Visual Grounding

金江林

Outline

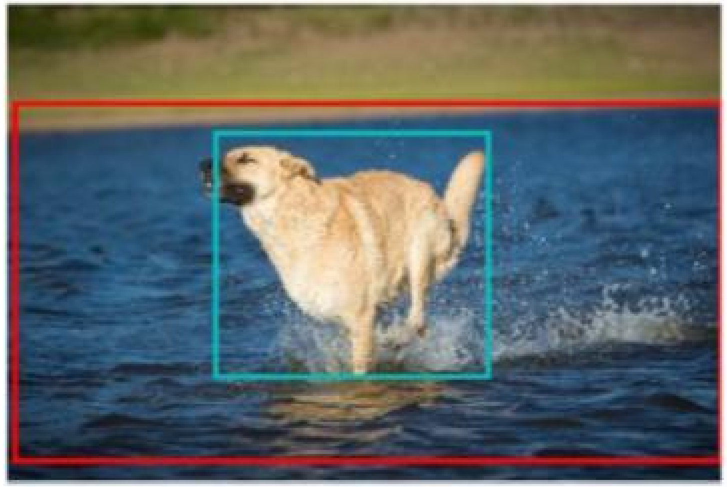
- Visual Grounding Background
    - Task
    - Dataset
- Fully Supervised
- Weakly Supervised
- Semi-Supervised
    - Consistency Regularization
    - Pseudo Label
    - Match
- Further Work

Task

- Phrase Grounding：对于给定的sentence，要定位其中提到的全部物体（phrase）

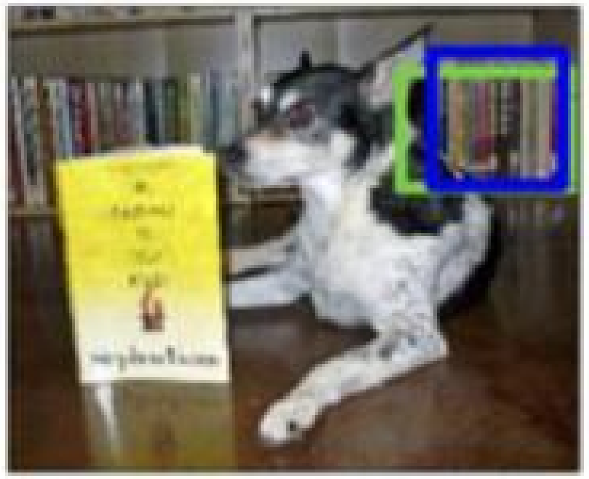
- Referring expression grounding：每个语言描述（expression）只指示一个物体

Phrases Grounding



(h) A white dog is running over the water.

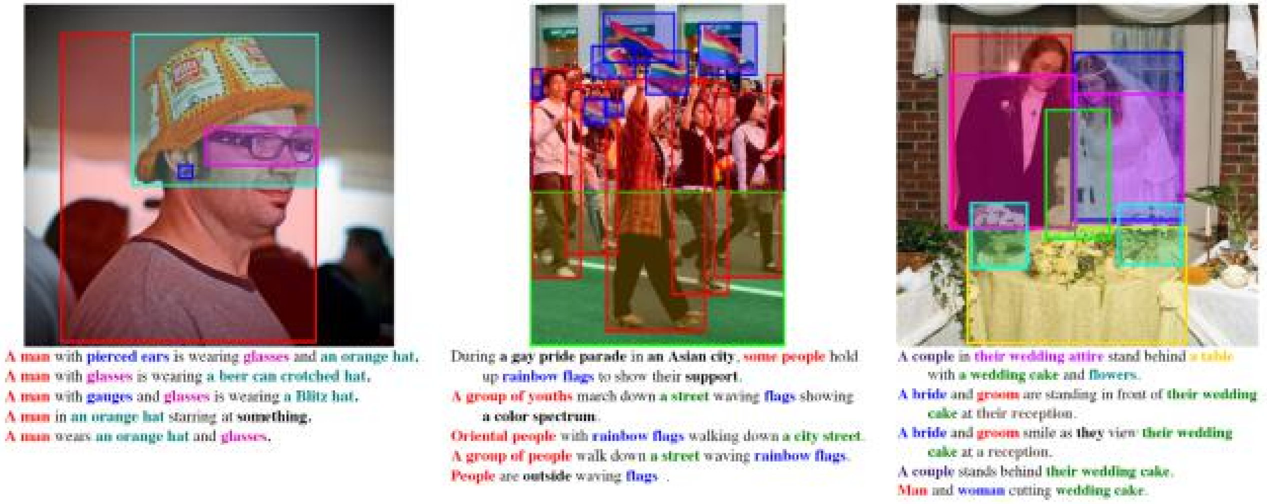
Refer expression Grounding



books about bears

Dataset

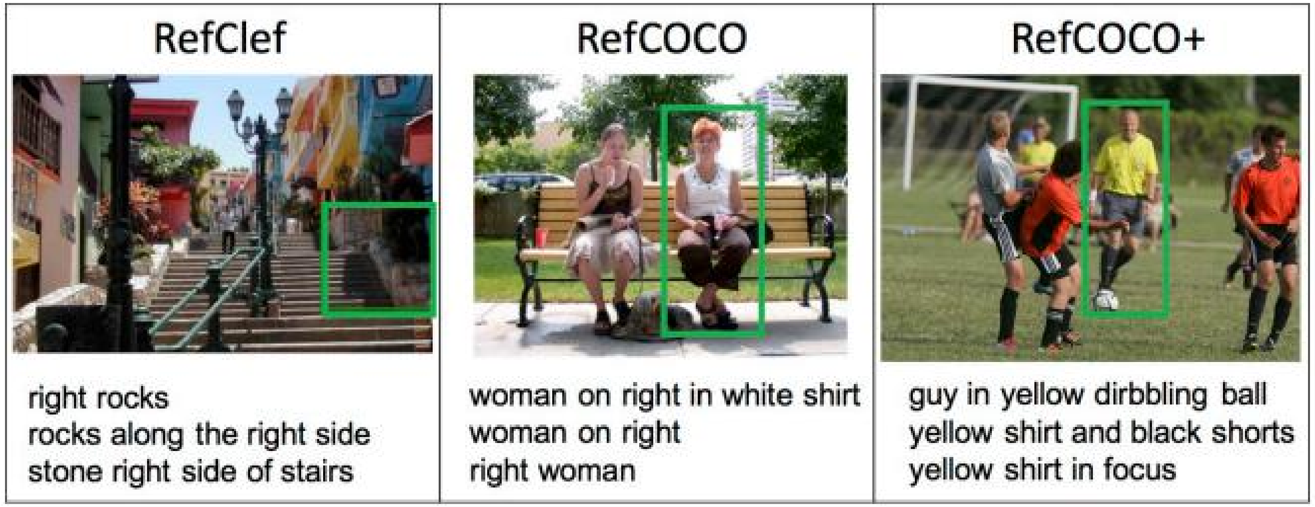
- Phrases Grounding
  - Flickr30k Entities
  - ReferItGame(RefClef)
- Refer expression Grounding
  - RefCOCO
  - RefCOCO+
  - RefCOCOg



A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

During a gay pride parade in an Asian city, some people hold up rainbow flags to show their support.
A group of youths march down a street waving flags showing a color spectrum.
Oriental people with rainbow flags walking down a city street.
A group of people walk down a street waving rainbow flags.
People are outside waving flags .

A couple in their wedding attire stand behind a table with a wedding cake and flowers.
A bride and groom are standing in front of their wedding cake at their reception.
A bride and groom smile as they view their wedding cake at a reception.
A couple stands behind their wedding cake.
Man and woman cutting wedding cake.



RefClef

right rocks
rocks along the right side
stone right side of stairs

RefCOCO

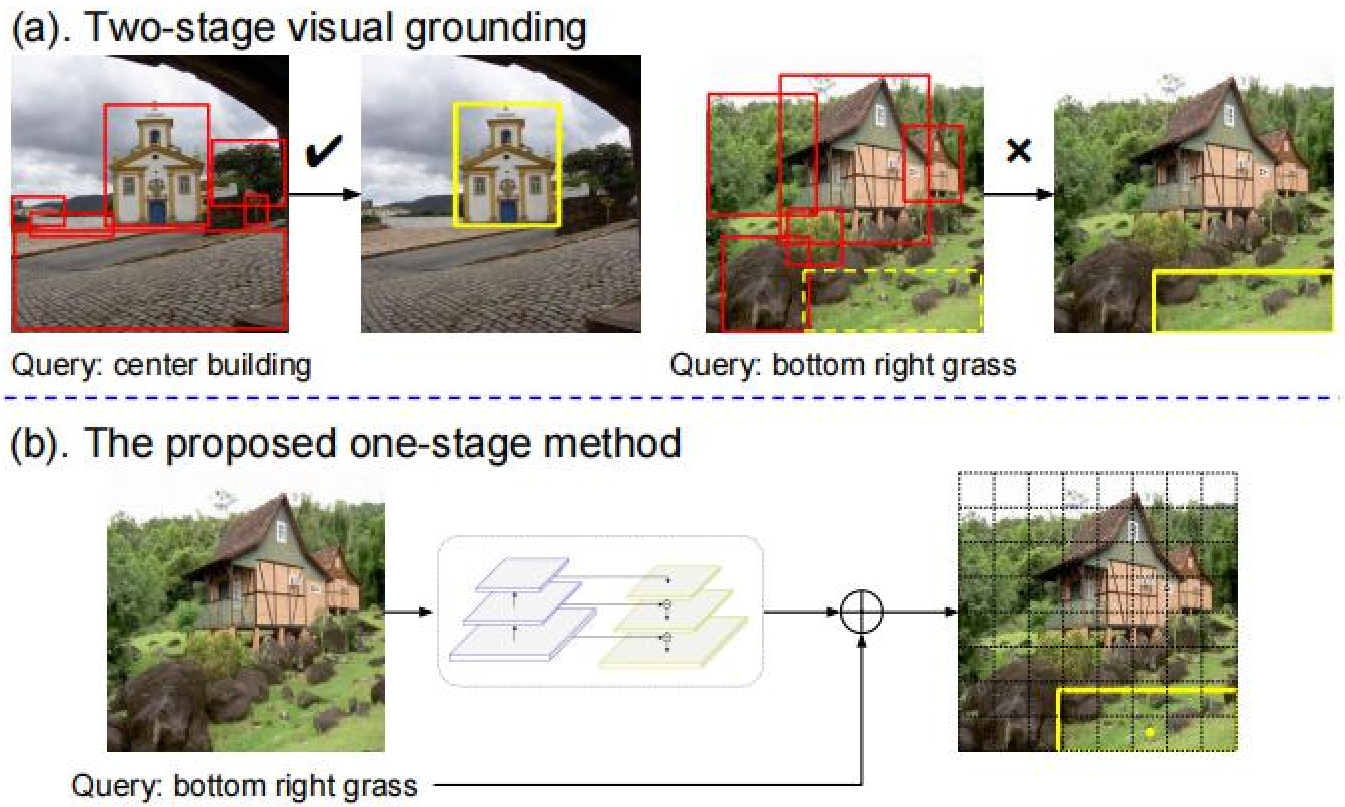
woman on right in white shirt
woman on right
right woman

RefCOCO+

guy in yellow dirbbling ball
yellow shirt and black shorts
yellow shirt in focus

Methods

□one stage or two stage

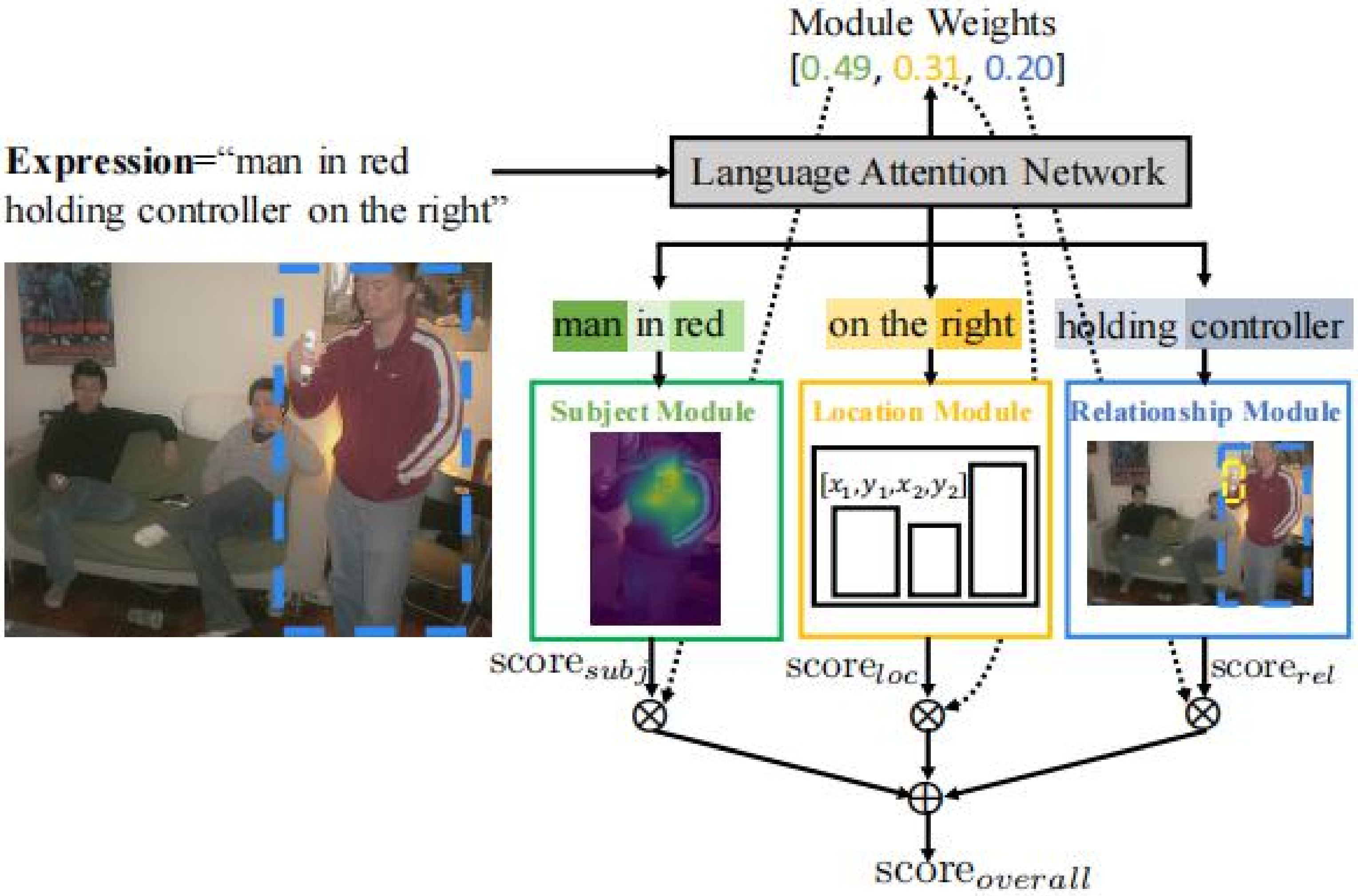

(a). Two-stage visual grounding
✔
×
Query: center building
Query: bottom right grass
(b). The proposed one-stage method
Query: bottom right grass

