Towards Few-Annotation Learning in Computer Vision: Application to Image Classification and Object Detection tasks

Quentin Bouniot

29/03/2023

Jury members:

Céline Hudelot, Professor, CentraleSupélec - Reviewer

Nicolas Thome, Professor, Sorbonne University - Reviewer

Diane Larlus, Research Scientist, Naver Labs Europe - Examiner

Devis Tuia, Associate Professor, EPFL - Examiner

David Filliat, Professor, ENSTA Paris - Examiner

Ievgen Redko, Principal Research Scientist, Huawei - Guest

Angélique Loesch, Research Scientist, CEA-List - Supervisor

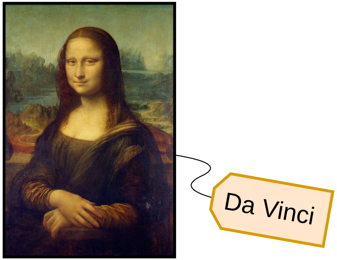
Romaric Audigier, Research Scientist, CEA-List - Supervisor

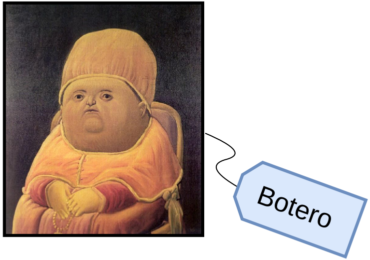
Amaury Habrard, Professor, Université Jean-Monnet - Director

cea list

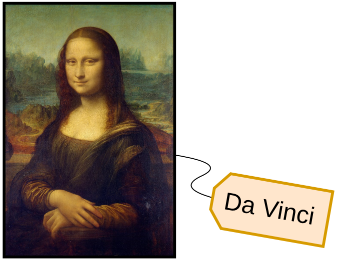
UNIVERSITÉ JEAN MONNET SAINT-ÉTIENNE

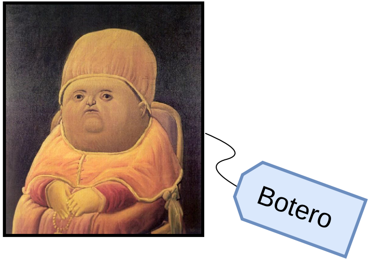
A Simple Problem ...

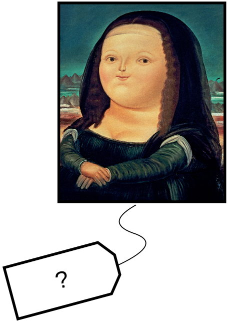




A Simple Problem ...

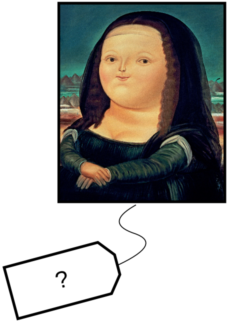






Who is the painter ?

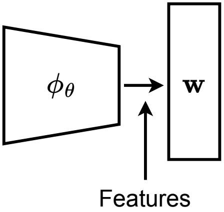
A Simple Problem ... for a Human !

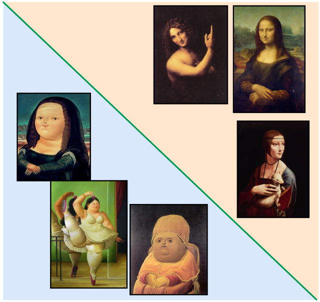


Who is the painter ?

- *Human* capacity to learn from few examples

Image Classification





- ϕ encoding function parametrized by θ
- Linear classifiers $\mathbf{w}$ (green line) separate each class

Learning from images

$$\mathcal{D}_{train} := \{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_N, \mathbf{y}_N)\} \sim P(\mathbf{X}, \mathbf{Y})$$

Model parameters

Data points

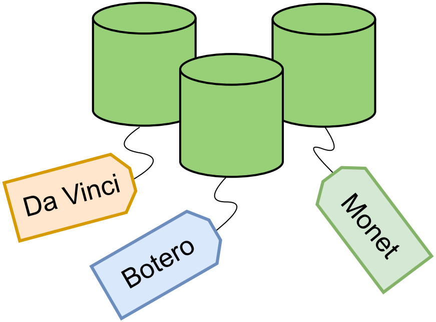
$$\underbrace{\hat{\theta}, \hat{\mathbf{w}}}_{\text{Model parameters}} := \arg\min_{\theta, \mathbf{w}} \sum_{i=1}^{N} \underbrace{\mathcal{L}}_{\text{Loss function}} (\underbrace{\mathbf{y}_i}_{\text{Label}} , \overbrace{\mathbf{x}_i}^{\text{Data points}} ; \theta, \mathbf{w})$$

Label

Loss function

- Learn parameters $\hat{\theta}$ and $\hat{\mathbf{w}}$ minimizing loss function $\mathcal{L}$ given data points $\mathbf{x}_i$ and labels $\mathbf{y}_i$.

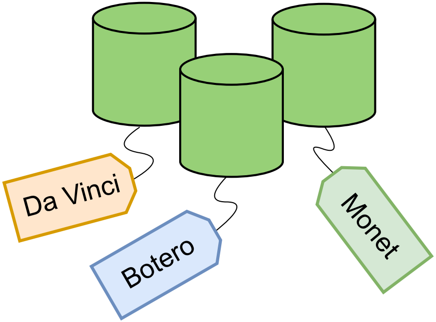
Practical Data Conditions



Expectations

- Many-Shot Learning: A lot of data and labels

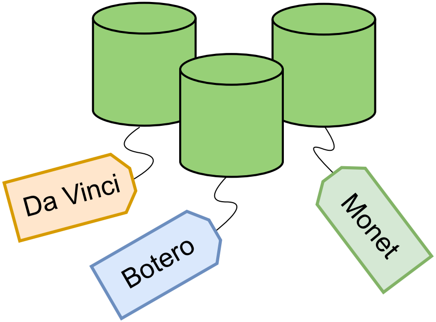
Practical Data Conditions

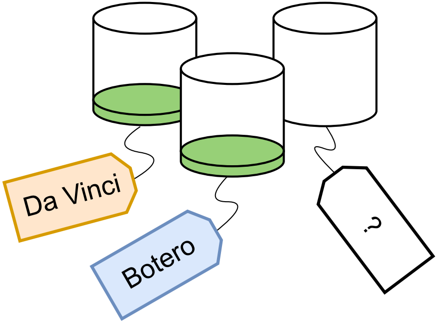


Expectations

- Many-Shot Learning: A lot of data and labels
- But labeling data is costly !

Practical Data Conditions





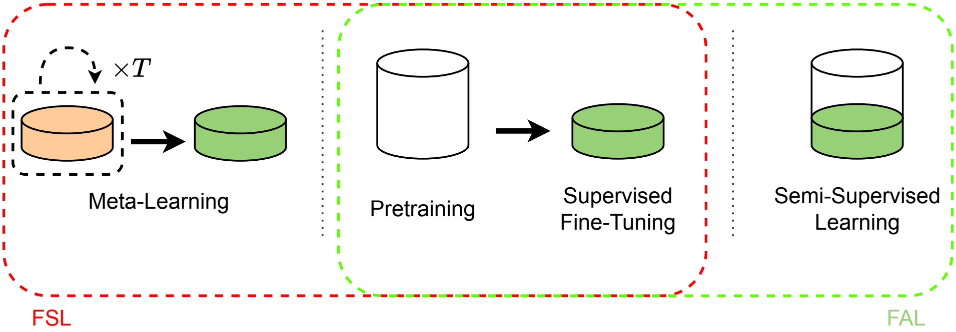
Expectations

- Many-Shot Learning: A lot of data and labels
- But labeling data is costly !

Reality

- Few Annotation Learning (FAL): A lot of data and few labels
- Few Shot Learning (FSL): Few data and labels

General Frameworks



- ▶ Contribution 1
- ▶ Contribution 2
- ▶ Contribution 3

Outline

1. Introduction
2. Improving Few-Shot Classification with Meta-Learning through Multi-Task Learning
   - Meta-Learning 101
   - Multi-Task Representation Learning Theory
   - Contrib 1: From Theory to Practice[1]
3. Improving Few-Annotation Learning for Object Detection
   - Background in Object Detection
   - Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation[2]
   - Contrib 3: Few Annotation Learning for Semi-Supervised Object Detection[3]
4. Conclusion and Broader Impacts

[1] Quentin Bouniot, Ievgen Redko, et al. "Improving Few-Shot Learning Through Multi-task Representation Learning Theory". In: *ECCV*. 2022.

[2] Quentin Bouniot, Romaric Audigier, et al. "Proposal-Contrastive Pretraining for Object Detection from Fewer Data". In: *ICLR*. 2023.

[3] Quentin Bouniot, Angélique Loesch, et al. "Towards Few-Annotation Learning for Object Detection: Are Transformer-Based Models More Efficient?" In: *WACV*. 2023.

Outline

1 Introduction

2 Improving Few-Shot Classification with Meta-Learning through Multi-Task Learning
- Meta-Learning 101
- Multi-Task Representation Learning Theory
- Contrib 1: From Theory to Practice[4]

3 Improving Few-Annotation Learning for Object Detection

4 Conclusion and Broader Impacts

[4] Quentin Bouniot, Ievgen Redko, et al. "Improving Few-Shot Learning Through Multi-task Representation Learning Theory". In: *ECCV*. 2022.

What is Meta-Learning ?

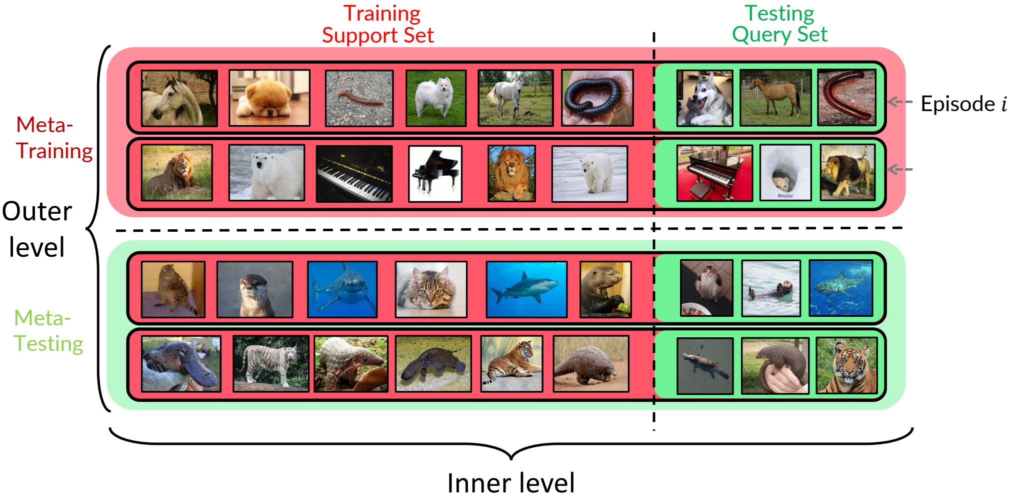
- Meta-Training: solve a set of *source tasks*.
- Meta-Testing: use knowledge from meta-training to solve *previously unseen tasks* more efficiently.

How is it related to Few-Shot Learning ?

The Meta-learner *learns to learn* a new task with few shots.

Introducing episodes

Meta-Learning 101



N-way k-shot episode: task with N different classes and k images for each class.

