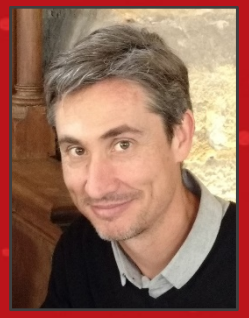





WACV WAIKOLOA HAWAII JAN 3-7 • 2023

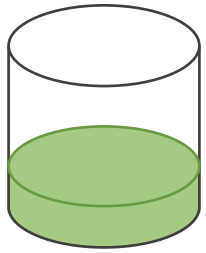
cea

UNIVERSITÉ JEAN MONNET SAINT-ÉTIENNE

Towards Few-Annotation Learning for Object Detection: Are Transformer-based Models More Efficient ?

Quentin Bouniot*, Angélique Loesch, Romaric Audigier, Amaury Habrard

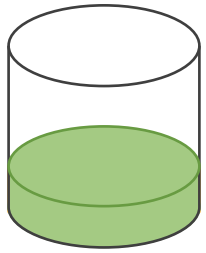
Few-Shot Learning



Few labeled images

Supervised Learning

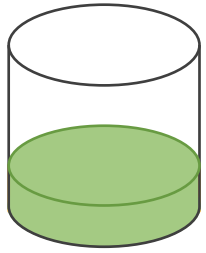
Few-Shot Learning



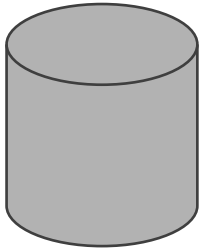
Few labeled images

Supervised Learning

Few-Annotation Learning



+

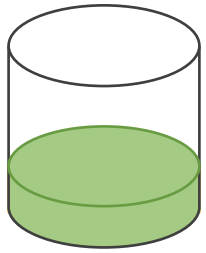


Few labeled images

A lot of unlabeled images

Semi-Supervised Learning

Few-Shot Learning

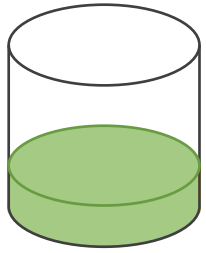


Few labeled images

Supervised Learning

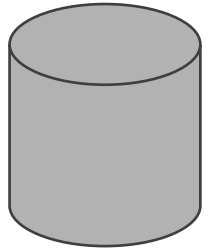
Stage 1
(Burn In)

Few-Annotation Learning



Few labeled images

+



A lot of unlabeled images

Semi-Supervised Learning

Stage 2

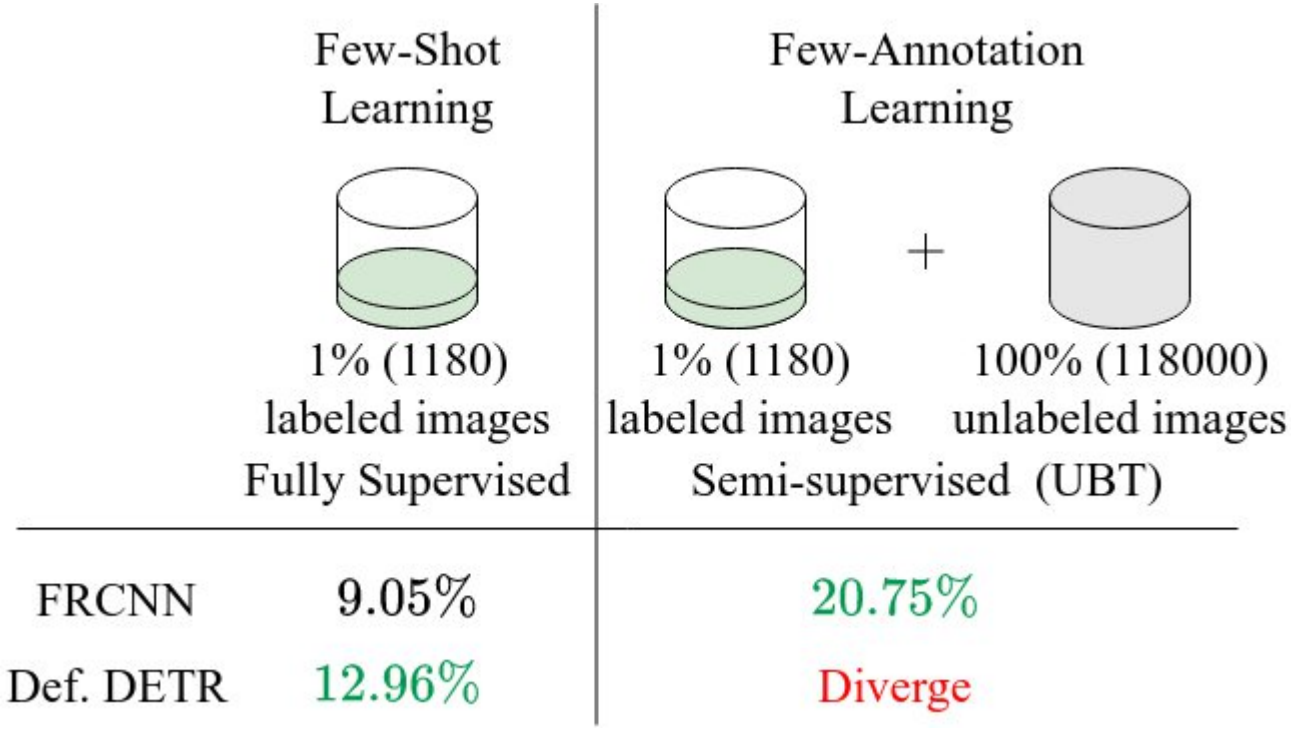
How do object detectors handle data scarcity ?

Method	Params.	COCO				VOC07	
		0.5% (590)	1% (1180)	5% (5900)	10% (11800)	5% (250)	10% (500)
FRCNN + FPN†	42M	6.83 ± 0.15	9.05 ± 0.16	18.47 ± 0.22	23.86 ± 0.81	18.47 ± 0.39	25.23 ± 0.22
Def. DETR	40M	$\mathbf{8.95 \pm 0.51}$	$\mathbf{12.96 \pm 0.08}$	$\mathbf{23.59 \pm 0.21}$	$\mathbf{28.55 \pm 0.08}$	$\mathbf{22.87 \pm 0.38}$	$\mathbf{29.03 \pm 0.46}$
Δ		+2.12	+3.91	+5.12	+4.69	+4.40	+3.80

- Performance on COCO and Pascal VOC 2007 datasets.
- Different percentages of labeled training data (with corresponding number of images).
- **Deformable DETR** performs better than Faster RCNN + FPN **with fewer labeled data**.

Ren et al., “Faster R-CNN: Towards real-time object detection with region proposal networks.” NeurIPS 2015
Lin et al., “Feature Pyramid Networks for Object Detection.” CVPR 2017.
Zhu et al., “Deformable DETR: Deformable transformers for end-to-end object detection.” ICLR 2021

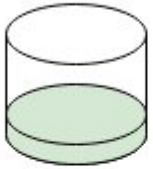
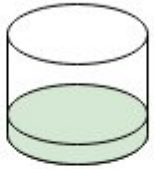
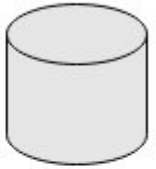
How do object detectors handle data scarcity ?



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

Liu et al. “Unbiased Teacher for Semi-Supervised Object Detection.” In ICLR 2021

How do object detectors handle data scarcity ?