✓ 训练效率高

✓ 推理更快

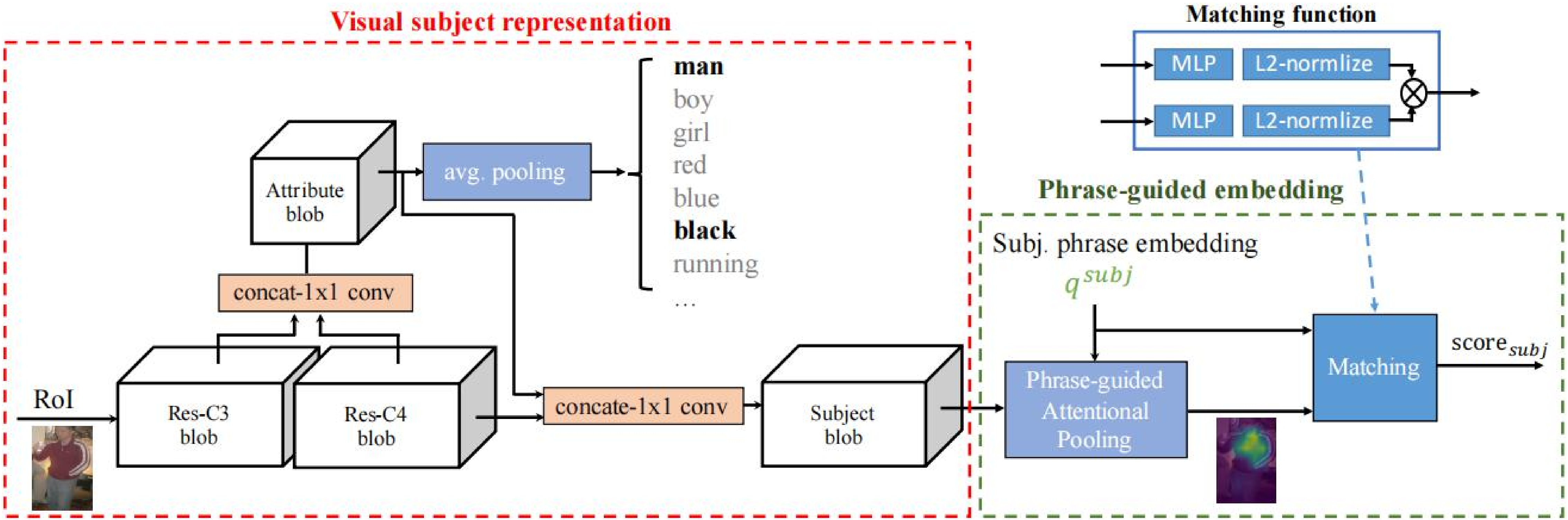
✓ 结果更准确

Fully Supervised



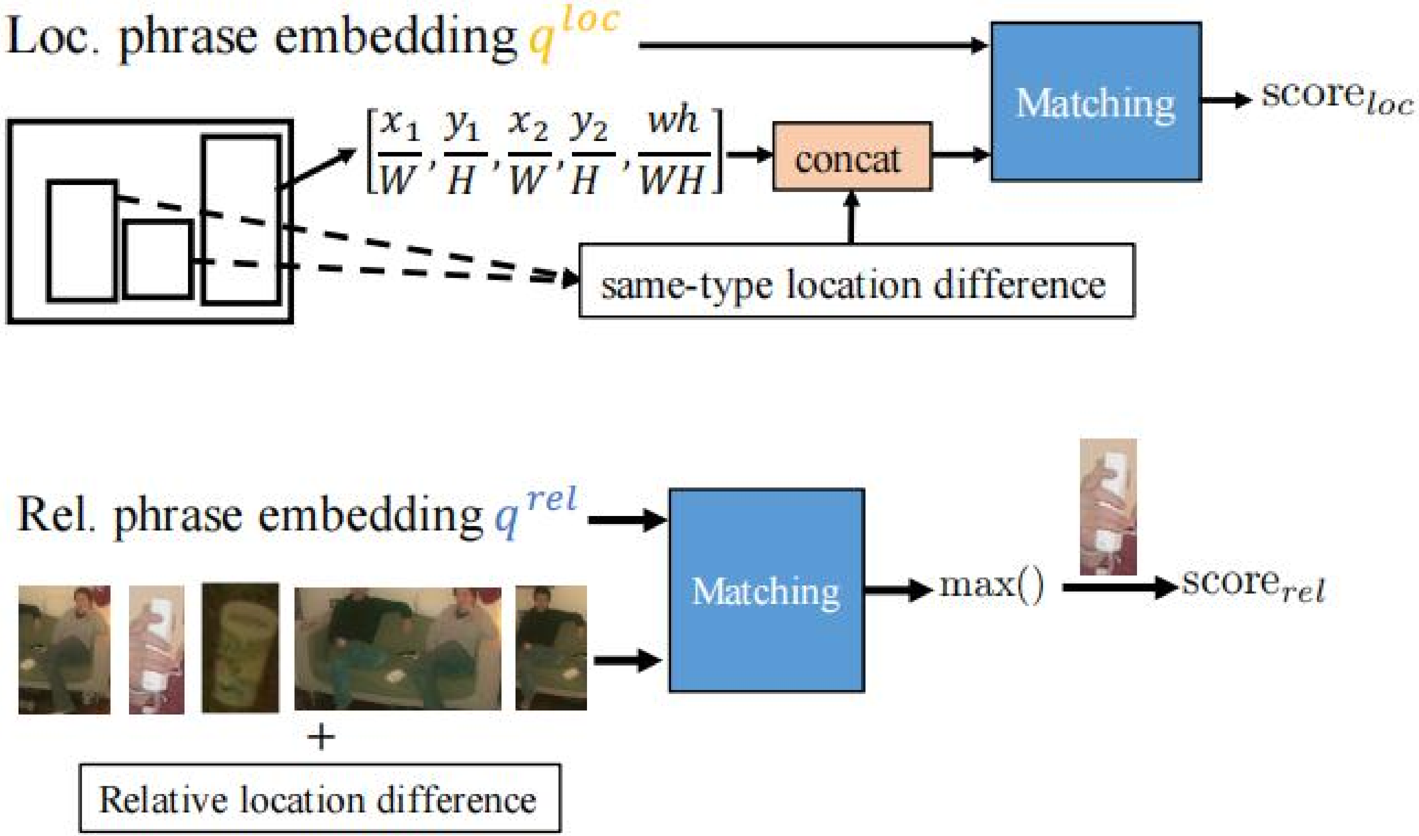
CVPR 2018, MAttNet: Modular Attention Network for Referring Expression Comprehension

Fully Supervised



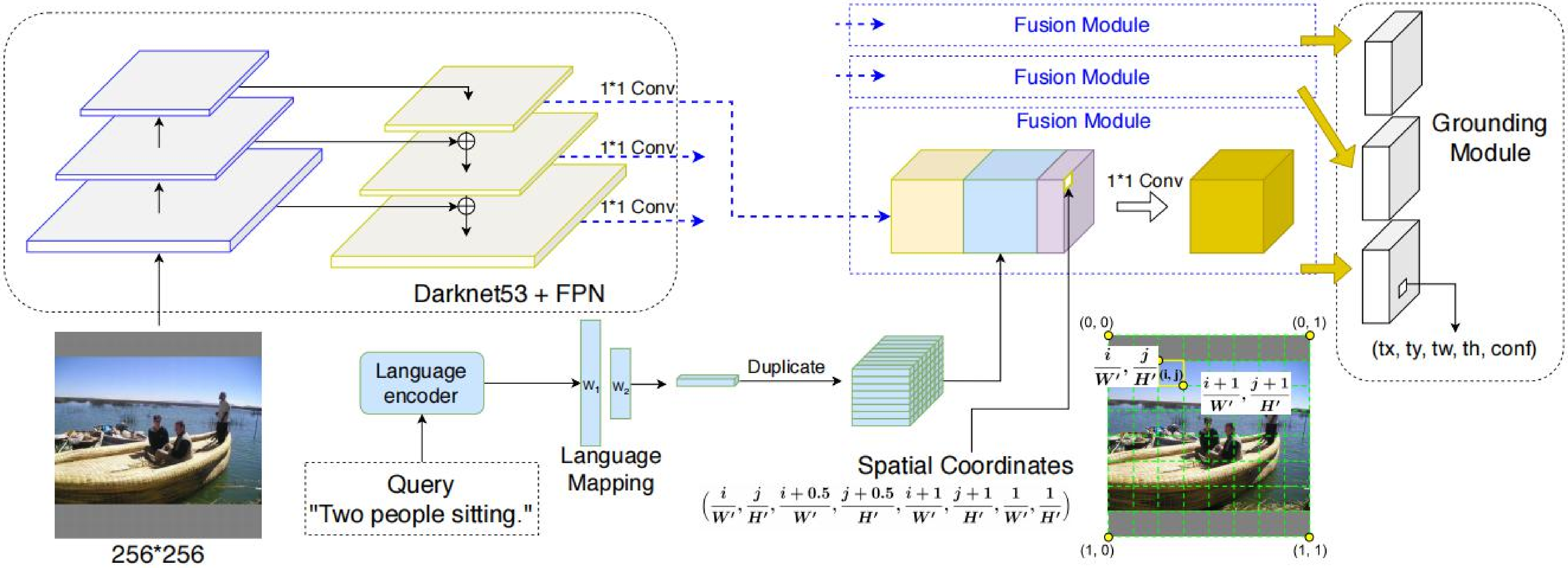
CVPR 2018, MAttNet: Modular Attention Network for Referring Expression Comprehension

Fully Supervised


Loc. phrase embedding q^{loc}
$\left[\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H}, \frac{wh}{WH}\right]$
concat
Matching
score_{loc}
same-type location difference
Rel. phrase embedding q^{rel}
Matching
max()
score_{rel}
+
Relative location difference

CVPR 2018, MAttNet: Modular Attention Network for Referring Expression Comprehension

