: komodor or a mop?

This should be a mop!

Visual info only



$$\mathbf{p}'_c = \lambda_c \cdot \mathbf{p}_c + (1 - \lambda_c) \cdot \mathbf{w}_c$$

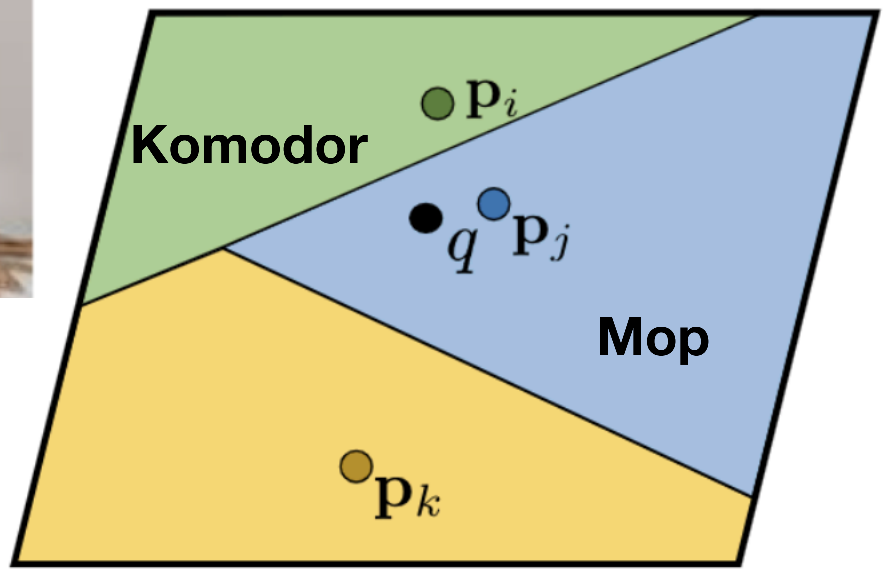


Adaptive Cross-Modal Few-Shot Learning,
Chen Xing, **Negar Rostamzadeh**, Boris N. Oreshkin, Pedro O. Pinheiro, **NeurIPS 2019**

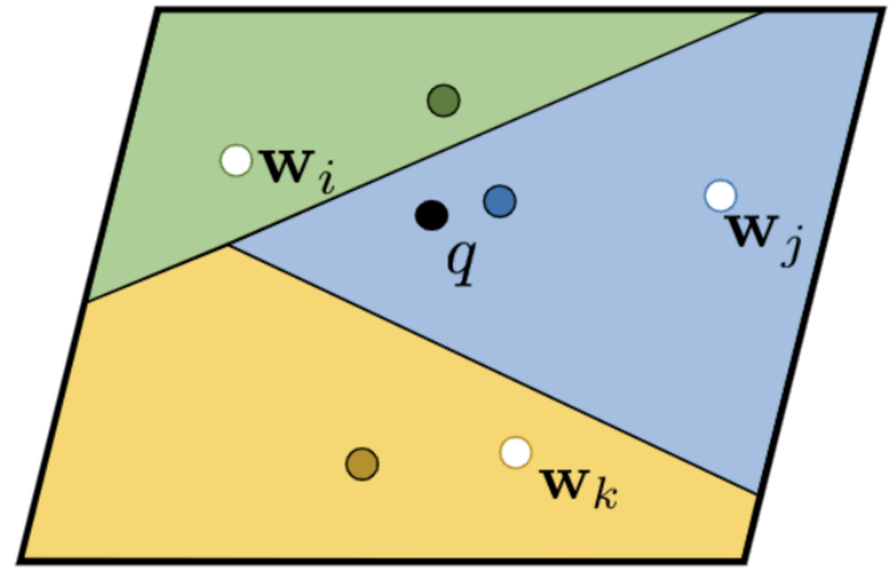
AM3: komodor or a mop?

