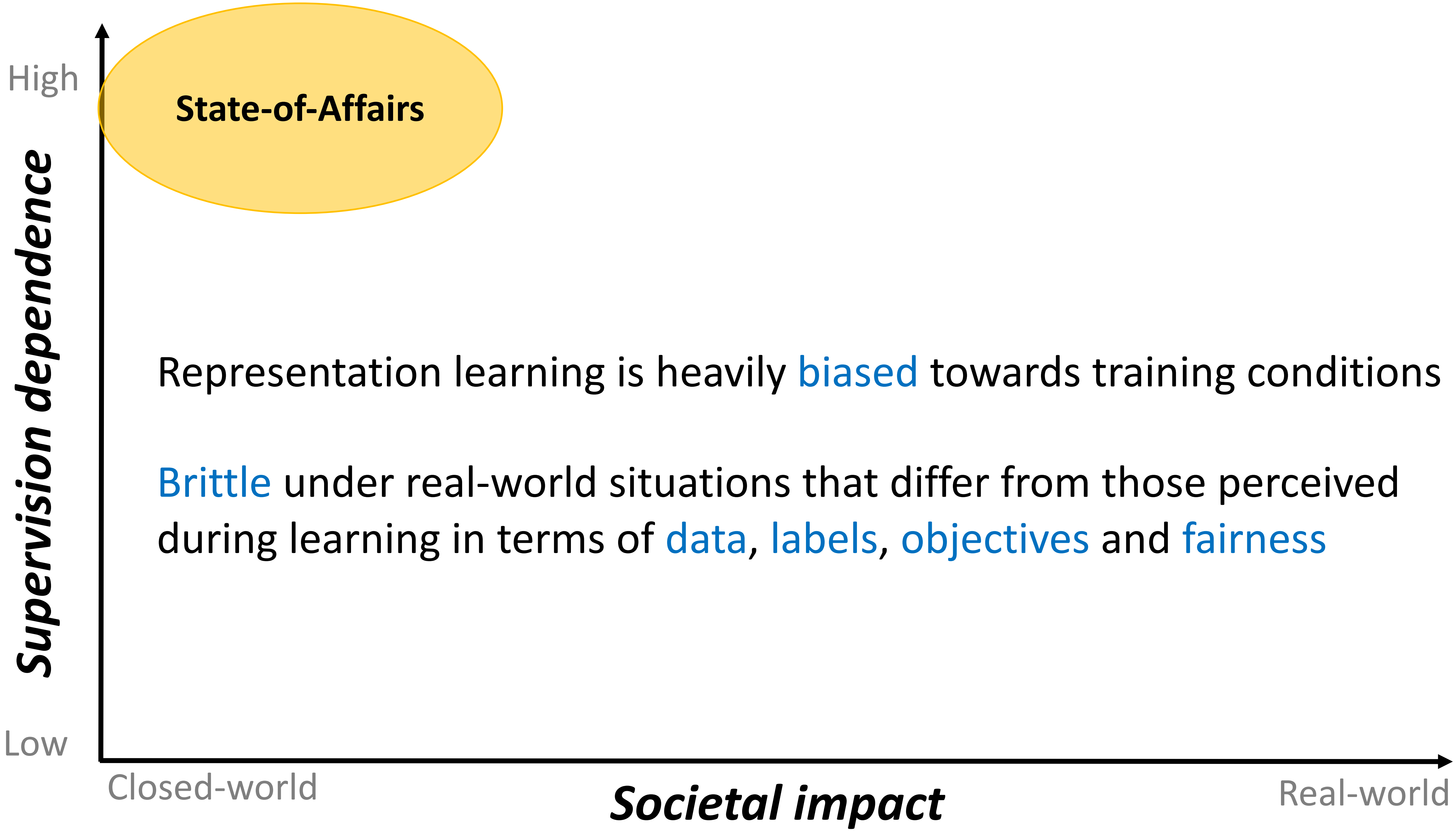
elise
European Network of AI Excellence Centres

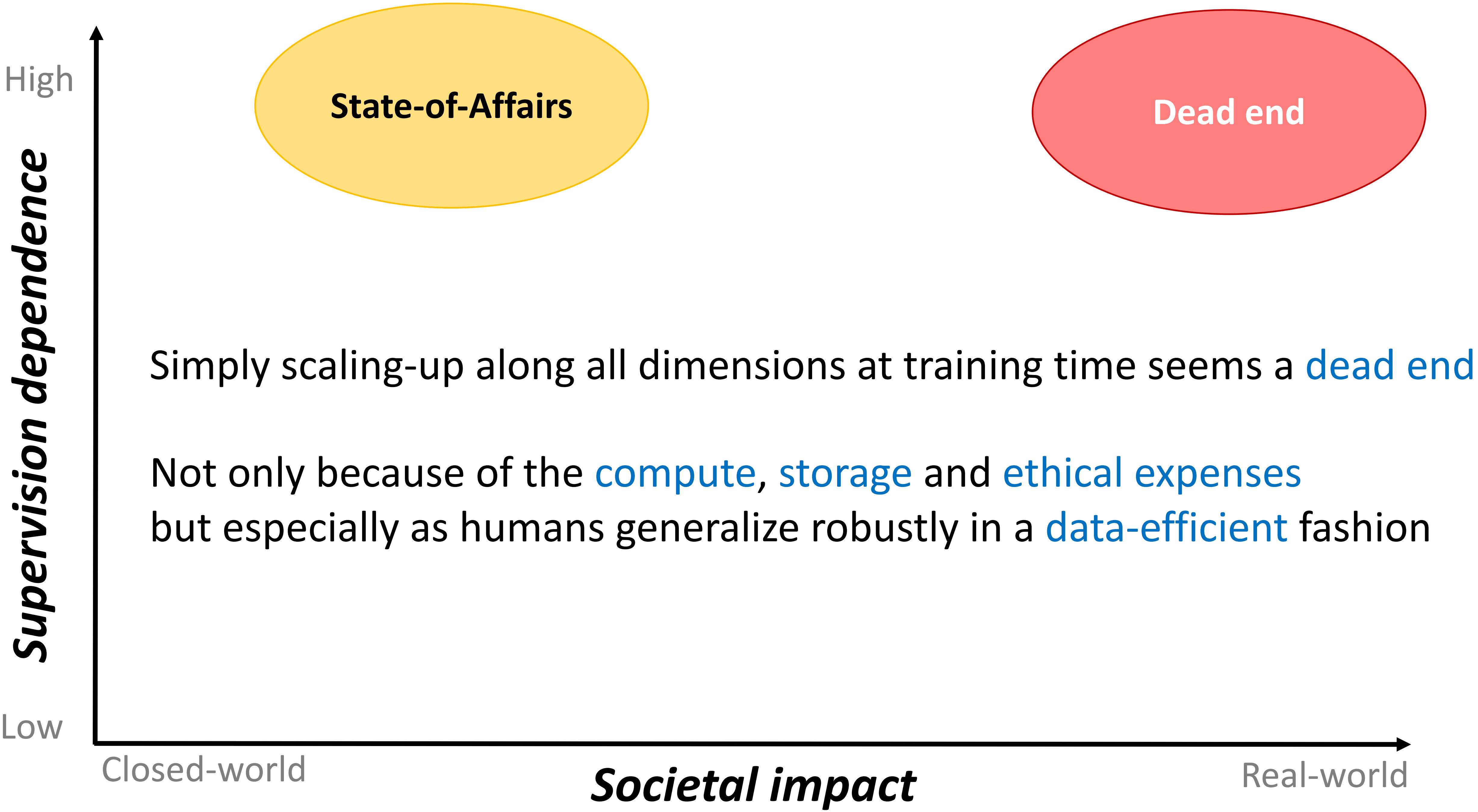
Real-World Learning

Prof. dr. Cees Snoek

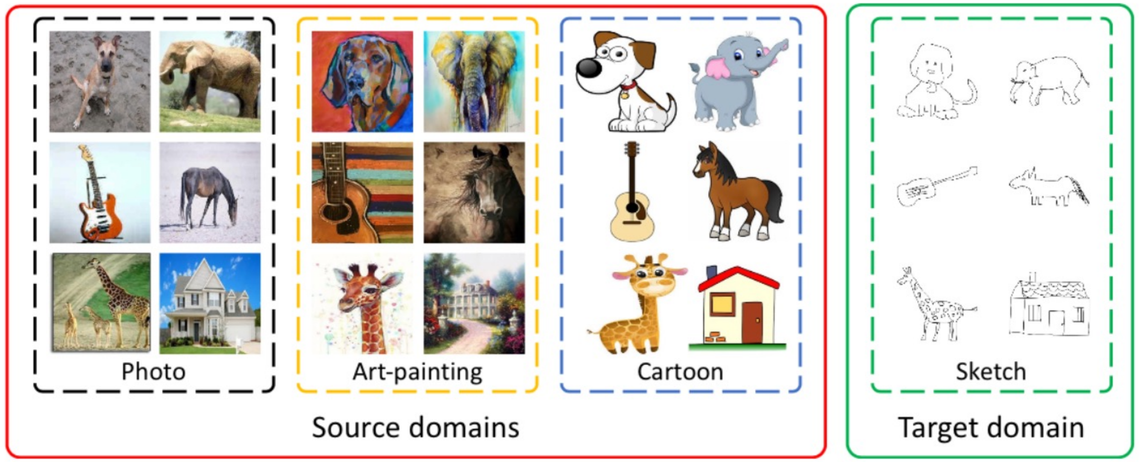
Video & Image Sense Lab

Awesome learning


High
State-of-Affairs
Supervision dependence
Representation learning is heavily biased towards training conditions
Brittle under real-world situations that differ from those perceived during learning in terms of data, labels, objectives and fairness
Low
Closed-world
Societal impact
Real-world

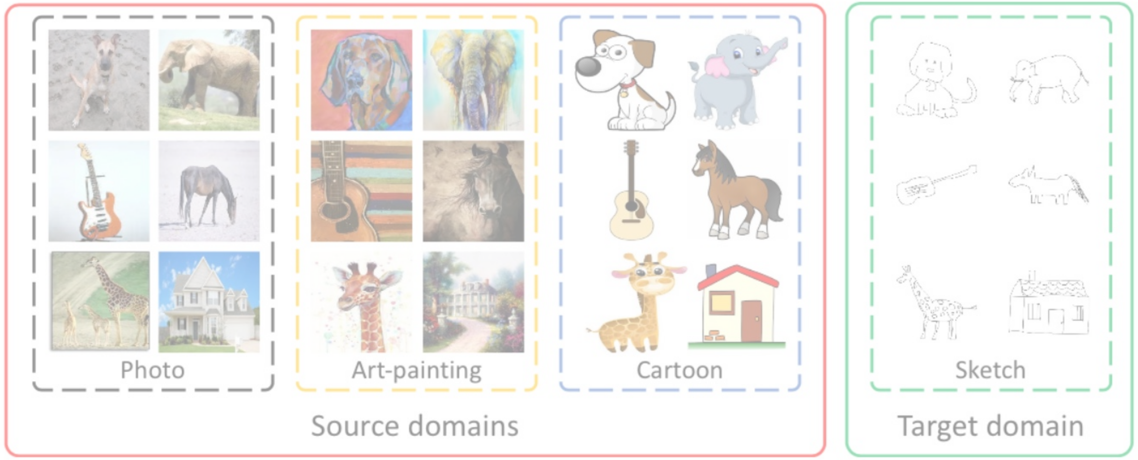

High
State-of-Affairs
Dead end
Supervision dependence
Simply scaling-up along all dimensions at training time seems a dead end
Not only because of the compute, storage and ethical expenses
but especially as humans generalize robustly in a data-efficient fashion
Low
Closed-world
Societal impact
Real-world