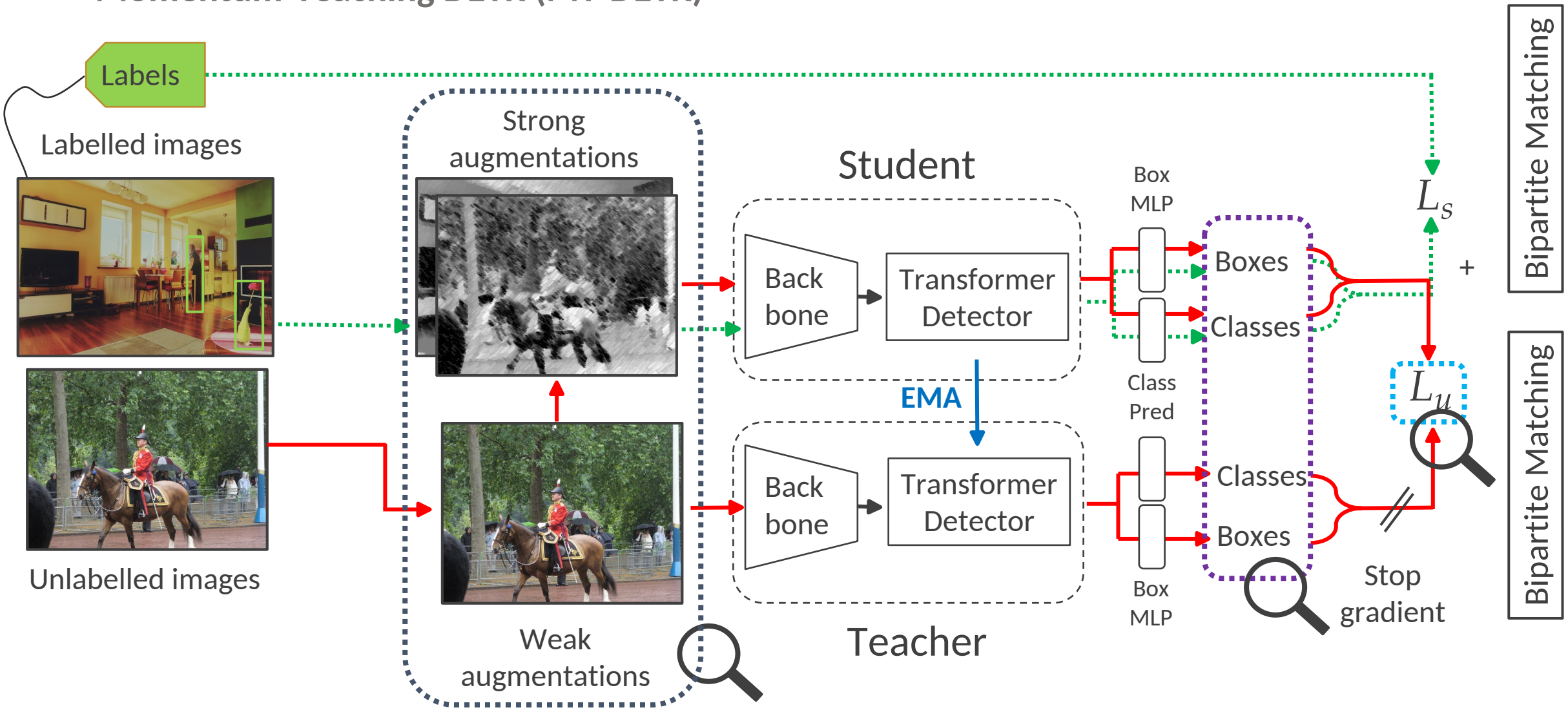
	Few-Shot Learning	Few-Annotation Learning	
		 + 	
	1% (1180) labeled images	1% (1180) labeled images	100% (118000) unlabeled images
	Fully Supervised	Semi-supervised (UBT)	
FRCNN	9.05%	20.75%	
Def. DETR	12.96%	Diverge	

Why ?

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

Liu et al. “Unbiased Teacher for Semi-Supervised Object Detection.” In ICLR 2021

Momentum-Teaching DETR (MT-DETR)



FAL-COCO

Method	OD Arch.	FAL-COCO			
		0.5% (590)	1% (1180)	5% (5900)	10% (11800)
STAC [127]	FRCNN + FPN	9.78 ± 0.53	13.97 ± 0.35	24.38 ± 0.12	28.64 ± 0.21
Instant-Teaching [159]	FRCNN + FPN	–	18.05 ± 0.15	26.75 ± 0.05	30.40 ± 0.05
Humble Teacher [129]	FRCNN + FPN	–	16.96 ± 0.38	27.70 ± 0.15	31.61 ± 0.28
Unbiased Teacher [91]	FRCNN + FPN	16.94 ± 0.23	20.75 ± 0.12	28.27 ± 0.11	31.50 ± 0.10
Soft Teacher [150]	FRCNN + FPN	–	20.46 ± 0.39	30.74 ± 0.08	34.04 ± 0.14
MT-DETR *(Ours)*	Def. DETR	**17.84** ± 0.54 (+8.89)	**22.03** ± 0.17 (+9.07)	**31.00** ± 0.11 (+7.41)	**34.52** ± 0.07 (+5.97)

- Evaluation on **different percentage of COCO labeled data** (with the corresponding number of images) and 100% of the dataset as unlabeled data.
- ✓ We achieve the **best performance** on all settings
- ✓ More **significant gap** when labeled data is scarce.

Take Home Message

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

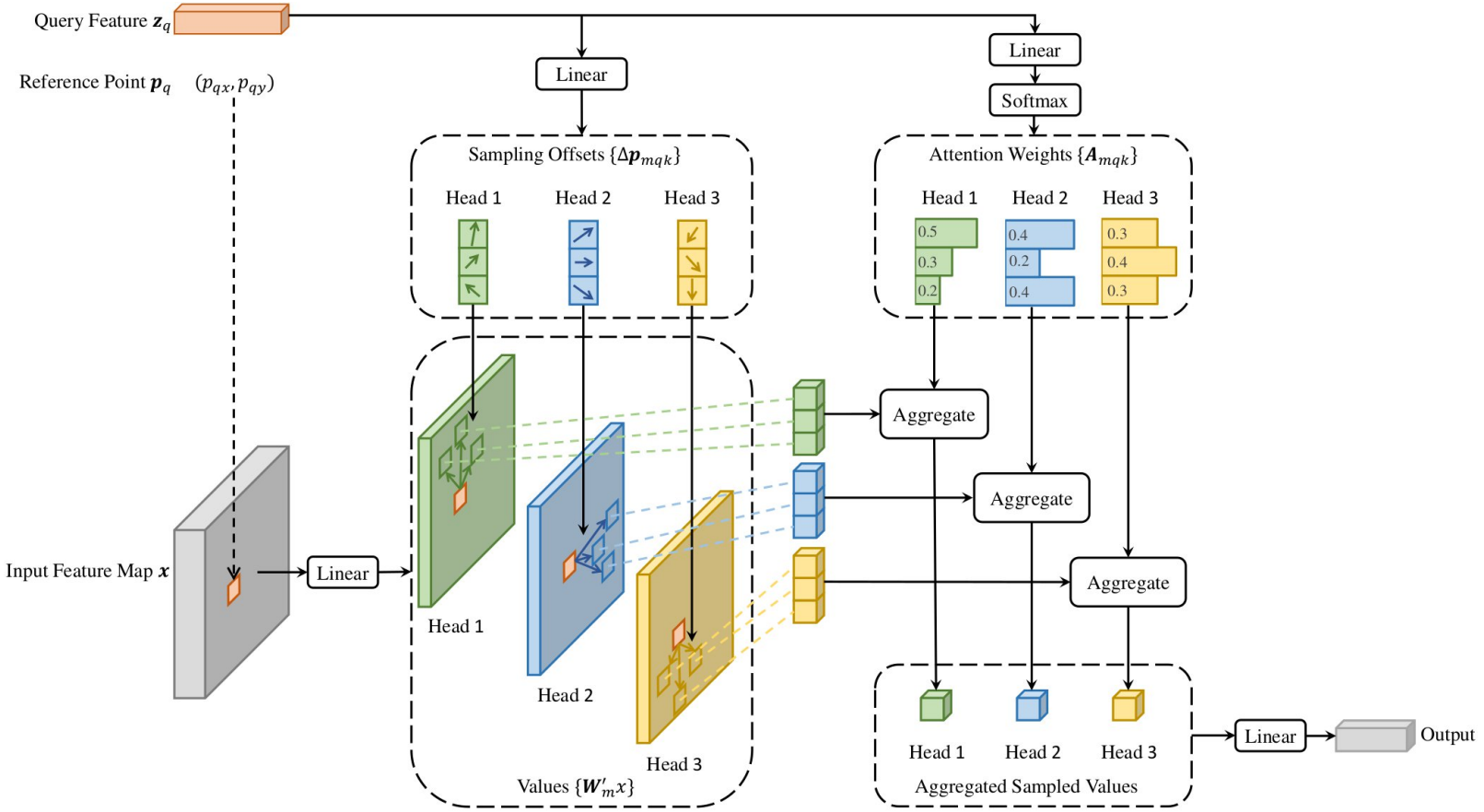
- Experiments with **transformer-based detector** with **scarce labeled data**
  - ✓ Better than convolutional detector when labels are limited
  - x Do not work with previous semi-supervised methods.