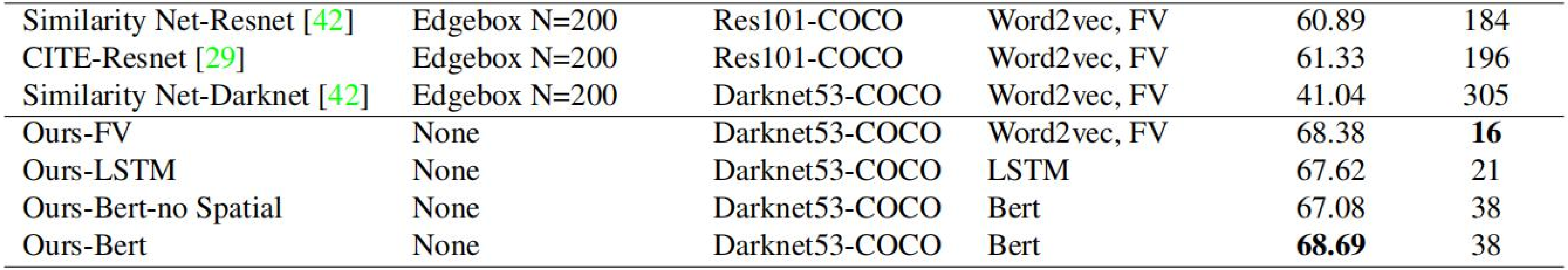
Fully Supervised



CVPR 2019, A Fast and Accurate One-Stage Approach to Visual Grounding

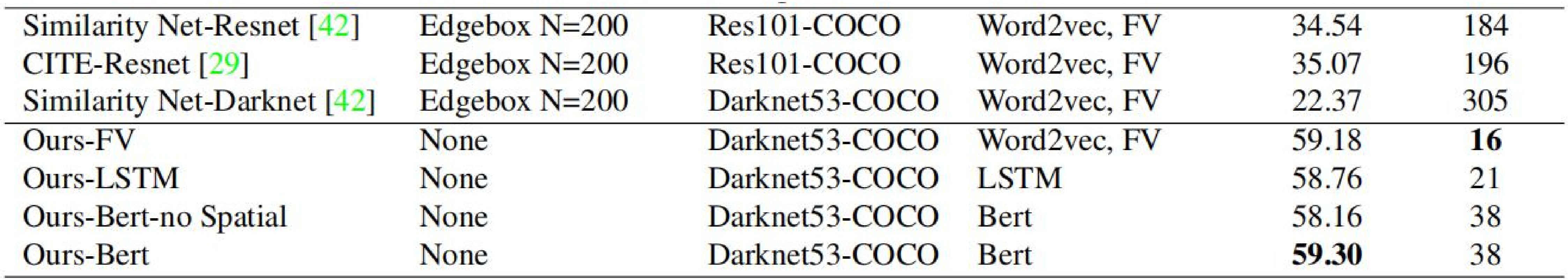
Results

Phrase localization results on the test set of Flickr30K Entities



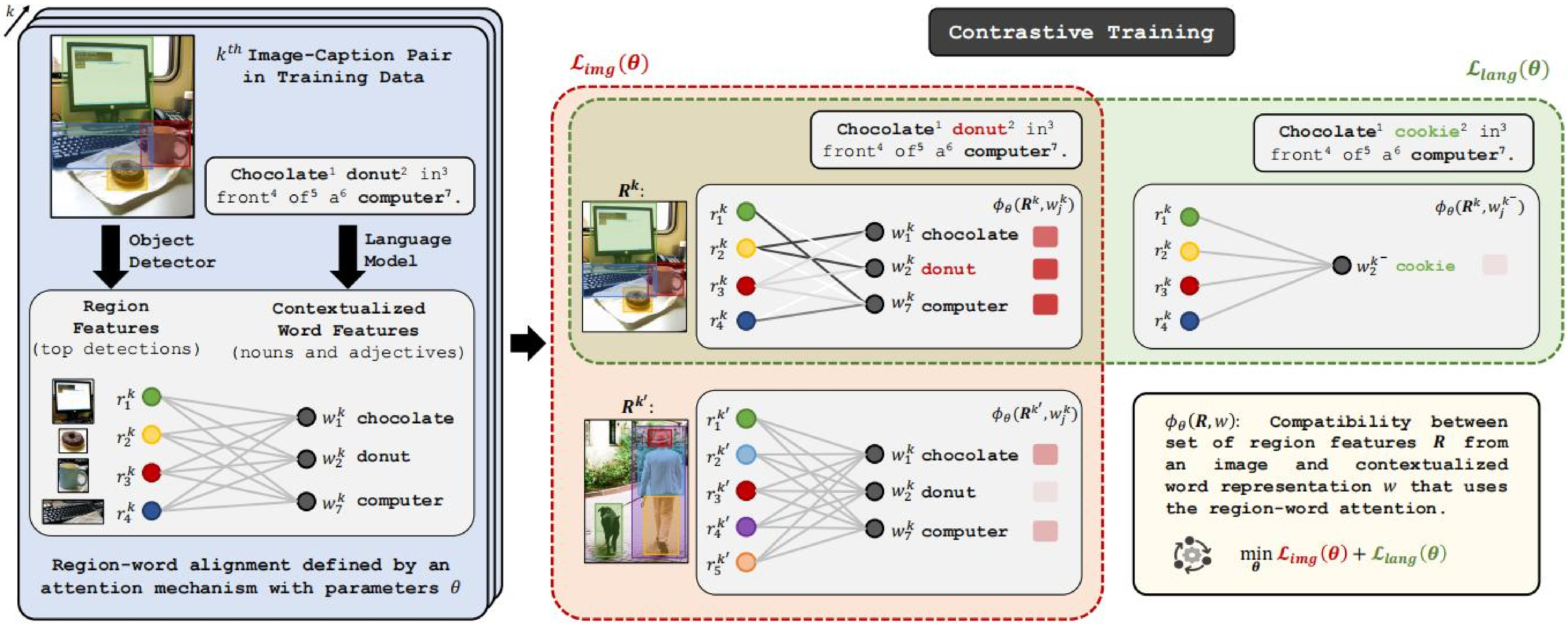
Similarity Net-Resnet [42]	Edgebox N=200	Res101-COCO	Word2vec, FV	60.89	184
CITE-Resnet [29]	Edgebox N=200	Res101-COCO	Word2vec, FV	61.33	196
Similarity Net-Darknet [42]	Edgebox N=200	Darknet53-COCO	Word2vec, FV	41.04	305
Ours-FV	None	Darknet53-COCO	Word2vec, FV	68.38	**16**
Ours-LSTM	None	Darknet53-COCO	LSTM	67.62	21
Ours-Bert-no Spatial	None	Darknet53-COCO	Bert	67.08	38
Ours-Bert	None	Darknet53-COCO	Bert	**68.69**	38

Referring expression comprehension results on the test set of ReferItGame



Similarity Net-Resnet [42]	Edgebox N=200	Res101-COCO	Word2vec, FV	34.54	184
CITE-Resnet [29]	Edgebox N=200	Res101-COCO	Word2vec, FV	35.07	196
Similarity Net-Darknet [42]	Edgebox N=200	Darknet53-COCO	Word2vec, FV	22.37	305
Ours-FV	None	Darknet53-COCO	Word2vec, FV	59.18	**16**
Ours-LSTM	None	Darknet53-COCO	LSTM	58.76	21
Ours-Bert-no Spatial	None	Darknet53-COCO	Bert	58.16	38
Ours-Bert	None	Darknet53-COCO	Bert	**59.30**	38

Weakly Supervised



k^{th} Image-Caption Pair in Training Data

Chocolate^1 donut^2 in^3 front^4 of^5 a^6 computer^7.

Object Detector

Language Model

Region Features (top detections)

Contextualized Word Features (nouns and adjectives)

Region-word alignment defined by an attention mechanism with parameters θ

Contrastive Training

Chocolate^1 donut^2 in^3 front^4 of^5 a^6 computer^7.

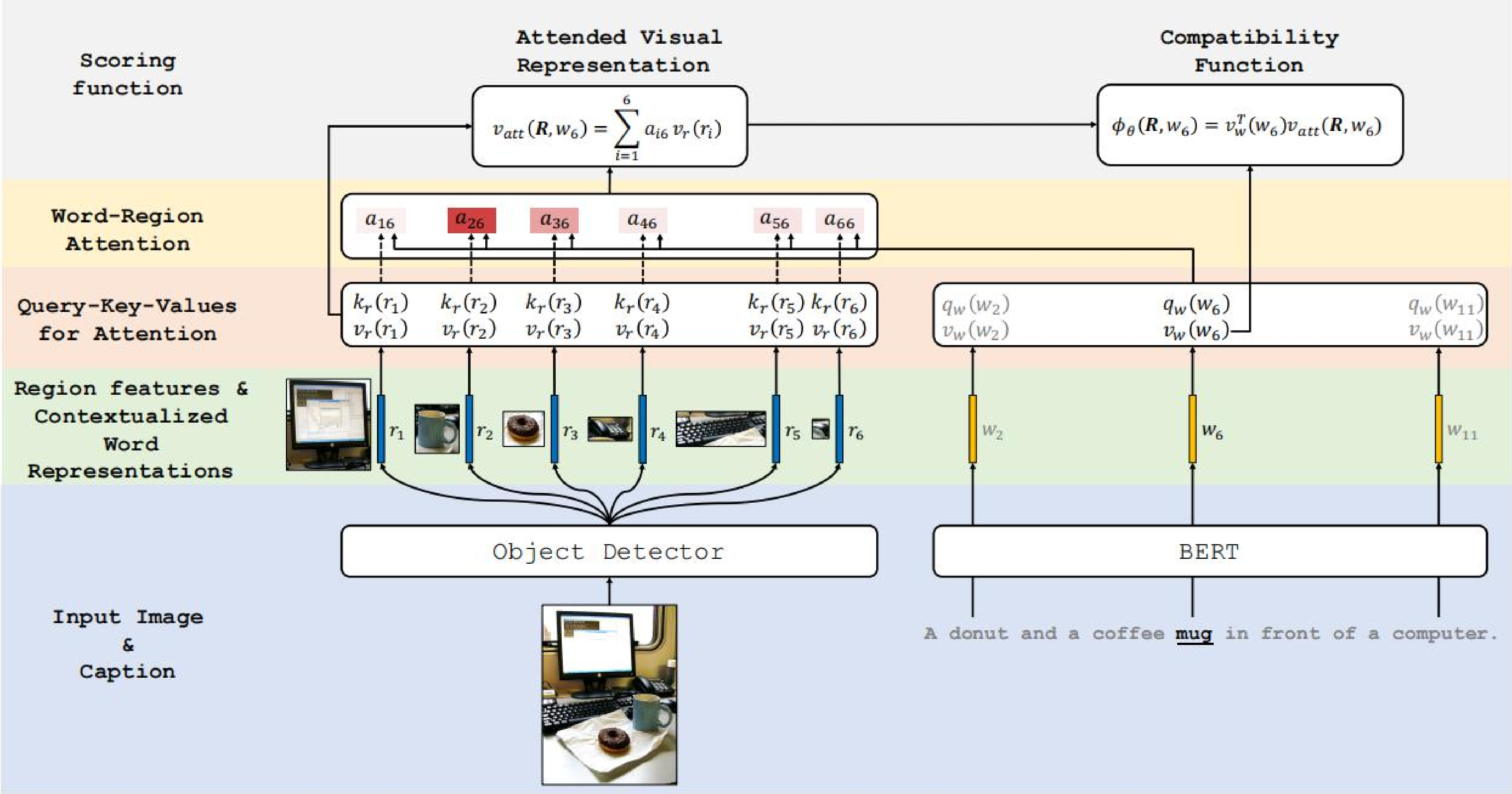
Chocolate^1 cookie^2 in^3 front^4 of^5 a^6 computer^7.

$\phi_\theta(\boldsymbol{R}, w)$: Compatibility between set of region features $\boldsymbol{R}$ from an image and contextualized word representation w that uses the region-word attention.

$\min_\theta \mathcal{L}_{img}(\theta) + \mathcal{L}_{lang}(\theta)$

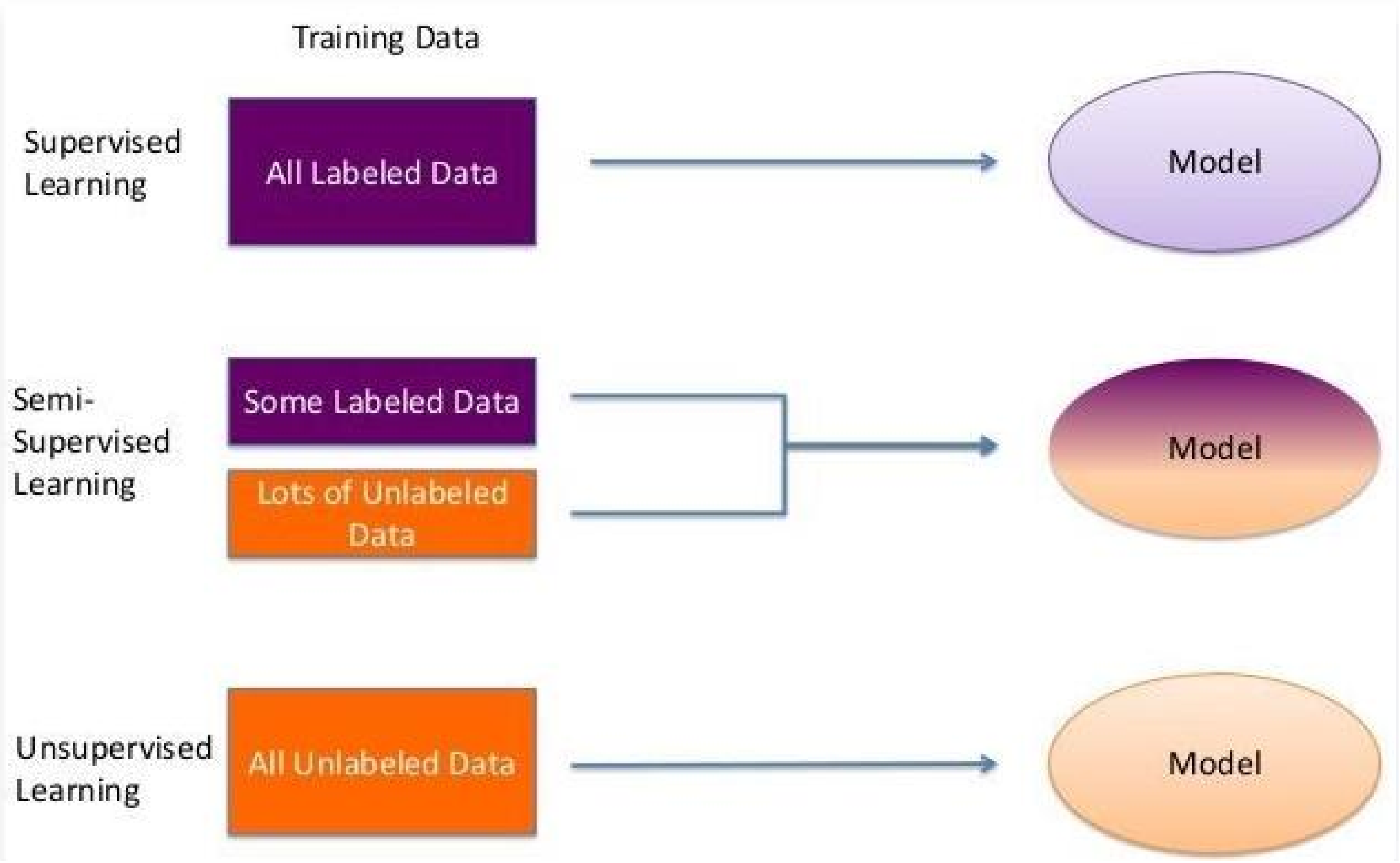
ECCV 2020 , Contrastive Learning for Weakly Supervised Phrase Grounding

Weakly Supervised


Scoring function
Attended Visual Representation
$v_{att}(\mathbf{R}, w_6) = \sum_{i=1}^{6} a_{i6}\, v_r(r_i)$
Compatibility Function
$\phi_\theta(\mathbf{R}, w_6) = v_w^T(w_6) v_{att}(\mathbf{R}, w_6)$
Word-Region Attention
a_{16} a_{26} a_{36} a_{46} a_{56} a_{66}
Query-Key-Values for Attention
$k_r(r_1)$ $k_r(r_2)$ $k_r(r_3)$ $k_r(r_4)$ $k_r(r_5)$ $k_r(r_6)$
$v_r(r_1)$ $v_r(r_2)$ $v_r(r_3)$ $v_r(r_4)$ $v_r(r_5)$ $v_r(r_6)$
$q_w(w_2)$ $q_w(w_6)$ $q_w(w_{11})$
$v_w(w_2)$ $v_w(w_6)$ $v_w(w_{11})$
Region features & Contextualized Word Representations
r_1 r_2 r_3 r_4 r_5 r_6
w_2 w_6 w_{11}
Object Detector
BERT
Input Image & Caption
A donut and a coffee mug in front of a computer.