Distribution gap



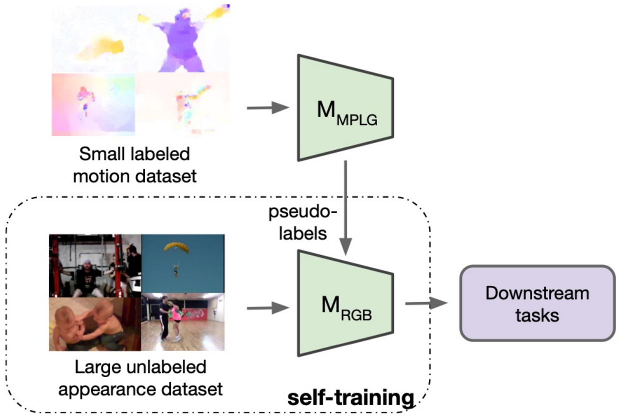
w/ Zehao Xiao *et al.*, ICML 2021

Distribution gap



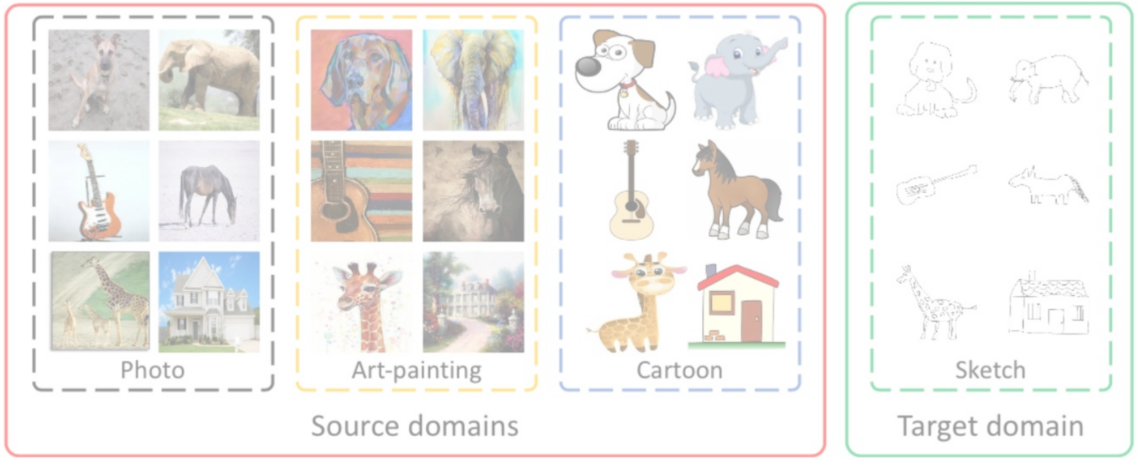
w/ Zehao Xiao *et al.*, ICML 2021

Label gap



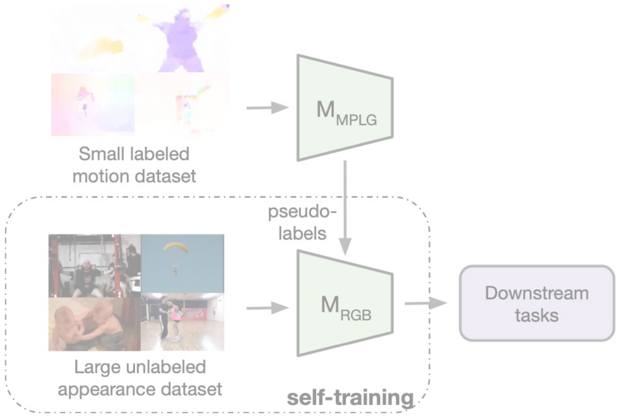
w/ Kirill Gavrilyuk *et al.*, ICCV 2021

Distribution gap



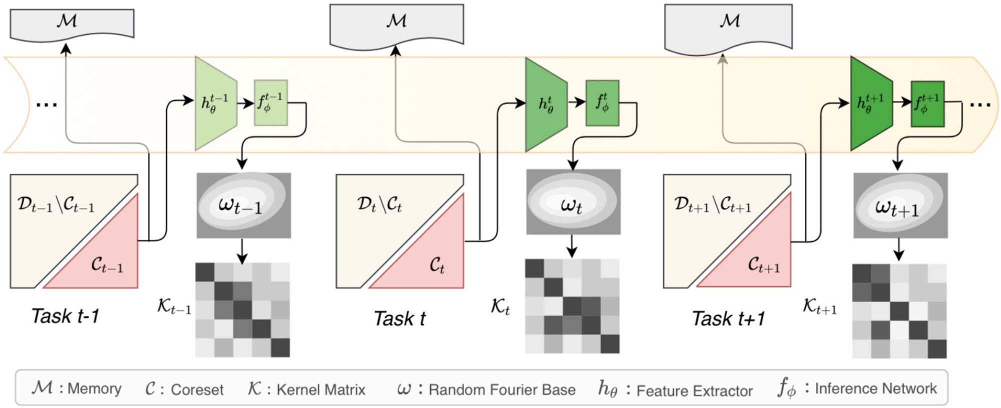
w/ Zehao Xiao *et al.*, ICML 2021

Label gap



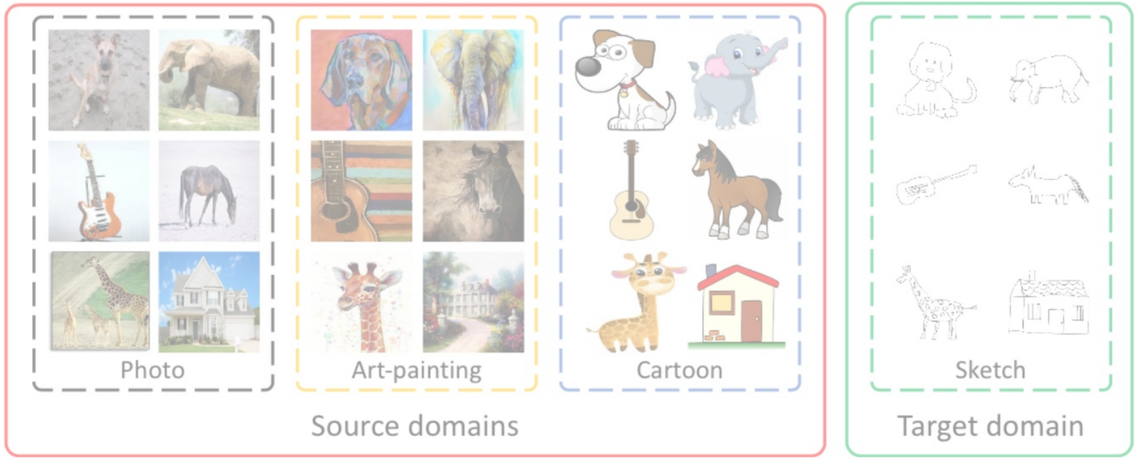
w/ Kirill Gavrilyuk *et al.*, ICCV 2021

Objectives gap



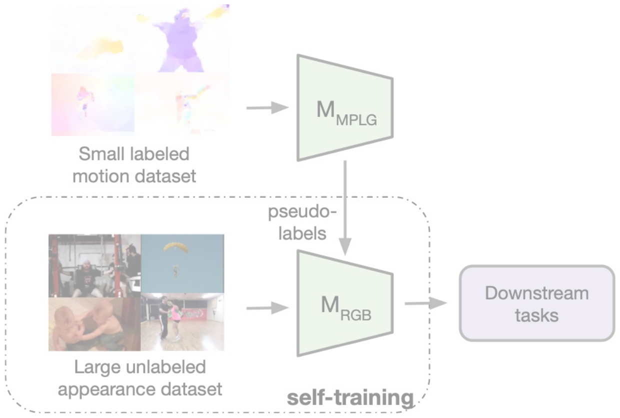
w/ Mohammad Mahdi Derakhshani *et al.*, ICML 2021

Distribution gap



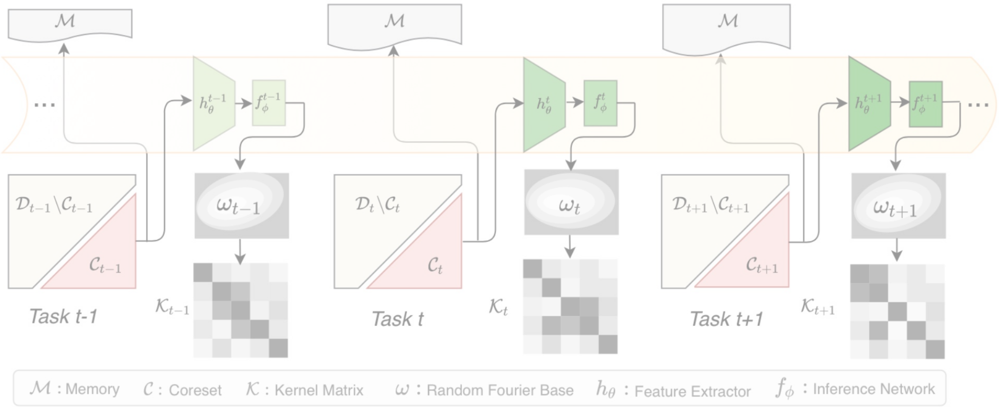
w/ Zehao Xiao *et al.*, ICML 2021

Label gap



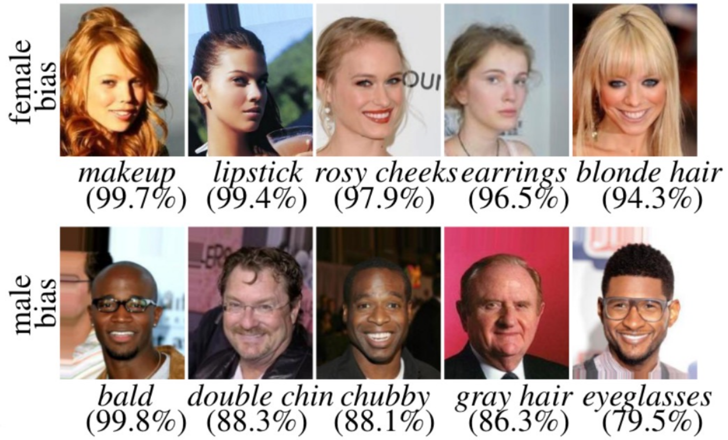
w/ Kirill Gavrilyuk *et al.*, ICCV 2021

Objectives gap

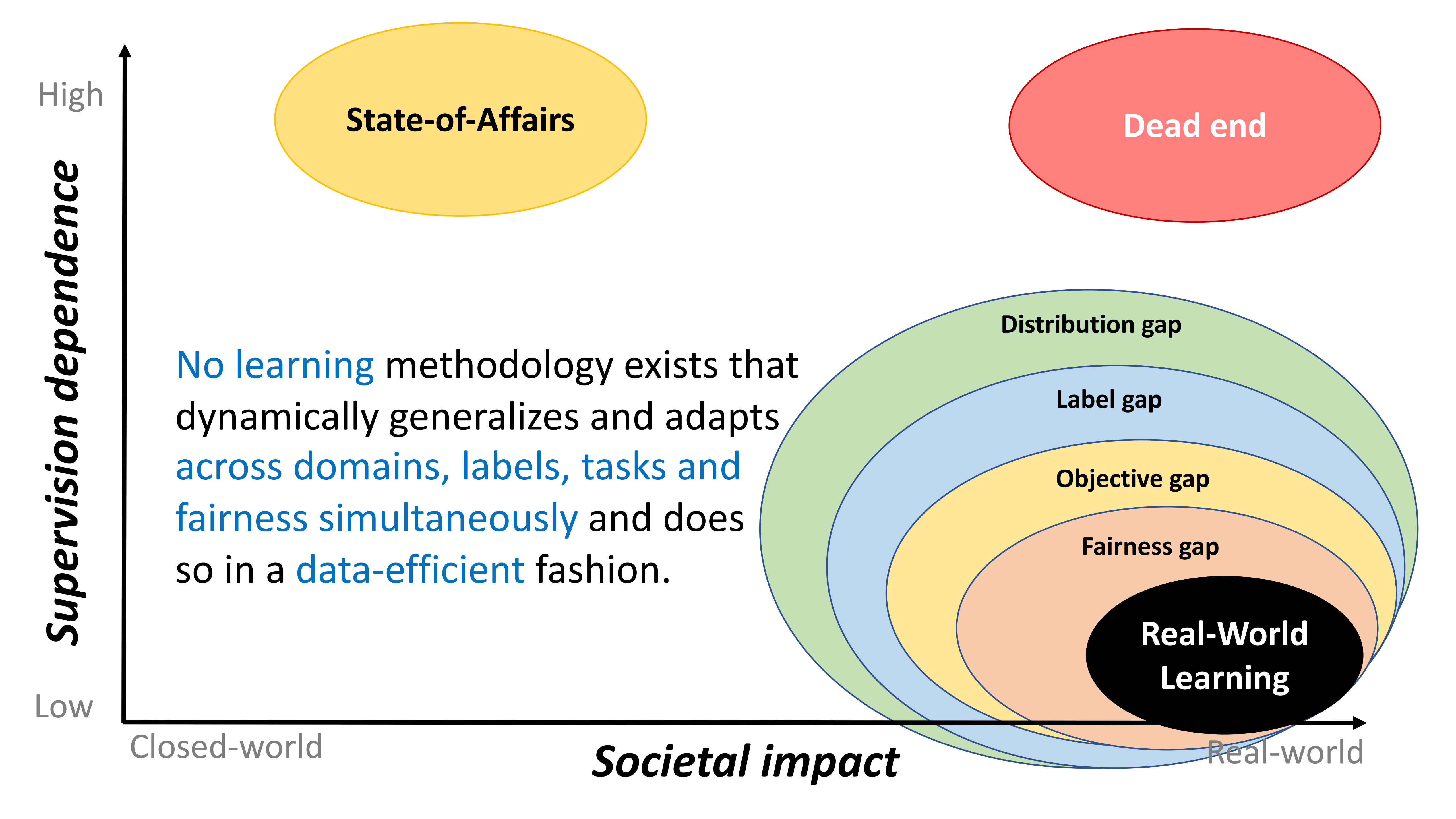


w/ Mohammad Mahdi Derakhshani *et al.*, ICML 2021

Fairness gap



w/ William Thong, BMVC 2021



High
Low
Supervision dependence
State-of-Affairs
Dead end
No learning methodology exists that dynamically generalizes and adapts across domains, labels, tasks and fairness simultaneously and does so in a data-efficient fashion.
Distribution gap
Label gap
Objective gap
Fairness gap
Real-World Learning
Closed-world
Societal impact
Real-world

This talk

We question common representation learning assumptions

i. Learning without **label** assumption

ii. Learning without **task** assumption

iii. Learning without **domain** assumption

I. Learning without label assumption

Pengwan Yang
University of Amsterdam

Pascal Mettes
University of Amsterdam

Cees Snoek
University of Amsterdam

Few-Shot Transformation of Common Actions into Time and Space. In *CVPR* 2021.