- **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**.

Thank you for your time !

Do not hesitate to contact us for questions !

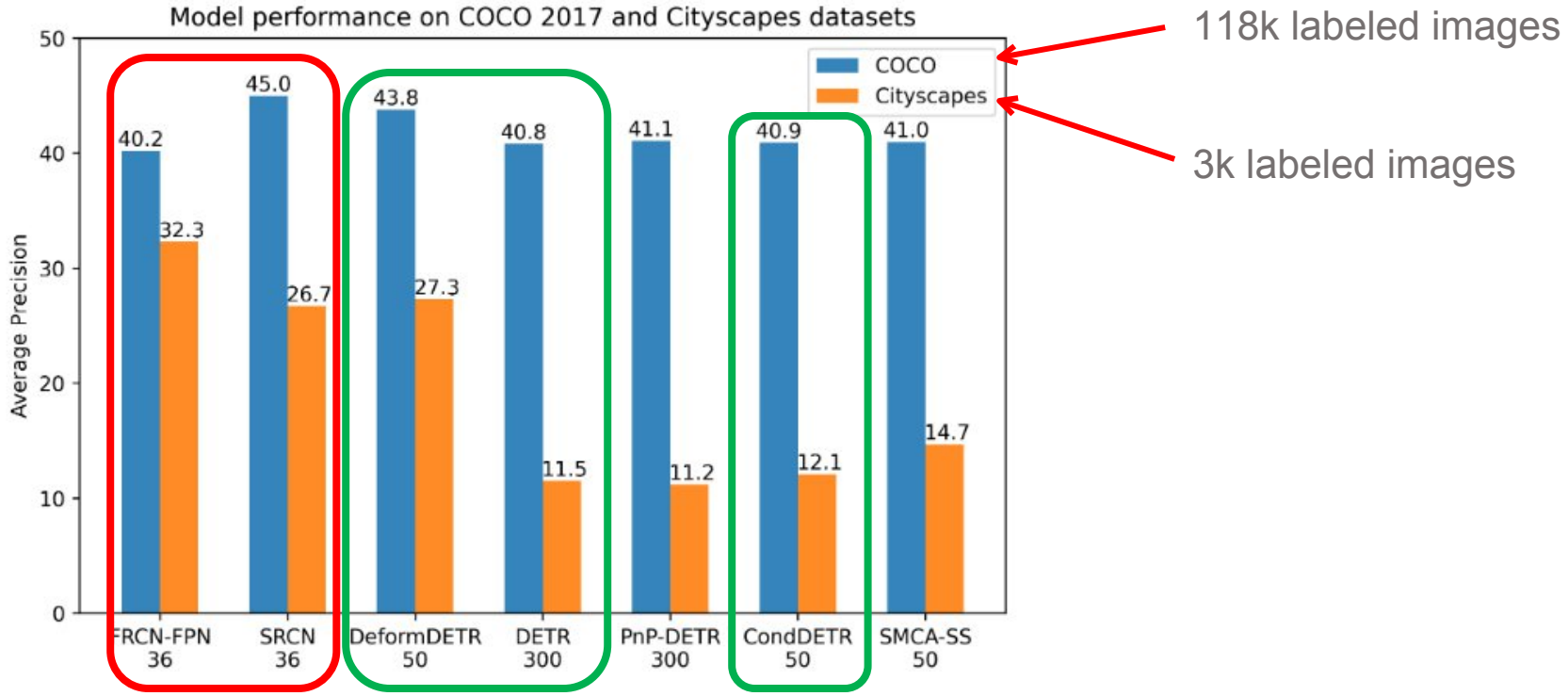
Transformer-based Detectors



- Deformable Attention operation tailored for dealing with input feature map.
- Learns **sampling offsets** for each **reference points** in the feature map.

Zhu et al., "Deformable DETR: Deformable transformers for end-to-end object detection." ICLR 2021

How do object detectors handle data scarcity ?



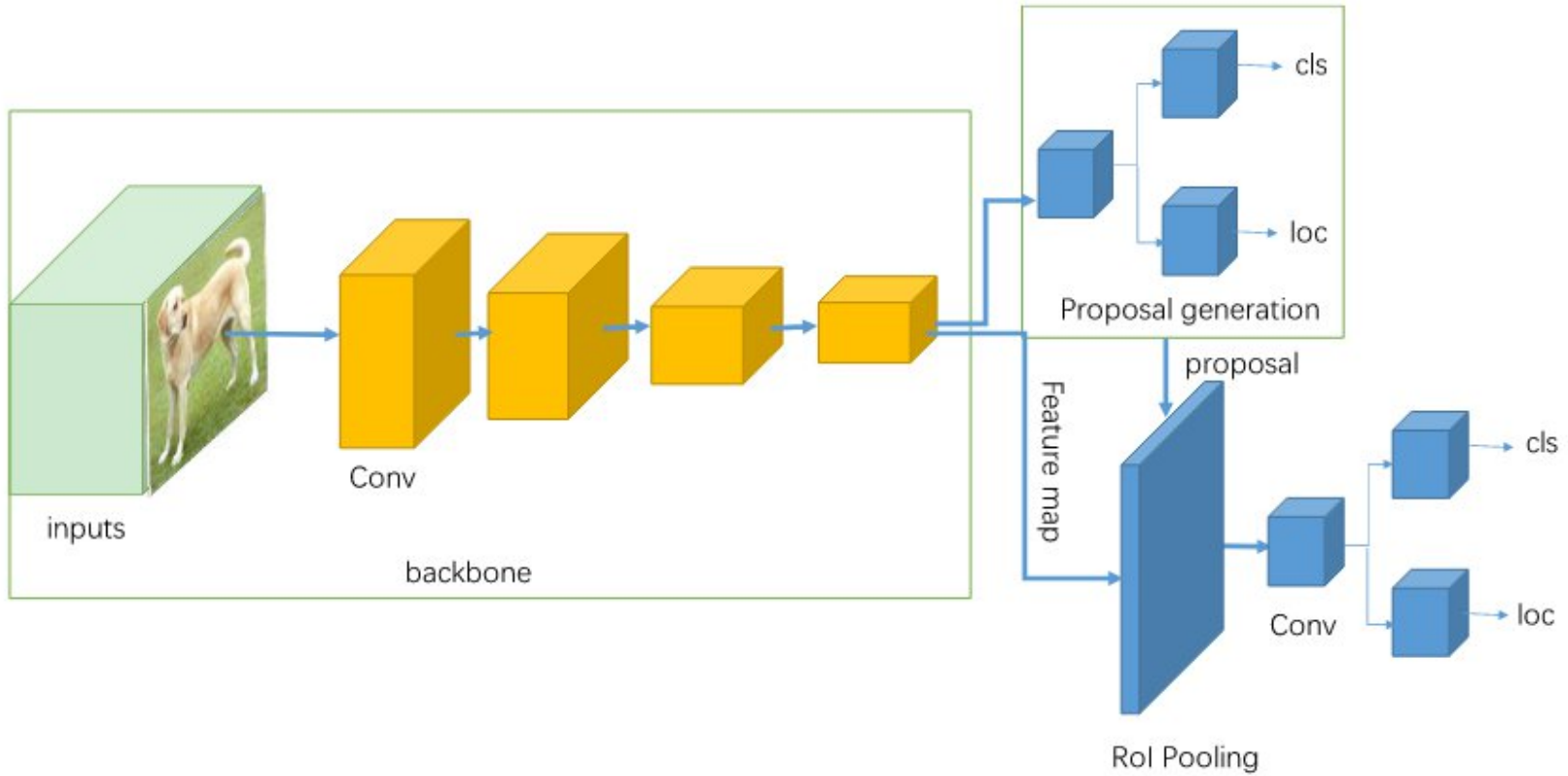
- Performance on **COCO** and **Cityscapes** datasets
- Comparison between **Convolutional-** and **Transformer-based** object detectors
- **Deformable DETR is strong with few labeled data**, but other transformer-based are **less data-efficient**.