Meta-Learning Problem Formulation

Meta-Learning 101

Data distributions:

Drawing N episodes

$$\forall t \in [1, \dots, N], \qquad \mathcal{T}_t \sim P(\mathcal{T}), \qquad \mathcal{T}_t := \mathcal{S}_t \cup \mathcal{Q}_t$$

Support sets

Query sets

Meta-Learning Problem Formulation

Meta-Learning 101

Data distributions:

Drawing N episodes

$$\forall t \in [1, \ldots, N], \qquad \mathcal{T}_t \sim P(\mathcal{T}), \qquad \mathcal{T}_t := \mathcal{S}_t \cup \mathcal{Q}_t$$

Support sets

Query sets

Inner-level:

Inner loss function

$$\boxed{\hat{\theta}_t, \hat{\mathbf{w}}_t} = \arg\min_{\theta, \mathbf{w}} \sum_{(x,y)\in\mathcal{S}_t} \boxed{\mathcal{L}_{\text{inner}}}(x, y; \theta, \mathbf{w})$$

Parameters specialized to each episode

Meta-Learning Problem Formulation

Meta-Learning 101

Data distributions:

Drawing N episodes

$$\forall t \in [1, \ldots, N], \qquad \mathcal{T}_t \sim P(\mathcal{T}), \qquad \mathcal{T}_t := \mathcal{S}_t \cup \mathcal{Q}_t$$

Support sets

Query sets

Inner-level:

Inner loss function

$$\boxed{\hat{\theta}_t, \hat{\mathbf{w}}_t} = \arg\min_{\theta, \mathbf{w}} \sum_{(x,y) \in \mathcal{S}_t} \boxed{\mathcal{L}_{\text{inner}}}(x, y; \theta, \mathbf{w})$$

Parameters specialized to each episode

Outer-level:

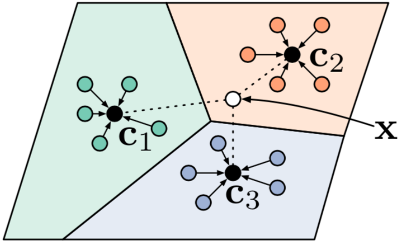
Initialization for new sets of episodes

Task-specific parameters learned

$$\boxed{\hat{\theta}, \hat{\mathbf{w}}} = \arg\min_{\theta, \mathbf{w}} \sum_{t=1}^{N} \sum_{(x,y) \in \mathcal{Q}_t} \boxed{\mathcal{L}_{\text{outer}}}(x, y; \hat{\theta}_t, \hat{\mathbf{w}}_t)$$

Outer loss function

Metric-based methods (ProtoNet [5])



- Support samples for each class i fused into **prototypes** $\mathbf{c}_i$.
- Probability distribution using **inverse of distances** to prototypes.

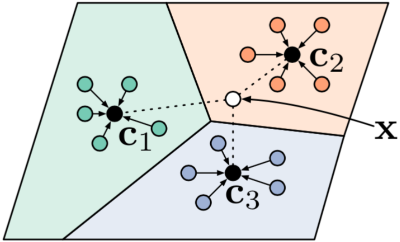
[5] Jake Snell, Kevin Swersky, and Richard S. Zemel. "Prototypical Networks for Few-shot Learning". In: *NeurIPS*. 2017

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks". In: *ICML*. 2017

Meta-Learning methods

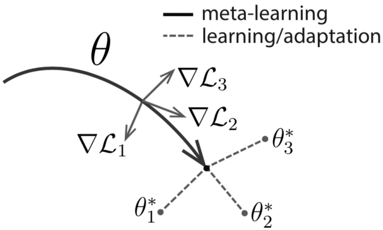
Meta-Learning 101

Metric-based methods (ProtoNet [5])



- Support samples for each class i fused into **prototypes** $\mathbf{c}_i$.
- Probability distribution using **inverse of distances** to prototypes.

Gradient-based methods (MAML [6])



- **End-to-end** bi-level optimization through **gradient descent**.

[5] Jake Snell, Kevin Swersky, and Richard S. Zemel. "Prototypical Networks for Few-shot Learning". In: *NeurIPS*. 2017

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks". In: *ICML*. 2017

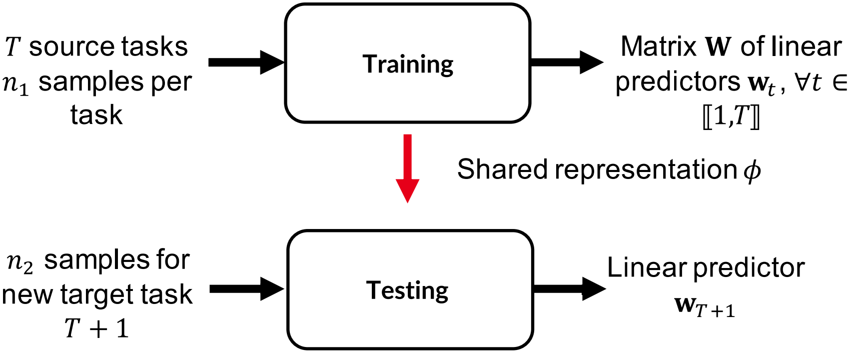
Introduction to MTR

Multi-Task Representation Learning Theory

T source tasks
n_1 samples per task → **Training** → Matrix $\mathbf{W}$ of linear predictors $\mathbf{w}_t$, $\forall t \in [\![1,T]\!]$

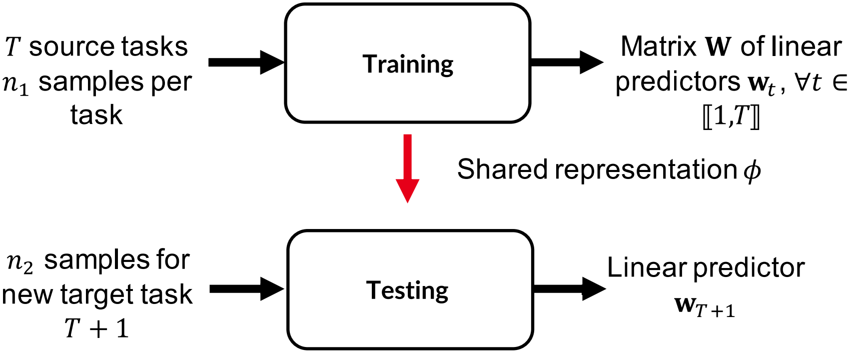
Introduction to MTR

Multi-Task Representation Learning Theory



Introduction to MTR

Multi-Task Representation Learning Theory

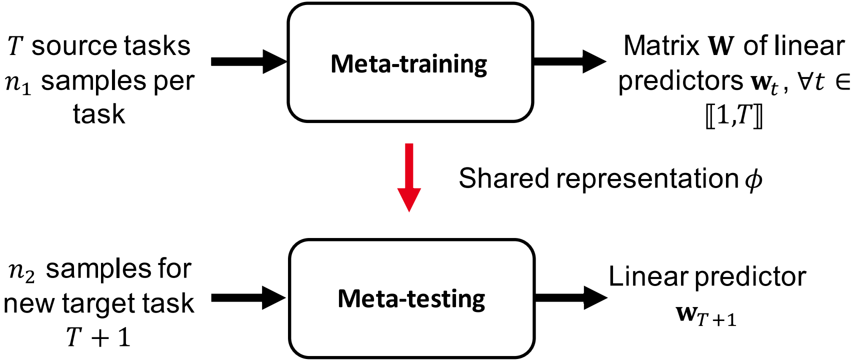


Goal: Minimize **excess risk** ER $= \mathcal{L}(\hat{\phi}, \hat{\mathbf{w}}_{T+1}) - \mathcal{L}(\phi^*, \mathbf{w}^*_{T+1})$,

- **True risk $\mathcal{L}$**
- **Optimal representation ϕ^***
- $\mathbf{w}^*_{T+1}$ **ideal target linear predictor.**

Link with Meta-Learning

Multi-Task Representation Learning Theory



Goal: Minimize **excess risk** ER $= \mathcal{L}(\hat{\phi}, \hat{\mathbf{w}}_{T+1}) - \mathcal{L}(\phi^*, \mathbf{w}^*_{T+1})$,

- **True risk $\mathcal{L}$**
- **Optimal representation ϕ^***
- $\mathbf{w}^*_{T+1}$ **ideal target linear predictor.**

Important Assumptions

Multi-Task Representation Learning Theory

Assumption 1: Diversity of the source tasks[7]

Condition Number $\kappa(\mathbf{W}^*) = \frac{\sigma_{\max}(\mathbf{W}^*)}{\sigma_{\min}(\mathbf{W}^*)}$ *should not increase* with T.

- Optimal predictors $\mathbf{W}^* = [\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*]$ **cover** all the **directions evenly**

[7] Simon S. Du et al. "Few-Shot Learning via Learning the Representation, Provably". In: *ICLR*. 2021; Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. "Provable Meta-Learning of Linear Representations". In: *arXiv*. 2020.

Important Assumptions

Multi-Task Representation Learning Theory

Assumption 1: Diversity of the source tasks[7]

Condition Number $\kappa(\mathbf{W}^*) = \frac{\sigma_{\max}(\mathbf{W}^*)}{\sigma_{\min}(\mathbf{W}^*)}$ *should not increase* with T.

- Optimal predictors $\mathbf{W}^* = [\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*]$ **cover** all the **directions evenly**

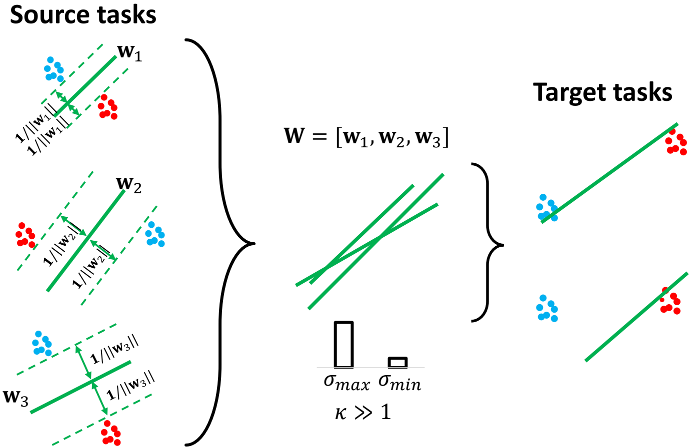
Assumption 2: Constant classification margin[7]

Norm of predictors $\|\mathbf{w}_t^*\|_{t \in [\![1,T]\!]}$ *should not increase* with T

[7] Simon S. Du et al. "Few-Shot Learning via Learning the Representation, Provably". In: *ICLR*. 2021; Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. "Provable Meta-Learning of Linear Representations". In: *arXiv*. 2020.

Illustration: Violated Assumptions

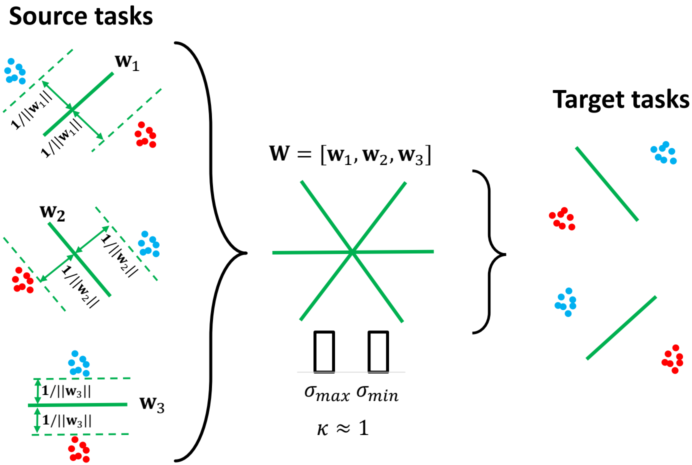
Multi-Task Representation Learning Theory



× Linear predictors cover **only part of the space** or **over-specialize** to the tasks

Illustration: Satisfied Assumptions

Multi-Task Representation Learning Theory



✓ **Assumption 1** makes sure that **linear predictors are complementary**

✓ **Assumption 2 avoids under- or over-specialization** to the tasks

Few-Shot Learning bound[8]

If assumptions are satisfied:

$$\mathrm{ER}(\phi, \mathbf{w}_{T+1}) \leq O\left(\frac{1}{n_1 T} + \frac{1}{n_2}\right)$$

Number of samples for target task

Number of samples per source tasks

Number of source tasks

- ✓ All *source* and *target* data are useful to decrease the bound of *excess risk*.
- ✓ Increasing **either** T or n_1 have an effect on the bound.

[8] Simon S. Du et al. "Few-Shot Learning via Learning the Representation, Provably". In: *ICLR*. 2021; Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. "Provable Meta-Learning of Linear Representations". In: *arXiv*. 2020.

Outline

1 Introduction

2 Improving Few-Shot Classification with Meta-Learning through Multi-Task Learning
- Meta-Learning 101
- Multi-Task Representation Learning Theory
- Contrib 1: From Theory to Practice[9]

3 Improving Few-Annotation Learning for Object Detection

4 Conclusion and Broader Impacts

[9] Quentin Bouniot, Ievgen Redko, et al. "Improving Few-Shot Learning Through Multi-task Representation Learning Theory". In: *ECCV*. 2022.