ECCV 2020 , Contrastive Learning for Weakly Supervised Phrase Grounding

Semi-Supervised



- 完整的大量注释难以获得
- 充分利用数据信息
- 提高模型的性能

- 有标签数据和无标签数据具有相同的分布
- 模型性能下降

Semi-Supervised

- Consistency Regularization:给定一个未标记的数据点及其扰动的形式，目标是最小化两个输出之间的距离。对于无标签图像，添加噪声之后模型预测也应该保持不变

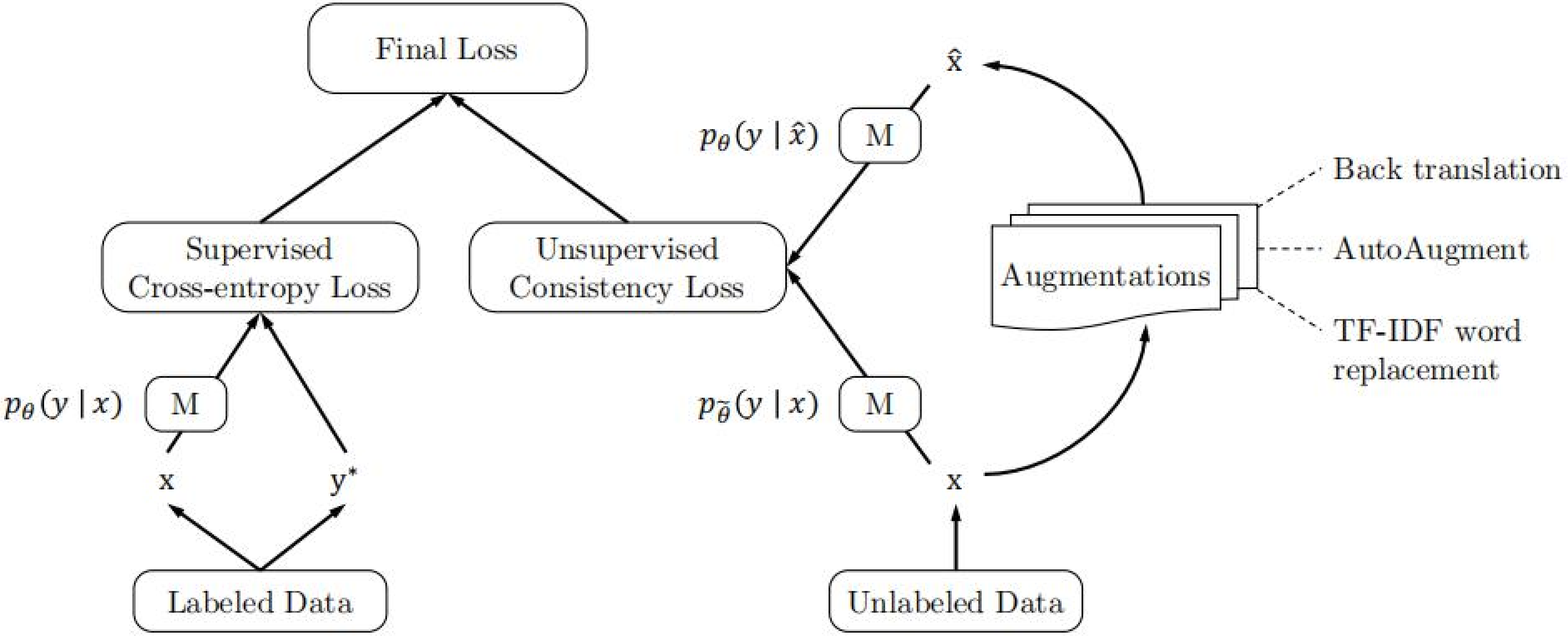
$$d(f_\theta(x)\ ,\ \ f(\hat{x})),$$

- Pseudo Label:先使用模型生成伪标签，然后与有标签的数据一起训练，提供额外的信息。

$$L=\frac{1}{n}\sum_{m=1}^{n}\sum_{i=1}^{C}L(y_i^m,f_i^m)+\alpha(t)\frac{1}{n'}\sum_{m=1}^{n'}\sum_{i=1}^{C}L(y_i'^m,f_i'^m),$$

- Match系列：将几种半监督方法进行混合

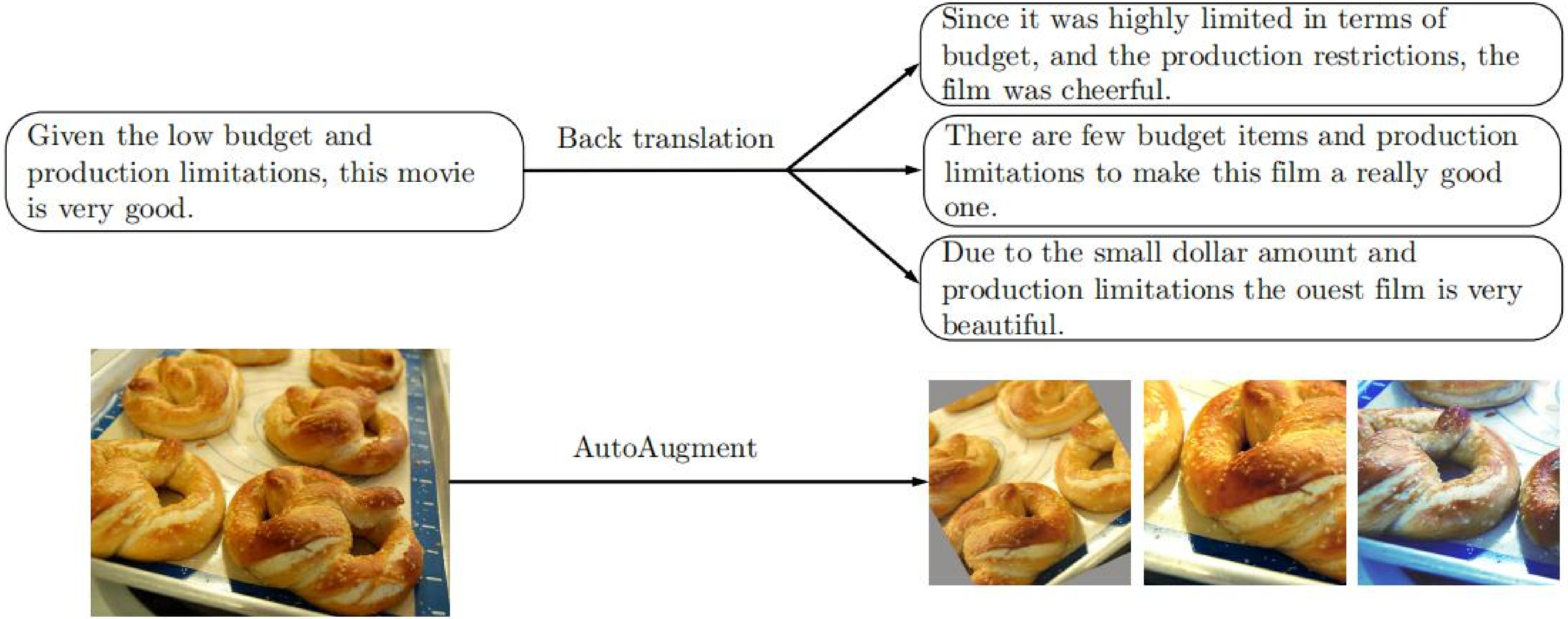
Semi-Supervised-Consistency Regularization



NeurIPS 2020 , Unsupervised Data Augmentationfor Consistency Training

Data Augmentation

- 图像分类:AutoAugment,在所有的图像处理转换方式中进行搜索，以便找到一个最优的增强策略
- 文本分类:Back translation,把一个样本（语言A）转换成另一个语言B再转换回来



Given the low budget and production limitations, this movie is very good.

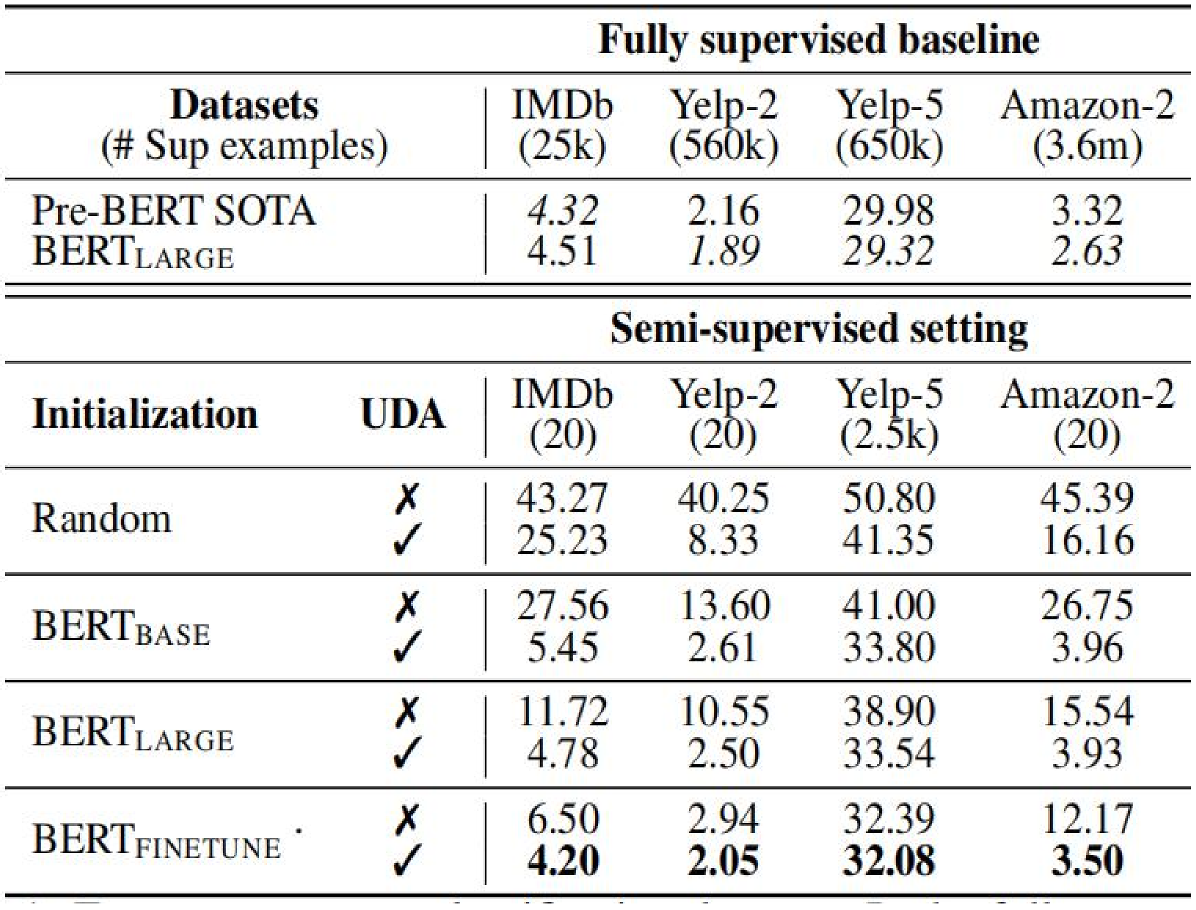
Since it was highly limited in terms of budget, and the production restrictions, the film was cheerful.

There are few budget items and production limitations to make this film a really good one.

Due to the small dollar amount and production limitations the ouest film is very beautiful.