Canonical Paradigm: few-shot classification

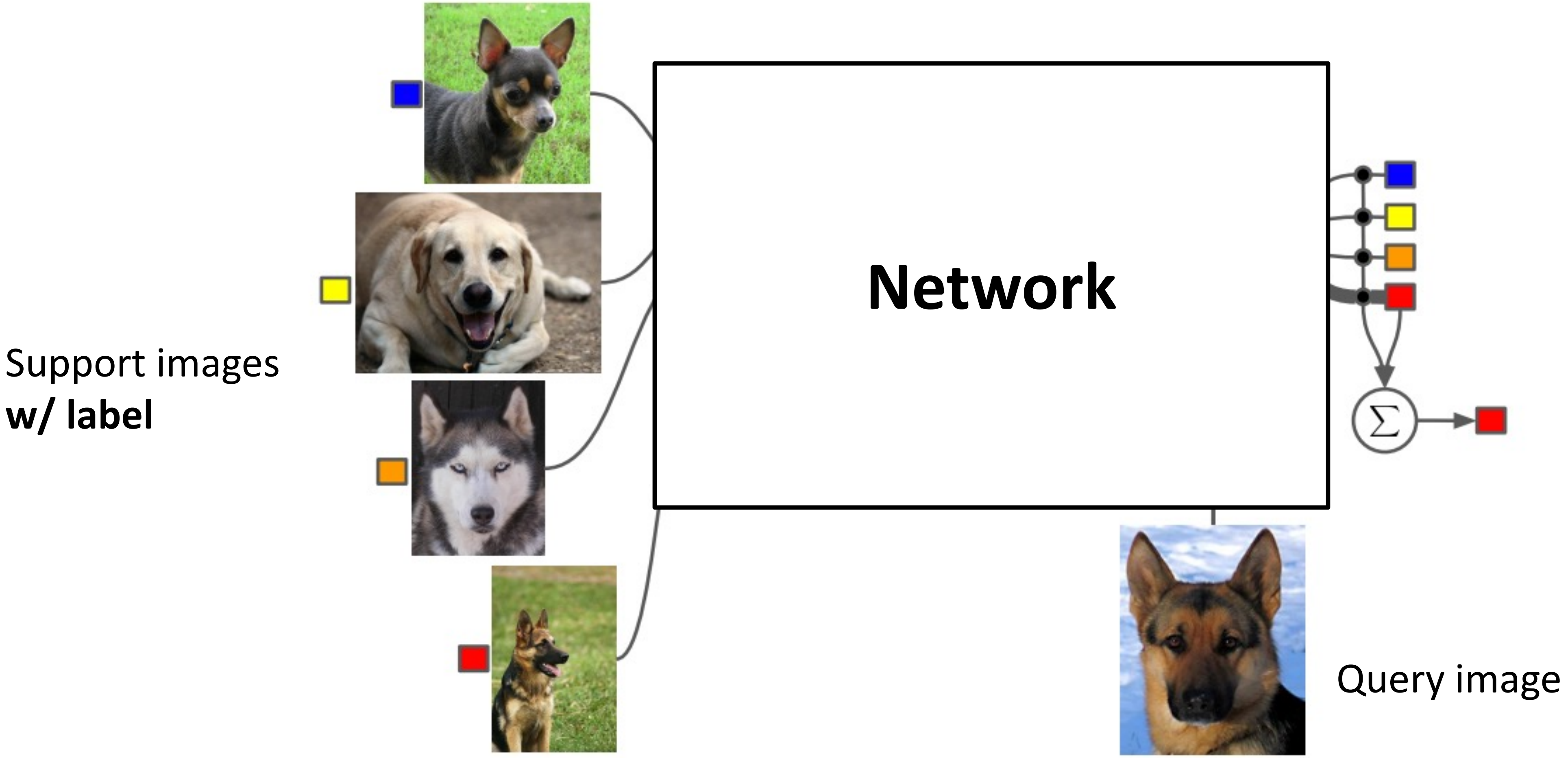


Figure credit: Vinyals et al. NeurIPS 2016

Canonical Paradigm: few-shot detection

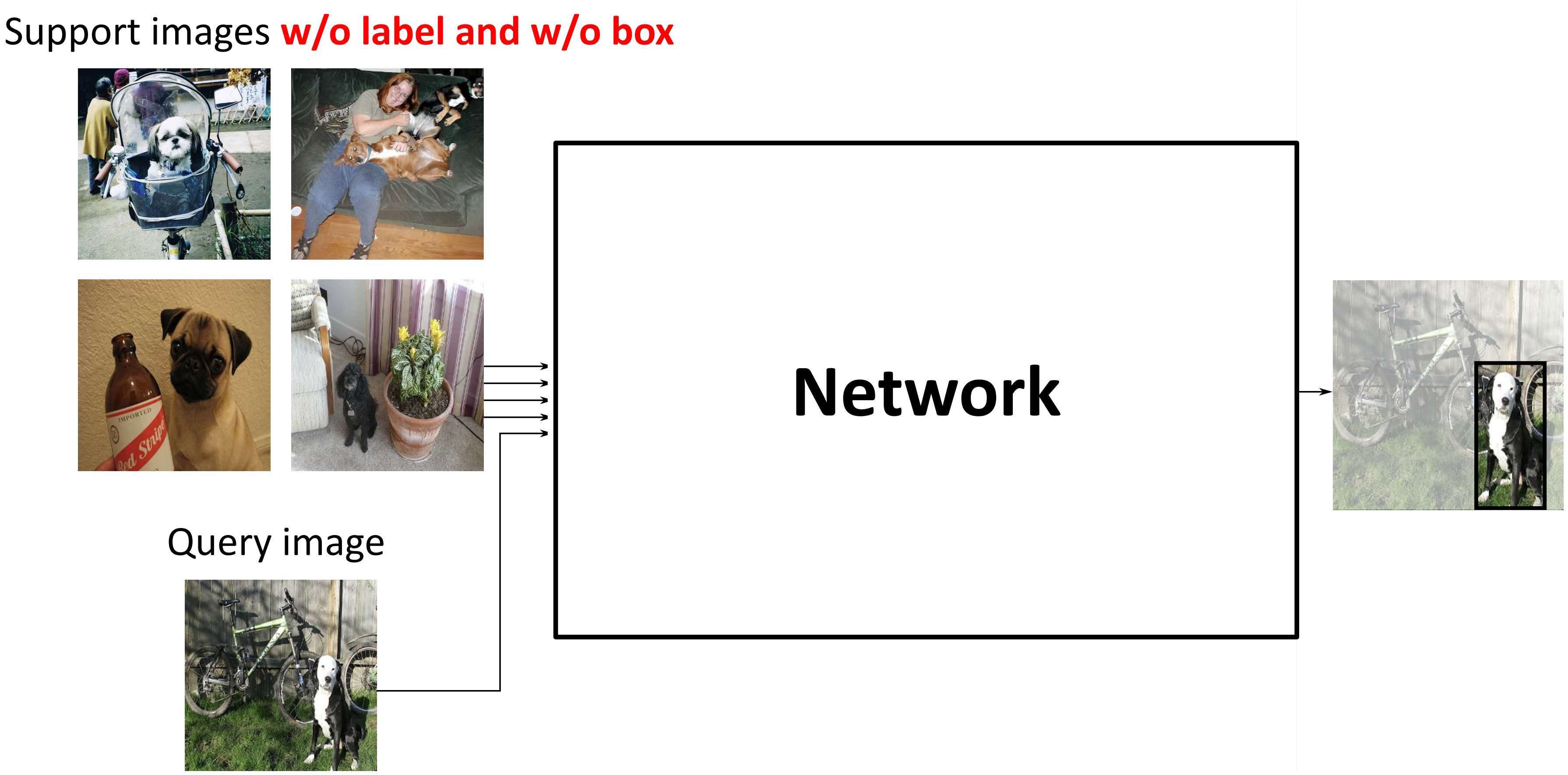
Support images
w/ label + box

Network

Query image

w/ Tao Hu et al. AAAI 2019

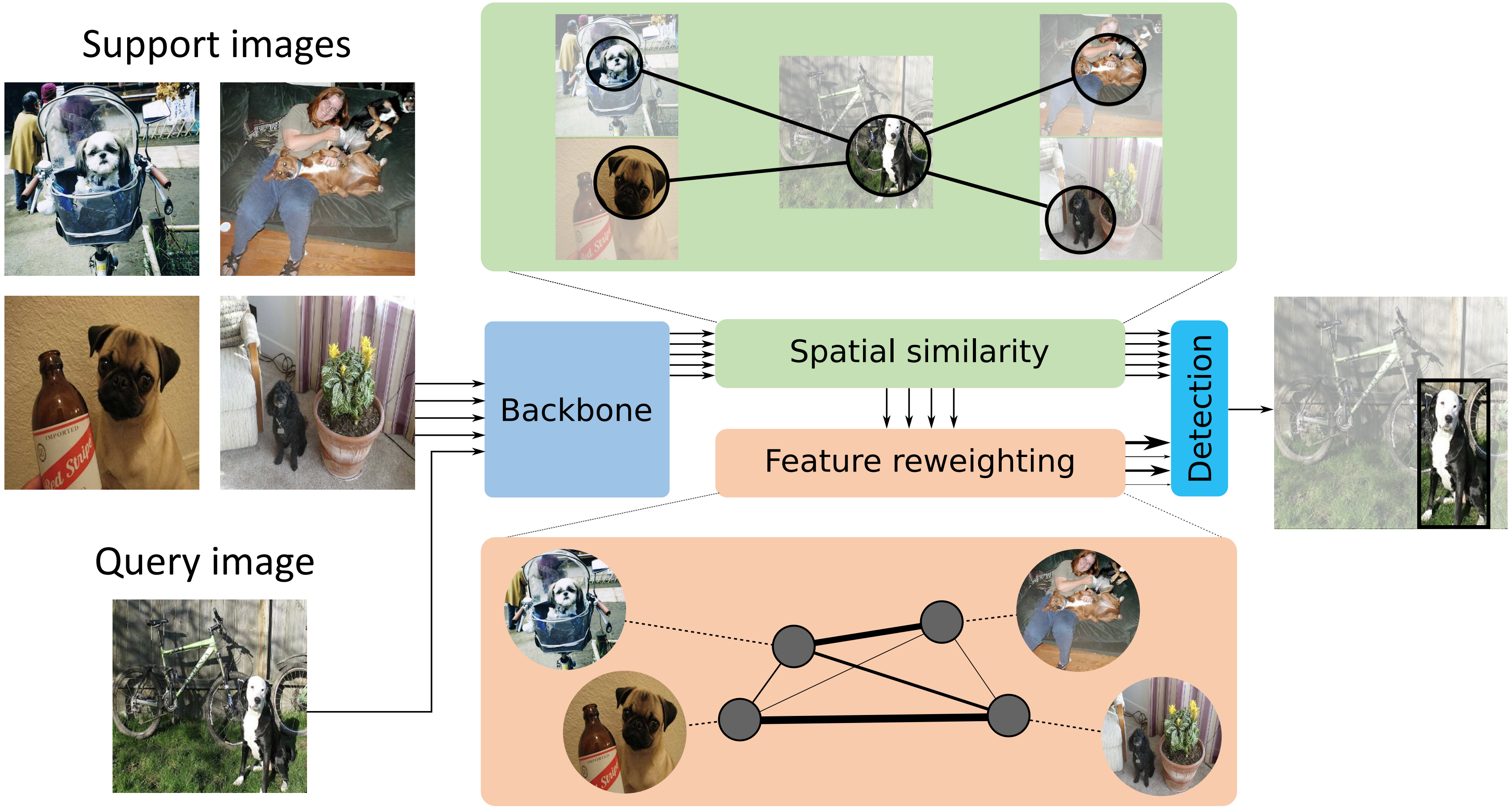
Few-shot common object localization



Localize the common object in the query image without any label and box annotation

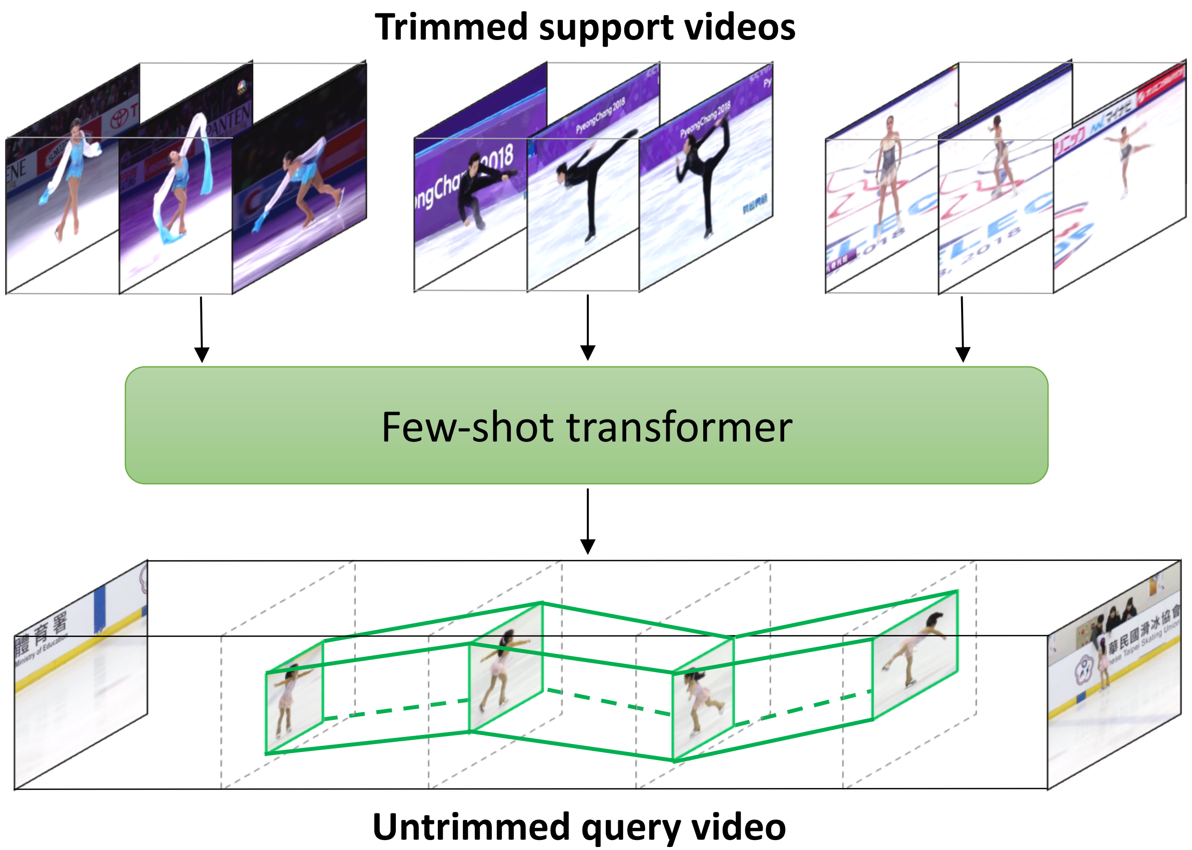
w/ Tao Hu et al. ICCV 2019

Few-shot common object localization



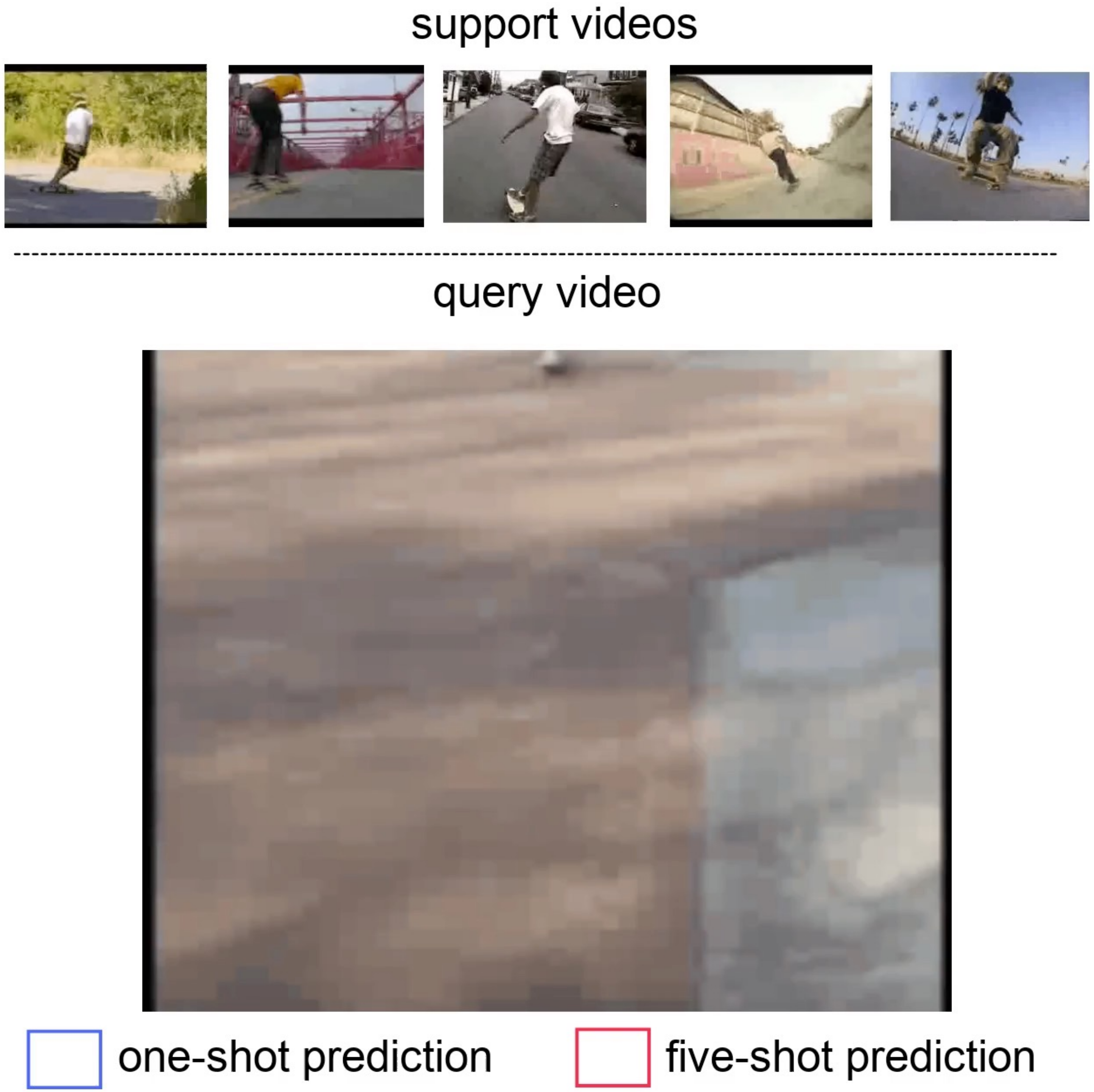
Localize the common object in the query image without any box annotation

Few-shot common action in video



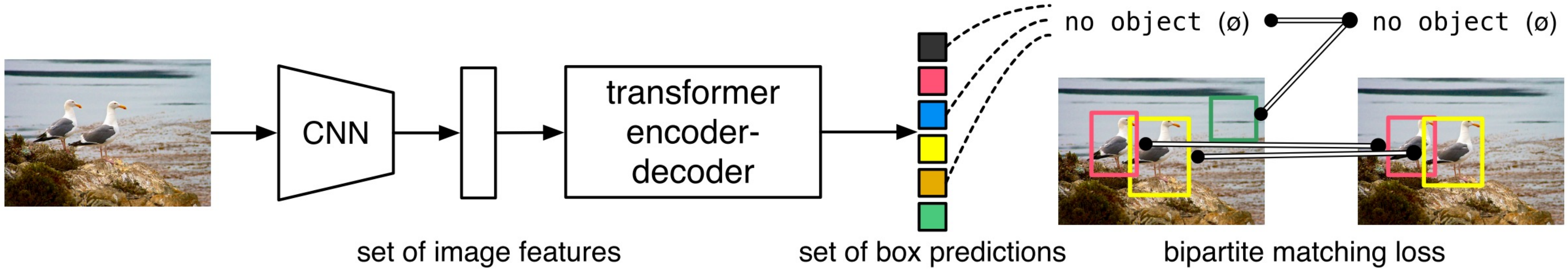
No need for action class label or any temporal and/or spatial annotation