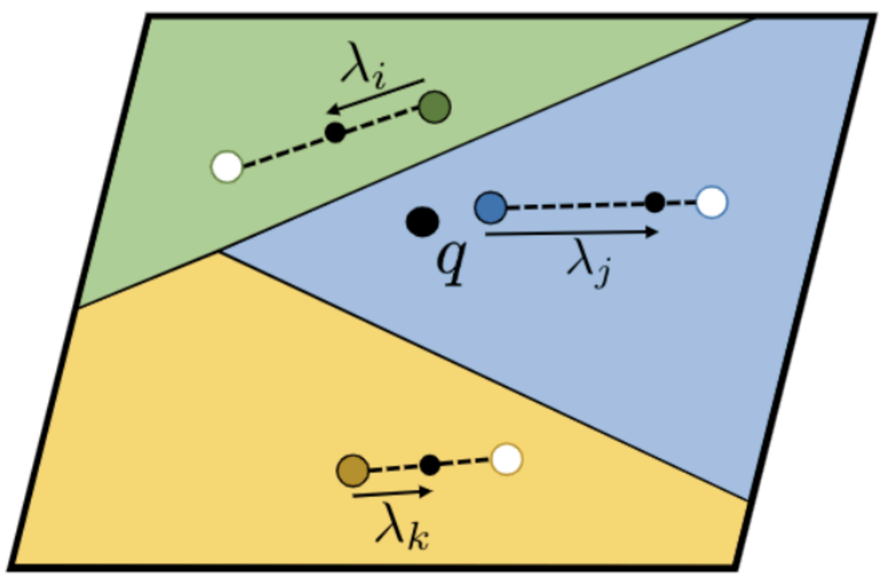
This should be a mop!

Visual info only



$$\mathbf{p}'_c = \lambda_c \cdot \mathbf{p}_c + (1 - \lambda_c) \cdot \mathbf{w}_c$$





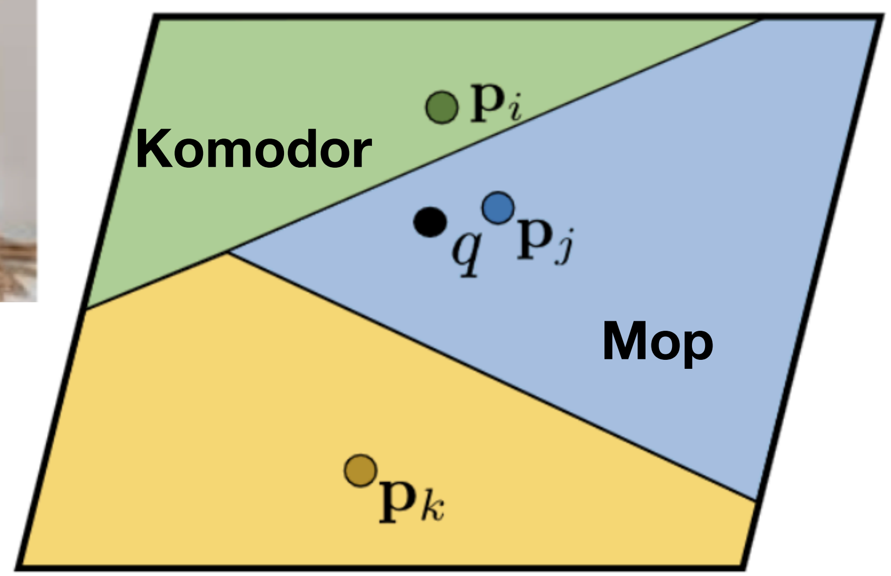
Adaptive Cross-Modal Few-Shot Learning,
Chen Xing, **Negar Rostamzadeh**, Boris N. Oreshkin, Pedro O. Pinheiro, **NeurIPS 2019**