What Happens in Practice ?

Contrib 1: From Theory to Practice

Idea:

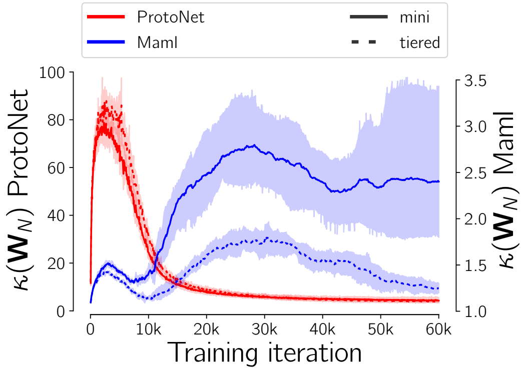
- Verify *assumptions 1 and 2* for meta-learning algorithms.

How ?

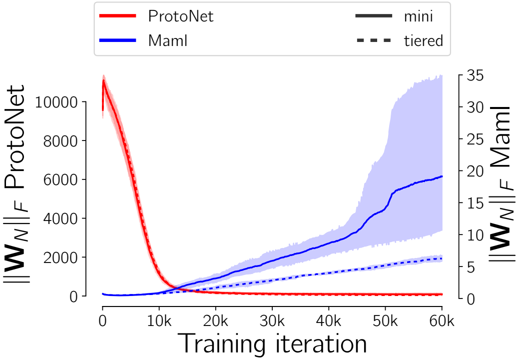
- Monitor *condition number* $\kappa(\mathbf{W}_N)$ and *norm of the predictors* $\|\mathbf{W}_N\|_F$ for the last N tasks

What Happens in Practice ?

Contrib 1: From Theory to Practice



Monitoring the *condition number*



Monitoring the *norm*

- ✓ ProtoNet *naturally verifies* the assumptions
- × MAML *does not verify* the assumptions

Why Does it Happen ?

Contrib 1: From Theory to Practice

Case of ProtoNet:

- Theorem (informal)

> If all prototypes are normalized,
> then all ProtoNet encoders verify Assumption 1.

✓ Norm minimization is *enough* to obtain well-behaved condition number for ProtoNet.

Case of MAML:

- Theorem (informal)

$$\boxed{\text{At iteration } i\text{, if } \sigma_{\text{min}} = 0 \text{ for last two tasks, then } \kappa(\hat{\mathbf{W}}_2^{i+1}) \geq \kappa(\hat{\mathbf{W}}_2^{i}).}$$

✓ The condition number for MAML can **increase** between iterations.

Ensuring Assumption 1: Spectral regularization

$$\boxed{\kappa(\mathbf{W}_N) = \frac{\sigma_{\text{max}}(\mathbf{W}_N)}{\sigma_{\text{min}}(\mathbf{W}_N)}}$$

✓ Regularizing with $\kappa(\mathbf{W}_N)$ leads to a better coverage of the searched space

Ensuring Assumption 1: **Spectral** regularization

$$\boxed{\kappa(\mathbf{W}_N) = \frac{\sigma_{\text{max}}(\mathbf{W}_N)}{\sigma_{\text{min}}(\mathbf{W}_N)}}$$

✓ Regularizing with $\kappa(\mathbf{W}_N)$ leads to a better coverage of the searched space

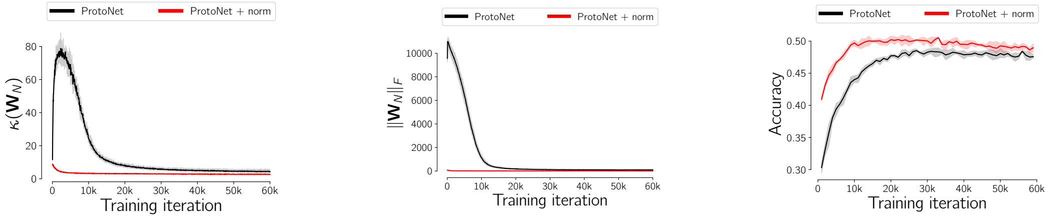
Ensuring Assumption 2: **Norm regularization or normalization** for linear predictors

✓ **Normalizing predictors** ensure **constant margin** that **does not change** with T

Experimental Results

Contrib 1: From Theory to Practice

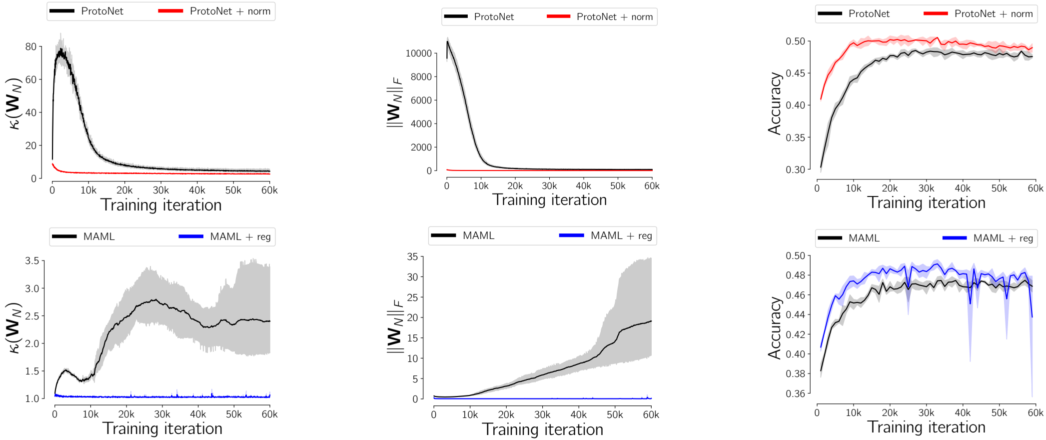
Experiments on mini-ImageNet 5-way 1-shot



Experimental Results

Contrib 1: From Theory to Practice

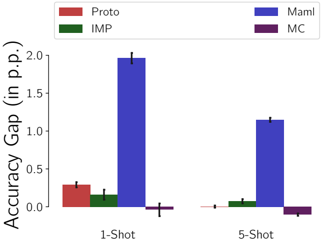
Experiments on mini-ImageNet 5-way 1-shot



✓ Our **regularization** and **normalization** have the intended effects.

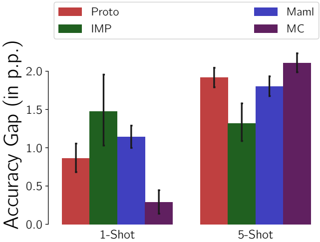
Experimental Results

Contrib 1: From Theory to Practice



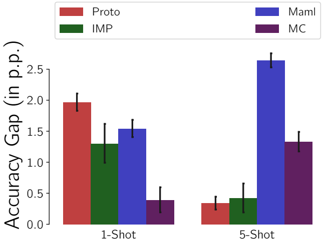
Omniglot

20-way



miniImageNet

5-way



tieredImageNet

5-way

- ✓ *Statistically significant* improvements with our regularization and normalization.
- ✓ *Better generalization* when the assumptions are not verified naturally.

Improving Few-Shot Learning Through Multi-Task Representation Learning Theory

- ✓ **Connection** between Meta-Learning and Multi-Task Representation Learning Theory
- ✓ Explaining why some meta-learning methods **naturally fulfill** theoretical assumptions of the best learning bounds.
- ✓ We prove that it is possible to enforce the assumptions and propose **practical ways** which leads to **significant** performance improvements.

Outline

1. Introduction

2. Improving Few-Shot Classification with Meta-Learning through Multi-Task Learning

3. Improving Few-Annotation Learning for Object Detection
   - Background in Object Detection
   - Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation[10]
   - Contrib 3: Few Annotation Learning for Semi-Supervised Object Detection[11]

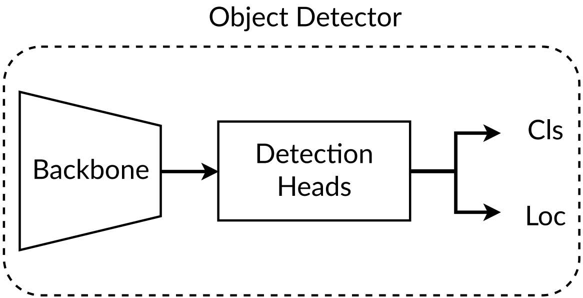
4. Conclusion and Broader Impacts

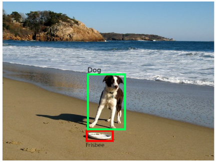
[10] Quentin Bouniot, Romaric Audigier, et al. "Proposal-Contrastive Pretraining for Object Detection from Fewer Data". In: *ICLR*. 2023.

[11] Quentin Bouniot, Angélique Loesch, et al. "Towards Few-Annotation Learning for Object Detection: Are Transformer-Based Models More Efficient?" In: *WACV*. 2023.

Object Detectors in a Nutshell

Background in Object Detection



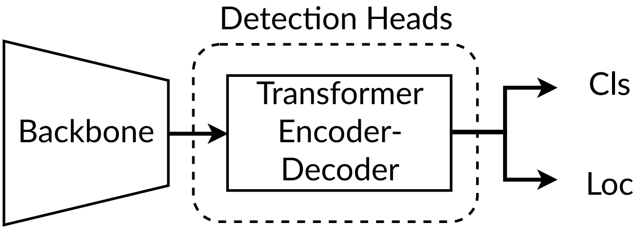


- Detectors composed of **backbone model** and **detection-specific heads**.
- Predict **class (Cls)** and **location (Loc)** for each objects in an image.

Object Detection 101

Background in Object Detection

Transformer-based methods (e.g., DETR[12])



- **Simpler** overall architecture, without **hand-crafted heuristics**.
- Increasingly popular architecture and **strong performance with few data**.

[12] Nicolas Carion et al. "End-to-end object detection with transformers". In: *ECCV*. 2020.

Few-Shot Learning Setting

Background in Object Detection

How do object detectors handle data scarcity ?

Method	Arch.	Mini-COCO			
		0.5% (590)	1% (1.2k)	5% (5.9k)	10% (11.8k)
FCOS[13]	Conv.	5.42 ± 0.01	8.43 ± 0.03	17.01 ± 0.01	20.98 ± 0.01
FRCNN + FPN[14]	Conv.	6.83 ± 0.15	9.05 ± 0.16	18.47 ± 0.22	23.86 ± 0.81
Def. DETR[15]	Trans.	$\mathbf{8.95 \pm 0.51}$	$\mathbf{12.96 \pm 0.08}$	$\mathbf{23.59 \pm 0.21}$	$\mathbf{28.55 \pm 0.08}$

- Performance on COCO with different **percentages** of labeled training data.
- **Def. DETR** stronger than FRCNN + FPN and FCOS **with fewer labeled data.**

[13] Zhi Tian et al. “Fcos: Fully convolutional one-stage object detection”. In: *ICCV*. 2019.

[14] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *NeurIPS*. 2015; Tsung-Yi Lin et al. “Feature pyramid networks for object detection”. In: *CVPR*. 2017.

[15] Xizhou Zhu et al. “Deformable DETR: Deformable Transformers for End-to-End Object Detection”. In: *ICLR*. 2021.

Outline

1 Introduction

2 Improving Few-Shot Classification with Meta-Learning through Multi-Task Learning

3 Improving Few-Annotation Learning for Object Detection
- Background in Object Detection
- Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation[16]
- Contrib 3: Few Annotation Learning for Semi-Supervised Object Detection[17]

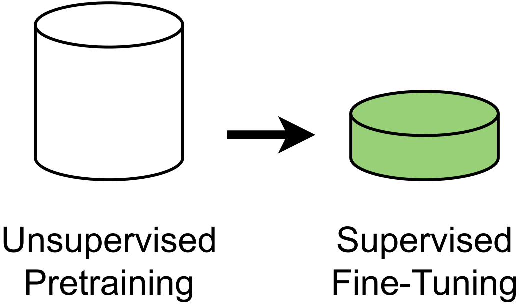
4 Conclusion and Broader Impacts

[16] Quentin Bouniot, Romaric Audigier, et al. "Proposal-Contrastive Pretraining for Object Detection from Fewer Data". In: *ICLR*. 2023.