Example


support videos
query video
one-shot prediction
five-shot prediction

Carion et al. ECCV 2020

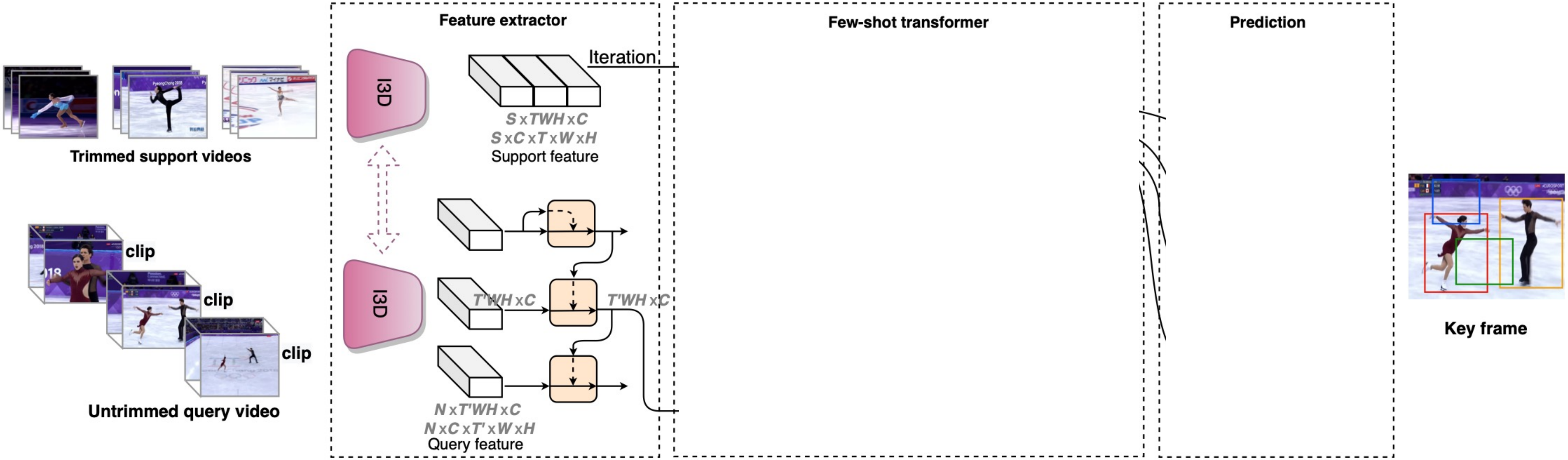
Inspiration: object detection transformers



Benefits of transformers:

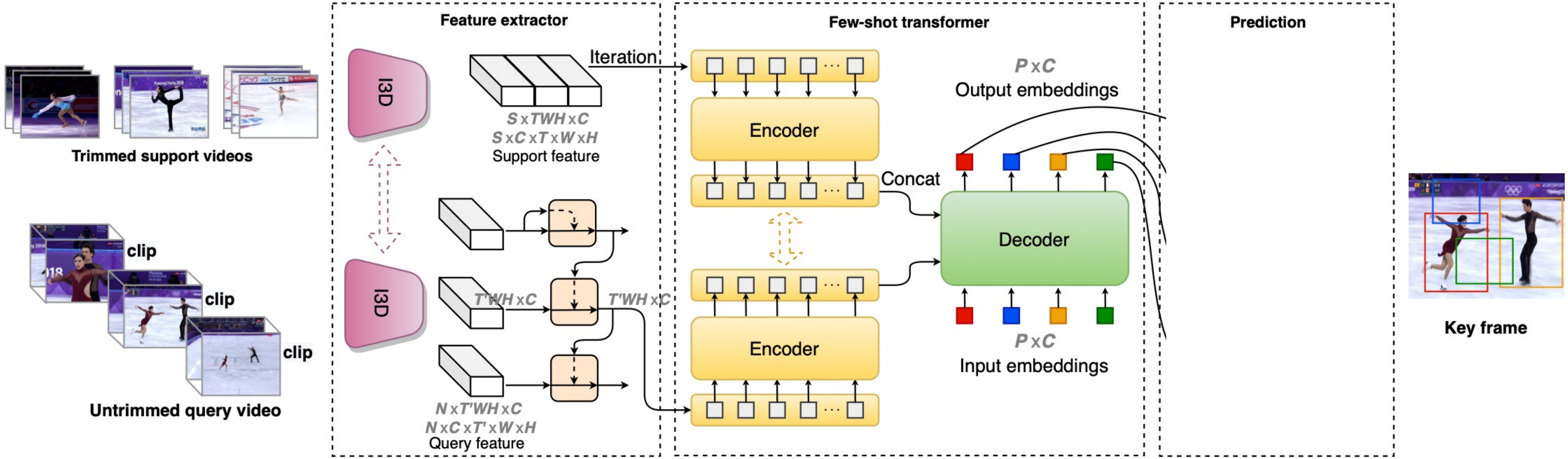
i) it avoids the needle-in-the-haystack problem with proposals
ii) it provides powerful relation modeling capability

Method



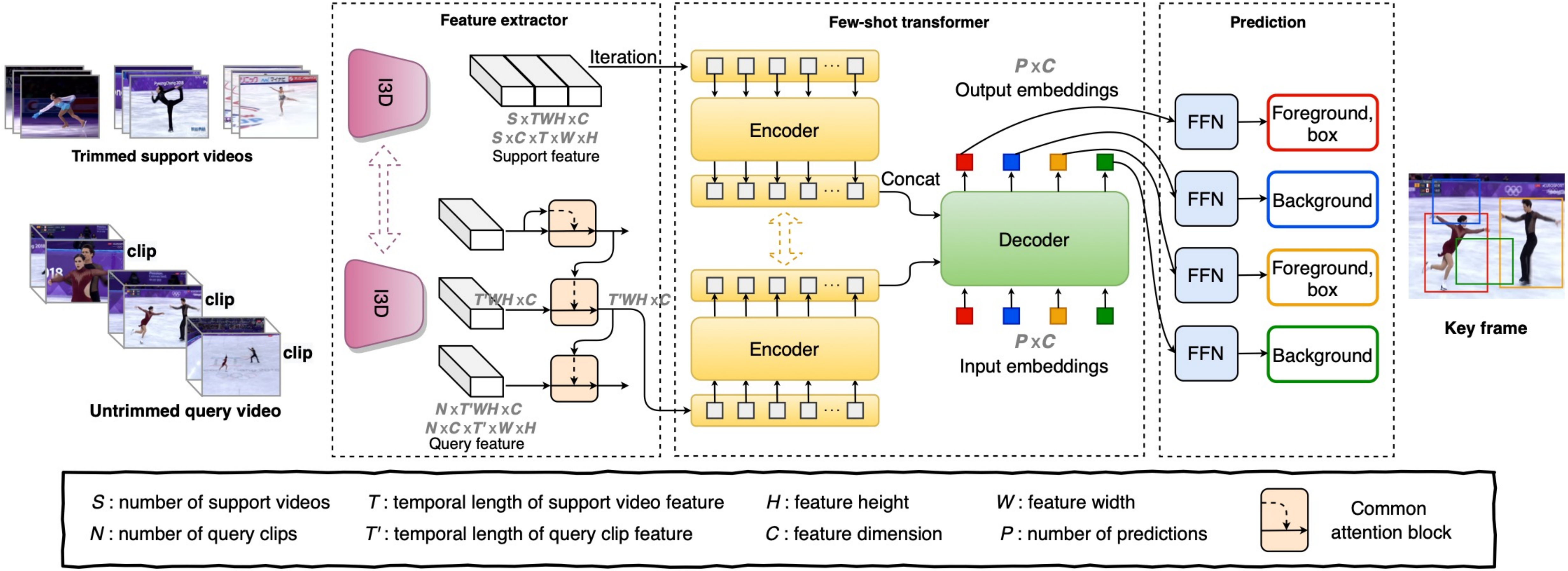
S : number of support videos
N : number of query clips
T : temporal length of support video feature
T' : temporal length of query clip feature
H : feature height
C : feature dimension
W : feature width
P : number of predictions
Common attention block

Method



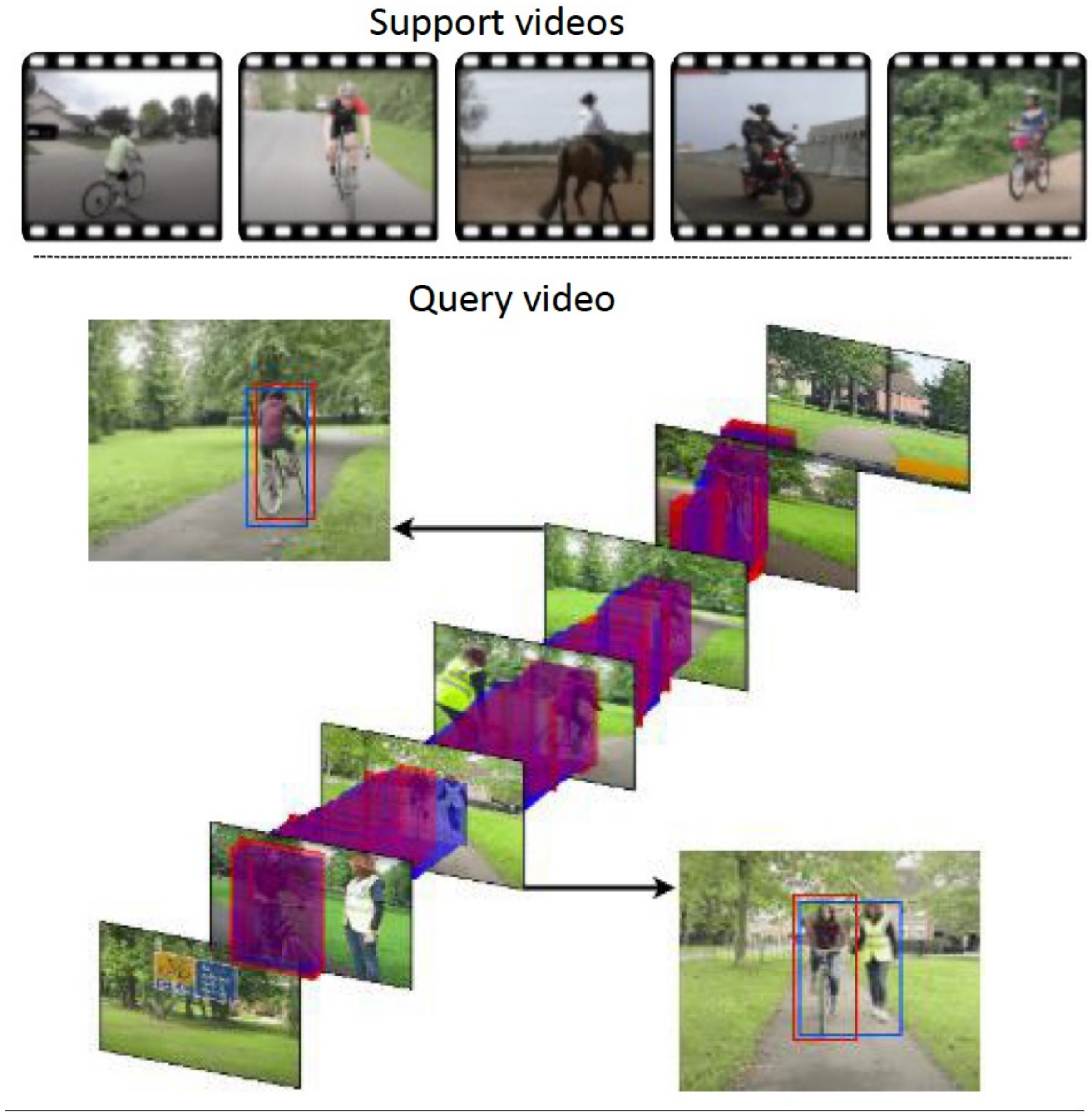
S : number of support videos
N : number of query clips
T : temporal length of support video feature
T' : temporal length of query clip feature
H : feature height
C : feature dimension
W : feature width
P : number of predictions
Common attention block

Method


Trimmed support videos
clip
clip
clip
Untrimmed query video
Feature extractor
I3D
I3D
Iteration
S x TWH x C
S x C x T x W x H
Support feature
T'WH x C
T'WH x C
N x T'WH x C
N x C x T' x W x H
Query feature
Few-shot transformer
Encoder
Encoder
Concat
P x C
Output embeddings
Decoder
P x C
Input embeddings
Prediction
FFN
FFN
FFN
FFN
Foreground, box
Background
Foreground, box
Background
Key frame
S : number of support videos
N : number of query clips
T : temporal length of support video feature
T' : temporal length of query clip feature
H : feature height
C : feature dimension
W : feature width
P : number of predictions
Common attention block

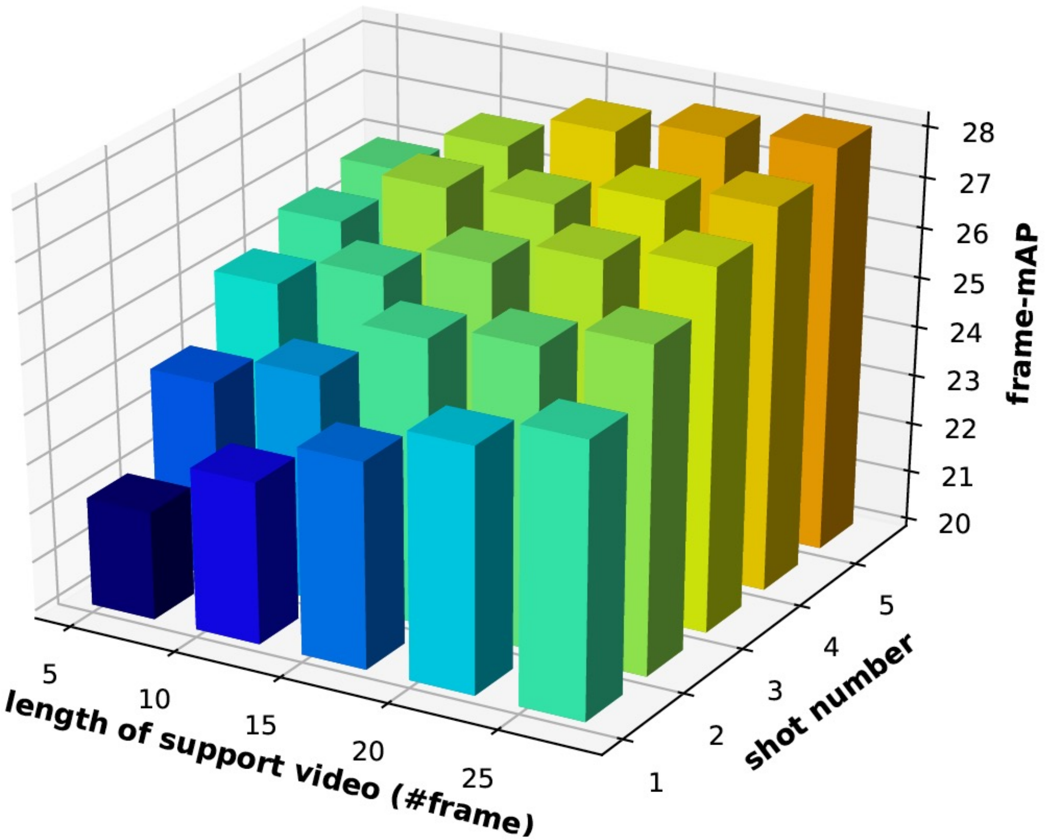
Results

one-shot (blue)
five-shot (red)



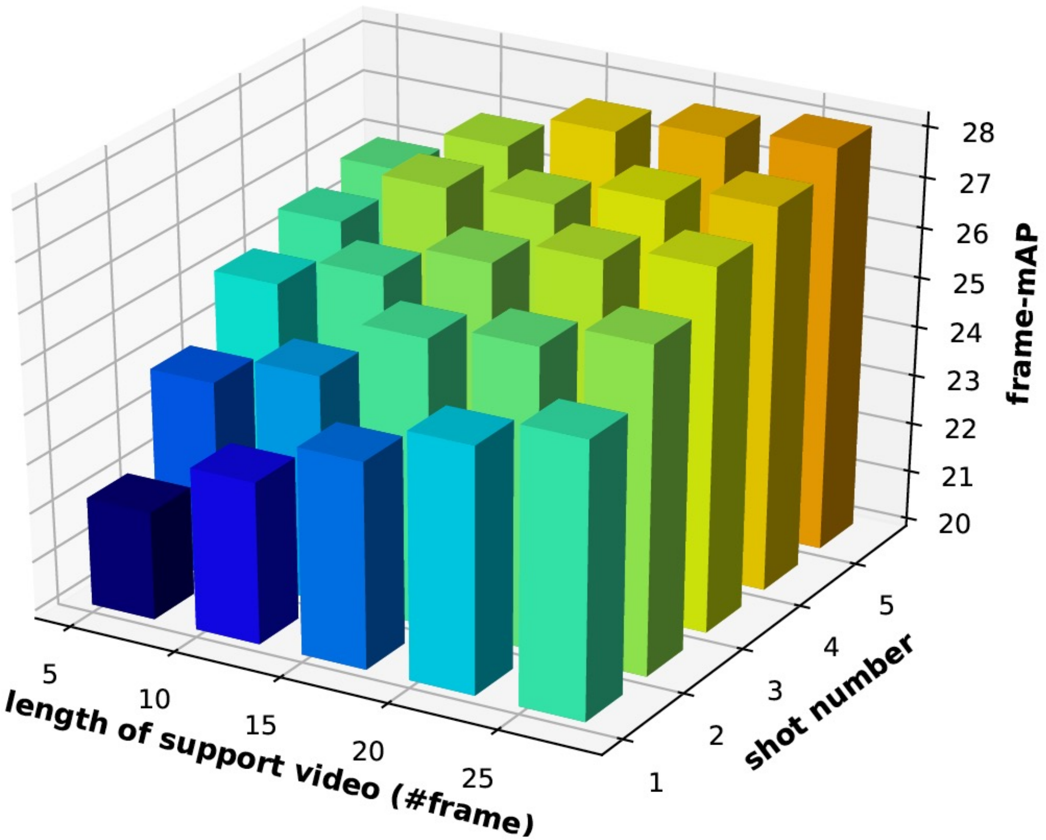
Common action localization in time and space

Ablations on Common-AVA



Influence of length and number of support videos.
We obtain a more precise common localization with more and longer support videos.