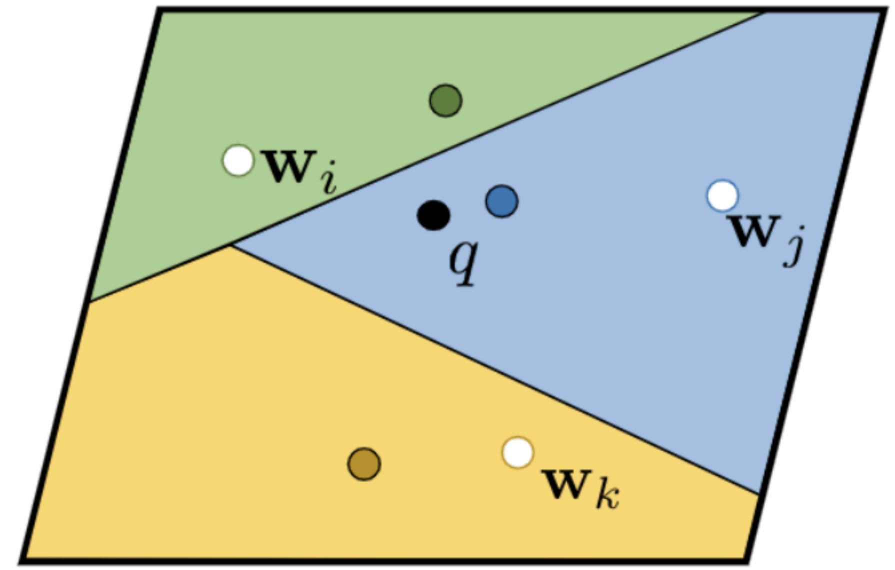
AM3: komodor or a mop?

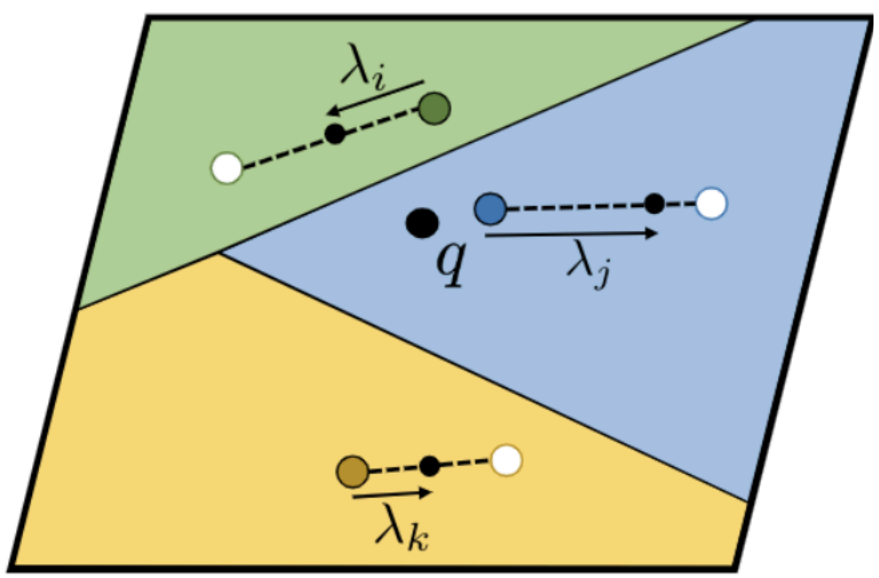
This should be a mop!