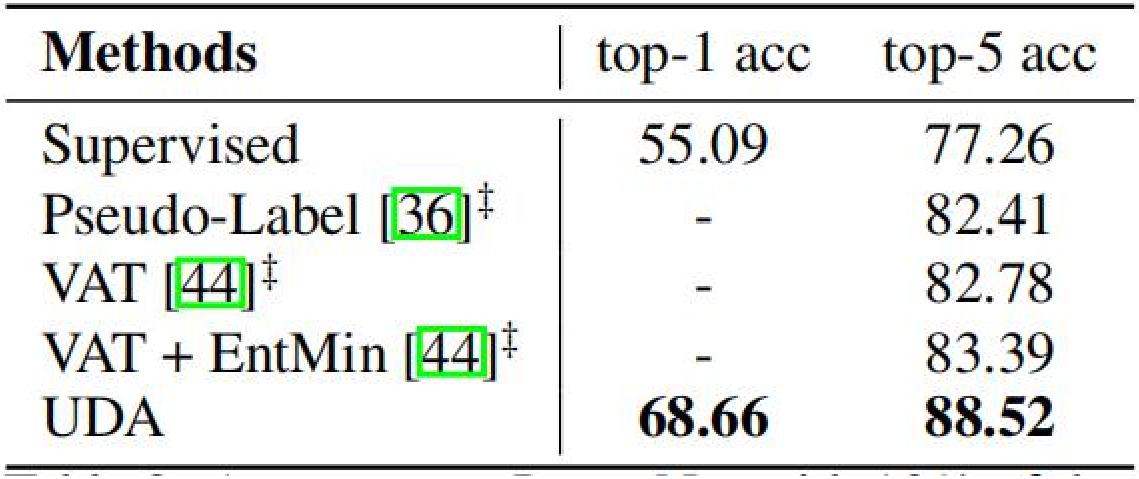
NeurIPS 2020 , Unsupervised Data Augmentationfor Consistency Training

Results



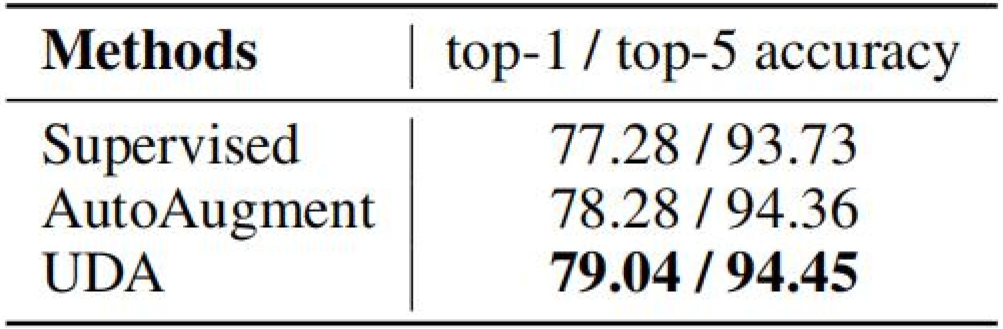
		Fully supervised baseline			
Datasets (# Sup examples)		IMDb (25k)	Yelp-2 (560k)	Yelp-5 (650k)	Amazon-2 (3.6m)
Pre-BERT SOTA		*4.32*	2.16	29.98	3.32
BERT$_{\text{LARGE}}$		4.51	*1.89*	*29.32*	*2.63*

		Semi-supervised setting			
Initialization	**UDA**	IMDb (20)	Yelp-2 (20)	Yelp-5 (2.5k)	Amazon-2 (20)
Random	✗	43.27	40.25	50.80	45.39
	✓	25.23	8.33	41.35	16.16
BERT$_{\text{BASE}}$	✗	27.56	13.60	41.00	26.75
	✓	5.45	2.61	33.80	3.96
BERT$_{\text{LARGE}}$	✗	11.72	10.55	38.90	15.54
	✓	4.78	2.50	33.54	3.93
BERT$_{\text{FINETUNE}}$	✗	6.50	2.94	32.39	12.17
	✓	**4.20**	**2.05**	**32.08**	**3.50**



Methods	top-1 acc	top-5 acc
Supervised	55.09	77.26
Pseudo-Label [36][‡]	-	82.41
VAT [44][‡]	-	82.78
VAT + EntMin [44][‡]	-	83.39
UDA	**68.66**	**88.52**

Accuracy on ImageNet with 10% of the labeled set

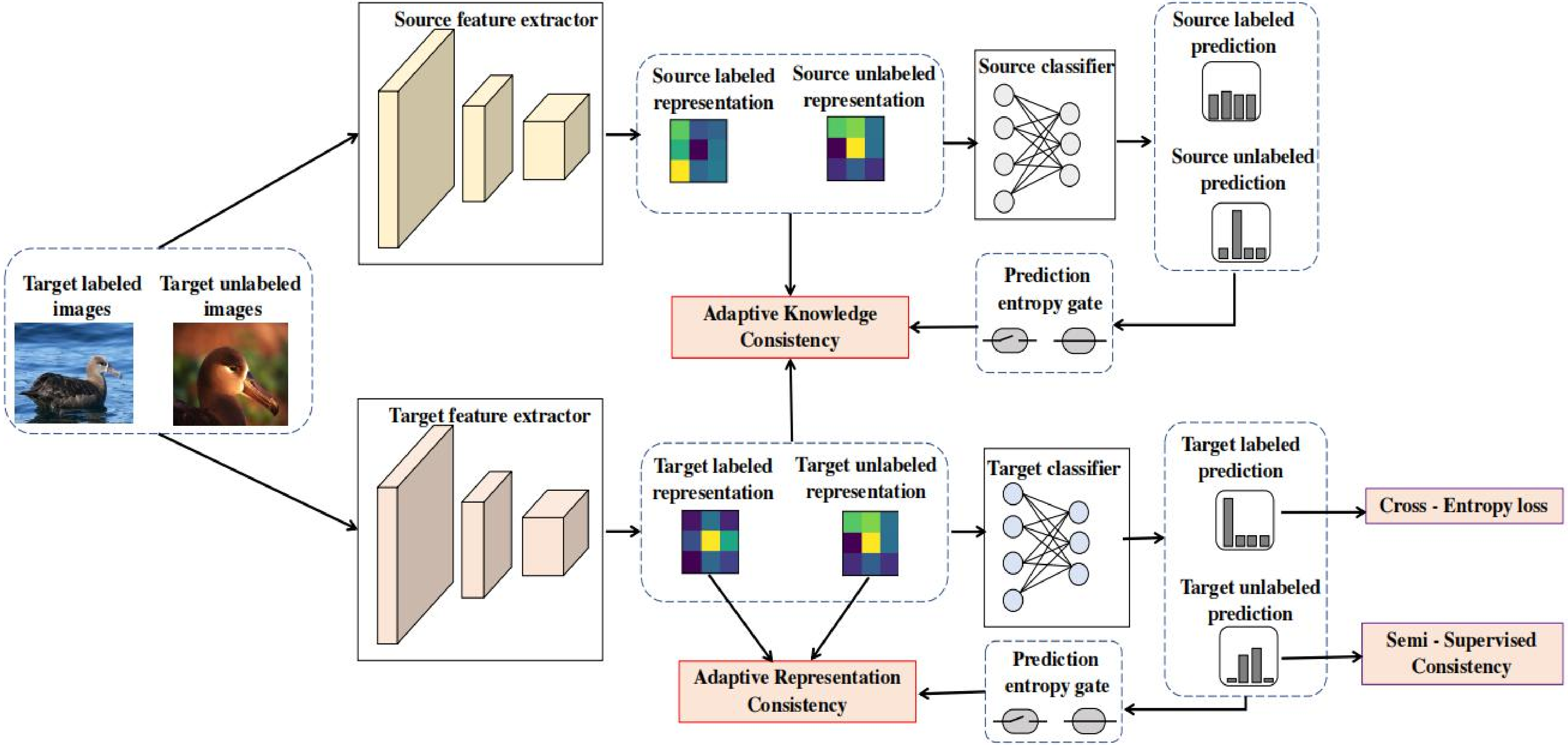


Methods	top-1 / top-5 accuracy
Supervised	77.28 / 93.73
AutoAugment	78.28 / 94.36
UDA	**79.04 / 94.45**

Accuracy on the full ImageNet

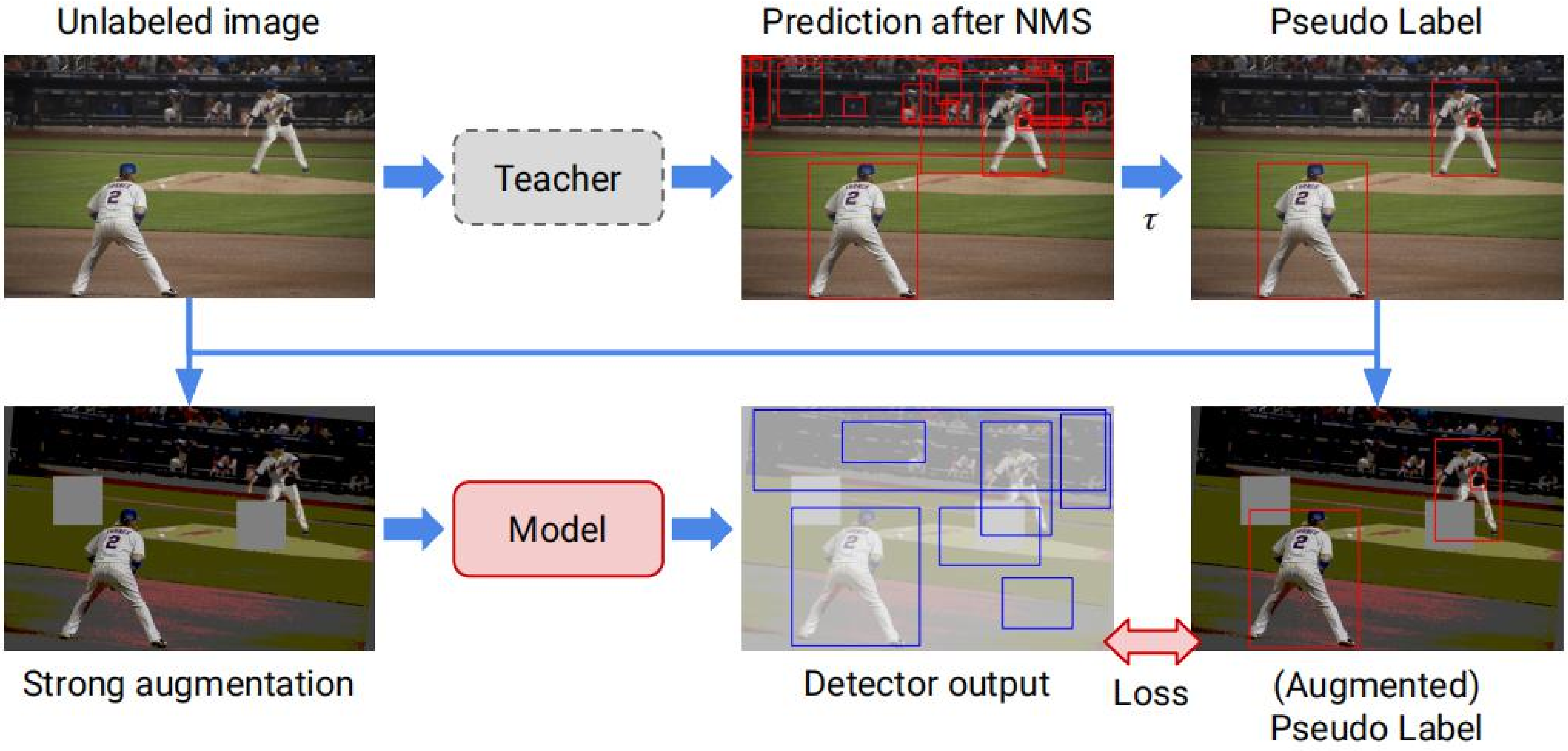
NeurIPS 2020 , Unsupervised Data Augmentationfor Consistency Training

Consistency Regularization


Source feature extractor
Source labeled representation
Source unlabeled representation
Source classifier
Source labeled prediction
Source unlabeled prediction
Target labeled images
Target unlabeled images
Prediction entropy gate
Adaptive Knowledge Consistency
Target feature extractor
Target labeled representation
Target unlabeled representation
Target classifier
Target labeled prediction
Cross - Entropy loss
Target unlabeled prediction
Semi - Supervised Consistency
Adaptive Representation Consistency
Prediction entropy gate

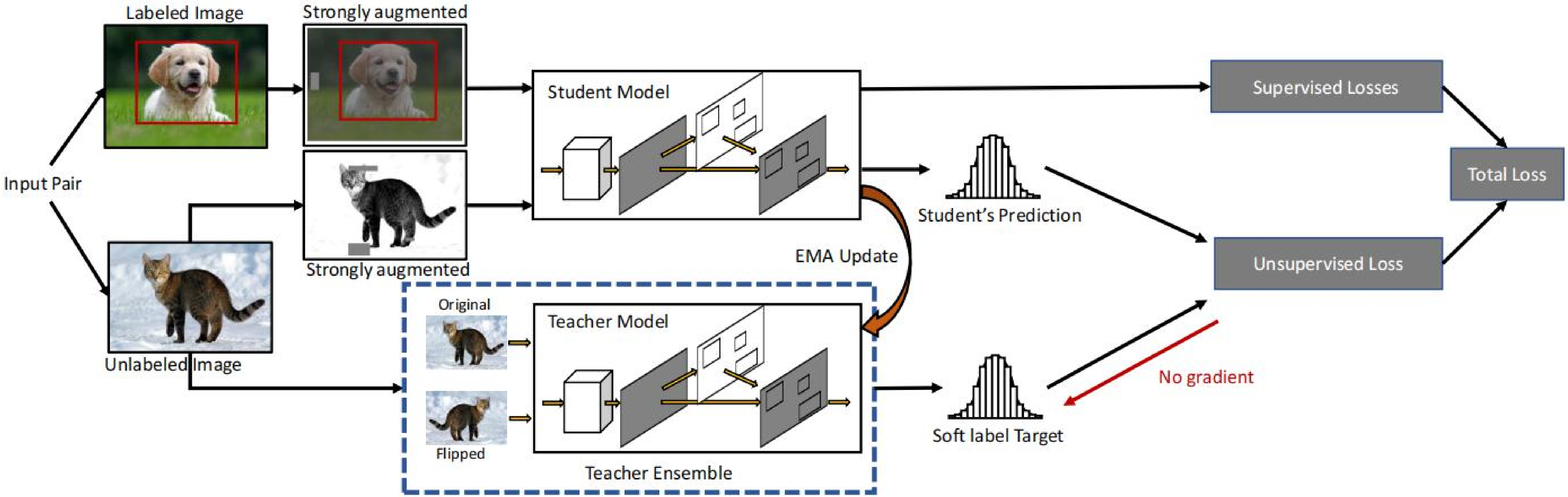
CVPR 2021, Adaptive Consistency Regularization for Semi-Supervised Transfer Learning