Ablations on Common-AVA

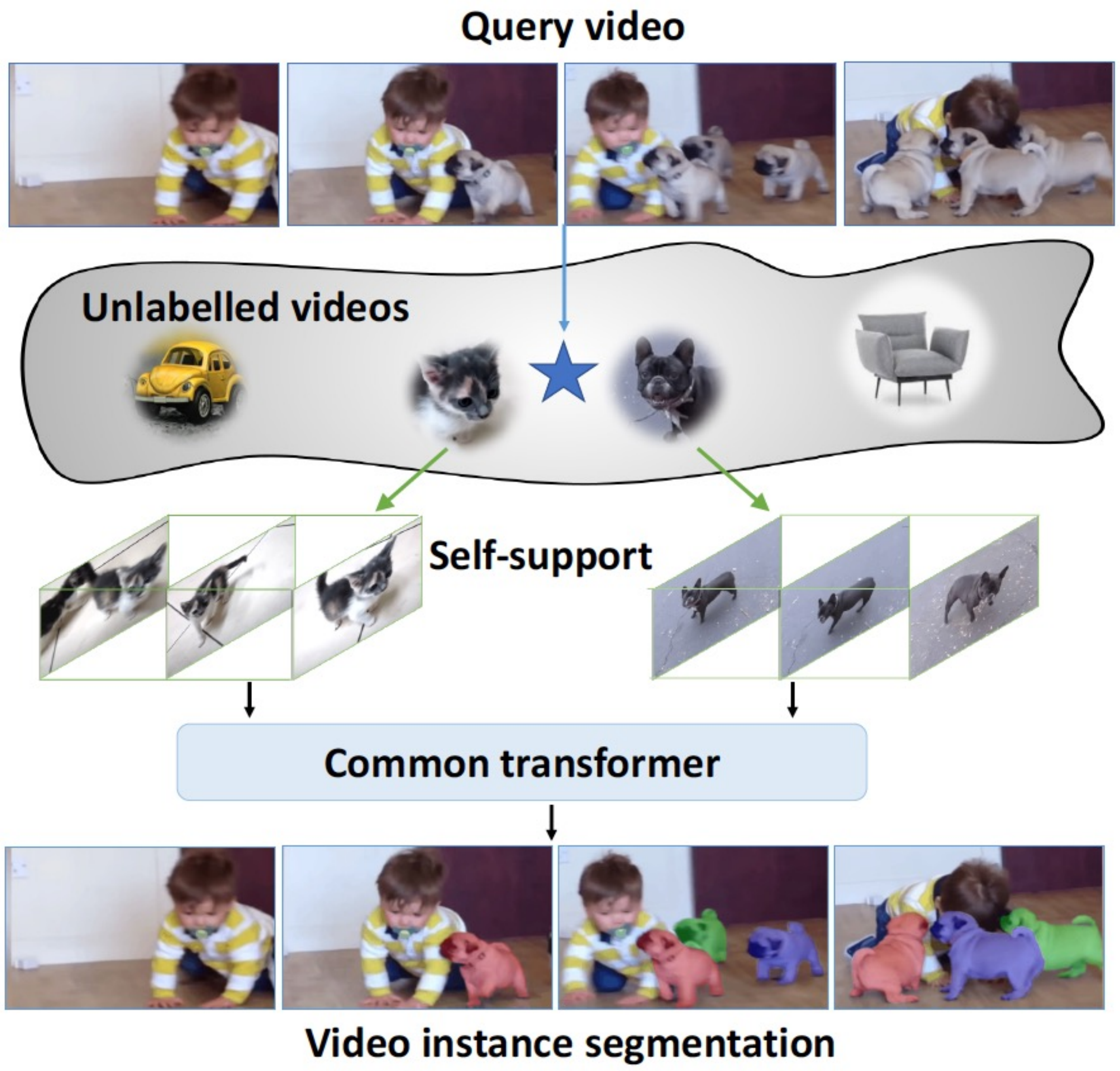


Influence of length and number of support videos. We obtain a more precise common localization with more and longer support videos.

No noise	28.1
Video-level noise	
1 noisy support video of other class	26.8
1 noisy support video without action	26.3
2 noisy support videos of different class	25.3
2 noisy support videos of same class	24.7
Frame-level noise	
2 noisy frames in each support video	27.9
4 noisy frames in each support video	27.4
6 noisy frames in each support video	26.1
8 noisy frames in each support video	24.5

Effect of noisy support videos for the five-shot setting. The result shows our robustness.

Self-support video instance segmentation



Find support videos using the query

Pool from unlabelled video in self-supervised fashion

Transformer enables instance segmentation

No labels, no masks.

Video instance segmentation results

Self-support examples

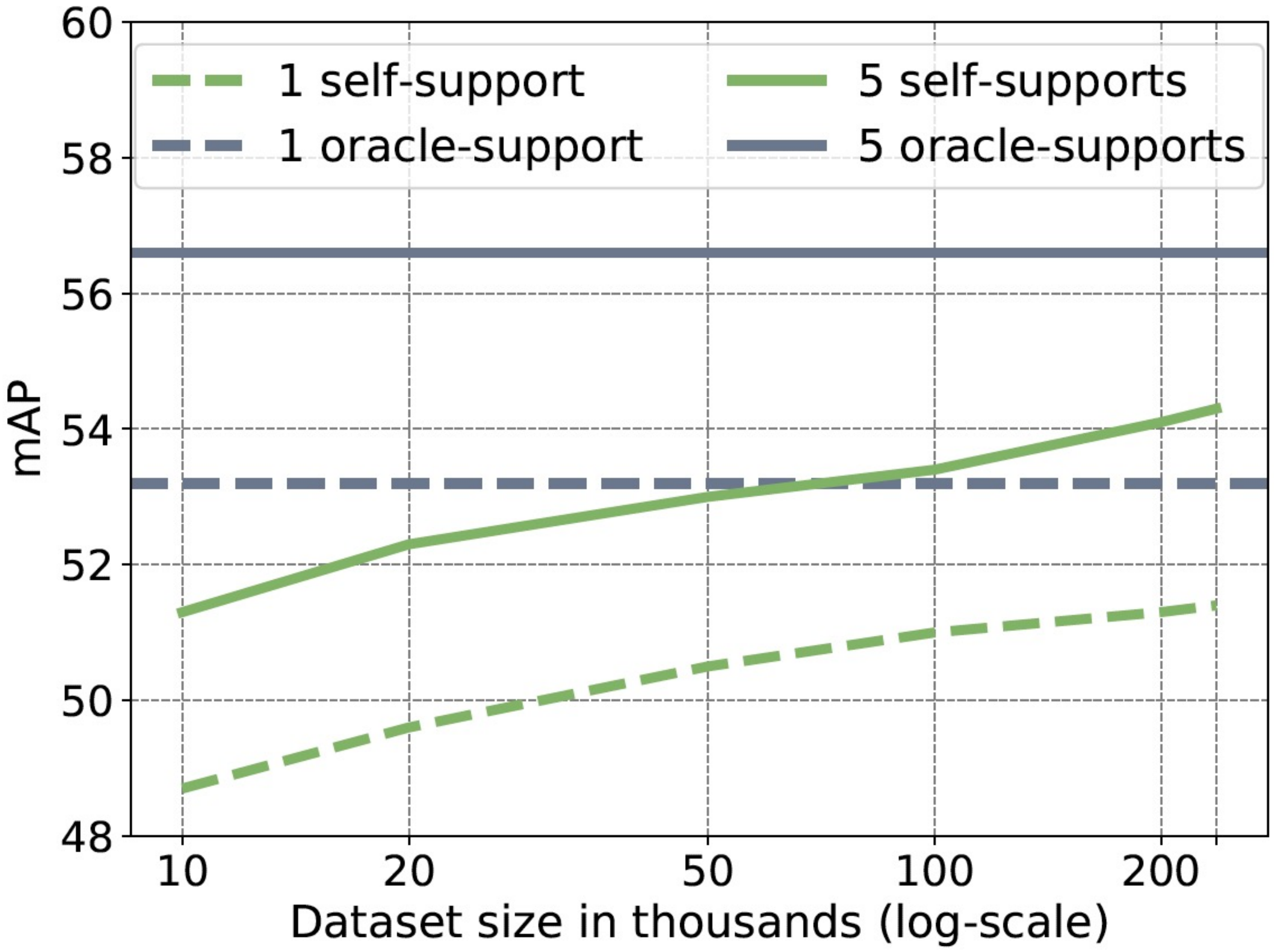
Query video

Video instance segmentation results

Self-support examples

Query video

Scalable video instance segmentation?



II. Learning without task assumption

Yingjun Du
University of Amsterdam

Xiantong Zhen
University of Amsterdam

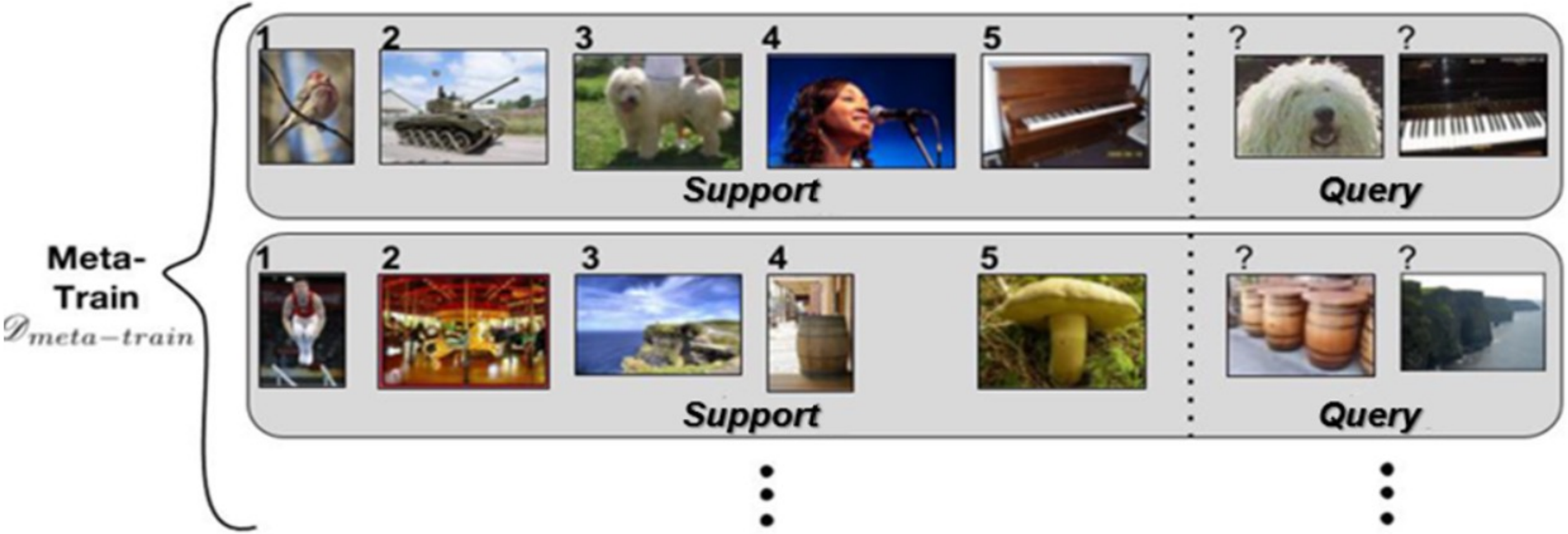
Ling Shao
Inception Institute of AI

Cees Snoek
University of Amsterdam

MetaNorm: Learning to Normalize Few-Shot Batches Across Domains. In *ICLR* 2021.

Few-shot meta-learning

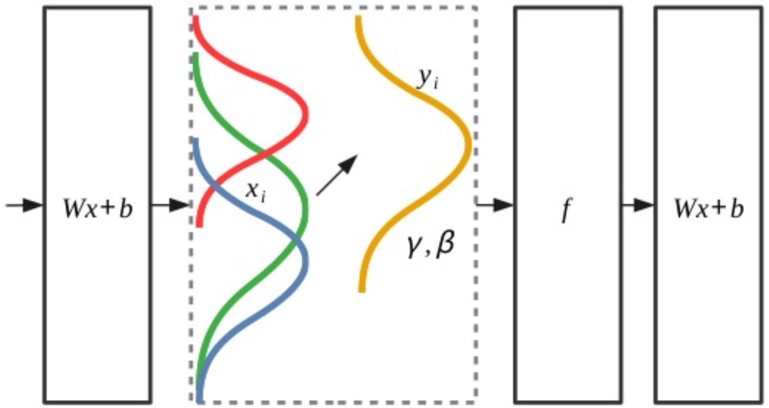
5-way, 1-shot



Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In ICLR 2017.

Deep learning work horse: batch normalization

Stabilize the distribution of internal activations during training



$$\mu_{\mathcal{B}} = \frac{1}{M}\sum_{i=1}^{M} a_i, \qquad \sigma_{\mathcal{B}}^2 = \frac{1}{M}\sum_{i=1}^{M}(a_i - \mu_{\mathcal{B}})^2$$

$$a_i' \leftarrow \mathbf{BN}(a_i) \equiv \gamma \hat{a}_i + \beta, \quad \text{where,} \quad \hat{a}_i = \frac{a_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}},$$

Challenge I: batch statistics become unstable with small batch sizes

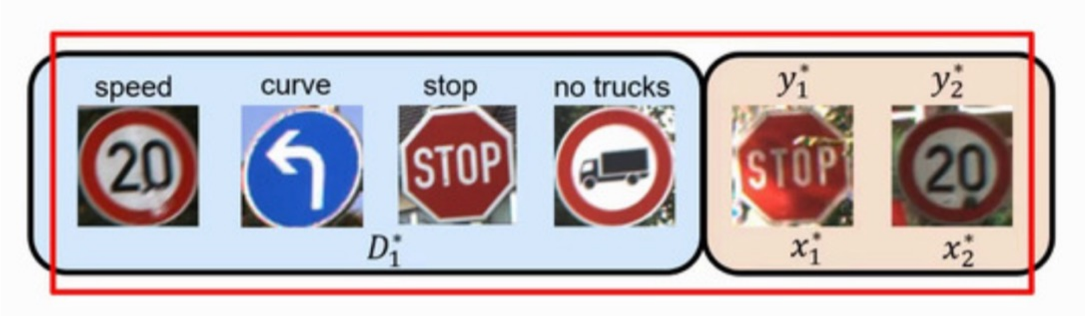
Challenge II: distribution shift between source and target domains

Sergey Ioffe & Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICML 2015.