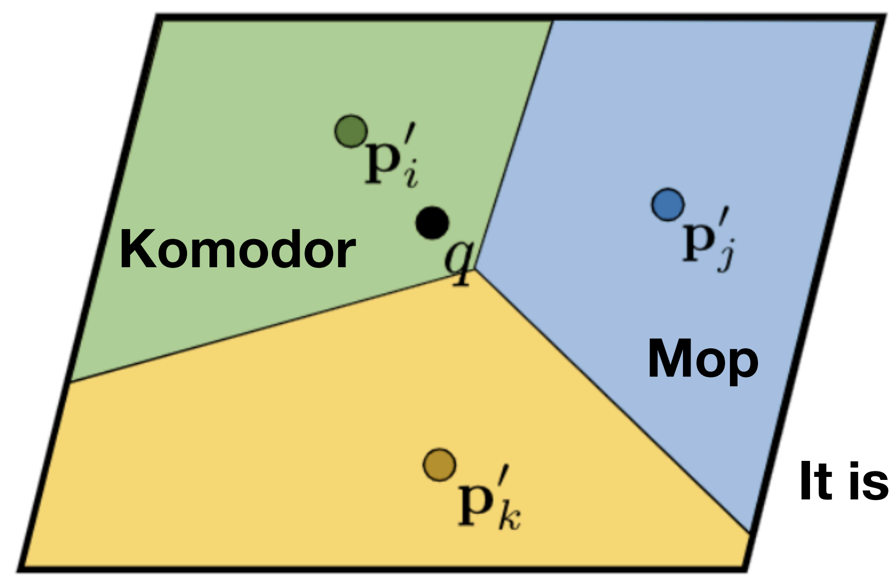
Visual info only

$$\mathbf{p}'_c = \lambda_c \cdot \mathbf{p}_c + (1 - \lambda_c) \cdot \mathbf{w}_c$$







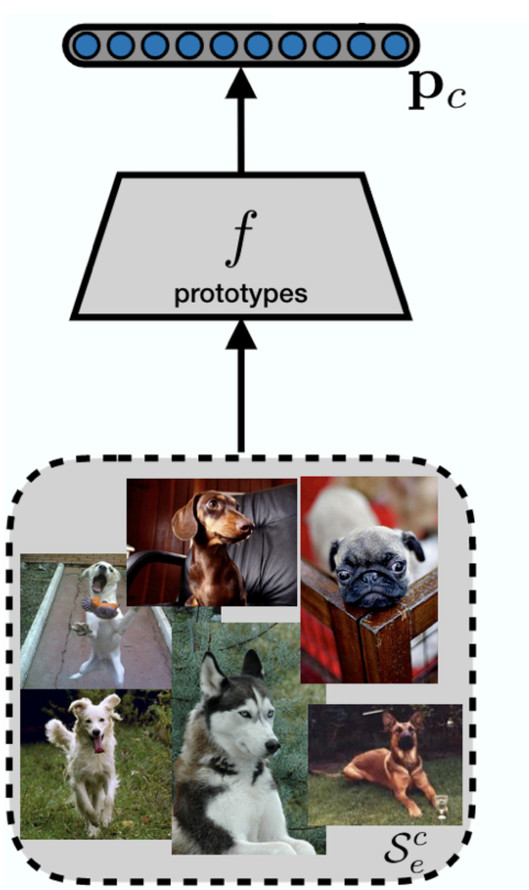


It is a Komondor!

Visual + Semantic

Adaptive Cross-Modal Few-Shot Learning,
Chen Xing, **Negar Rostamzadeh**, Boris N. Oreshkin, Pedro O. Pinheiro, **NeurIPS 2019**