Wang et al. "Towards Data-efficient Detection Transformers." In ECCV 2022
Meng et al. "Conditional DETR for fast training convergence." ICCV 2021

Augmentations used

Augmentations	Probability	Parameters	Supervised branch	Unsupervised branch	
				Weak	Strong
Horizontal Flip	0.5	–	✓	✓	✓
Resize	1.0	short edge $\in$ range(480,801,32)	✓	✓	✓
Color Jitter	0.8	(brightness, contrast, saturation, hue) $=$ (0.4, 0.4, 0.4, 0.1)	✓		✓
Grayscale	0.2	–	✓		✓
Gaussian Blur	0.5	$\sigma \in [0.1, 2.0]$	✓		✓
CutOut	0.7	scale $\in [0.05, 0.2]$, ratio $\in [0.3, 3.3]$	✓		✓
	0.5	scale $\in [0.02, 0.2]$, ratio $\in [0.1, 6]$	✓		✓
	0.3	scale $\in [0.02, 0.2]$, ratio $\in [0.05, 8]$	✓		✓
Rotate	0.3	degrees $\in [-30, 30]$			✓
Shear	0.3	shear$_x \in [-30, 30]$, shear$_y \in [-30, 30]$			✓
Rescale + Pad + Translation	0.5	translate$_x \in [0, 0.25]$, translate$_y \in [0, 0.25]$ scale$_x \in [0.25, 0.75]$, scale$_y \in [0.25, 0.75]$			✓

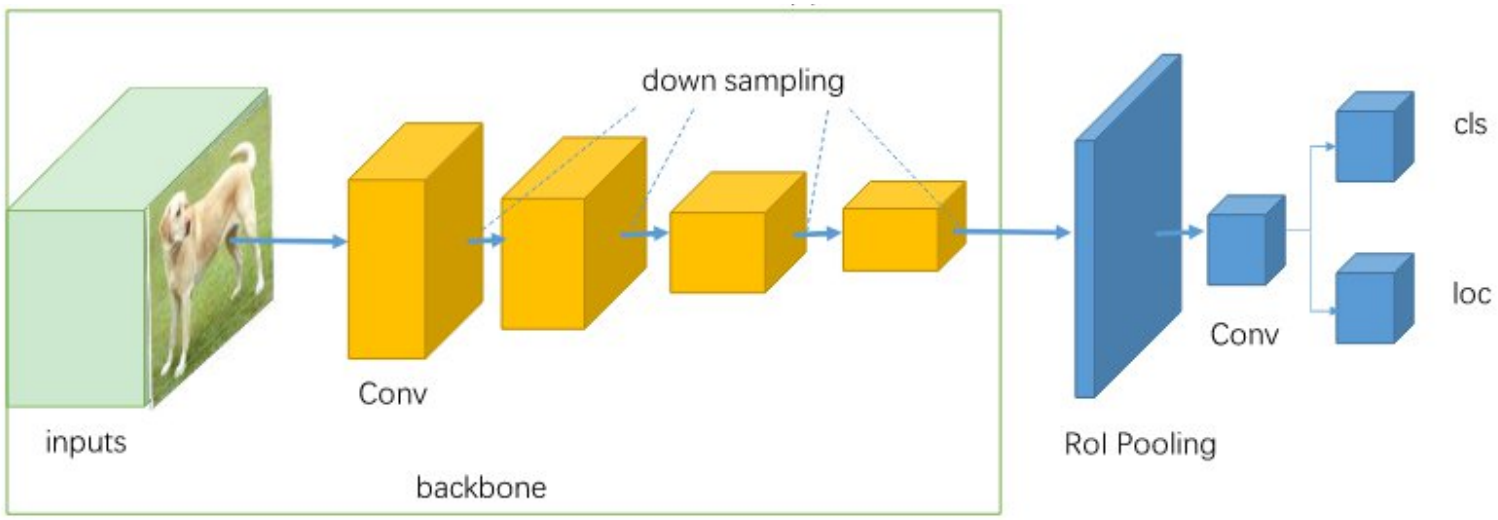
Two-stage Detectors



- First stage proposes **candidate object** bounding boxes
- Second stage extracts features from each candidate for **classification** and **regression** tasks.
- Most representative detector is **Faster-RCNN**

Jiao et al., "A survey of deep learning-based object detection." IEEE Access, 2019.
Ren et al., "Faster R-CNN: Towards real-time object detection with region proposal networks". NeurIPS 2015

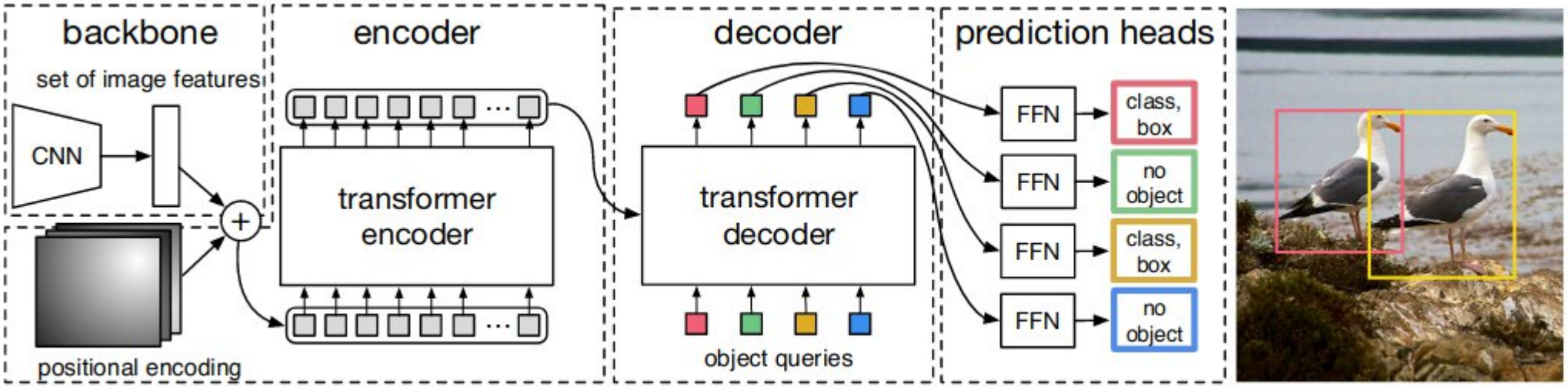
One-stage Detectors



- Classification and localization in a single shot using a **dense sampling**.
- Predefined **anchors** of various scales are refined for localization.
- Simpler design, real-time inference speed but lower performance.