Transductive Batch Normalization

Compute batch statistics by using all available query data

Transductive



support query

Requirement to have test set samples available limits real-world use

TaskNorm

Identified the limiting assumption of the transductive setting

Leverages statistics from both layer and instance normalization

Better than batch norm, sometimes better than transductive.

John Bronskill *et al.* TaskNorm: Rethinking batch normalization for meta-learning. In ICML 2020.

Our proposal: MetaNorm

Leverage the meta-learning setting

Infer statistics from the support set that better match the query set

$$D_{\mathrm{KL}}\left[q(m|S)||p(m|Q)\right]$$

Distribution inferred from support set

Distribution inferred from query set

Achieve adaptive batch normalization

Meta-training optimization

Support

$$\mu_S = \frac{1}{|\mathcal{S}|}\sum_{i=1}^{|\mathcal{S}|} f_\mu^\ell(\mathbf{a}_i),$$

$$\sigma_S = \frac{1}{|\mathcal{S}|}\sum_{i=1}^{|\mathcal{S}|} f_\sigma^\ell\big((\mathbf{a}_i - \mu_S)^2\big).$$

Query

$$\mu_Q = \frac{1}{|\mathcal{Q}|}\sum_{i=1}^{|\mathcal{Q}|} f_\mu^\ell(\mathbf{a}_i),$$

$$\sigma_Q = \frac{1}{|\mathcal{Q}|}\sum_{i=1}^{|\mathcal{Q}|} f_\sigma^\ell\big((\mathbf{a}_i - \mu_Q)^2\big).$$

$$D_{\mathrm{KL}}\left[\boxed{\mathcal{N}(\mu_S,\sigma_S)} \,\|\, \boxed{\mathcal{N}(\mu_Q,\sigma_Q)}\right],$$

Hypernetworks $f_\mu^\ell, f_\sigma^\ell$, generate (μ_S, σ_S) and (μ_Q, σ_Q) from the support and query sets, for calculating the KL term during meta-training optimization.

Meta-testing

Given a test task, the learned hypernetworks $f_{\mu}^{\ell}, f_{\sigma}^{\ell}$ take the support set as input to generate normalization statistics directly for the query set.

$$\mu_S = \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} f_{\mu}^{\ell}(\mathbf{a}_i), \qquad \sigma_S = \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} f_{\sigma}^{\ell}\left((\mathbf{a}_i - \mu_S)^2\right).$$

$$a' = \gamma \left(\frac{a - \mu_S}{\sqrt{\sigma_S^2 + \epsilon}} \right) + \beta,$$

Effect of the KL term

	Label gap		*Distribution gap*				
	Few-shot classification		**Domain generalization**				
MetaNorm	5-way, 1-shot	5-way, 5-shot	*Photo*	*Art*	*Cartoon*	*Sketch*	***Mean***
w/o KL	34.3 $\pm$ 1.5	50.7 $\pm$ 0.8	88.96	71.25	65.37	69.28	73.72
w/ KL	**46.8** $\pm$ **1.6**	**60.1** $\pm$ **0.8**	**95.99**	**85.01**	**78.63**	**83.17**	**85.70**

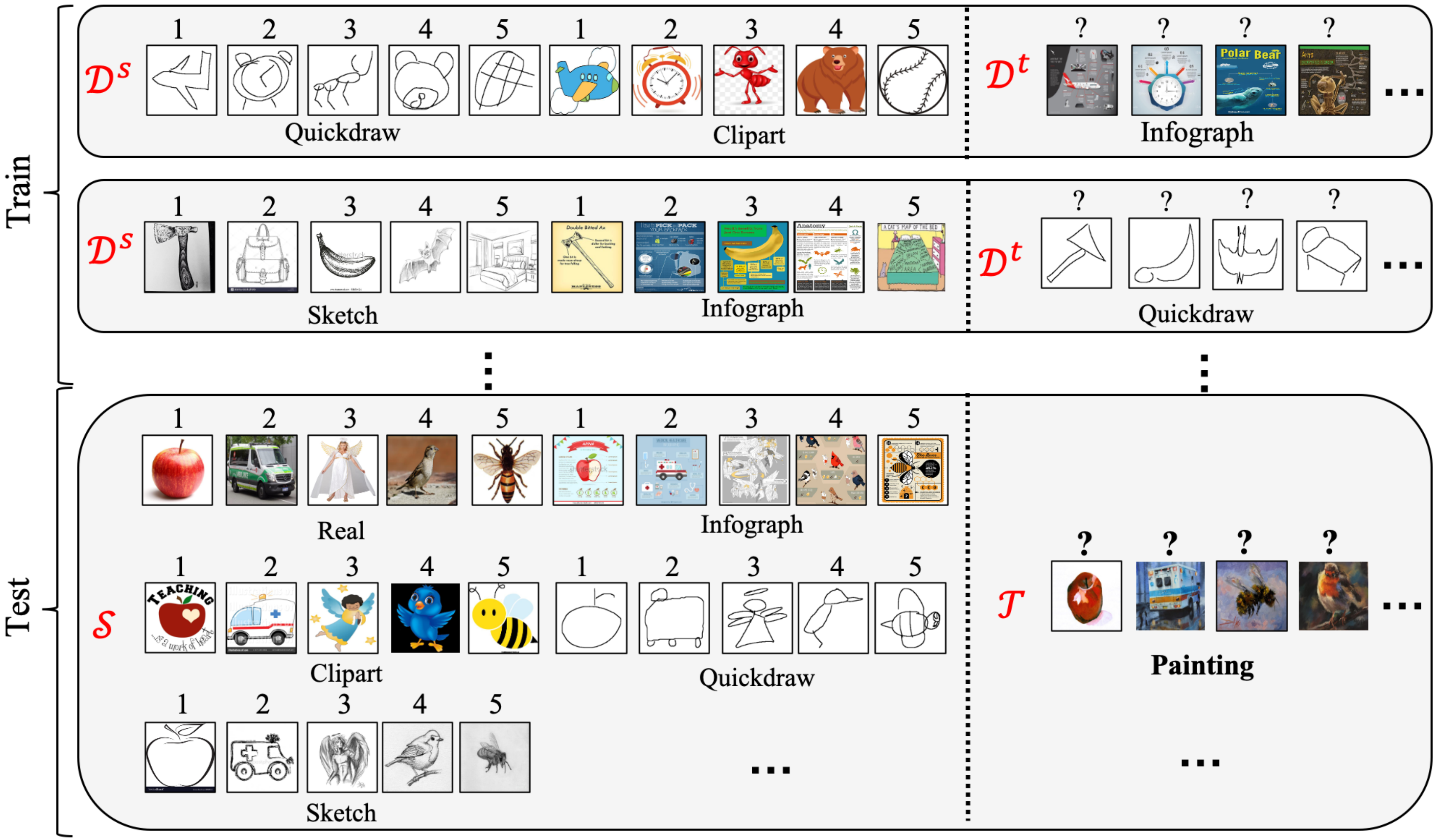
Effective for both few-shot classification and many-shot domain generalization

Comparison with other batch norms

	ProtoNets		MAML	
	5-way, 1-shot	5-way, 5-shot	5-way, 1-shot	5-way, 5-shot
TBN	45.9 $\pm$ 0.6	65.5 $\pm$0.9	45.5 $\pm$1.8	59.7 $\pm$ 0.9
CBN	47.8 $\pm$ 0.6	66.7 $\pm$ 0.5	20.1 $\pm$ 0.0	20.2 $\pm$ 0.2
TaskNorm	47.5 $\pm$ 0.6	65.3 $\pm$ 0.5	42.0 $\pm$ 1.7	58.1 $\pm$ 0.9
MetaNorm	**48.1** $\pm$ **1.6**	**65.9** $\pm$ **0.9**	**46.8** $\pm$ **1.6**	**60.1** $\pm$ **0.8**

MetaNorm outperforms transductive and non-transductive normalizations

Real-world learning: few-shot domain generalization



Few-shot domain generalization

	MAML	
	5-way, 1-shot	5-way, 5-shot
TBN	28.7 ± 1.8	49.3 ±0.8
CBN	20.0 ± 0.0	20.1 ± 0.2
TaskNorm	26.9 ± 1.7	47.4 ± 0.8
MetaNorm	**32.7** ± **1.7**	**51.9** ± **0.9**

MetaNorm allows for batch normalization of small batches across domains.

III. Learning without domain assumption




Zehao Xiao
University of Amsterdam

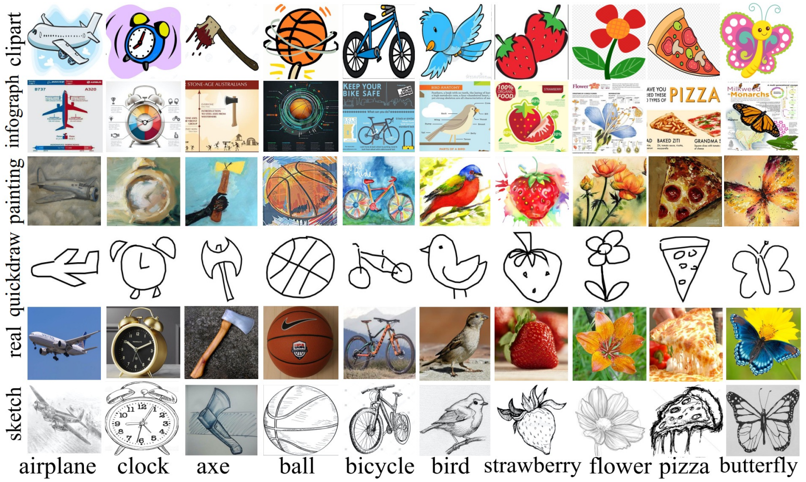
Xiantong Zhen
University of Amsterdam

Ling Shao
Inception Institute of AI

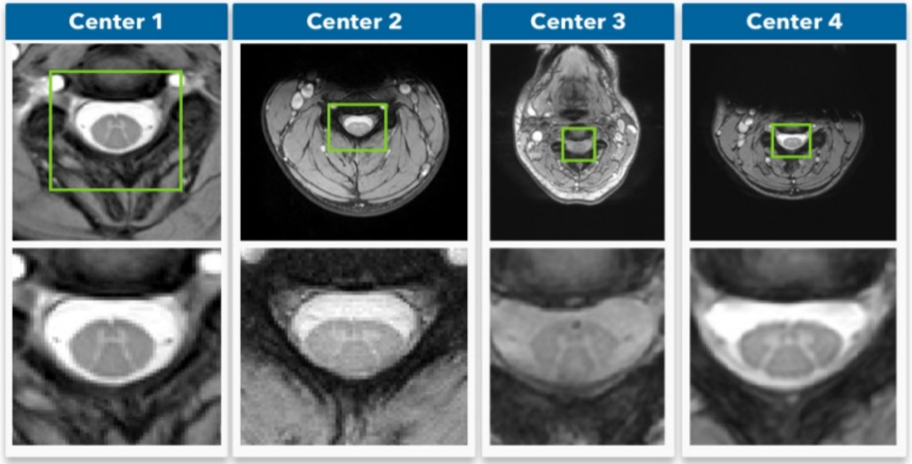
Cees Snoek
University of Amsterdam

Learning to Generalize across Domains on Single Test Samples. *Submitted*.

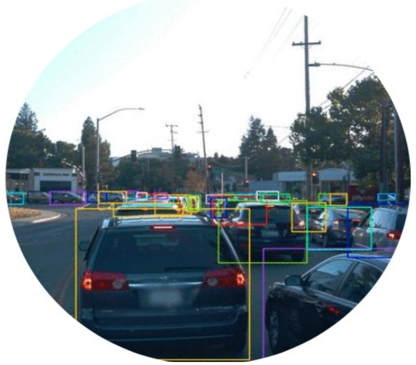
Distribution gaps are a fact of life



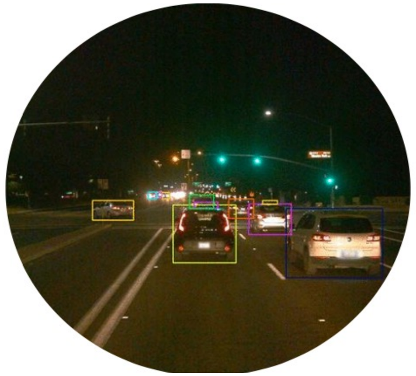
Images with different style



Medical images from different devices



Daylight



Night

Downtown

Suburban

Autopilot data in different environments

Test-time training

Update model parameters by self-supervision before prediction

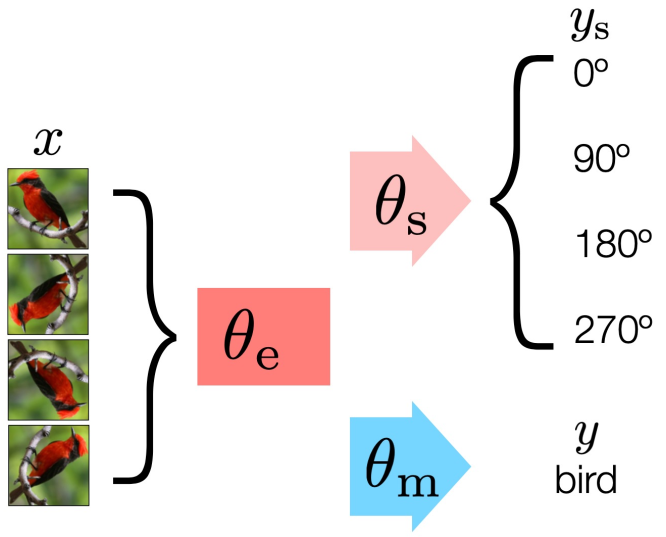
training

$$\min_{\theta_e,\theta_s,\theta_m} \mathbb{E}_P \left[\begin{array}{l} \ell_m(x, y; \theta_e, \theta_m) \\ + \ell_s(x, y_s; \theta_e, \theta_s) \end{array} \right]$$

testing

$$\min_{\theta_e,\theta_s} \mathbb{E}_Q \left[\ell_s(x, y_s; \theta_e, \theta_s) \right]$$

$\rightarrow \theta(x)$: make prediction on x

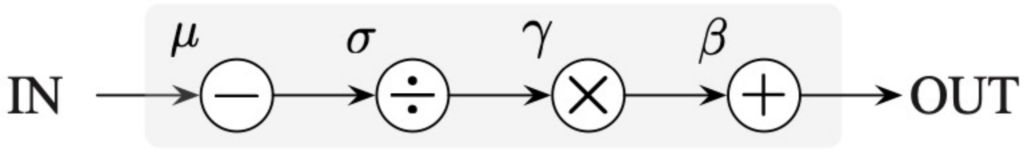


Needs additional self-supervised model, plus fine-tuning

Yu Sun *et al.* Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In ICML 2020.