[17] Quentin Bouniot, Angélique Loesch, et al. "Towards Few-Annotation Learning for Object Detection: Are Transformer-Based Models More Efficient?" In: *WACV*. 2023.

Setting considered

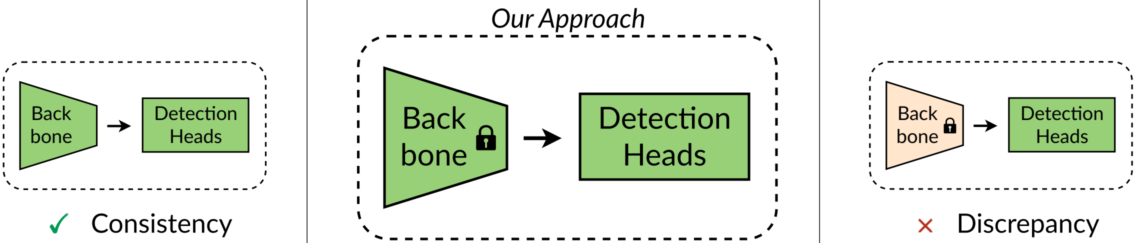
Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation



Pretraining in Object Detection

Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation

Overall Pretraining



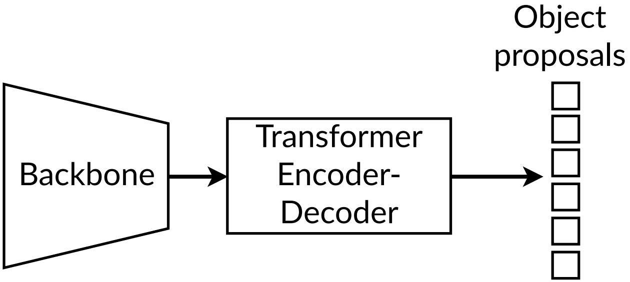
- ✓ Consistency
- × Costly

- ✓ Consistency
- ✓ Less costly

- × Discrepancy
- ✓ Less costly

Transformer-based Detectors

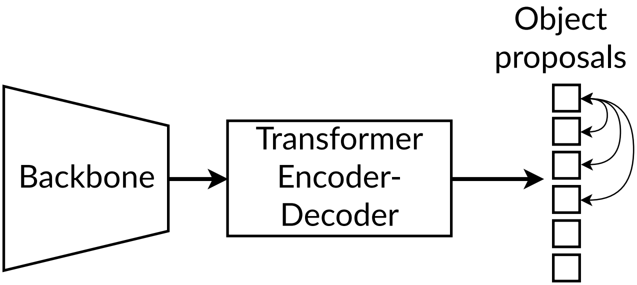
Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation



- Transformer-based detectors generates N proposals $\gg k$ objects in images.

Transformer-based Detectors

Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation

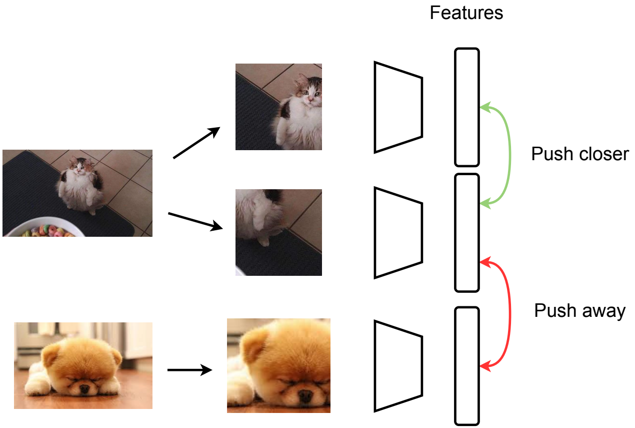


- Transformer-based detectors generates N proposals $\gg k$ objects in images.

Contribution: Contrastive learning between proposals.

Classical Contrastive Learning

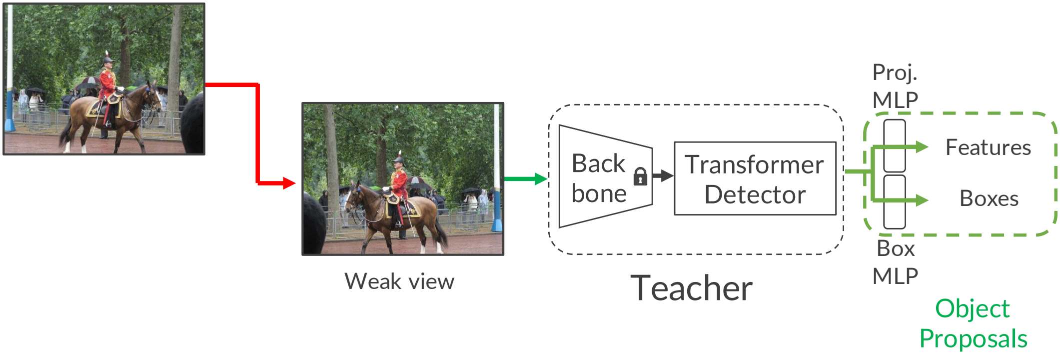
Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation



- Push closer positive examples and push away negative examples.

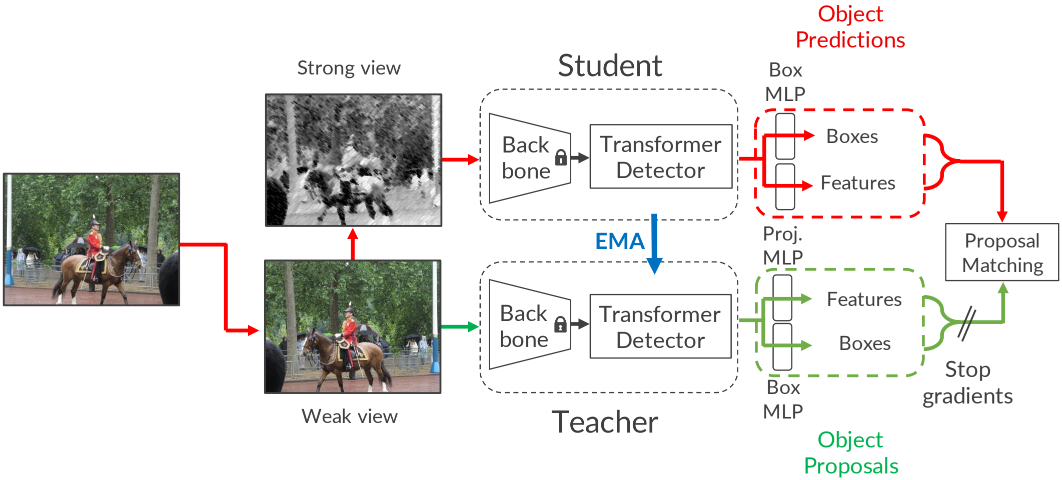
Proposal-Contrastive Learning

Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation



Proposal-Contrastive Learning

Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation



- Object Proposals from Teacher are matched with Predictions from Student.

Proposal-Contrastive Learning

Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation

Unsupervised Proposal Matching

Object Proposals

$$\hat{\sigma}_i^{\text{prop}} = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_{j=1}^{N} \mathcal{L}_{\text{prop\_match}}\left(\mathbf{y}_{(i,j)}, \hat{\mathbf{y}}_{(i,\sigma(j))}\right)$$

Permutations of N elements

Object Predictions

- Proposal j found by the teacher associated to prediction $\hat{\sigma}_i^{\text{prop}}(j)$ of the student.

Proposal-Contrastive Learning

Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation

Unsupervised Proposal Matching

Object Proposals

$$\hat{\sigma}_i^{\text{prop}} = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_{j=1}^{N} \mathcal{L}_{\text{prop\_match}}\left(\mathbf{y}_{(i,j)}, \hat{\mathbf{y}}_{(i,\sigma(j))}\right)$$

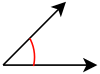
Permutations of N elements

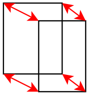
Object Predictions

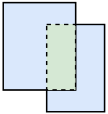
- Proposal j found by the teacher associated to prediction $\hat{\sigma}_i^{\text{prop}}(j)$ of the student.

Matching Cost $\mathcal{L}_{\text{prop\_match}}$ depends on:

- features similarity
- L_1 loss of box coordinates
- generalized IoU loss



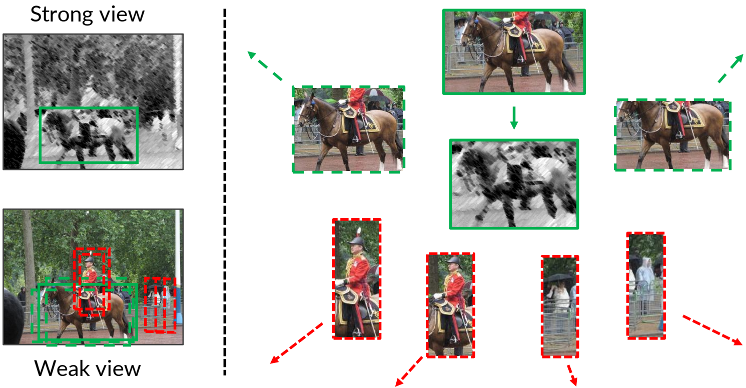




Proposal-Contrastive Learning

Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation

Naive way

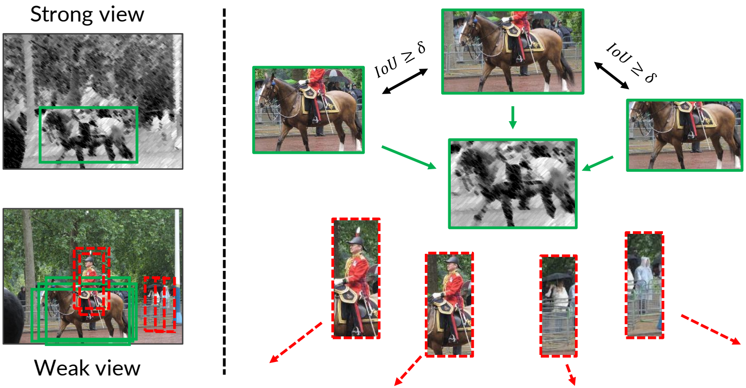


× Close proposals considered as negative examples.

Proposal-Contrastive Learning

Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation

Localization-aware Contrastive loss



✓ Overlapping proposals are considered as positive examples.

Soft Contrastive Estimation (SCE) loss function[18]

Relations between proposals

Temperature

$$\boxed{p'_{(in,jm)}} = \frac{\mathbb{1}_{i\neq n}\mathbb{1}_{j\neq m}\exp(\mathbf{z}_{(i,j)} \cdot \mathbf{z}_{(n,m)}/\tau_t)}{\sum_{k=1}^{N_b}\sum_{l=1}^{N}\mathbb{1}_{i\neq k}\mathbb{1}_{j\neq l}\exp(\mathbf{z}_{(i,j)} \cdot \mathbf{z}_{(k,l)}/\tau_t)}$$

Features of Object Proposals

[18] Julien Denize et al. "Similarity contrastive estimation for self-supervised soft contrastive learning". In: *WACV*. 2023.

Soft Contrastive Estimation (SCE) loss function[18]

Relations between proposals

Temperature

$$\boxed{p'_{(in,jm)}} = \frac{\mathbb{1}_{i\neq n}\mathbb{1}_{j\neq m}\exp(\mathbf{z}_{(i,j)}\cdot\mathbf{z}_{(n,m)}/\tau_t)}{\sum_{k=1}^{N_b}\sum_{l=1}^{N}\mathbb{1}_{i\neq k}\mathbb{1}_{j\neq l}\exp(\mathbf{z}_{(i,j)}\cdot\mathbf{z}_{(k,l)}/\tau_t)}$$

Features of Object Proposals

Features of Object Predictions

$$\boxed{p''_{(in,jm)}} = \frac{\exp(\mathbf{z}_{(i,j)}\cdot\hat{\mathbf{z}}_{(n,m)}/\tau)}{\sum_{k=1}^{N_b}\sum_{l=1}^{N}\exp(\mathbf{z}_{(i,j)}\cdot\hat{\mathbf{z}}_{(k,l)}/\tau)}$$

Contrastive aspect between predictions and proposals

[18] Julien Denize et al. "Similarity contrastive estimation for self-supervised soft contrastive learning". In: *WACV*. 2023.