Semi-Supervised-Pseudo Label



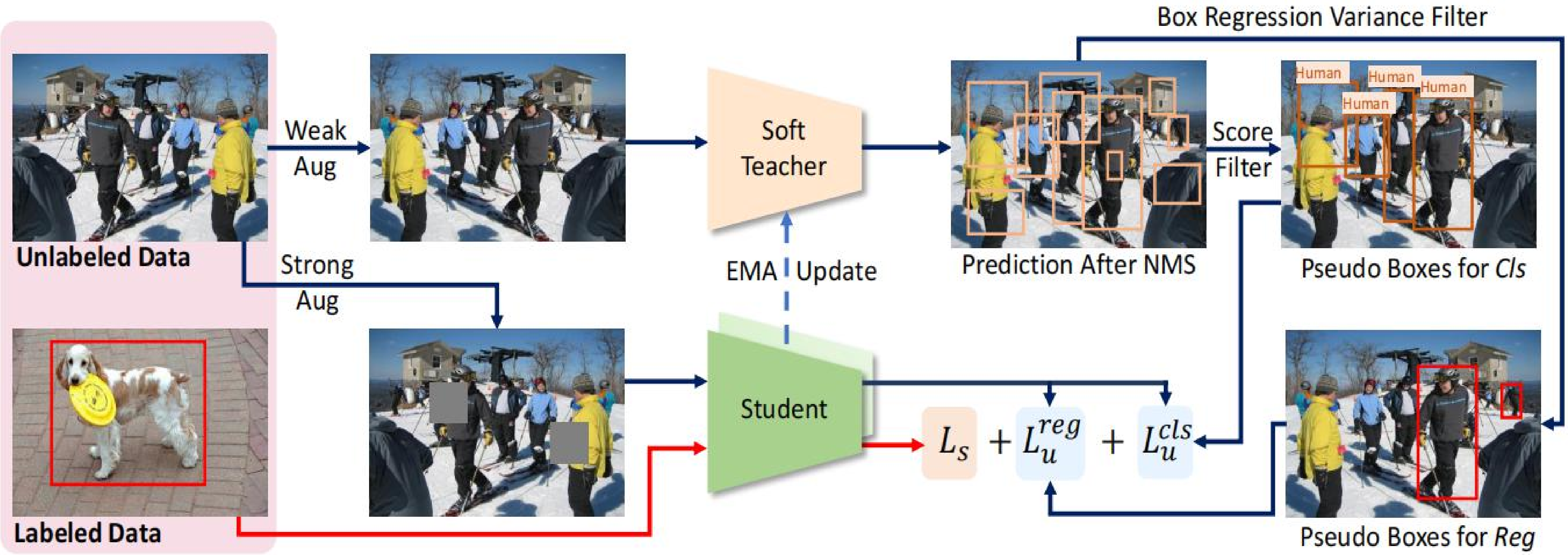
2020, A Simple Semi-Supervised Learning Framework for Object Detection

Pseudo Label



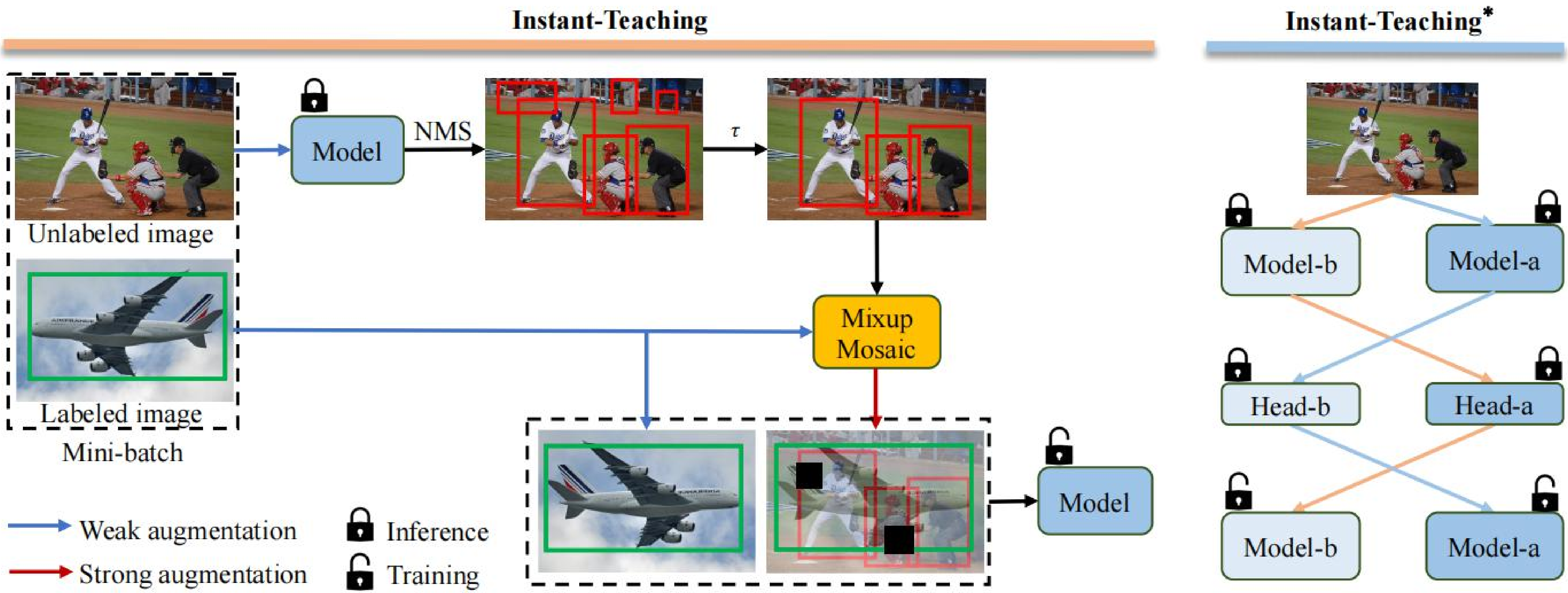
CVPR 2021, Humble Teachers Teach Better Students for Semi-Supervised Object Detection

Pseudo Label



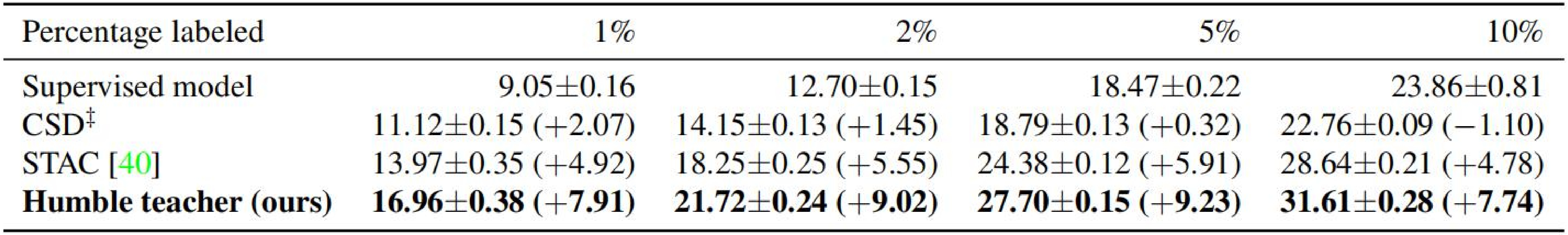
ICCV 2021, End-to-End Semi-Supervised Object Detection with Soft Teacher

Pseudo Label

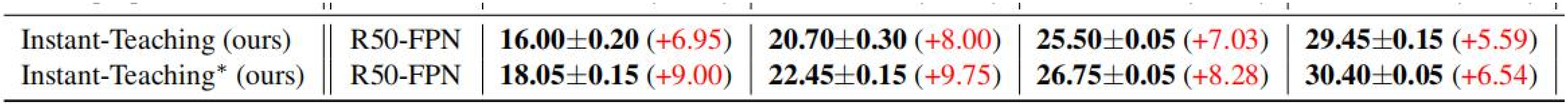


CVPR 2021, Instant-Teaching: An End-to-End Semi-Supervised Object Detection Framework

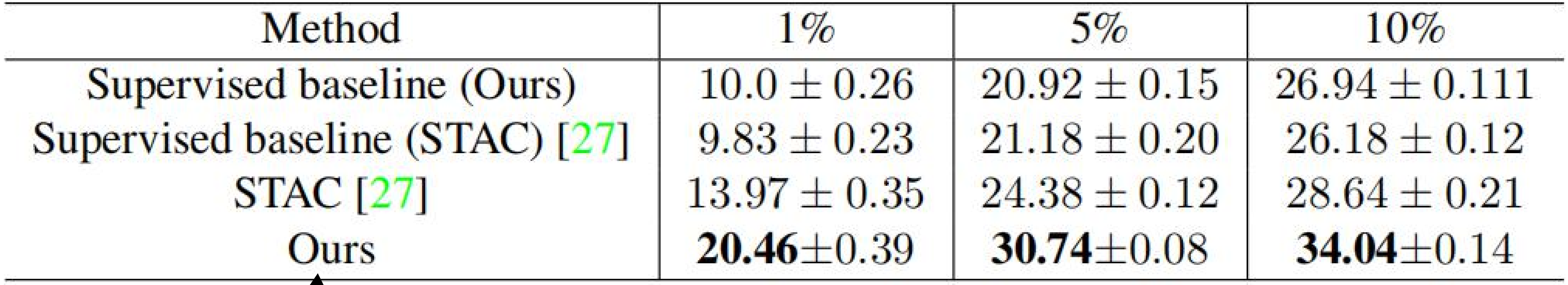
Results



Percentage labeled	1%	2%	5%	10%
Supervised model	9.05±0.16	12.70±0.15	18.47±0.22	23.86±0.81
CSD[‡]	11.12±0.15 (+2.07)	14.15±0.13 (+1.45)	18.79±0.13 (+0.32)	22.76±0.09 (−1.10)
STAC [40]	13.97±0.35 (+4.92)	18.25±0.25 (+5.55)	24.38±0.12 (+5.91)	28.64±0.21 (+4.78)
Humble teacher (ours)	**16.96±0.38** (+7.91)	**21.72±0.24** (+9.02)	**27.70±0.15** (+9.23)	**31.61±0.28** (+7.74)



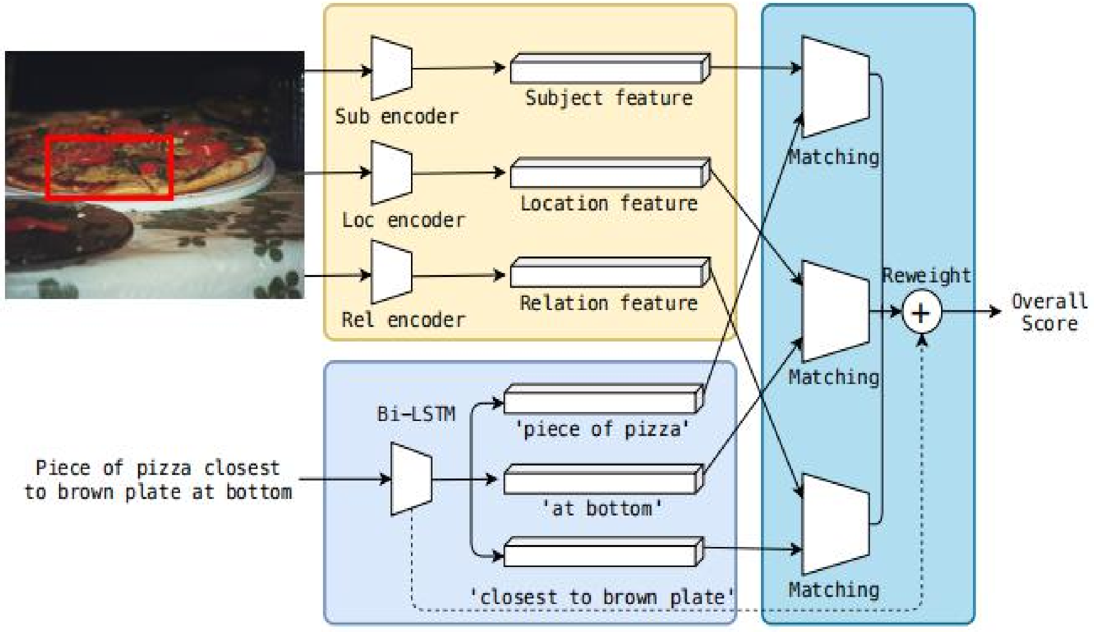
Instant-Teaching (ours)	R50-FPN	**16.00±0.20** (+6.95)	**20.70±0.30** (+8.00)	**25.50±0.05** (+7.03)	**29.45±0.15** (+5.59)
Instant-Teaching* (ours)	R50-FPN	**18.05±0.15** (+9.00)	**22.45±0.15** (+9.75)	**26.75±0.05** (+8.28)	**30.40±0.05** (+6.54)

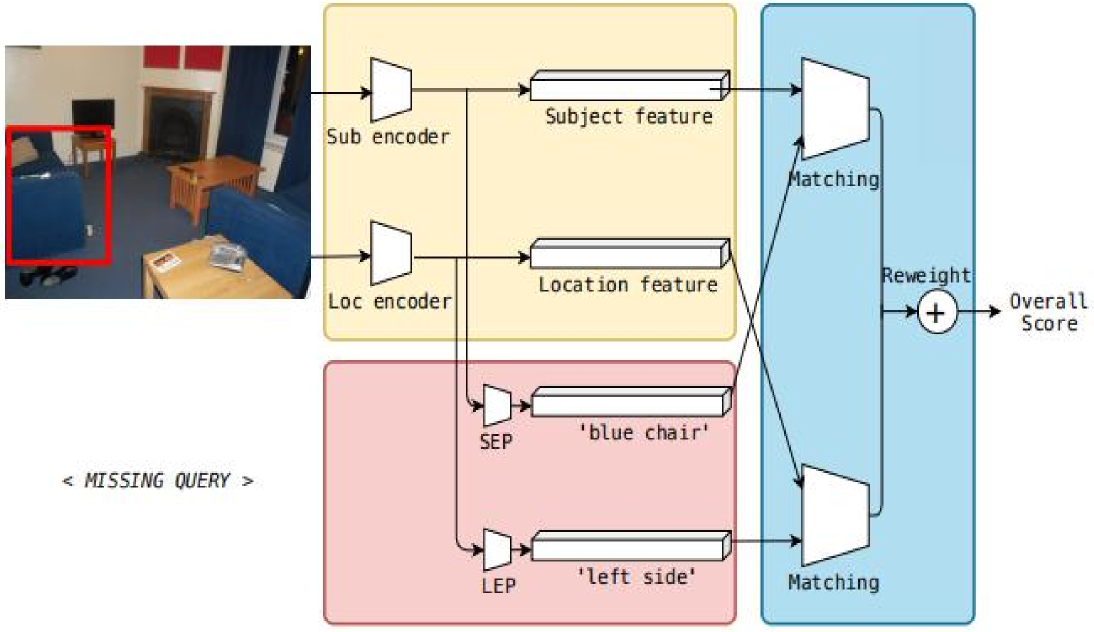


Method	1%	5%	10%
Supervised baseline (Ours)	10.0 ± 0.26	20.92 ± 0.15	26.94 ± 0.111
Supervised baseline (STAC) [27]	9.83 ± 0.23	21.18 ± 0.20	26.18 ± 0.12
STAC [27]	13.97 ± 0.35	24.38 ± 0.12	28.64 ± 0.21
Ours	**20.46**±0.39	**30.74**±0.08	**34.04**±0.14