AM3: Adaptive Modality Mixture Model

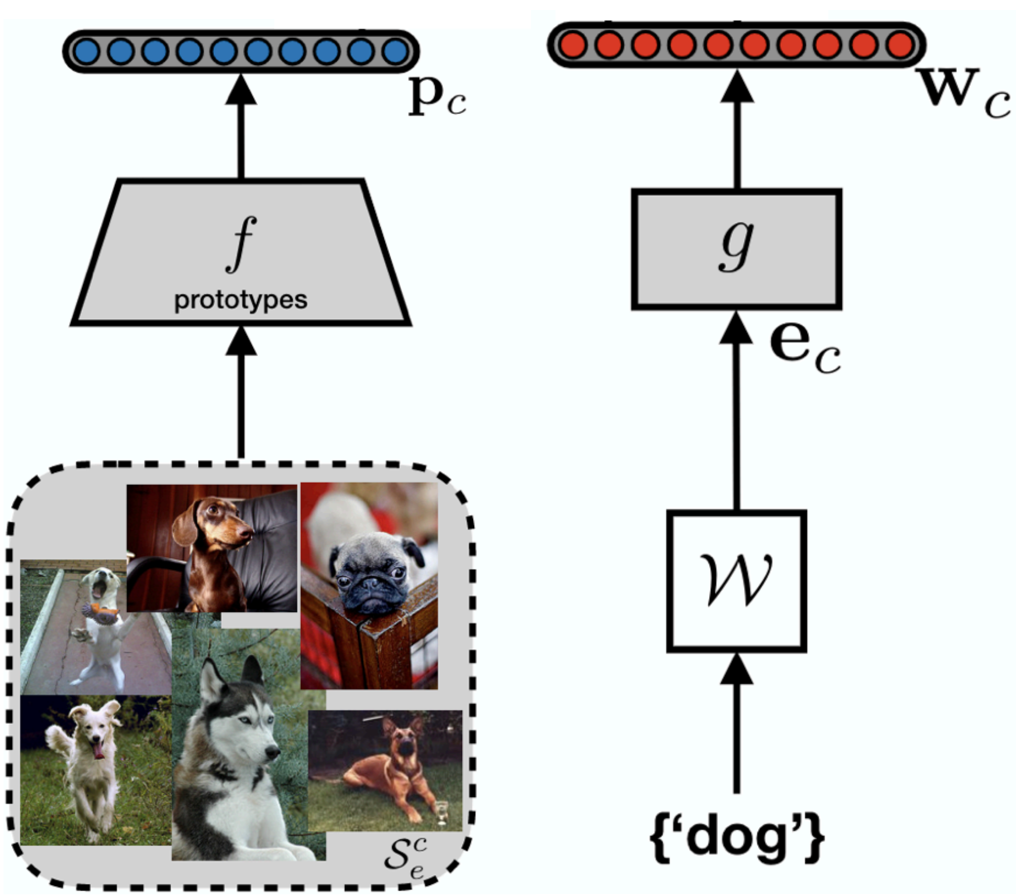


Adaptive Cross-Modal Few-Shot Learning,
Chen Xing, **Negar Rostamzadeh**, Boris N. Oreshkin, Pedro O. Pinheiro, **NeurIPS 2019**