Proposal-Contrastive Learning

Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation

Localization-aware similarity distribution

$$w^{\text{Loc}}_{(in,jm)} = \lambda_{\text{SCE}} \cdot \mathbb{1}_{i=n} \mathbb{1}_{IoU_i(j,m) \geq \delta} + (1 - \lambda_{\text{SCE}}) \cdot p'_{(in,jm)}$$

IoU between proposals in same image above threshold δ

Proposal-Contrastive Learning

Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation

Localization-aware similarity distribution

$$w^{\text{Loc}}_{(in,jm)} = \lambda_{\text{SCE}} \cdot \mathbb{1}_{i=n}\mathbb{1}_{IoU_i(j,m)\geq\delta} + (1-\lambda_{\text{SCE}}) \cdot p'_{(in,jm)}$$

IoU between proposals in same image above threshold δ

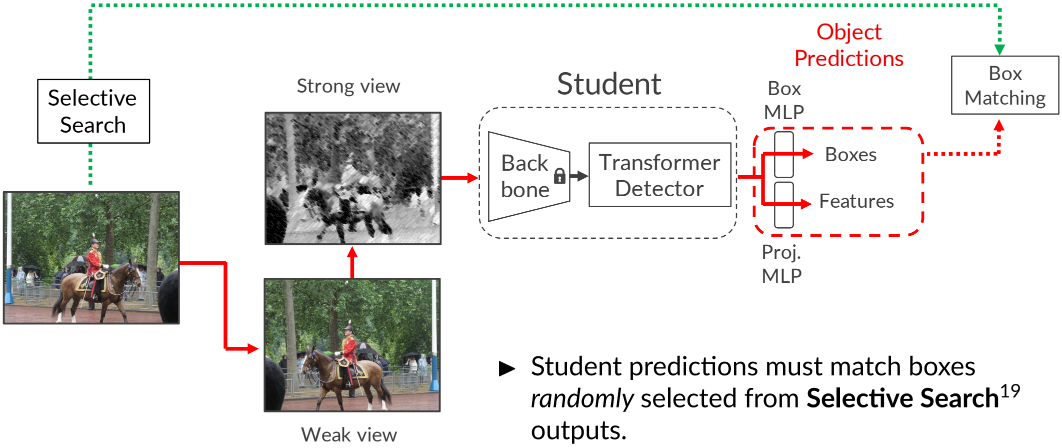
Localized SCE (LocSCE) function

$$\mathcal{L}_{\text{LocSCE}}(\mathbf{y}, \hat{\mathbf{y}}, \hat{\sigma}^{\text{prop}}) = -\frac{1}{N_b N} \sum_{i=1}^{N_b} \sum_{n=1}^{N_b} \sum_{j=1}^{N} \sum_{m=1}^{N} w^{\text{Loc}}_{(in,jm)} \log(p''_{(in,j\hat{\sigma}^{\text{prop}}_n(m))})$$

Effective batch size

Avoiding Collapse

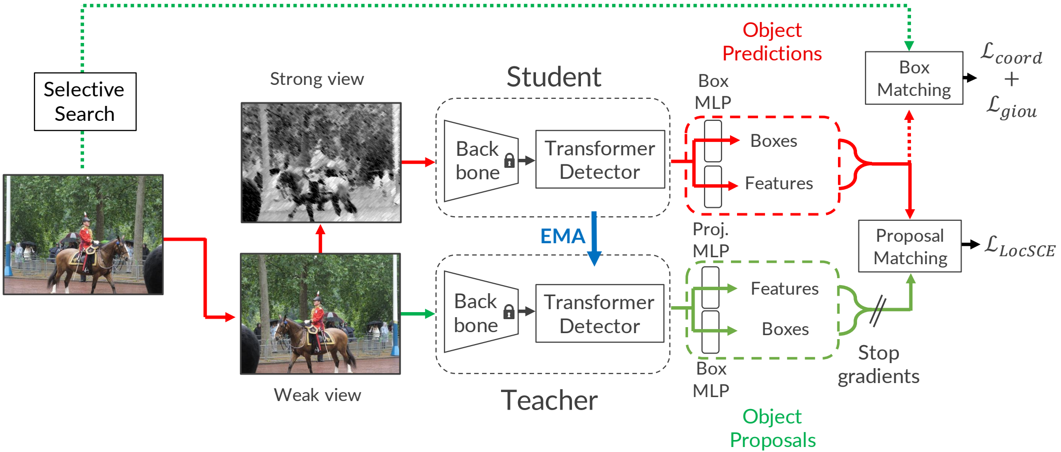
Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation



- Student predictions must match boxes *randomly* selected from **Selective Search**[19] outputs.

[19] Jasper RR Uijlings et al. “Selective search for object recognition”. In: *IJCV*. 2013.

Proposal Selection Contrast (ProSeCo)



- Full pretraining procedure with both contrastive and localization learning.

Experimental Results

Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation

Pretraining on ImageNet, finetuning on Mini-COCO

Pretraining	Arch.	Mini-COCO		
		1% (1.2k)	5% (5.9k)	10% (11.8k)
Supervised	Trans.	13.0	23.6	28.6
SwAV[20]	Trans.	13.3	24.5	29.5
SCRL[21]	Trans.	16.4	26.2	30.6
DETReg[22]	Trans.	15.9	26.1	30.9
Supervised	Conv.	–	19.4	24.7
SoCo*[23]	Conv.	–	26.8	31.1
ProSeCo (Ours)	Trans.	**18.0**	**28.8**	**32.8**

[20] Mathilde Caron et al. "Unsupervised learning of visual features by contrasting cluster assignments". In: *NeurIPS*. 2020.

[21] Byungseok Roh et al. "Spatially consistent representation learning". In: *CVPR*. 2021.

[22] Amir Bar et al. "Detreg: Unsupervised pretraining with region priors for object detection". In: *CVPR*. 2022.

[23] Fangyun Wei et al. "Aligning pretraining for detection via object-level contrastive learning". In: *NeurIPS*. 2021.

Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation

Finetuning on other datasets

Pretraining	FSOD-test	FSOD-train	PASCAL VOC	Mini-VOC	
	100% (11k)	100% (42k)	100% (16k)	5% (0.8k)	10% (1.6k)
Supervised	39.3	42.6	59.5	33.9	40.8
DETReg[24]	43.2	43.3	63.5	43.1	48.2
ProSeCo (Ours)	**46.6**	**47.2**	**65.1**	**46.1**	**51.3**

✓ Improvements of about **2 points over SOTA** on all datasets considered.

[24] Amir Bar et al. "Detreg: Unsupervised pretraining with region priors for object detection". In: *CVPR*. 2022.

Take Home Message

Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation

We propose ProSeCo, a Proposal-Contrastive Pretraining strategy for Object Detection with Transformers.

- ✓ Leverage high number of Object Proposals for **Proposal-Contrastive Learning**.
- ✓ Our **ProSeCo improves performance** when training with limited labeled data.
- ✓ **Consistency** with object-level features is important for Object Detection.
- ✓ **Location information** helps for Proposal-Contrastive learning.

Outline

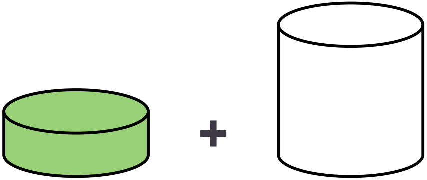
1. Introduction
2. Improving Few-Shot Classification with Meta-Learning through Multi-Task Learning
3. Improving Few-Annotation Learning for Object Detection
   - Background in Object Detection
   - Contrib 2: Unsupervised Pretraining for Object Detection with Fewer Annotation[25]
   - Contrib 3: Few Annotation Learning for Semi-Supervised Object Detection[26]
4. Conclusion and Broader Impacts

[25] Quentin Bouniot, Romaric Audigier, et al. "Proposal-Contrastive Pretraining for Object Detection from Fewer Data". In: *ICLR*. 2023.

[26] Quentin Bouniot, Angélique Loesch, et al. "Towards Few-Annotation Learning for Object Detection: Are Transformer-Based Models More Efficient?" In: *WACV*. 2023.

Setting considered

Contrib 3: Few Annotation Learning for Semi-Supervised Object Detection



Semi-Supervised
Learning

Few-Annotation Learning Setting

Contrib 3: Few Annotation Learning for Semi-Supervised Object Detection

How do object detectors handle label scarcity ?

	Few-Shot Learning	Few-Annotation Learning
	1% (1180) labeled images	1% (1180) labeled images + 100% (118000) unlabeled images
	Fully Supervised	Semi-supervised (UBT)
FRCNN + FPN	9.05%	20.75%
Def. DETR	12.96%	Diverge

- Performance on COCO with 1% labeled training data.
- Unbiased Teacher (UBT)[27] with Def. DETR **does not converge.**

[27] Yen-Cheng Liu et al. "Unbiased Teacher for Semi-Supervised Object Detection". In: *ICLR*. 2021.

Few-Annotation Learning Setting

Contrib 3: Few Annotation Learning for Semi-Supervised Object Detection

How do object detectors handle label scarcity ?

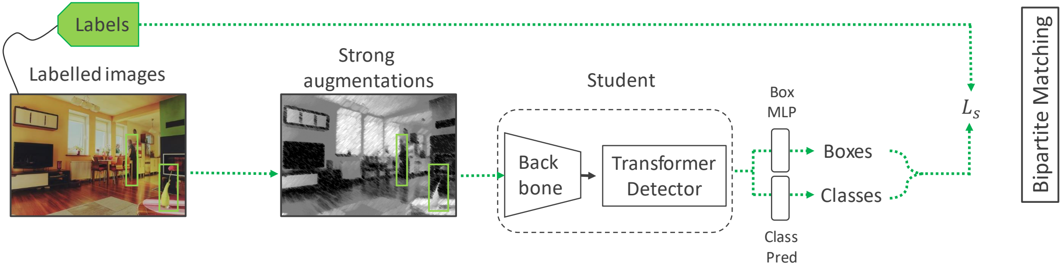
	Few-Shot Learning 1% (1180) labeled images Fully Supervised	Few-Annotation Learning 1% (1180) labeled images + 100% (118000) unlabeled images Semi-supervised (UBT)
FRCNN + FPN	9.05%	20.75%
Def. DETR	12.96%	Diverge

Why ?

- Performance on COCO with 1% labeled training data.
- Unbiased Teacher (UBT)[28] with Def. DETR **does not converge.**

[28] Yen-Cheng Liu et al. "Unbiased Teacher for Semi-Supervised Object Detection". In: *ICLR*. 2021.

Supervised branch

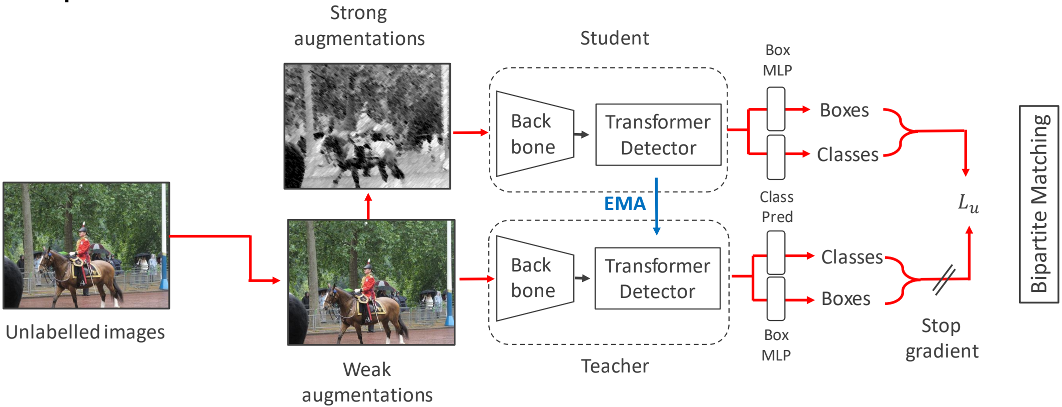


- ▶ Supervised training of the student model with supervised Hungarian algorithm.