Test-time adaptation

Normalize test-batch predictions by entropy minimization



normalization $\mu \leftarrow \mathbb{E}[x_t], \sigma^2 \leftarrow \mathbb{E}[(\mu - x_t)^2]$

transformation $\gamma \leftarrow \gamma + \partial H / \partial \gamma, \beta \leftarrow \beta + \partial H / \partial \beta$

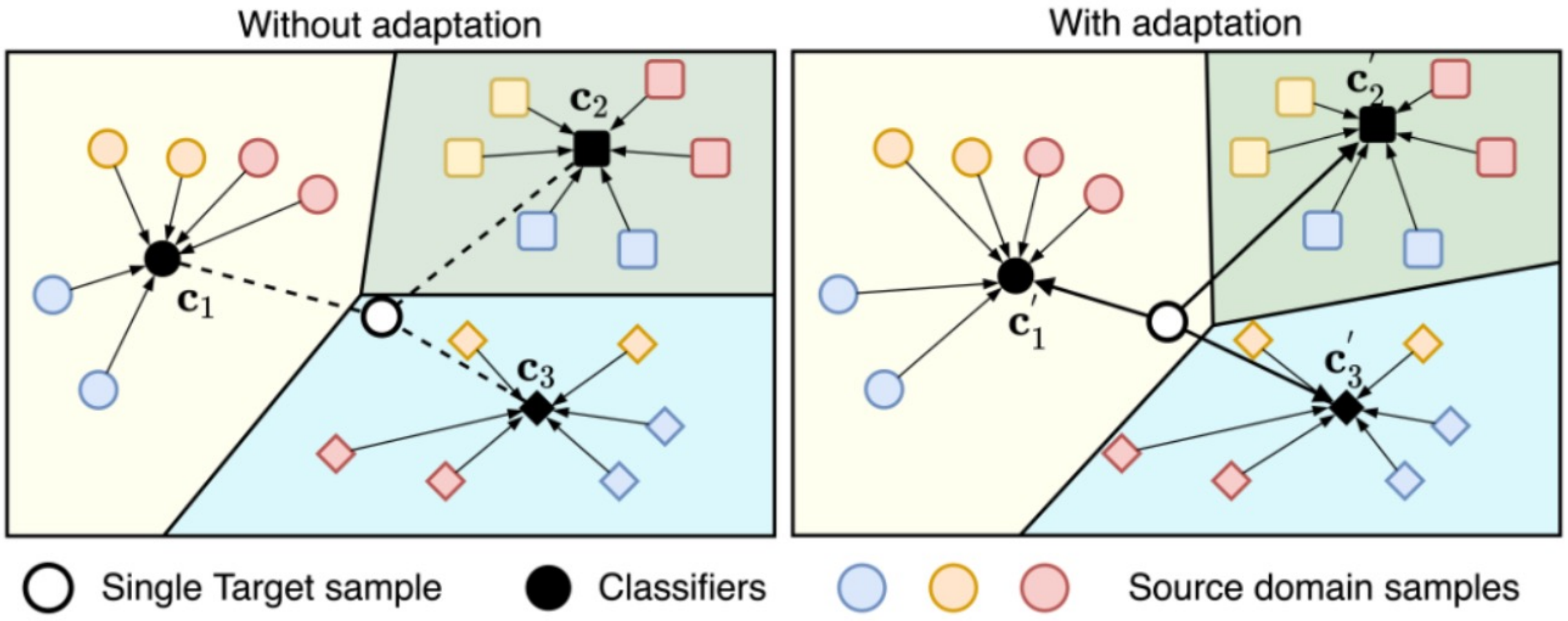
Needs a batch to be from the same domain, plus fine-tuning

Outperforms test-time training

Dequan Wang *et al.* Tent: Fully Test-time Adaptation by Entropy Minimization. In ICLR 2021.

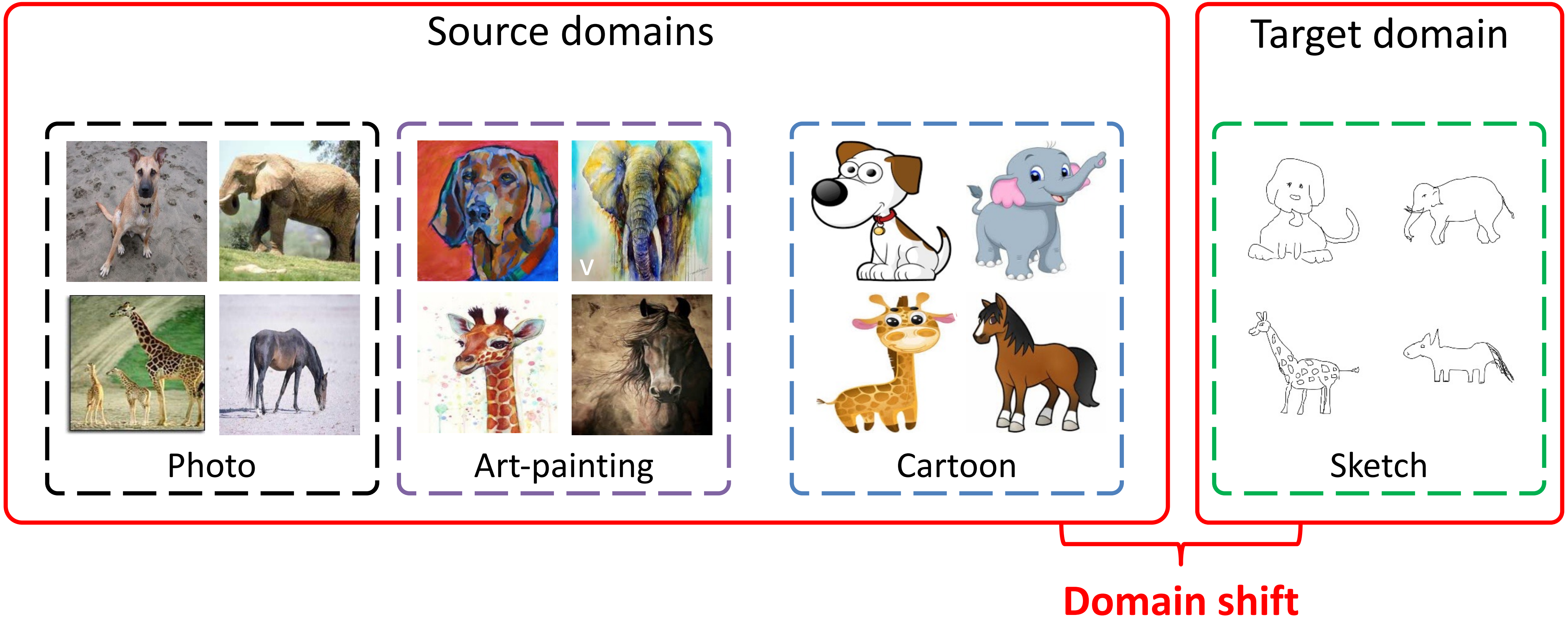
Key idea

Adapt source domain classifiers to each individual target sample



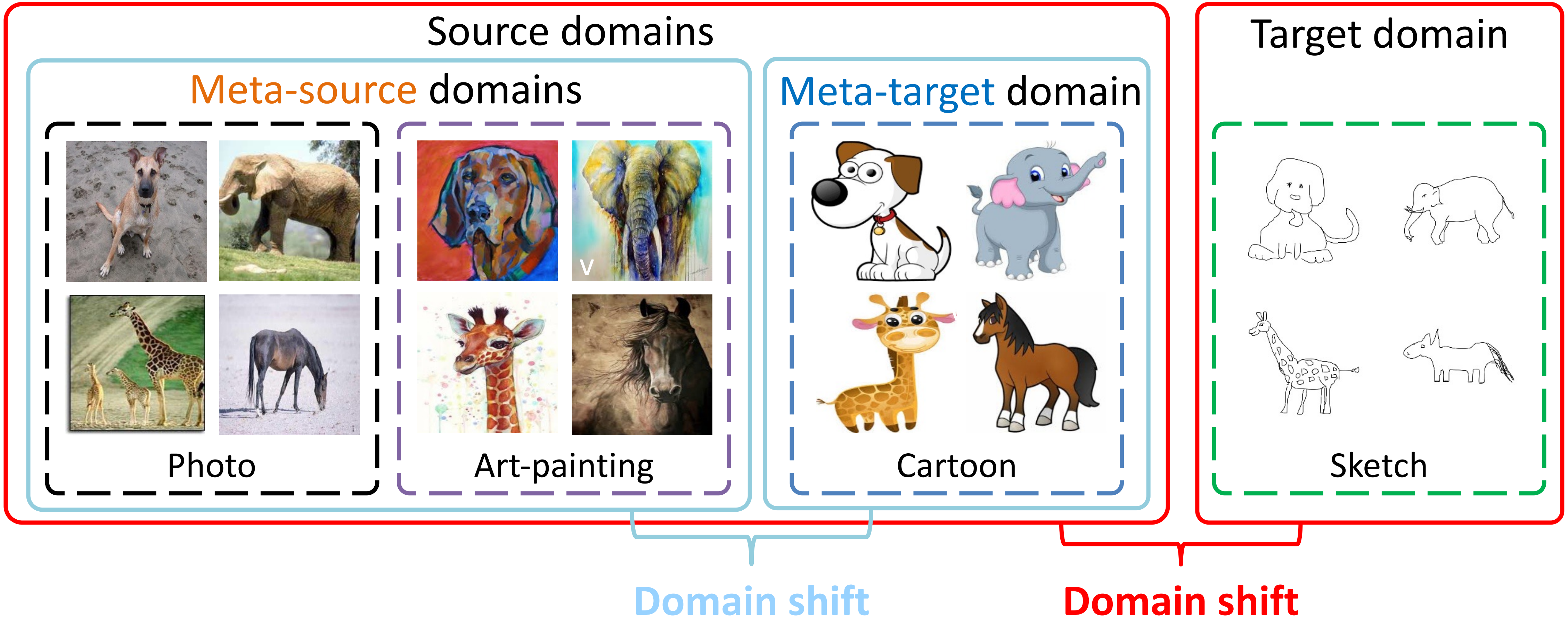
Meta-learning framework

Mimic shift between source and target by shift among source domains



Meta-learning framework

Mimic shift between source and target by shift among source domains



Adaptation as variational inference

Incorporate the test sample as a conditional for generating model parameters

$$\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \mathcal{T}') = \log \int p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'}) p(\boldsymbol{\theta}_{t'}|\mathcal{T}') d\boldsymbol{\theta}_{t'},$$

Meta-target

Adaptation as variational inference

Incorporate the test sample as a conditional for generating model parameters

$$\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \mathcal{T}') = \log \int p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'})p(\boldsymbol{\theta}_{t'}|\mathcal{T}')d\boldsymbol{\theta}_{t'},$$

Meta-target

Intractable during inference, so we approximate by source domain similarity

$$\geq \mathbb{E}_{q(\boldsymbol{\theta}_{t'})}[\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'})] - \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'}, \mathcal{S}')||p(\boldsymbol{\theta}_{t'}|\mathcal{T}')].$$

Meta-source

Adaptation as variational inference

Incorporate the test sample as a conditional for generating model parameters

$$\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \mathcal{T}') = \log \int p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'}) p(\boldsymbol{\theta}_{t'}|\mathcal{T}') d\boldsymbol{\theta}_{t'},$$

Meta-target

Intractable during inference, so we approximate by source domain similarity

$$\geq \mathbb{E}_{q(\boldsymbol{\theta}_{t'})}[\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'})] - \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'}, \mathcal{S}')||p(\boldsymbol{\theta}_{t'}|\mathcal{T}')].$$

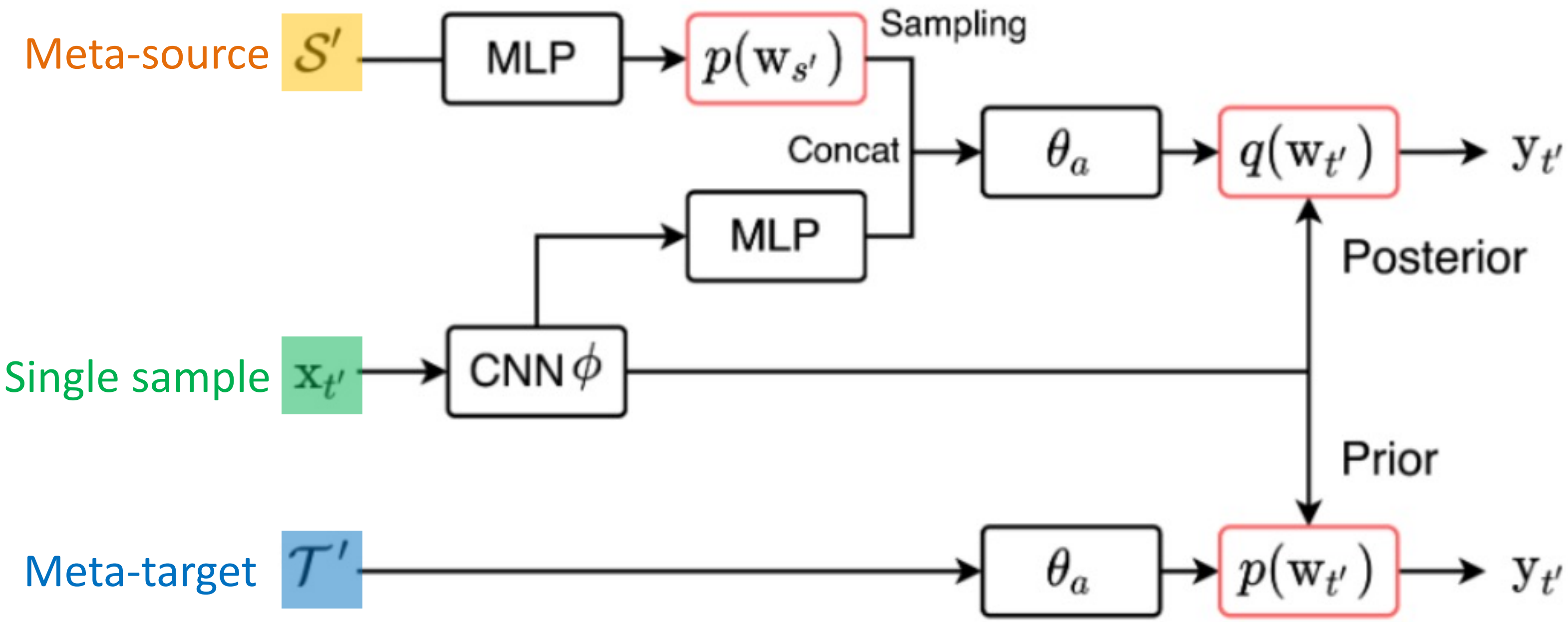
Meta-source

Our model **learns the ability to adapt** the meta-source model to each meta-target instance across different domain shifts

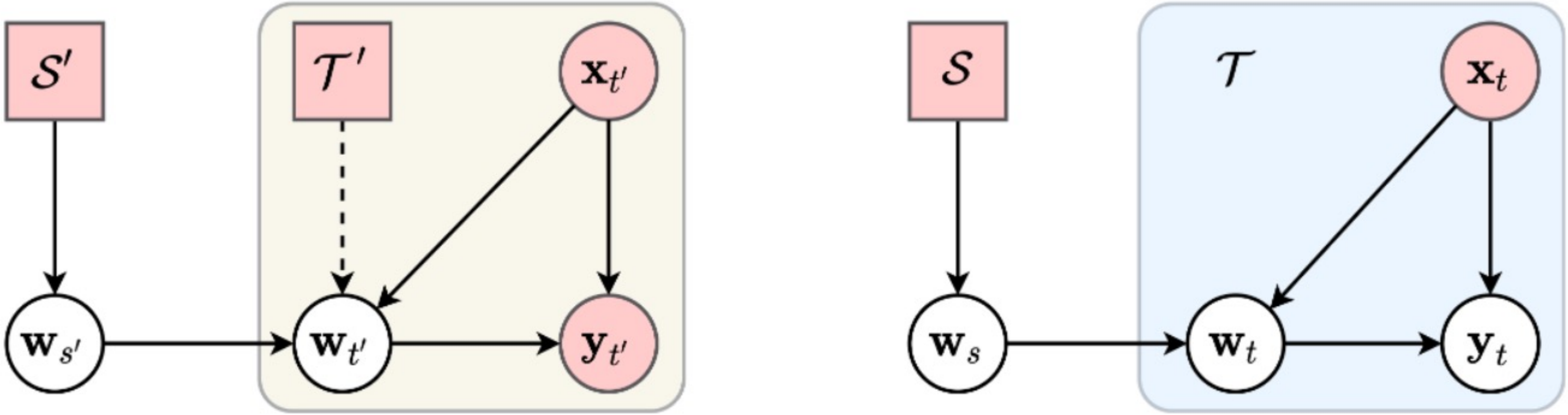
Computational feasability

We divide the model θ into a feature extractor ϕ and a classifier w.