Jiao et al., "A survey of deep learning-based object detection." IEEE Access, 2019.

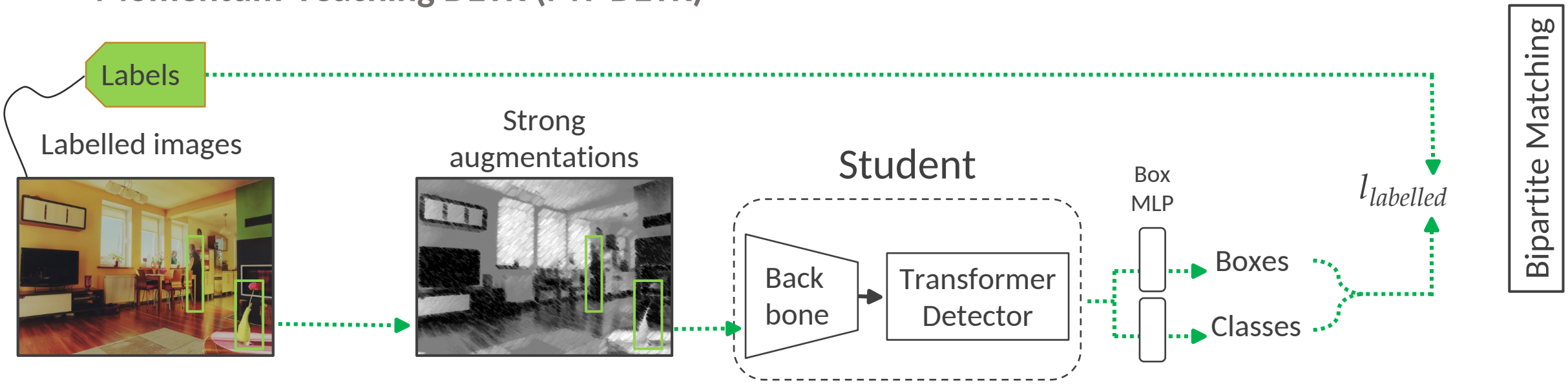
Transformer-based Detectors



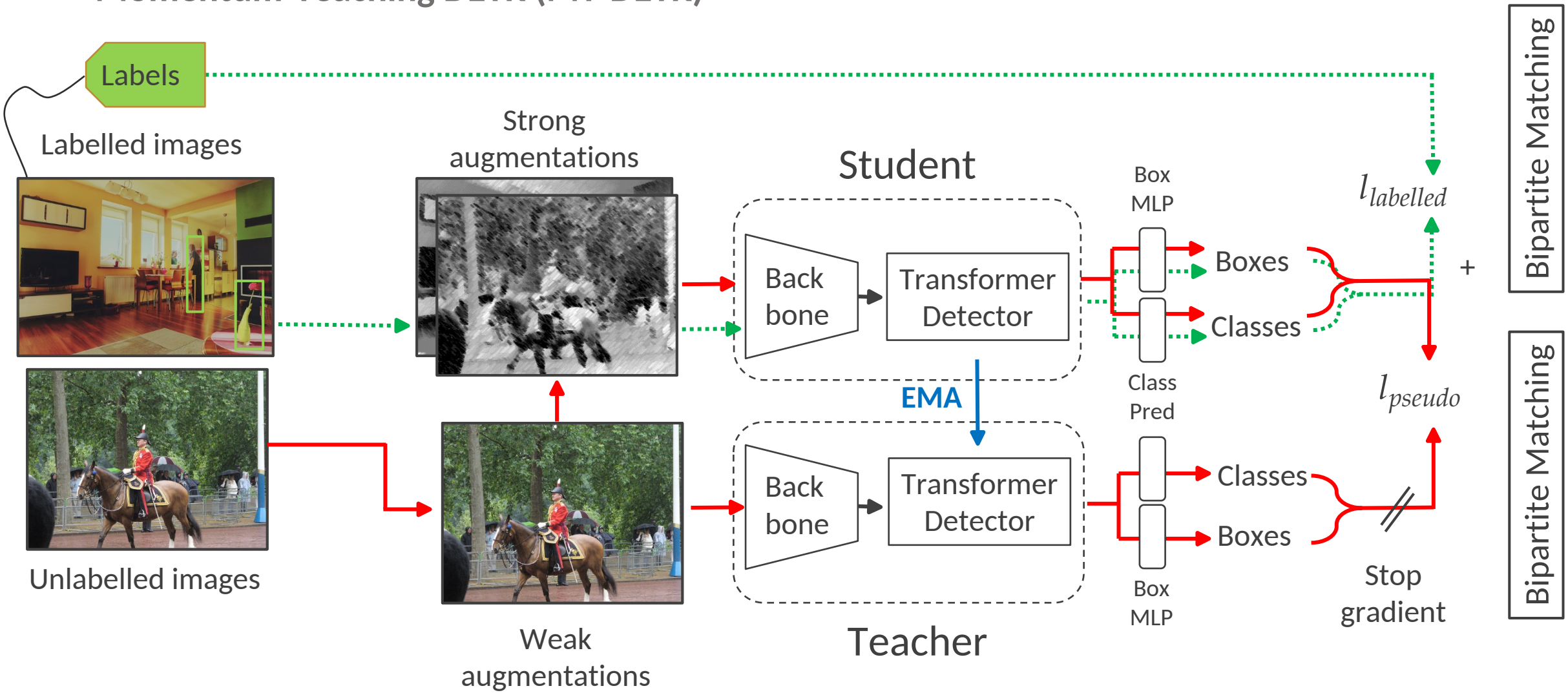
- Increasingly popular architecture for end-to-end detection.
- Simpler overall architecture, **without hand-crafted heuristics** (NMS, RoI Pooling or anchors).
- Rely on **Hungarian algorithm** for optimal matching between predictions and ground truths.
- Now strong contenders for SOTA performance.

Carion et al., "End-to-end object detection with transformers." ECCV, 2020.
Zhu et al., "Deformable DETR: Deformable transformers for end-to-end object detection." ICLR 2021