Momentum-Teaching DETR

Contrib 3: Few Annotation Learning for Semi-Supervised Object Detection

Unsupervised branch

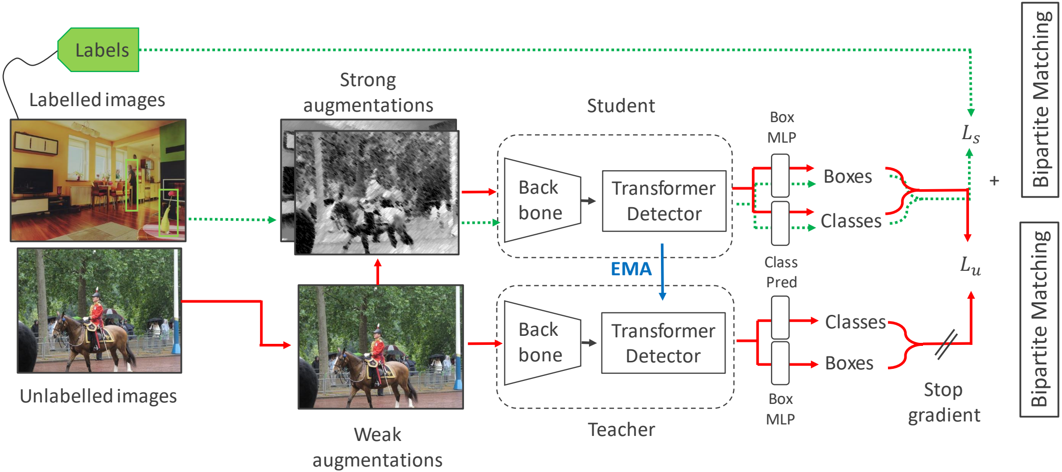


- Teacher model provides **pseudo-label** for Student model.
- Difference with ProSeCo: Reusing **class** information + **supervised** information.

Momentum-Teaching DETR

Contrib 3: Few Annotation Learning for Semi-Supervised Object Detection

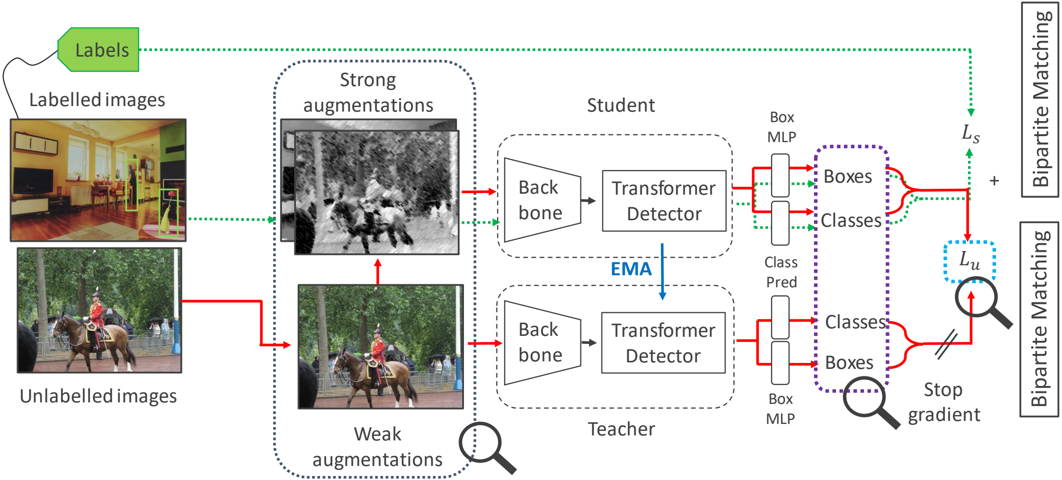
Semi-supervised Learning with Momentum-Teaching DETR (MT-DETR)



Momentum-Teaching DETR

Contrib 3: Few Annotation Learning for Semi-Supervised Object Detection

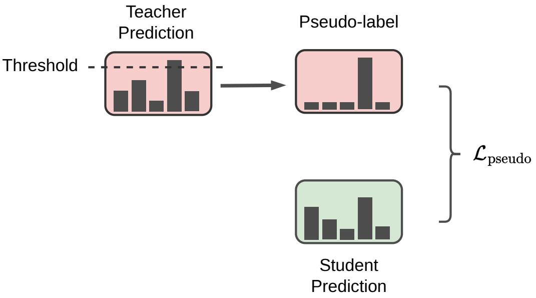
Semi-supervised Learning with Momentum-Teaching DETR (MT-DETR)



Hard vs Soft Pseudo-labeling

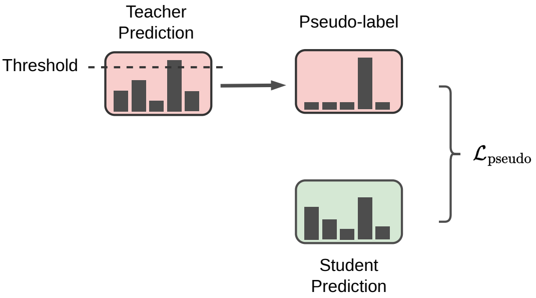
Contrib 3: Few Annotation Learning for Semi-Supervised Object Detection

Hard Pseudo-labeling



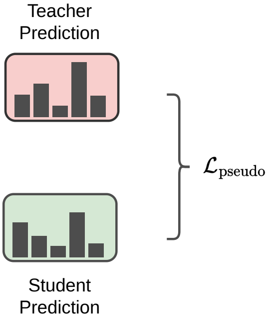
- ✕ Encourage **high confidence** predictions
- ✕ Focus on **prevailing** class
- ✕ **Additional hyperparameter** with the threshold

Hard Pseudo-labeling



- × Encourage **high confidence** predictions
- × Focus on **prevailing** class
- × **Additional hyperparameter** with the threshold

Soft Pseudo-labeling



- ✓ **Preserves relations** between classes
- ✓ More **diversity** in prevailing class

Performance Comparison with State of the Art

Contrib 3: Few Annotation Learning for Semi-Supervised Object Detection

Method	Arch.	FAL-COCO			
		0.5% (590)	1% (1180)	5% (5900)	10% (11800)
FRCNN + FPN	Conv.	6.83 ± 0.15	9.05 ± 0.16	18.47 ± 0.22	23.86 ± 0.81
STAC[29]	Conv.	9.78 ± 0.53	13.97 ± 0.35	24.38 ± 0.12	28.64 ± 0.21
Instant-Teaching[30]	Conv.	–	18.05 ± 0.15	26.75 ± 0.05	30.40 ± 0.05
Humble Teacher[31]	Conv.	–	16.96 ± 0.38	27.70 ± 0.15	31.61 ± 0.28
Unbiased Teacher[32]	Conv.	16.94 ± 0.23	20.75 ± 0.12	28.27 ± 0.11	31.50 ± 0.10
Soft Teacher[33]	Conv.	–	20.46 ± 0.39	30.74 ± 0.08	34.04 ± 0.14
Def. DETR	Trans.	8.95 ± 0.51	12.96 ± 0.08	23.59 ± 0.21	28.55 ± 0.08
MT-DETR *(Ours)*	Trans.	**17.84** ± 0.54	**22.03** ± 0.17	**31.00** ± 0.11	**34.52** ± 0.07

[29] Kihyuk Sohn et al. “A simple semi-supervised learning framework for object detection”. In: *arXiv*. 2020.
[30] Qiang Zhou et al. “Instant-teaching: An end-to-end semi-supervised object detection framework”. In: *CVPR*. 2021.
[31] Yihe Tang et al. “Humble teachers teach better students for semi-supervised object detection”. In: *CVPR*. 2021.
[32] Yen-Cheng Liu et al. “Unbiased Teacher for Semi-Supervised Object Detection”. In: *ICLR*. 2021.
[33] Mengde Xu et al. “End-to-end semi-supervised object detection with soft teacher”. In: *ICCV*. 2021.

Performance Comparison with State of the Art

Contrib 3: Few Annotation Learning for Semi-Supervised Object Detection

Method	Arch.	FAL-VOC 07-12		
		5% (250)	10% (500)	100% (5000)
FRCNN + FPN	Conv.	18.47 ± 0.39	25.23 ± 0.22	42.13
STAC[34]	Conv.	–	–	44.64
Instant-Teaching[35]	Conv.	–	–	50.00
Humble Teacher[36]	Conv.	–	–	53.04
Unbiased Teacher[37]	Conv.	35.98 ± 0.71	40.34 ± 0.95	54.61
Def. DETR	Trans.	22.87 ± 0.38	29.03 ± 0.46	51.34
MT-DETR *(Ours)*	Trans.	$\mathbf{36.95} \pm 0.53$	$\mathbf{43.15} \pm 1.10$	**56.2**

- ✓ We achieve the **best performance** on all settings
- ✓ More **significant gap** when labeled data is scarce
- ✓ Ablation study to find the **best combination** of training hyperparameters.

[34] Kihyuk Sohn et al. “A simple semi-supervised learning framework for object detection”. In: *arXiv*. 2020.

[35] Qiang Zhou et al. “Instant-teaching: An end-to-end semi-supervised object detection framework”. In: *CVPR*. 2021.

[36] Yihe Tang et al. “Humble teachers teach better students for semi-supervised object detection”. In: *CVPR*. 2021.

[37] Yen-Cheng Liu et al. “Unbiased Teacher for Semi-Supervised Object Detection”. In: *ICLR*. 2021.

Take Home Message

Contrib 3: Few Annotation Learning for Semi-Supervised Object Detection

Leverage few annotated data and unlabeled data for strong object detectors.

- Experiments with transformer-based detector with **scarce labeled data**
  - ✓ **Better** than convolutional detector when labels are **limited**
  - × **Do not work** with previous semi-supervised methods
- Our proposed **MT-DETR**:
  - ✓ **MT-DETR** is a semi-supervised approach for Transformer-based detectors
  - ✓ **Outperforms state-of-the-art** semi-supervised object detectors in **few-annotation learning**

Outline

1. Introduction
2. Improving Few-Shot Classification with Meta-Learning through Multi-Task Learning
3. Improving Few-Annotation Learning for Object Detection
4. Conclusion and Broader Impacts

Summary of contributions

- Contribution 1: Improving Meta-Learning algorithms through **Multi-Task Representation Learning theory**.[38]
- Contribution 2: **ProSeCo**, a Proposal-Contrastive Pretraining strategy for Object Detection with Transformers.[39]
- Contribution 3: **MT-DETR**, first semi-supervised approach tailored for Transformer-based detectors.[40]

[38] Quentin Bouniot, Ievgen Redko, et al. "Improving Few-Shot Learning Through Multi-task Representation Learning Theory". In: *ECCV*. 2022.

[39] Quentin Bouniot, Romaric Audigier, et al. "Proposal-Contrastive Pretraining for Object Detection from Fewer Data". In: *ICLR*. 2023.

[40] Quentin Bouniot, Angélique Loesch, et al. "Towards Few-Annotation Learning for Object Detection: Are Transformer-Based Models More Efficient?" In: *WACV*. 2023.

Perspectives

Towards bridging the gap between MTR theory and Meta-learning in practice.

- Take into account similarity between source and test tasks for *cross-domain generalization*.

Perspectives

Towards bridging the gap between MTR theory and Meta-learning in practice.

- Take into account similarity between source and test tasks for *cross-domain generalization*.

Towards leveraging unlabeled data for Object Detection using Transformers.

- Update the backbone during pretraining to further improve consistency.
- Improvements from self- and semi-supervision are less significant than for convolutional methods. Consider *more suited* unsupervised tasks ?

Broader Impacts

Computational Costs

- *Few annotations does not imply few computations !*
- Meta-learning is computationally expensive because of episodic training and bilevel optimization.
- Learning with unlabeled data requires a large number of training iterations.

Broader Impacts

Computational Costs

- *Few annotations does not imply few computations !*
- Meta-learning is computationally expensive because of episodic training and bilevel optimization.
- Learning with unlabeled data requires a large number of training iterations.

Environmental Costs

- *A lot of computations implies a high carbon footprint !*
- But can reduce costly annotation phases for large-scale datasets: about *12 tCO2eq* for annotating COCO dataset !
- For comparison: *4.6 tCO2eq* for all experiments in this thesis (about 180 000 GPU hours), *4 tCO2eq* for a round trip to Hawaii.

Broader Impacts

Computational Costs

- *Few annotations does not imply few computations !*
- Meta-learning is computationally expensive because of episodic training and bilevel optimization.
- Learning with unlabeled data requires a large number of training iterations.