AM3: Adaptive Modality Mixture Model

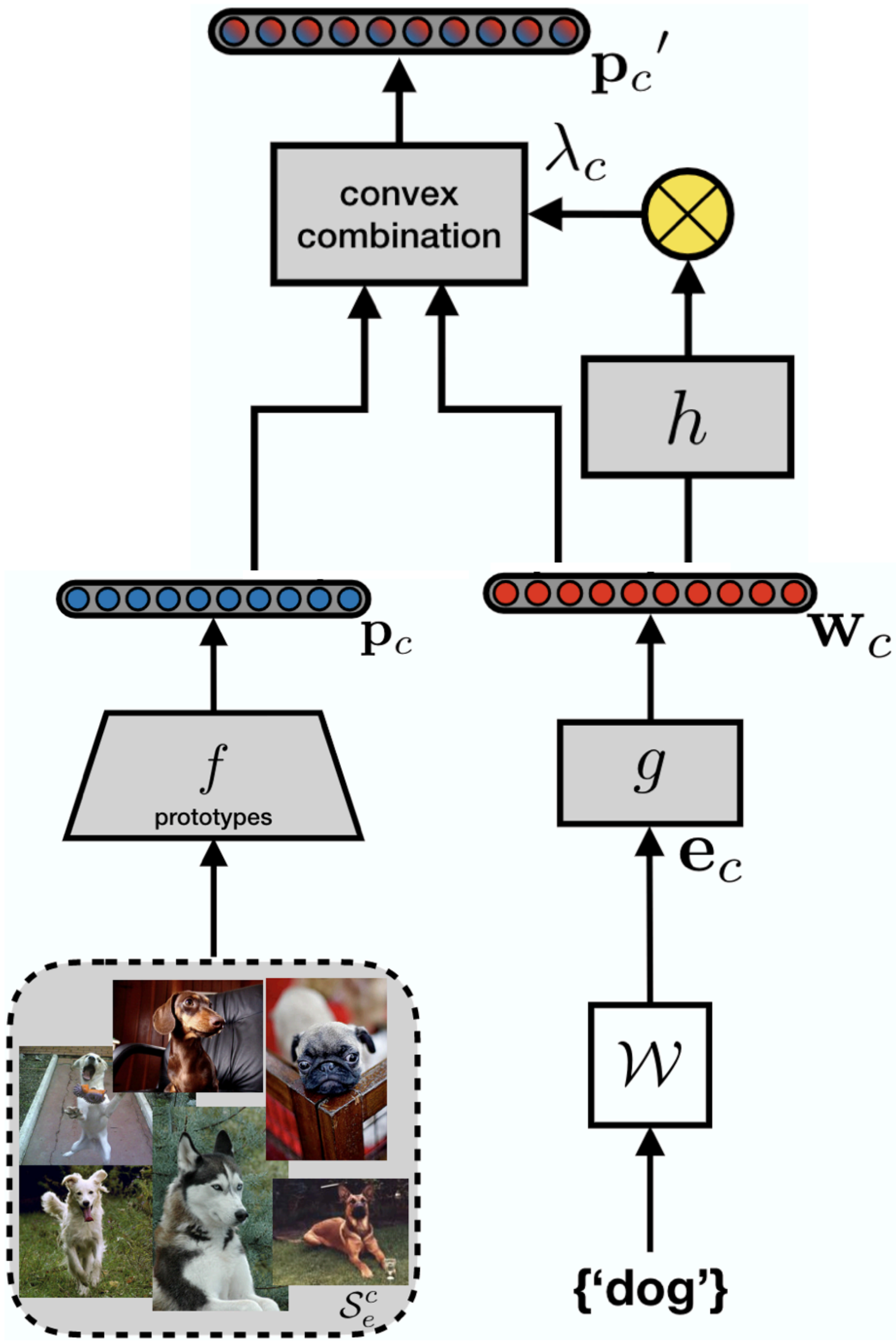
- e_c is the label embedding for category c pre-trained on unsupervised large text corpora