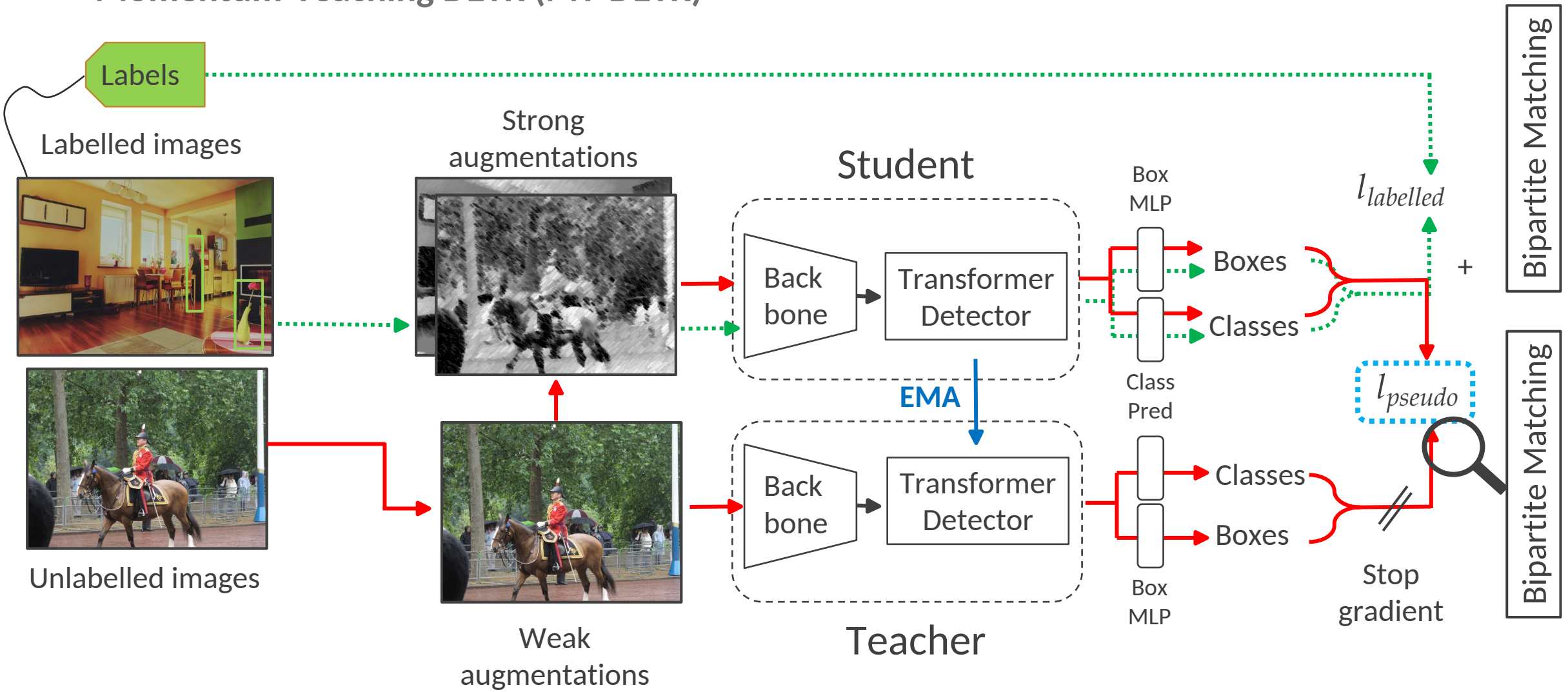
Momentum-Teaching DETR (MT-DETR)



Momentum-Teaching DETR (MT-DETR)

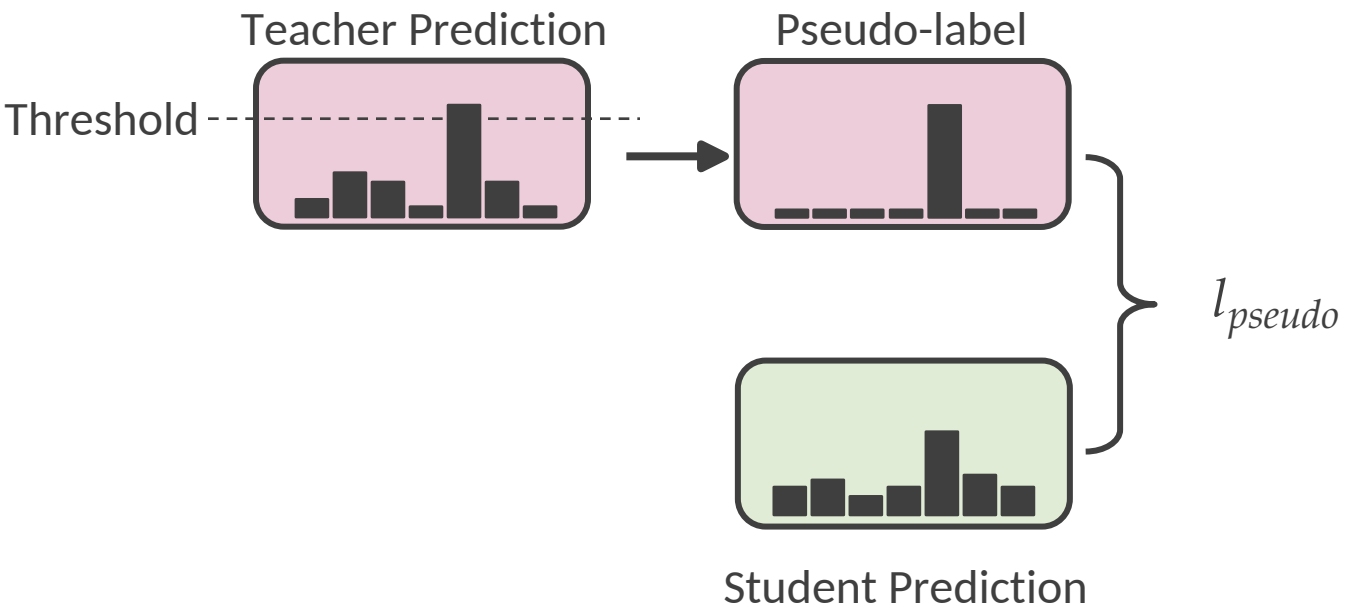


Momentum-Teaching DETR (MT-DETR)



Hard vs *Soft* Pseudo-labeling

Hard Pseudo-labeling:

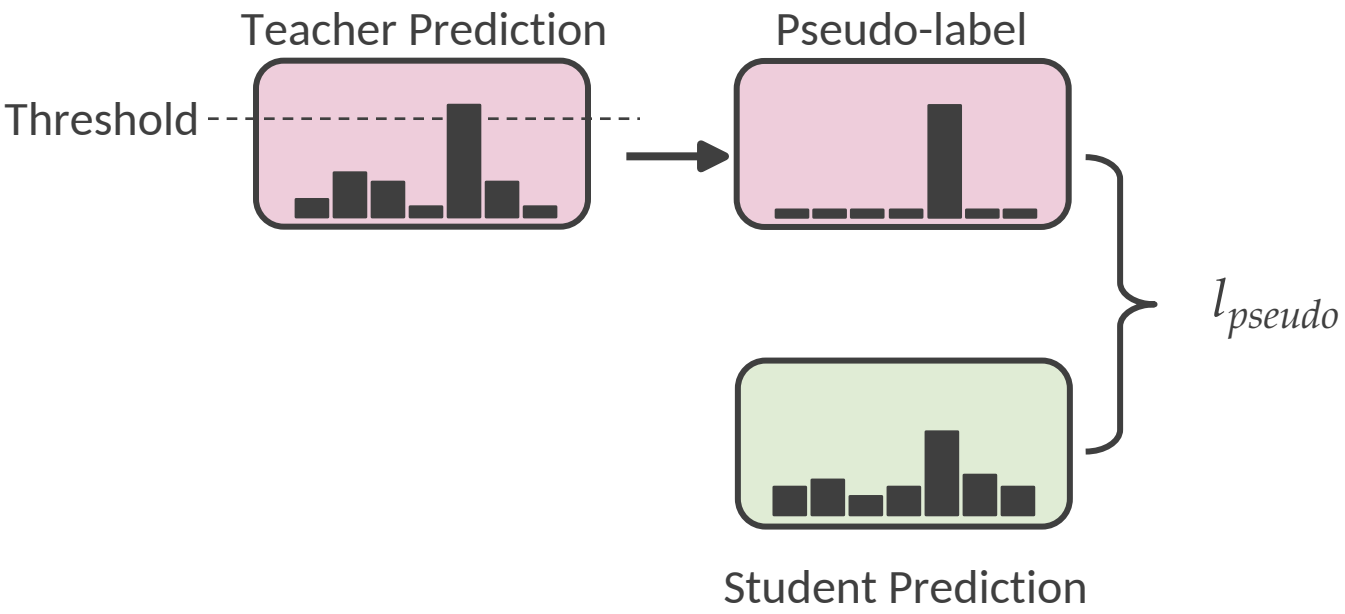


- × Encourage **high confidence** predictions
- × Focus on **prevailing** class
- × **Additional** hyperparameter with the threshold