Environmental Costs

- *A lot of computations implies a high carbon footprint !*
- But can reduce costly annotation phases for large-scale datasets: about *12 tCO2eq* for annotating COCO dataset !
- For comparison: *4.6 tCO2eq* for all experiments in this thesis (about 180 000 GPU hours), *4 tCO2eq* for a round trip to Hawaii.

Accessibility

- *Reduces the need for labels !*
- Can be crucial for a lot of applications.

Thank you for listening !

International publications:

- **Quentin Bouniot**, Ievgen Redko, Romaric Audigier, Angélique Loesch, Amaury Habrard. ”Improving Few-Shot Learning through Multi-task Representation Learning Theory”. In *ECCV*, 2022.
- **Quentin Bouniot**, Angélique Loesch, Romaric Audigier, Amaury Habrard. ”Towards Few-Annotation Learning for Object Detection: Are Transformer-based Models More Efficient ?”. In *WACV*, 2023.
- **Quentin Bouniot**, Romaric Audigier, Angélique Loesch, Amaury Habrard. ”Proposal-Contrastive Pretraining for Object Detection from Fewer Data”. In *ICLR*, 2023.

Workshops and communications:

- **Quentin Bouniot**, Ievgen Redko, Romaric Audigier, Angélique Loesch, Amaury Habrard. ”Putting Theory to Work : From Learning Bounds to Meta-Learning Algorithms”. In *NeurIPS Workshop on Meta-Learning (MetaLearn)*, 2020.
- **Quentin Bouniot**, Ievgen Redko, Romaric Audigier, Angélique Loesch, Amaury Habrard. ”Vers une meilleure compréhension des méthodes de méta-apprentissage à travers la théorie de l’apprentissage de représentations multi-tâches”. In *CAp*, 2021.
- **Quentin Bouniot**, Ievgen Redko, Romaric Audigier, Angélique Loesch, Amaury Habrard. ”Improving Few-Shot Learning through Multi-task Representation Learning Theory”. In *GdR ISIS*, 2021.
- **Quentin Bouniot** & Ievgen Redko, ”Understanding Few-Shot Multi-Task Representation Learning Theory”. In *ICLR Blog Track*, 2022.

Contributions

- Quentin Bouniot, Ievgen Redko, et al. "Improving Few-Shot Learning Through Multi-task Representation Learning Theory". In: *ECCV*. 2022.
- Quentin Bouniot, Romaric Audigier, et al. "Proposal-Contrastive Pretraining for Object Detection from Fewer Data". In: *ICLR*. 2023.
- Quentin Bouniot, Angélique Loesch, et al. "Towards Few-Annotation Learning for Object Detection: Are Transformer-Based Models More Efficient?" In: *WACV*. 2023.

References I

- Jake Snell, Kevin Swersky, and Richard S. Zemel. "Prototypical Networks for Few-shot Learning". In: *NeurIPS*. 2017.
- Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks". In: *ICML*. 2017.
- Simon S. Du et al. "Few-Shot Learning via Learning the Representation, Provably". In: *ICLR*. 2021.
- Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. "Provable Meta-Learning of Linear Representations". In: *arXiv*. 2020.
- Nicolas Carion et al. "End-to-end object detection with transformers". In: *ECCV*. 2020.
- Zhi Tian et al. "Fcos: Fully convolutional one-stage object detection". In: *ICCV*. 2019.
- Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *NeurIPS*. 2015.
- Tsung-Yi Lin et al. "Feature pyramid networks for object detection". In: *CVPR*. 2017.

References II

- Xizhou Zhu et al. "Deformable DETR: Deformable Transformers for End-to-End Object Detection". In: *ICLR*. 2021.
- Julien Denize et al. "Similarity contrastive estimation for self-supervised soft contrastive learning". In: *WACV*. 2023.
- Jasper RR Uijlings et al. "Selective search for object recognition". In: *IJCV*. 2013.
- Mathilde Caron et al. "Unsupervised learning of visual features by contrasting cluster assignments". In: *NeurIPS*. 2020.
- Byungseok Roh et al. "Spatially consistent representation learning". In: *CVPR*. 2021.
- Amir Bar et al. "Detreg: Unsupervised pretraining with region priors for object detection". In: *CVPR*. 2022.
- Fangyun Wei et al. "Aligning pretraining for detection via object-level contrastive learning". In: *NeurIPS*. 2021.

References III

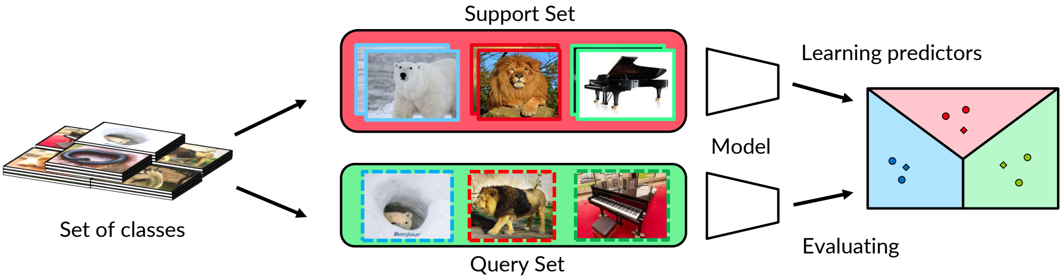
- Yen-Cheng Liu et al. “Unbiased Teacher for Semi-Supervised Object Detection”. In: *ICLR*. 2021.
- Kihyuk Sohn et al. “A simple semi-supervised learning framework for object detection”. In: *arXiv*. 2020.
- Qiang Zhou et al. “Instant-teaching: An end-to-end semi-supervised object detection framework”. In: *CVPR*. 2021.
- Yihe Tang et al. “Humble teachers teach better students for semi-supervised object detection”. In: *CVPR*. 2021.
- Mengde Xu et al. “End-to-end semi-supervised object detection with soft teacher”. In: *ICCV*. 2021.
- Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. “The Benefit of Multitask Representation Learning”. In: *JMLR*. 2016.
- Haoxiang Wang, Han Zhao, and Bo Li. “Bridging Multi-Task Learning and Meta-Learning: Towards Efficient Training and Effective Adaptation”. In: *ICML*. 2021.

References IV

- Yunhui Guo et al. "A Broader Study of Cross-Domain Few-Shot Learning". In: *ECCV*. 2020.
- Joseph Redmon et al. "You only look once: Unified, real-time object detection". In: *CVPR*. 2016.

Appendix

Episodic Training



- **Disjoint** sets of classes between meta-training and meta-testing classes.
- Construction of ***episodes*** from dataset.

Multi-Task Representation Learning Theory

Traditional PAC-bounds[41]

$$\boxed{\mathrm{ER}(\phi, \mathbf{w}_{T+1}) \leq O\left(\frac{1}{n_1} + \frac{1}{T}\right)}$$

- × Requires n_1 *and* T to tend to infinity.
- × Doesn't explain the success in *few data regime*.

[41] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. "The Benefit of Multitask Representation Learning". In: *JMLR*. 2016.

Mismatch in settings

Multi-task training $\neq$ Episodic training

- Mismatch in problem formulation and objectives

But shared optimization formulation, with some simplification

- The differences are empirically negligible[42]

[42] Haoxiang Wang, Han Zhao, and Bo Li. "Bridging Multi-Task Learning and Meta-Learning: Towards Efficient Training and Effective Adaptation". In: *ICML*. 2021.

Mismatch in settings

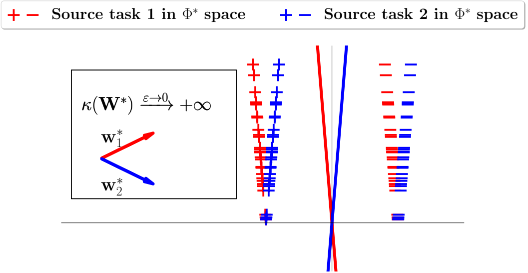
Inner Learning Rate λ
x
Adaptation Steps τ

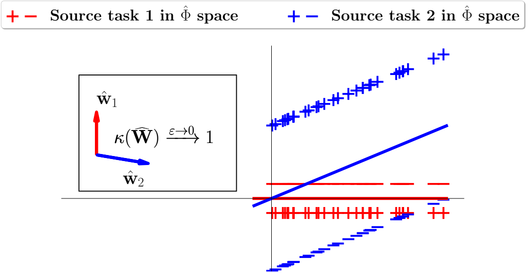
$$||\text{ANIL predictions} - \text{MTR predictions}||_2 \leq O\left(\boxed{\lambda\tau} + \frac{1}{\boxed{L}}\right)$$

Network
Depth L

Can we force the assumptions ?

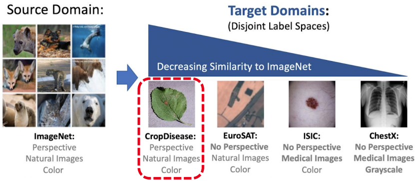
Given $\mathbf{W}^*$ such that $\kappa(\mathbf{W}^*) \gg 1$, can we learn $\hat{\mathbf{W}}$ with $\kappa(\hat{\mathbf{W}}) \approx 1$ while solving the underlying classification problems equally well ?





✓ Even when $\mathbf{W}^*$ does not satisfy the assumptions, it is possible to learn $\hat{\phi}$ to respect them.

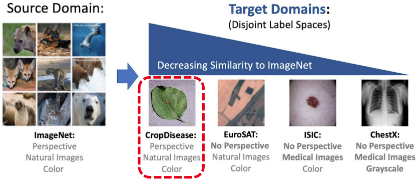
Experimental Results



5-way
1-shot

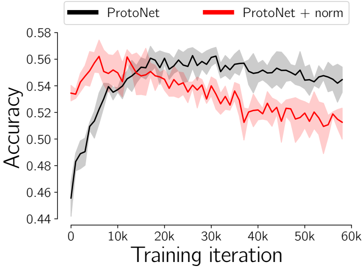
Guo et al., “A Broader Study of Cross-Domain Few-Shot Learning”

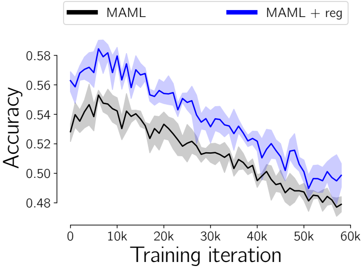
Experimental Results



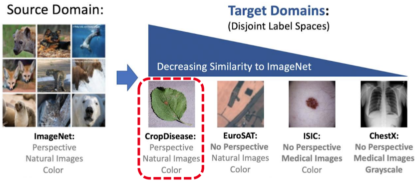
5-way
1-shot

Guo et al., "A Broader Study of Cross-Domain Few-Shot Learning"



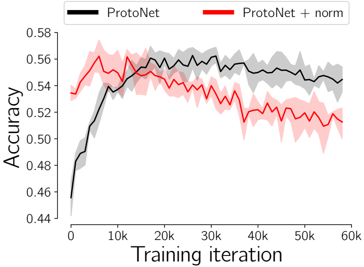


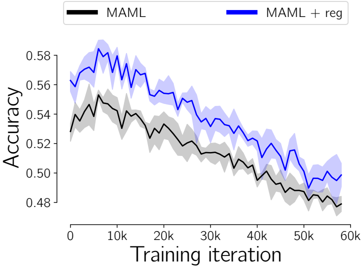
Experimental Results



5-way
1-shot

Guo et al., "A Broader Study of Cross-Domain Few-Shot Learning"

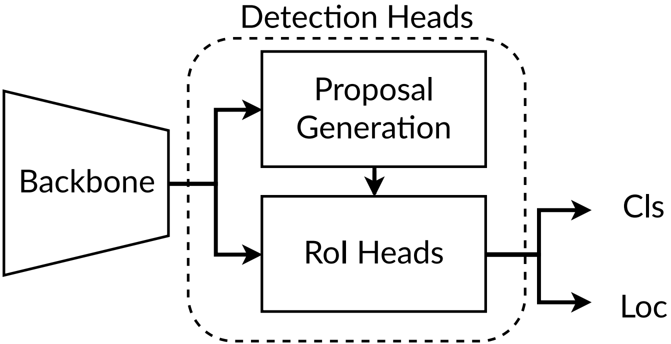




- ✗ Improvement does not translate to cross-domain for *metric-based methods*.
- ✓ *Gradient-based methods* keep their accuracy gains.