↑
Soft Teacher

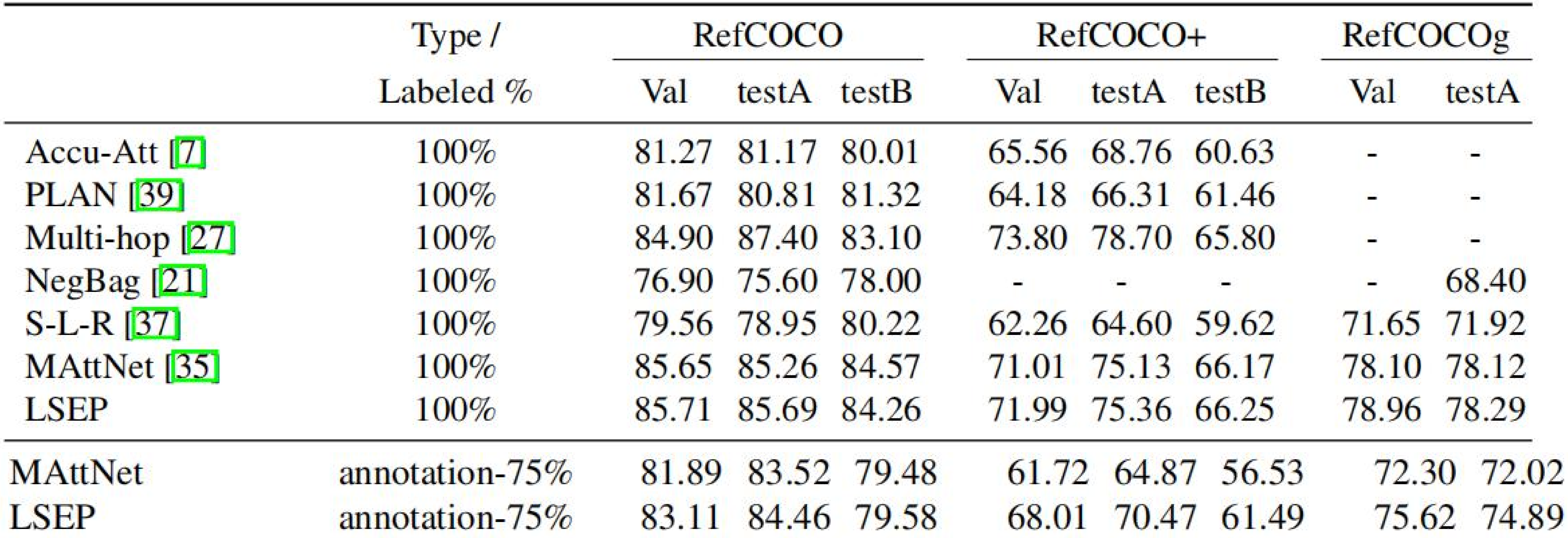
Pseudo Label



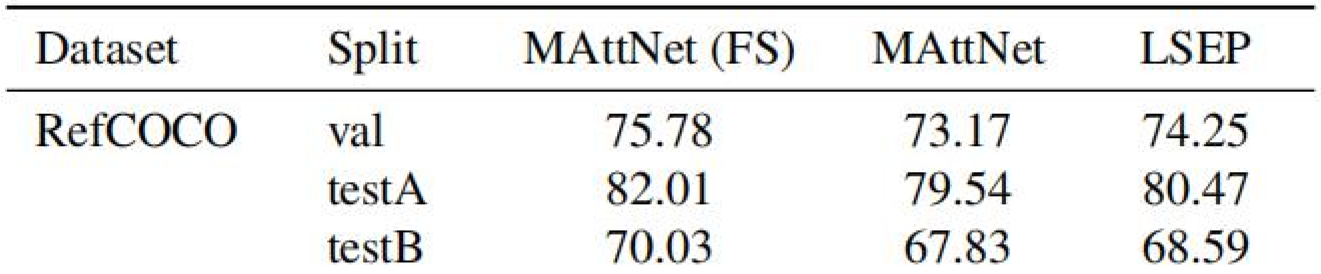


WACV 21, Utilizing Every Image Object for Semi-supervised Phrase Grounding

Pseudo Label



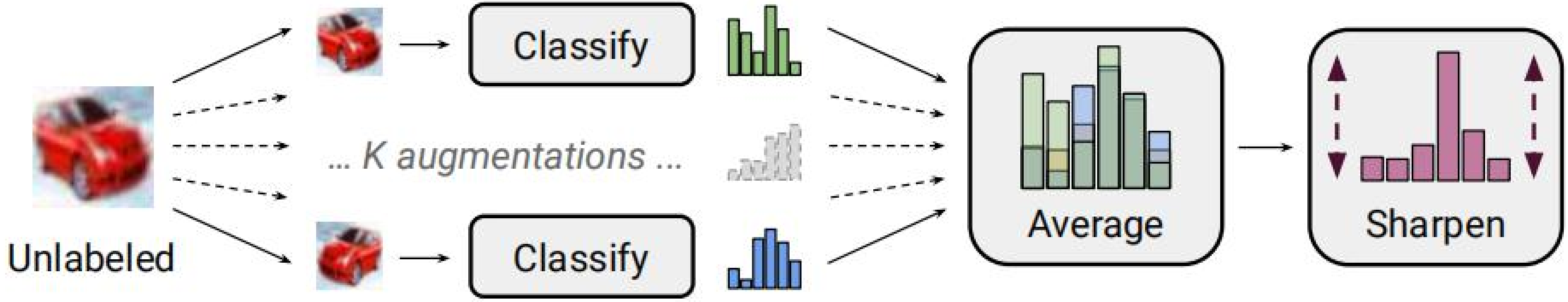
	Type /	RefCOCO			RefCOCO+			RefCOCOg	
	Labeled %	Val	testA	testB	Val	testA	testB	Val	testA
Accu-Att [7]	100%	81.27	81.17	80.01	65.56	68.76	60.63	-	-
PLAN [39]	100%	81.67	80.81	81.32	64.18	66.31	61.46	-	-
Multi-hop [27]	100%	84.90	87.40	83.10	73.80	78.70	65.80	-	-
NegBag [21]	100%	76.90	75.60	78.00	-	-	-	-	68.40
S-L-R [37]	100%	79.56	78.95	80.22	62.26	64.60	59.62	71.65	71.92
MAttNet [35]	100%	85.65	85.26	84.57	71.01	75.13	66.17	78.10	78.12
LSEP	100%	85.71	85.69	84.26	71.99	75.36	66.25	78.96	78.29
MAttNet	annotation-75%	81.89	83.52	79.48	61.72	64.87	56.53	72.30	72.02
LSEP	annotation-75%	83.11	84.46	79.58	68.01	70.47	61.49	75.62	74.89



Dataset	Split	MAttNet (FS)	MAttNet	LSEP
RefCOCO	val	75.78	73.17	74.25
	testA	82.01	79.54	80.47
	testB	70.03	67.83	68.59

WACV 21, Utilizing Every Image Object for Semi-supervised Phrase Grounding

Semi-Supervised-Match



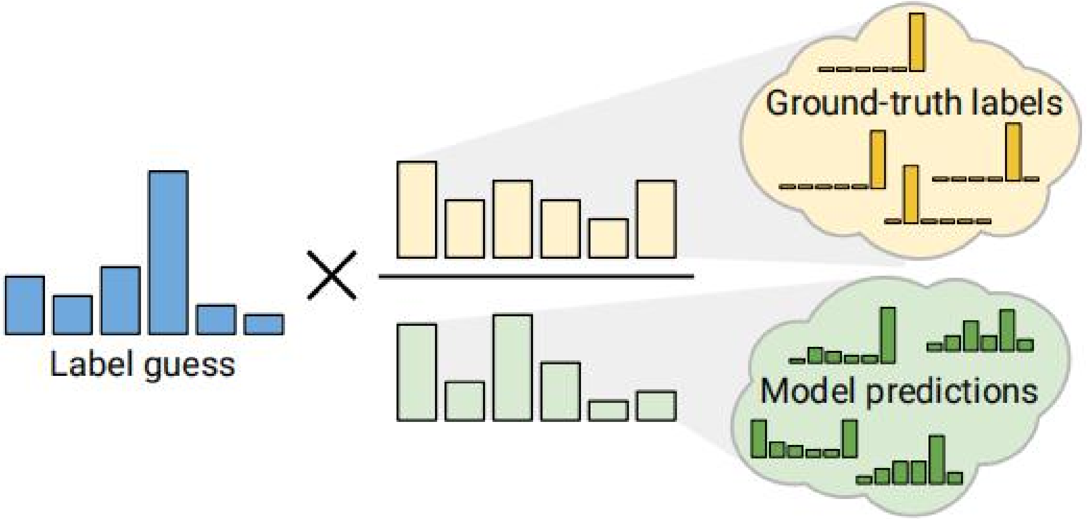
Mixup

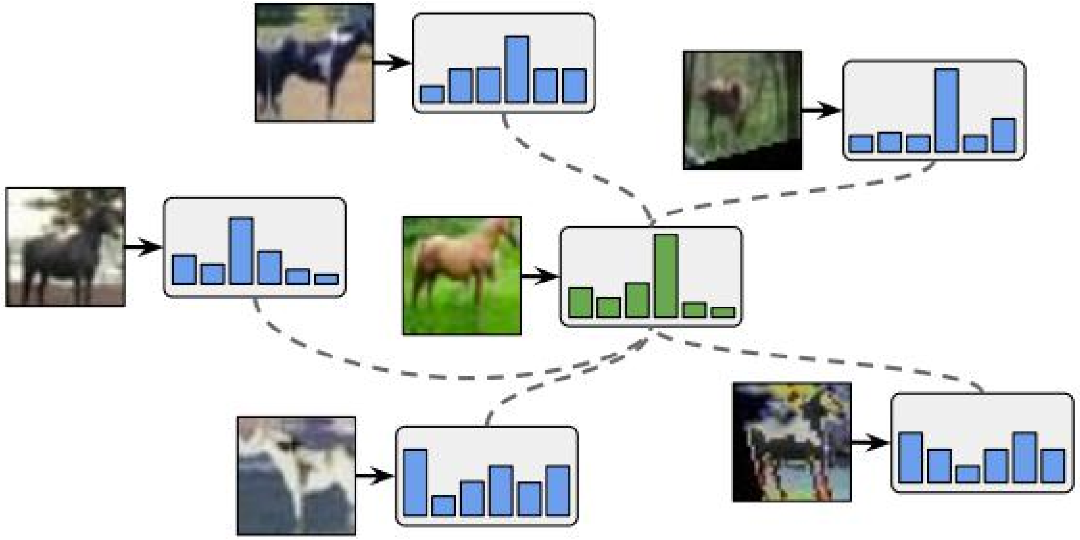
$$\tilde{x} = \lambda x_i + (1-\lambda)x_j$$
$$\tilde{y} = \lambda y_i + (1-\lambda)y_j$$

$$L_X = \frac{1}{|X'|}\sum_{x,p\in X'} H(p, p_{\text{model}}(y|x;\theta))$$
$$L_U = \frac{1}{L|U'|}\sum_{u,q\in U'} \|q, p_{\text{model}}(y|u;\theta)\|_2^2$$
$$L = L_x + \lambda_U L_U.$$