- $w_c = g(e_c)$ is a transformed version of the label embedding for category c



Adaptive Cross-Modal Few-Shot Learning,
Chen Xing, **Negar Rostamzadeh**, Boris N. Oreshkin, Pedro O. Pinheiro, **NeurIPS 2019**

AM3: Adaptive Modality Mixture Model



- e_c is the label embedding for category c pre-trained on unsupervised large text corpora
- $w_c = g(e_c)$ is a transformed version of the label embedding for category c
- h is the adaptive mixing network with parameters θ_h
- λ_c is calculated *w.r.t.* the transformed label embedding

$$\lambda_c = \frac{1}{1 + \exp(-h(\mathbf{w}_c))}$$

$$\mathbf{p}'_c = \lambda_c \cdot \mathbf{p}_c + (1 - \lambda_c) \cdot \mathbf{w}_c$$

Adaptive Cross-Modal Few-Shot Learning,
Chen Xing, **Negar Rostamzadeh**, Boris N. Oreshkin, Pedro O. Pinheiro, **NeurIPS 2019**

AM3: Comparison to the SOTA

Model	Test Accuracy		
	5-way 1-shot	5-way 5-shot	5-way 10-shot
Uni-modality few-shot learning baselines			
Matching Network (Vinyals et al., 2016)	$43.56 \pm 0.84\%$	$55.31 \pm 0.73\%$	-
Prototypical Network (Snell et al., 2017)	$49.42 \pm 0.78\%$	$68.20 \pm 0.66\%$	$74.30 \pm 0.52\%$
Discriminative k-shot (Bauer et al., 2017)	$56.30 \pm 0.40\%$	$73.90 \pm 0.30\%$	$78.50 \pm 0.00\%$
Meta-Learner LSTM (Ravi & Larochelle, 2017)	$43.44 \pm 0.77\%$	$60.60 \pm 0.71\%$	-
MAML (Finn et al., 2017)	$48.70 \pm 1.84\%$	$63.11 \pm 0.92\%$	-
ProtoNets w Soft k-Means (Ren et al., 2018)	$50.41 \pm 0.31\%$	$69.88 \pm 0.20\%$	-
SNAIL (Mishra et al., 2018)	$55.71 \pm 0.99\%$	$68.80 \pm 0.92\%$	-
CAML (Jiang et al., 2019)	$59.23 \pm 0.99\%$	$72.35 \pm 0.71\%$	-
LEO (Rusu et al., 2019)	$61.76 \pm 0.08\%$	$77.59 \pm 0.12\%$	-
Modality alignment baselines			
DeViSE (Frome et al., 2013)	$37.43\pm0.42\%$	$59.82\pm0.39\%$	$66.50\pm0.28\%$
ReViSE (Hubert Tsai et al., 2017)	$43.20\pm0.87\%$	$66.53\pm0.68\%$	$72.60\pm0.66\%$
CBPL (Lu et al., 2018)	$58.50\pm0.82\%$	$75.62\pm0.61\%$	-
f-CLSWGAN (Xian et al., 2018)	$53.29\pm0.82\%$	$72.58\pm0.27\%$	$73.49\pm0.29\%$
CADA-VAE (Schönfeld et al., 2018)	$58.92\pm1.36\%$	$73.46\pm1.08\%$	$76.83\pm0.98\%$
Modality alignment baselines extended to metric-based FSL framework			
DeViSE-FSL	$56.99 \pm 1.33\%$	$72.63 \pm 0.72\%$	$76.70 \pm 0.53\%$
ReViSE-FSL	$57.23 \pm 0.76\%$	$73.85 \pm 0.63\%$	$77.21 \pm 0.31\%$
f-CLSWGAN-FSL	$58.47 \pm 0.71\%$	$72.23 \pm 0.45\%$	$76.90 \pm 0.38\%$
CADA-VAE-FSL	$61.59 \pm 0.84\%$	$75.63 \pm 0.52\%$	$79.57 \pm 0.28\%$
AM3 and its backbones			
ProtoNets++	$56.52 \pm 0.45\%$	$74.28 \pm 0.20\%$	$78.31 \pm 0.44\%$
AM3-ProtoNets++	$65.21 \pm 0.30\%$	$75.20 \pm 0.27\%$	$78.52 \pm 0.28\%$
TADAM (Oreshkin et al., 2018)	$58.56 \pm 0.39\%$	$76.65 \pm 0.38\%$	$80.83 \pm 0.37\%$
AM3-TADAM	$\mathbf{65.30 \pm 0.49\%}$	$\mathbf{78.10 \pm 0.36\%}$	$\mathbf{81.57 \pm 0.47\ \%}$