ϕ is shared across domains, while w is trained to be adapted



Generalization at test-time

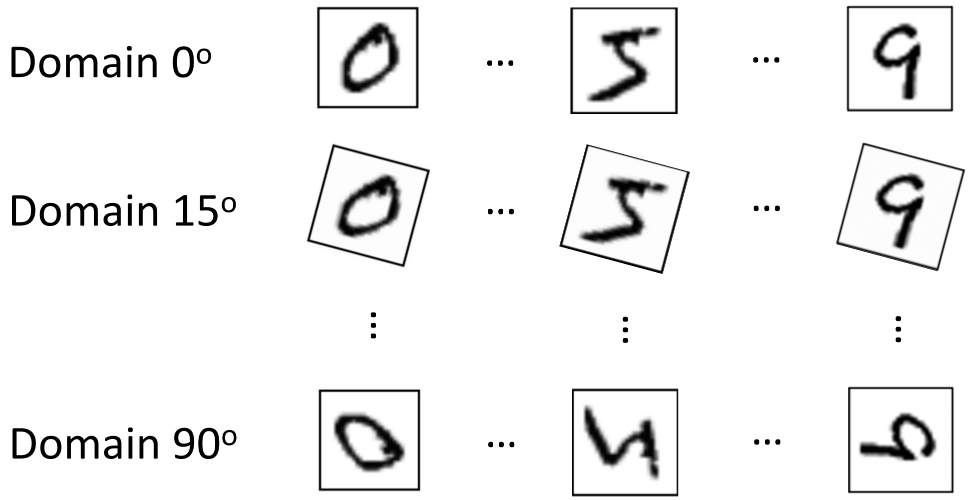


(a) Training on meta-source ($\mathcal{S}'$) and meta-target ($\mathcal{T}'$) domains (b) Testing on the unseen domain ($\mathcal{T}$)

Adaptation is achieved by generating w_t for each target sample with only one forward pass using an amortization inference network

Comparison with test-time adaptation (Tent)

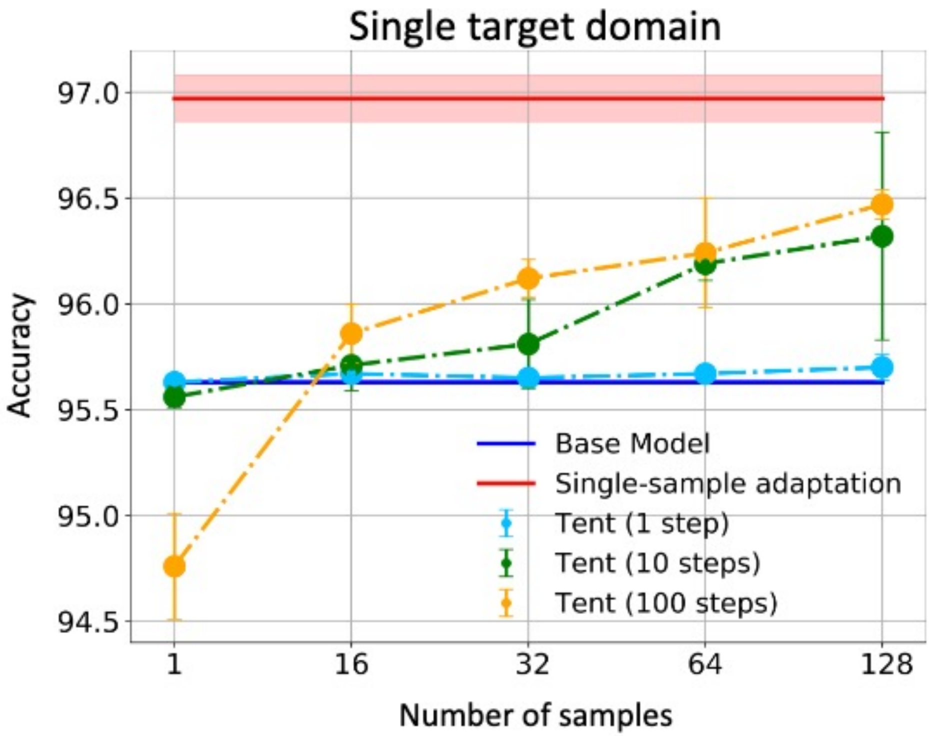
Results on Rotated-MNIST



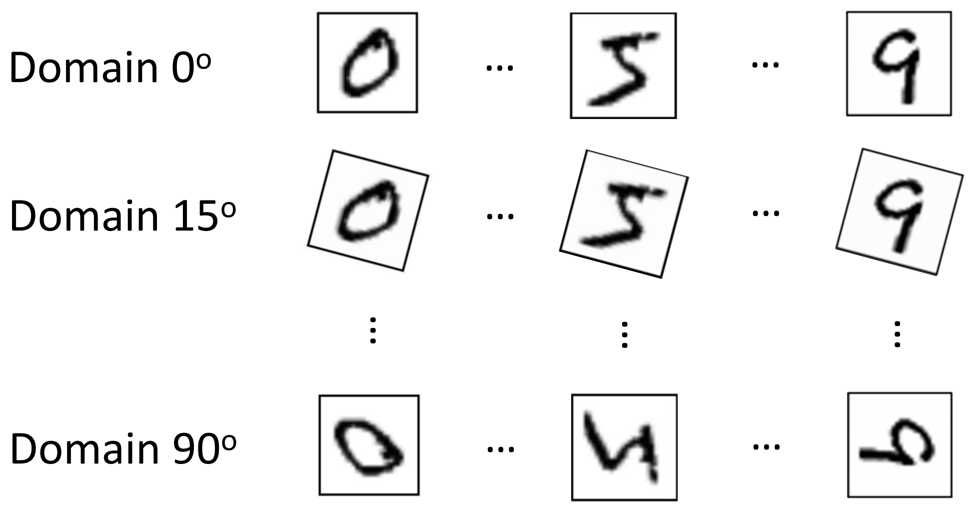
Dequan Wang *et al.* Tent: Fully Test-time Adaptation by Entropy Minimization. In ICLR 2021.

Comparison with test-time adaptation (Tent)

Tent works well with a large batch of samples from a single target domain



Results on Rotated-MNIST

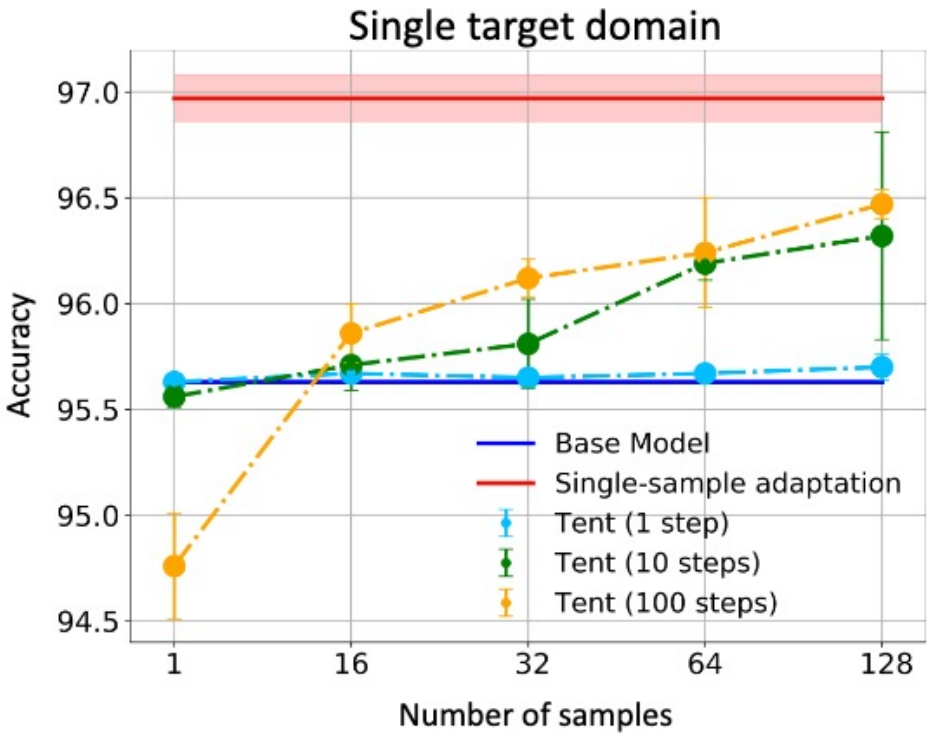


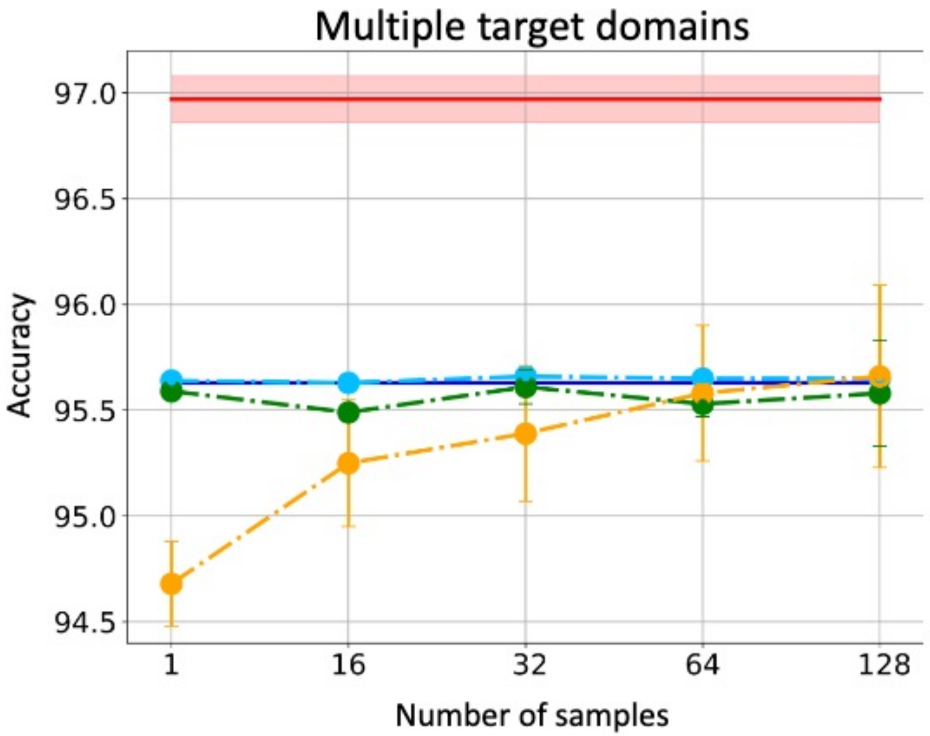
Dequan Wang *et al.* Tent: Fully Test-time Adaptation by Entropy Minimization. In ICLR 2021.

Comparison with test-time adaptation (Tent)

Tent works well with a large batch of samples from a single target domain

We outperform with a single sample, especially for multiple target domains





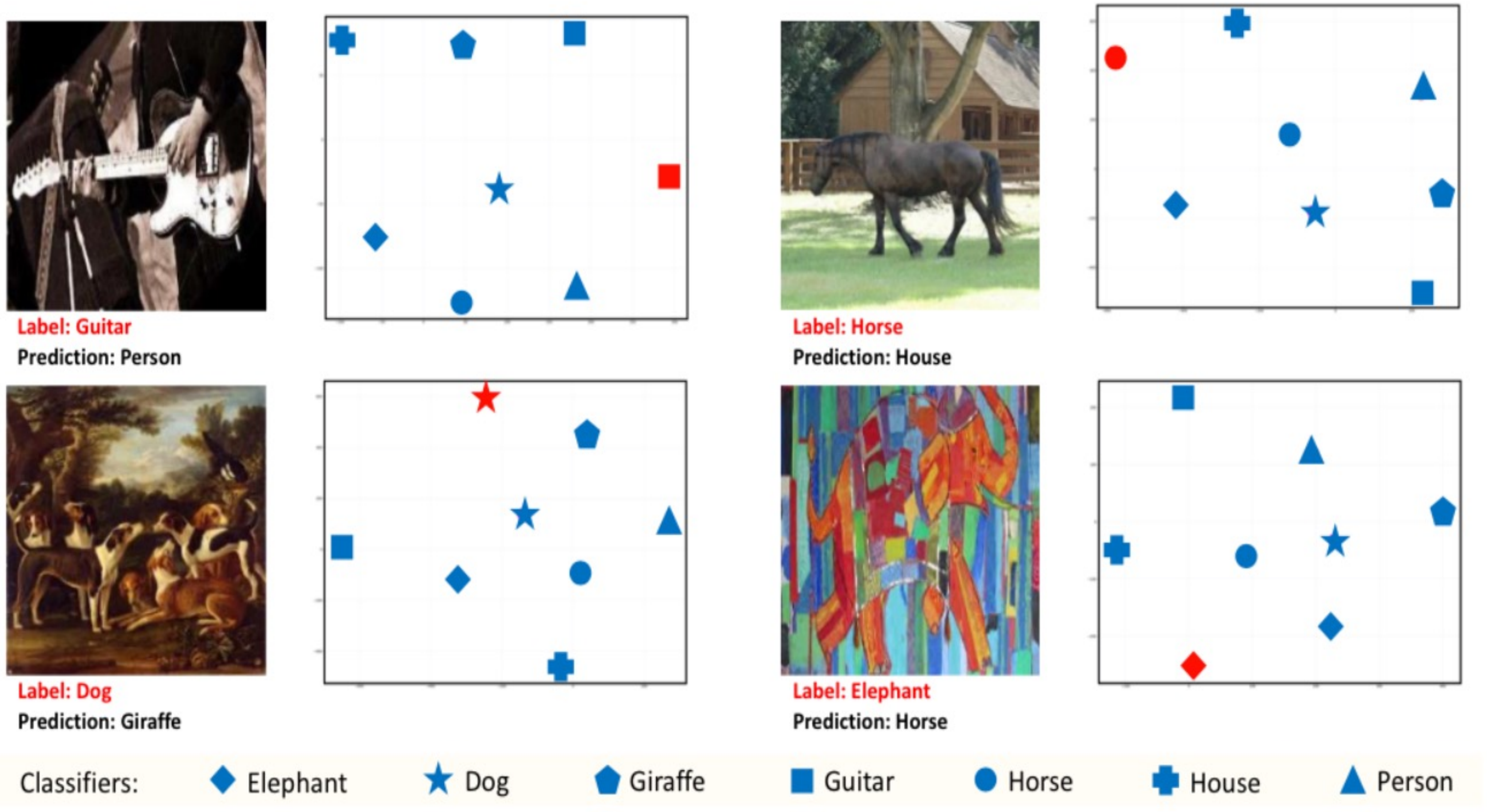
Dequan Wang *et al.* Tent: Fully Test-time Adaptation by Entropy Minimization. In ICLR 2021.

More comparisons

The better the base network, the more we gain.

	PACS benchmark		Office-Home benchmark		
	ResNet-18	**ResNet-50**	**ResNet-18**	**ResNet-50**	
Wang et al. ICLR 2021	83.09	86.23	64.13	67.99	Test-time adaptation
Dubey et al. CVPR 2021	--	84.50		68.90	
Zhou et al. ECCV 2020	83.70	84.90	65.63	67.66	Domain generation
Seo et al. ECCV 2020	**85.11**	86.64	62.90	--	Normalization
Ours	84.15	**87.51**	**66.02**	**71.07**	

Failure cases



Conclusions

Need for **real-world learning** across domains, labels, tasks and with fairness.

Need to **question** common learning **assumptions.**

Label, task and **domain** assumptions can be relaxed during learning.

Thank you

https://ivi.fnwi.uva.nl/vislab/

Contact info

Prof. dr. Cees Snoek
https://www.ceessnoek.info

cgmsnoek@uva.nl

@cgmsnoek