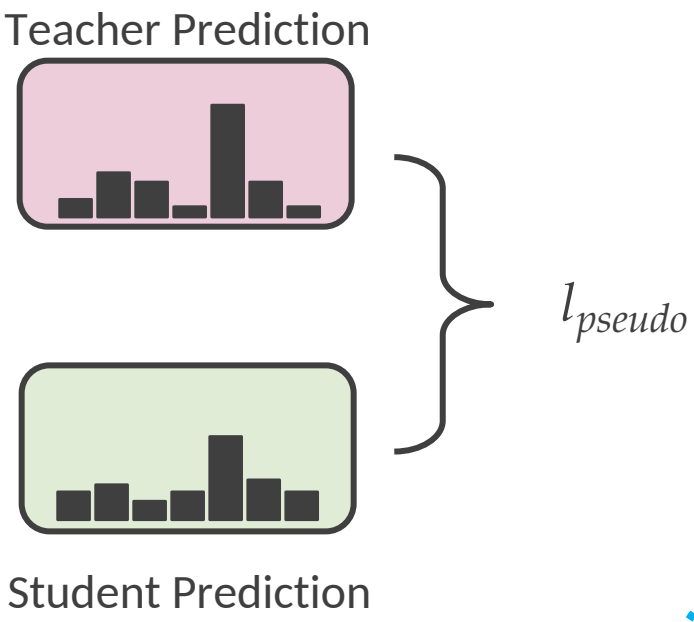
Sohn et al. "FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence." NeurIPS 2020
Tang et al. "Humble Teachers teach better students for semi-supervised object detection." In CVPR 2021

Hard vs *Soft* Pseudo-labeling

Hard Pseudo-labeling:



Soft Pseudo-labeling:

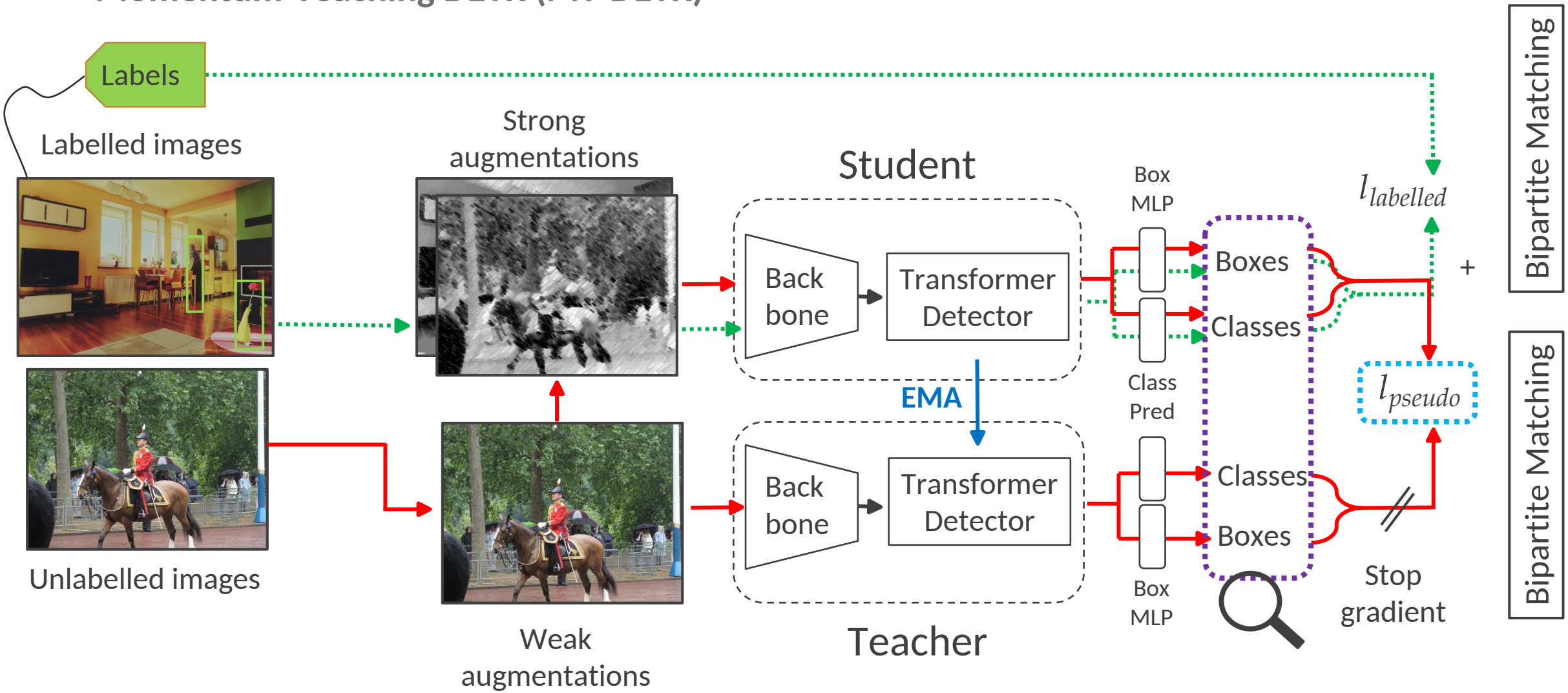


- ✕ Encourage **high confidence** predictions
- ✕ Focus on **prevailing** class
- ✕ **Additional** hyperparameter with the threshold

- ✓ **Relations** between classes
- ✓ More **diversity** in prevailing classes

Sohn et al. "FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence." NeurIPS 2020
Tang et al. "Humble Teachers teach better students for semi-supervised object detection." In CVPR 2021

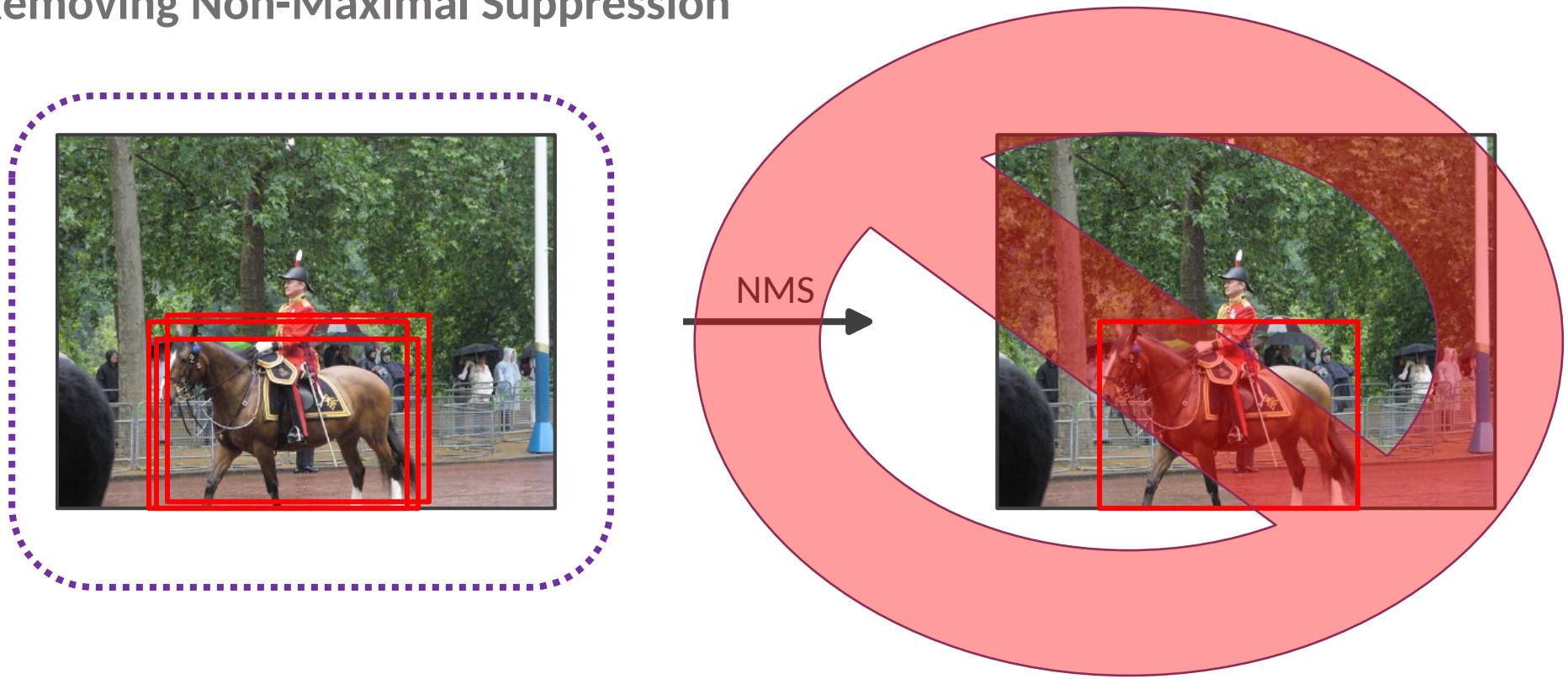
Momentum-Teaching DETR (MT-DETR)