Conclusion on AM3

ping-pong ball

egg

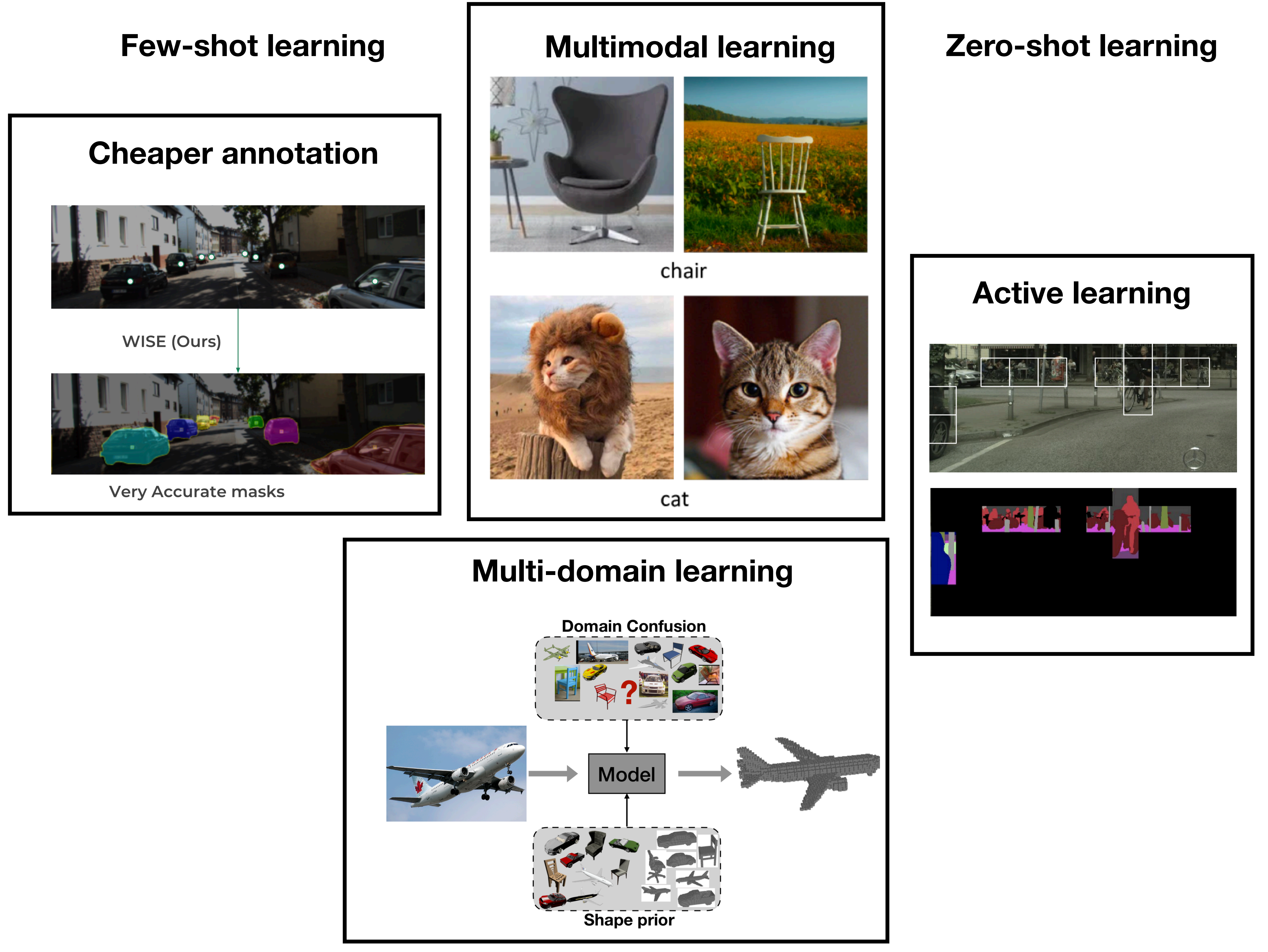
chair

Komondor

mop

cat

Adaptive Cross-Modal Few-Shot Learning,
Chen Xing, **Negar Rostamzadeh**, Boris N. Oreshkin, Pedro O. Pinheiro, **NeurIPS 2019**



Few-shot learning
Multimodal learning
Zero-shot learning
Cheaper annotation
WISE (Ours)
Very Accurate masks
chair
cat
Active learning
Multi-domain learning
Domain Confusion
Model
Shape prior

Check out this paper in the main conference, presented by Arantxa Casanova

Reinforced Active Learning for Semantic Segmentation

Arantxa Casanova, Pedro O. Pinheiro, Negar Rostamzdeh, Chris Pal

Thanks to all my co-authors!

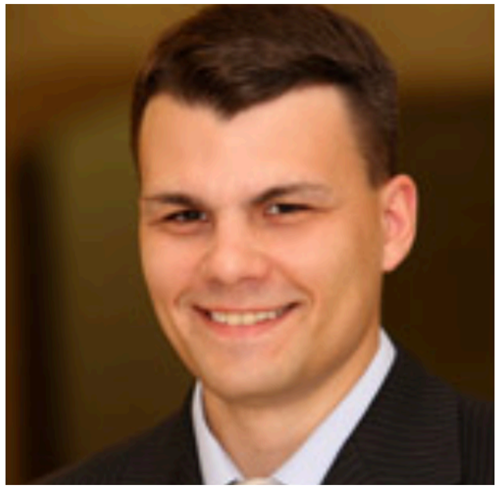
Pedro Pinheiro

Sugjin Ahn

Chen Xing

Arantxa Casanova

Chris Pal



Boris Oreshkin

Issam Laradji

David Vazquez

Mark Schmidt

Thanks for listening to me!