Object Detection 101

Two-stage methods (e.g., Faster-RCNN[43])

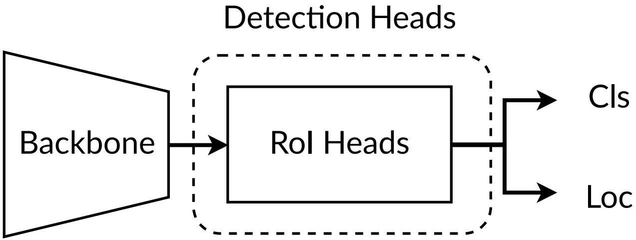


- First stage proposes candidate object bounding boxes (proposals).
- Second stage refines each proposal.

[43] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *NeurIPS*. 2015.

Object Detection 101

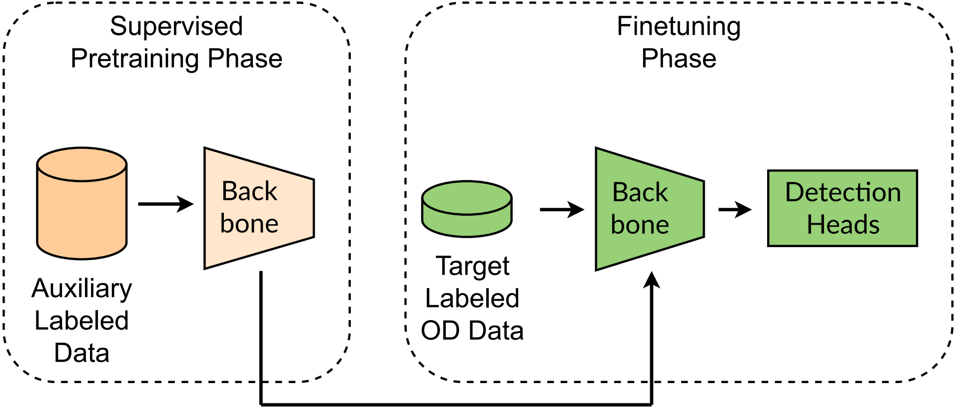
One-stage methods (e.g., YOLO[44])



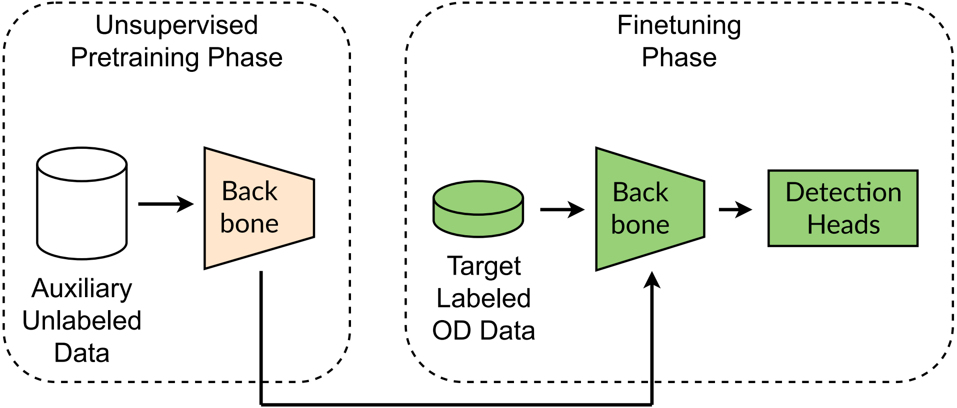
- Classification and localization in a single shot using a dense sampling.
- Predefined anchors or reference points are refined for localization.
- Simpler design, real-time inference speed but lower performance.

44 Joseph Redmon et al. "You only look once: Unified, real-time object detection". In: *CVPR*. 2016.

Pretraining in Object Detection

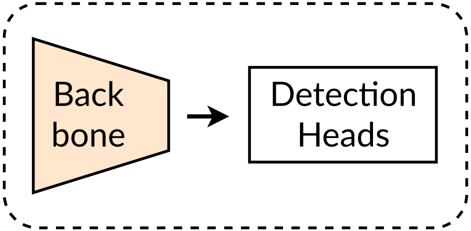


Pretraining in Object Detection



Pretraining in Object Detection

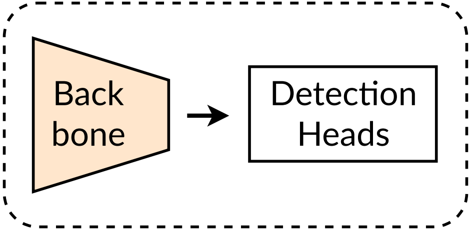
Backbone Pretraining



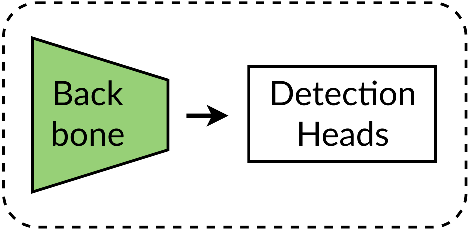
× Image-level pretraining task

Pretraining in Object Detection

Backbone Pretraining



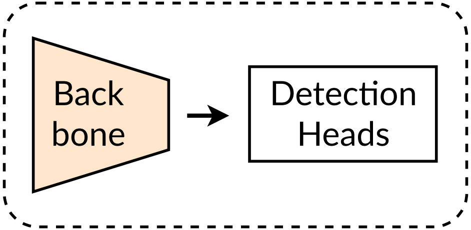
× Image-level pretraining task



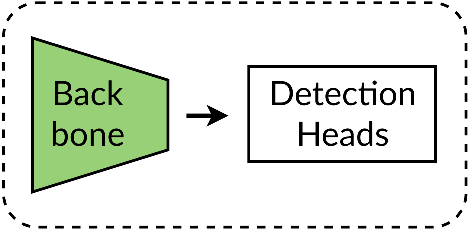
✓ Object-level pretraining task

Pretraining in Object Detection

Backbone Pretraining



× Image-level pretraining task



✓ Object-level pretraining task

× Pretraining limited to the **backbone**

Experimental Results

Ablation Studies

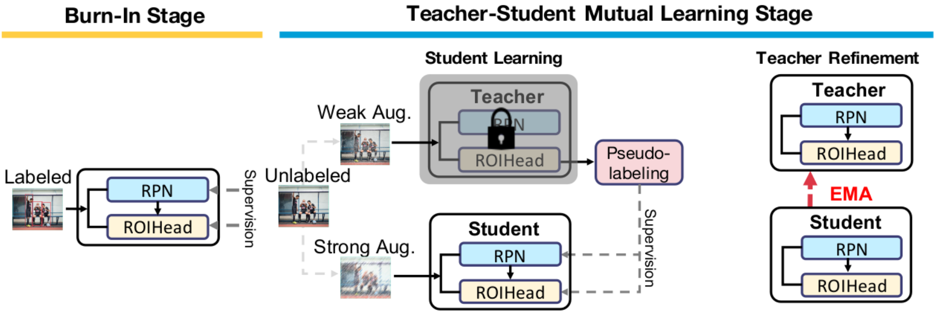
Pretraining	Dataset	mAP
ProSeCo w/ SwAV	COCO	27.4
ProSeCo w/ SwAV	IN	27.8
DETReg w/ SCRL	IN	28.0
ProSeCo w/ SCRL	IN	**28.8**

Loss	δ	mAP
SCE	1.0	26.1
LocSCE (Ours)	0.2	27.0
LocSCE (Ours)	0.7	27.1
LocSCE (Ours)	0.5	**27.8**

- ▶ Comparisons on Mini-COCO 5%
- ▶ Dataset diversity more important than closeness to downstream task
- ✓ Consistency in the features improves performance
- ✓ Location of proposals helps for introducing easy positives for contrastive learning

Semi-Supervised Object Detection

Unbiased Teacher (UBT)[45]



- Burn-in stage: Teacher model trained on labeled data.
- Weak and strong augmentations for unlabeled data.
- Teacher provides pseudo-labels for student model.
- Teacher updated with Exponential Moving Average (EMA).

[45] Yen-Cheng Liu et al. "Unbiased Teacher for Semi-Supervised Object Detection". In: *ICLR*. 2021.

Ablation Studies

Name	Augmentations
Basic	Horizontal Flip Resize
Photo.	Color Jitter Grayscale Gaussian Blur
CutOut	CutOut
Geom.	Rotate Shear Rescale + Pad

Augmentations used	mAP (in %)
Basic + Photo.	17.8
Basic + Photo. + CutOut	
| w/ NMS + Hard PL [46]	Div.
| w/o NMS + Soft PL *(Ours)*	21.1
Basic + Photo. + CutOut + Geom.	**21.6**
Basic + Photo. + CutOut + Geom. + Augmentations in Supervised branch	**22.3**

- ✓ Adding more augmentations leads to the best results
- ✓ Removing post-processing of proposals solves the diverging issue

[46] Liu et al., "Unbiased Teacher for Semi-Supervised Object Detection".

Ablation Studies

Contrib 3: Few Annotation Learning for Semi-Supervised Object Detection

Augmentations used	mAP (in %)
Basic + Photo.	17.8
Basic + Photo. + CutOut	
| w/ NMS + Hard PL[47]	Div.
| w/o NMS + Soft PL *(Ours)*	21.1
Basic + Photo. + CutOut + Geom.	**21.6**
Basic + Photo. + CutOut + Geom. + Augmentations in Supervised branch	**22.3**

- ✓ Adding more augmentations leads to the best results
- ✓ Removing post-processing of proposals solves the diverging issue

[47] Yen-Cheng Liu et al. "Unbiased Teacher for Semi-Supervised Object Detection". In: *ICLR*. 2021.

Ablation Studies

Contrib 3: Few Annotation Learning for Semi-Supervised Object Detection

Ablative Variant	EMA Scheduling		Initialization		NMS	Confidence Thresholding				mAP (in %)
	Cosine	Constant	After FT	From scratch		ø	0.5	0.7	0.9	
Best	✓		✓			✓				**22.25**
Abl. Sched.		✓	✓			✓				21.48
Abl. Init.	✓			✓		✓				16.51
Abl. NMS	✓		✓		✓	✓				19.85
Abl. Thresh.	✓		✓				✓			10.26
	✓		✓					✓		17.34
	✓		✓						✓	12.37

Best combination found:

- ✓ Cosine scheduling
- ✓ Initialization after fine-tuning
- ✓ No post-processing of pseudo-labels

Comparing Pretraining Strategy

Method	Pretrain.	FAL-COCO			
		0.5% (590)	1% (1180)	5% (5900)	10% (11800)
Supervised	Sup.	8.95 ± 0.51	12.96 ± 0.08	23.59 ± 0.21	28.55 ± 0.08
Supervised	ProSeCo	11.37 ± 0.40	17.90 ± 0.08	28.33 ± 0.33	32.60 ± 0.28
MT-DETR *(Ours)*	Sup.	**17.84** ± 0.54	**22.03** ± 0.17	**31.00** ± 0.11	**34.52** ± 0.07
MT-DETR *(Ours)*	ProSeCo	14.33 ± 0.17	21.73 ± 0.12	**32.00** ± 0.16	**35.83** ± 0.17

- Our ProSeCo also improves performance with MT-DETR.
- However less effective with very few labels.

Environmental Footprint of this Thesis

Carbon emissions come from electricity required for running experiments.

- About *150 000 GPU hours (17 years)* on CEA HPC cluster.
- About *30 000 GPU hours (3.5 years)* on Jean-Zay HPC cluster.
- Assuming *400Wh* for CEA HPC cluster, *259Wh* [48] for Jean-Zay, with an emission of *68 gCO2eq/kWh*.
- Total of about *4.6 tons of CO2eq*.

And going to conferences:

- 1.1 (ECCV) + 3.9 (WACV) + 2.4 (ICLR incoming)
- Total of about *7.4 tons of CO2eq*
- But important to meet other researchers in the domain and better experience than virtual !

Overall of 12 tons of CO2eq, equivalent to the annotation of the whole COCO dataset !

[48] http://www.idris.fr/media/jean-zay/jean-zay-conso-heure-calcul.pdf