Non-Maximal Suppression (NMS)

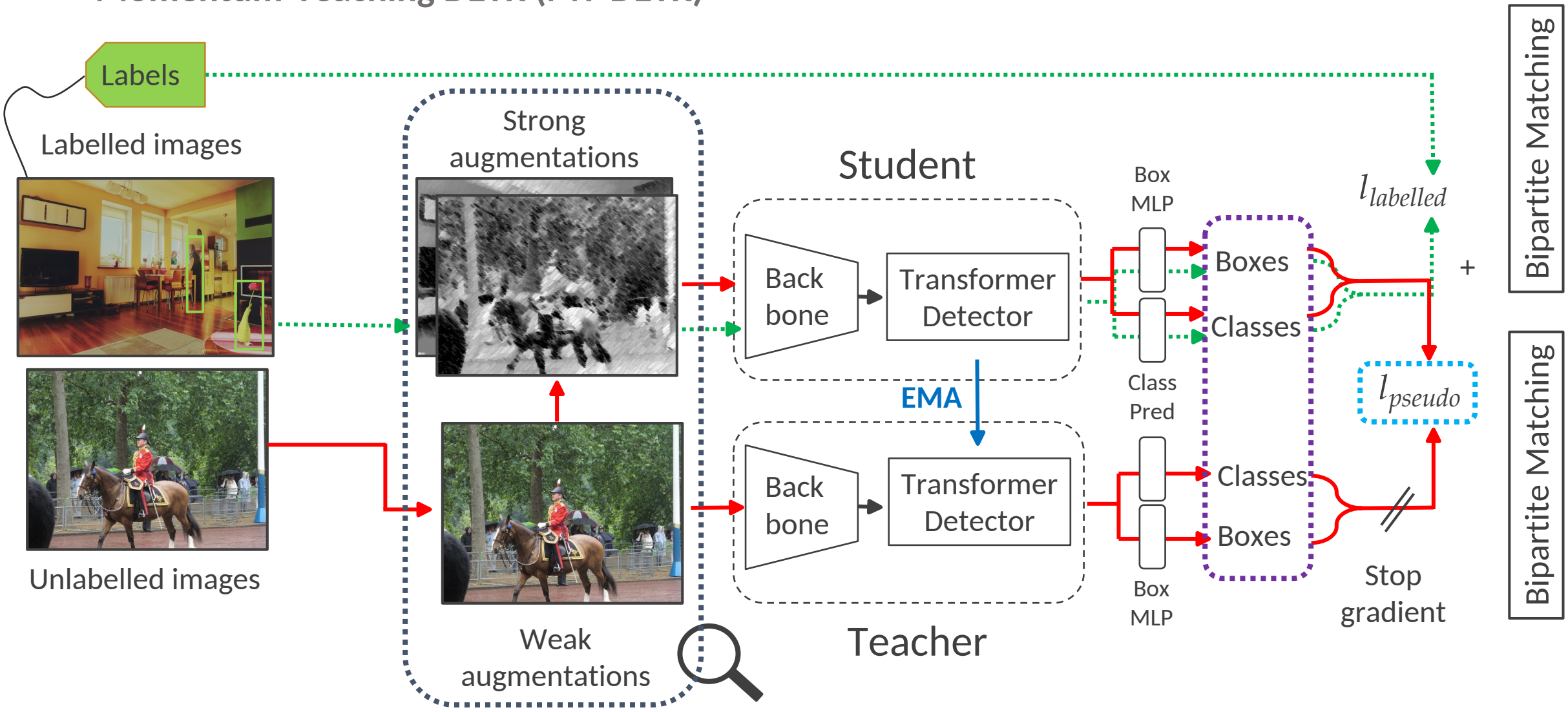
NMS

Removing Non-Maximal Suppression



- ✓ More **diverse** set of proposals without NMS
- ✓ Encourage proposals with **better localization** instead of more confidence
- ✓ **Less** hyperparameters

Momentum-Teaching DETR (MT-DETR)



Data augmentation

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 pseudo-labels	Div.
| w/o NMS + Soft pseudo-labels *(Ours)*	21.1
Basic + Photo. + CutOut + Geom.	**21.6**
Basic + Photo. + CutOut + Geom. + Augmentations in Supervised branch	**22.3**

Setting of UBT

Ours

- Common augmentations used in previous work
- ✓ **More augmentations** leads to the best results
- ✓ **Removing post-processing of proposals** solves the diverging issue

Ablation Studies

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 NMS
  - ✓ Soft labels without confidence thresholding

FAL-COCO

Method	OD Arch.	FAL-COCO			
		0.5% (590)	1% (1180)	5% (5900)	10% (11800)
STAC [127]	FRCNN + FPN	9.78 ± 0.53	13.97 ± 0.35	24.38 ± 0.12	28.64 ± 0.21
Instant-Teaching [159]	FRCNN + FPN	–	18.05 ± 0.15	26.75 ± 0.05	30.40 ± 0.05
Humble Teacher [129]	FRCNN + FPN	–	16.96 ± 0.38	27.70 ± 0.15	31.61 ± 0.28
Unbiased Teacher [91]	FRCNN + FPN	16.94 ± 0.23	20.75 ± 0.12	28.27 ± 0.11	31.50 ± 0.10
Soft Teacher [150]	FRCNN + FPN	–	20.46 ± 0.39	30.74 ± 0.08	34.04 ± 0.14
MT-DETR *(Ours)*	Def. DETR	**17.84** ± 0.54 (+8.89)	**22.03** ± 0.17 (+9.07)	**31.00** ± 0.11 (+7.41)	**34.52** ± 0.07 (+5.97)

- Evaluation on **different percentage of COCO labeled data** (with the corresponding number of images) and 100% of the dataset as unlabeled data.
- ✓ We achieve the **best performance** on all settings
- ✓ More **significant gap** when labeled data is scarce.

FAL-VOC

Method	OD Arch.	FAL-VOC 07-12		
		5% (250)	10% (500)	100% (5000)
STAC [127]	FRCNN + FPN	–	–	44.64
Instant-Teaching [159]	FRCNN + FPN	–	–	50.00
Humble Teacher [129]	FRCNN + FPN	–	–	53.04
Unbiased Teacher [github]	FRCNN + FPN	–	–	54.48
Unbiased Teacher*	FRCNN + FPN	35.98 ± 0.71	40.34 ± 0.95	54.61
MT-DETR *(Ours)*	Def. DETR	$\mathbf{36.95} \pm 0.53$ (+14.08)	$\mathbf{43.15} \pm 1.10$ (+14.12)	**56.2**

- Evaluation on **different percentage of VOC 2007 labeled data** (with the corresponding number of images) and 100% of VOC 2012 as unlabeled data.
- ✓ We achieve the **best performance** on all settings
- ✓ More **significant gap** when labeled data is scarce.