NeurIPS 2019 , MixMatch: A Holistic Approach to Semi-Supervised Learning

Match



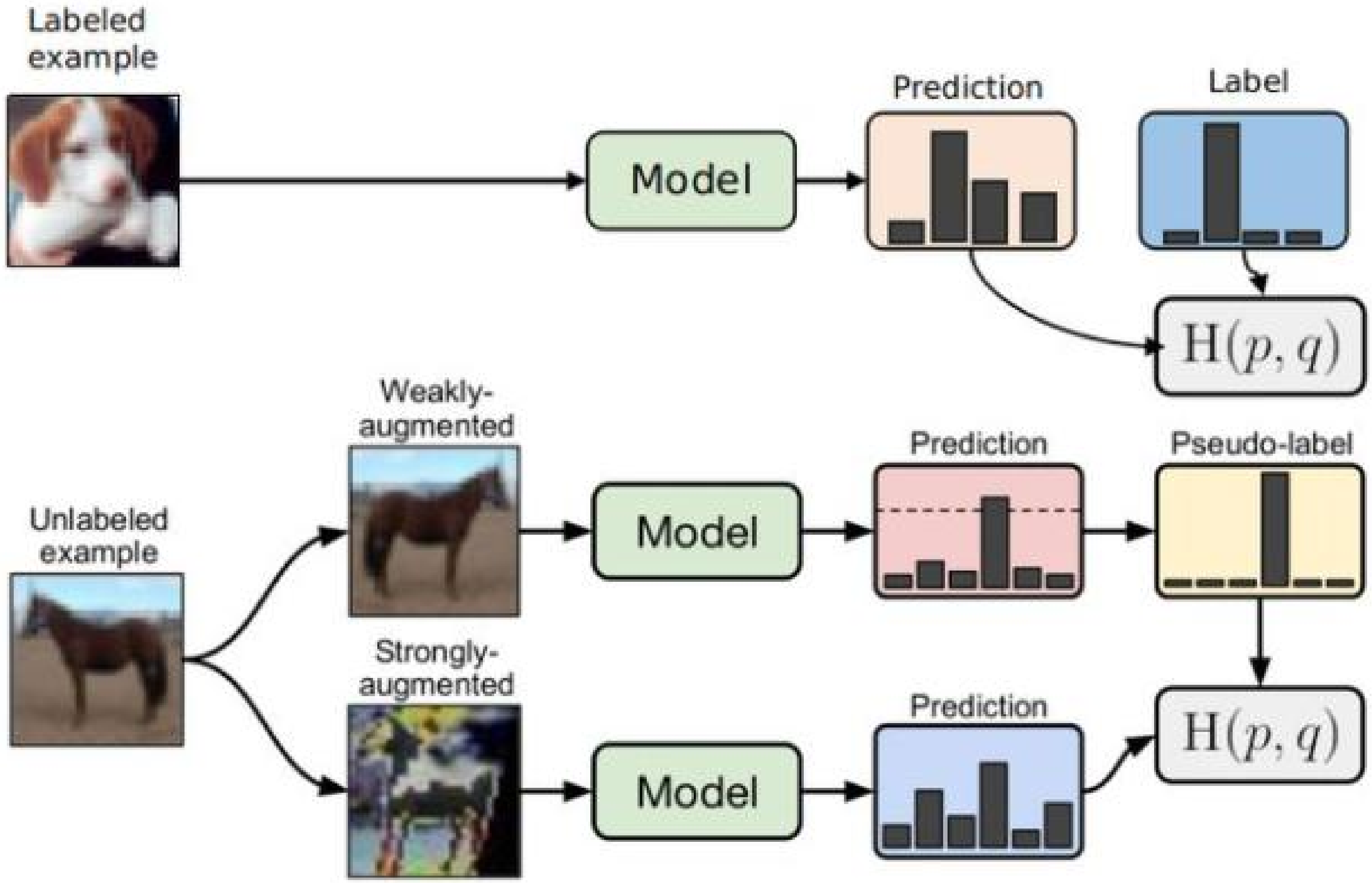


$$\tilde{q} = \text{Normalize}\left(q \times \frac{p(y)}{\tilde{p}(y)}\right).$$

- q: 对当前无标签数据的标签猜测
- $\tilde{p}(y)$:平均版本的无标签猜测
- p(y):有标签数据的标签分布

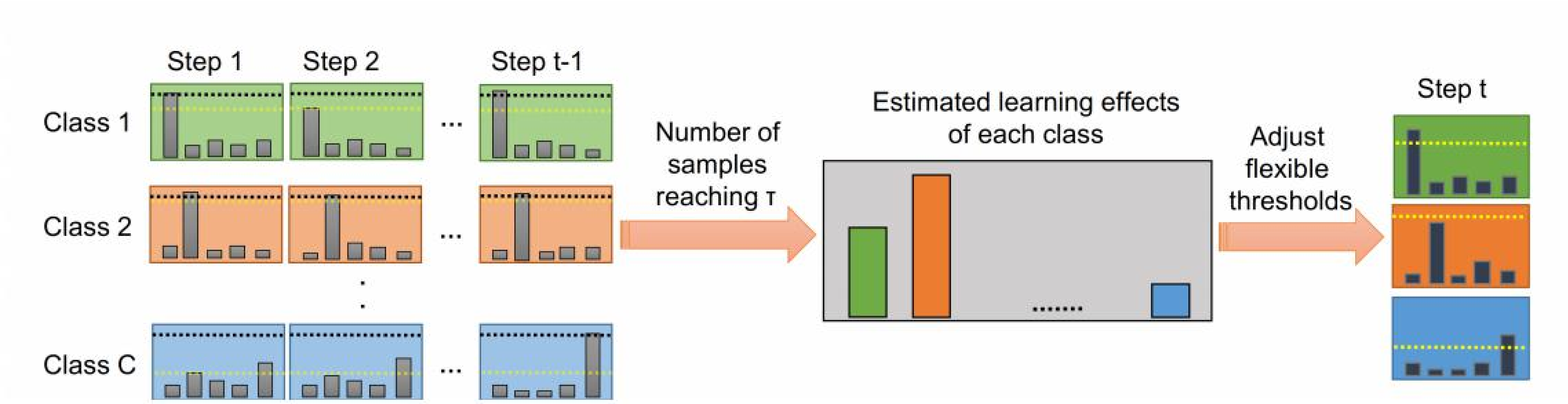
ICLR 2020 ,ReMixMatch: Semi-Supervised Learning with Distribution Alignment and Augmentation Anchoring

Match



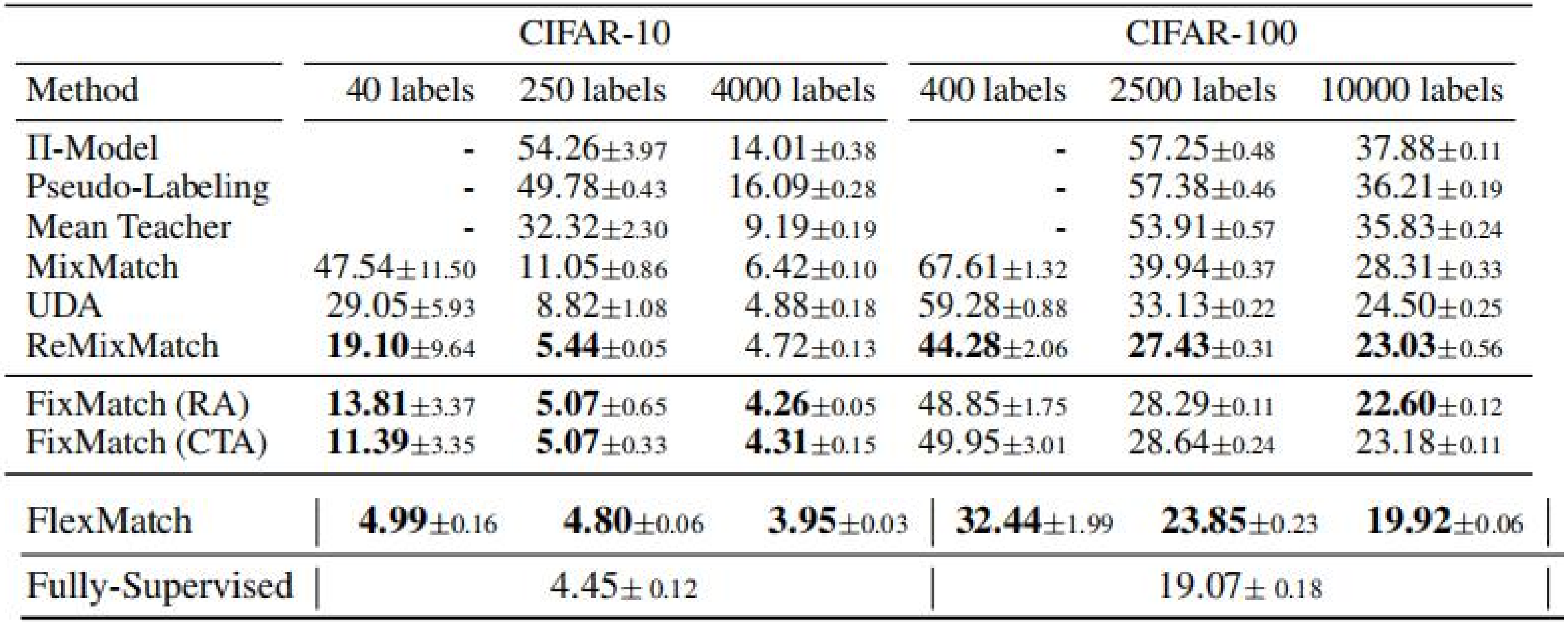
NeurIPS 2020 , FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence

Match


Step 1
Step 2
Step t-1
Class 1
Class 2
Class C
...
Number of samples reaching τ
Estimated learning effects of each class
....
Adjust flexible thresholds
Step t

NeurIPS 2021 , FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling

Results



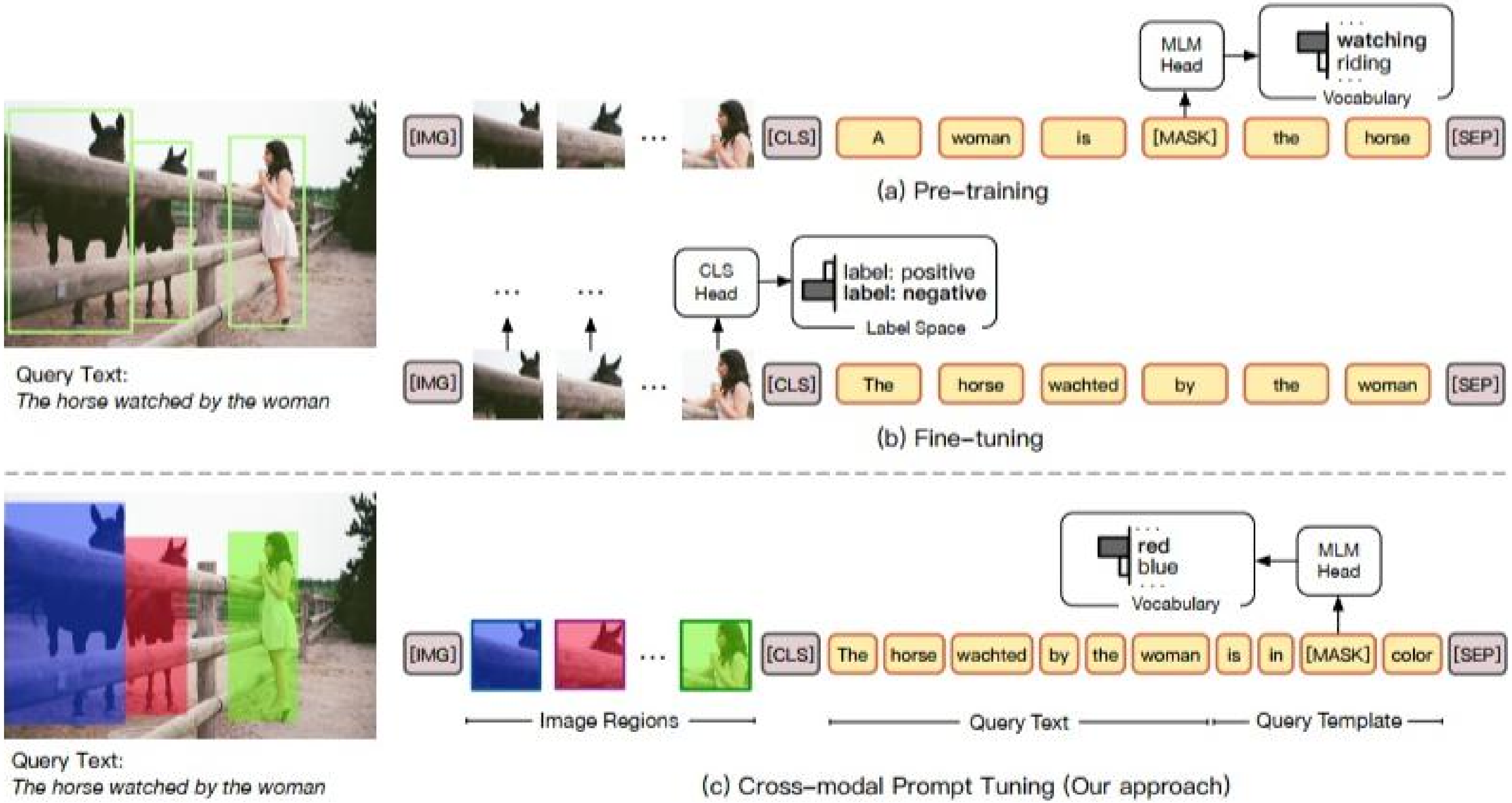
	CIFAR-10			CIFAR-100		
Method	40 labels	250 labels	4000 labels	400 labels	2500 labels	10000 labels
Π-Model	-	54.26±3.97	14.01±0.38	-	57.25±0.48	37.88±0.11
Pseudo-Labeling	-	49.78±0.43	16.09±0.28	-	57.38±0.46	36.21±0.19
Mean Teacher	-	32.32±2.30	9.19±0.19	-	53.91±0.57	35.83±0.24
MixMatch	47.54±11.50	11.05±0.86	6.42±0.10	67.61±1.32	39.94±0.37	28.31±0.33
UDA	29.05±5.93	8.82±1.08	4.88±0.18	59.28±0.88	33.13±0.22	24.50±0.25
ReMixMatch	**19.10**±9.64	**5.44**±0.05	4.72±0.13	**44.28**±2.06	**27.43**±0.31	**23.03**±0.56
FixMatch (RA)	**13.81**±3.37	**5.07**±0.65	**4.26**±0.05	48.85±1.75	28.29±0.11	**22.60**±0.12
FixMatch (CTA)	**11.39**±3.35	**5.07**±0.33	**4.31**±0.15	49.95±3.01	28.64±0.24	23.18±0.11
FlexMatch	**4.99**±0.16	**4.80**±0.06	**3.95**±0.03	**32.44**±1.99	**23.85**±0.23	**19.92**±0.06
Fully-Supervised	4.45± 0.12			19.07± 0.18		

NeurIPS 2021 , FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling

Further Work

- Semi-Supervised visual grounding
  - methods
    - one stage
    - two stage
  - strategy
    - teacher
    - data augmentation
- baseline model
  - mattnet
  - lesp

Further Work



(a) Pre-training

(b) Fine-tuning

(c) Cross-modal Prompt Tuning (Our approach)