Dual Mean-Teacher: An Unbiased Semi-Supervised Framework for Audio-Visual Source Localization

Yuxin Guo, Shijie Ma, Hu Su, Zhiqing Wang, Yuhao Zhao, Wei Zou, Siyang Sun, Yun Zheng
School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS),
Institute of Automation of Chinese Academy of Sciences, Beijing, China
DAMO Academy, Alibaba Group
{guoyuxin2021, wei.zou}@ia.ac.cn

What is Audio-Visual Source Localization?

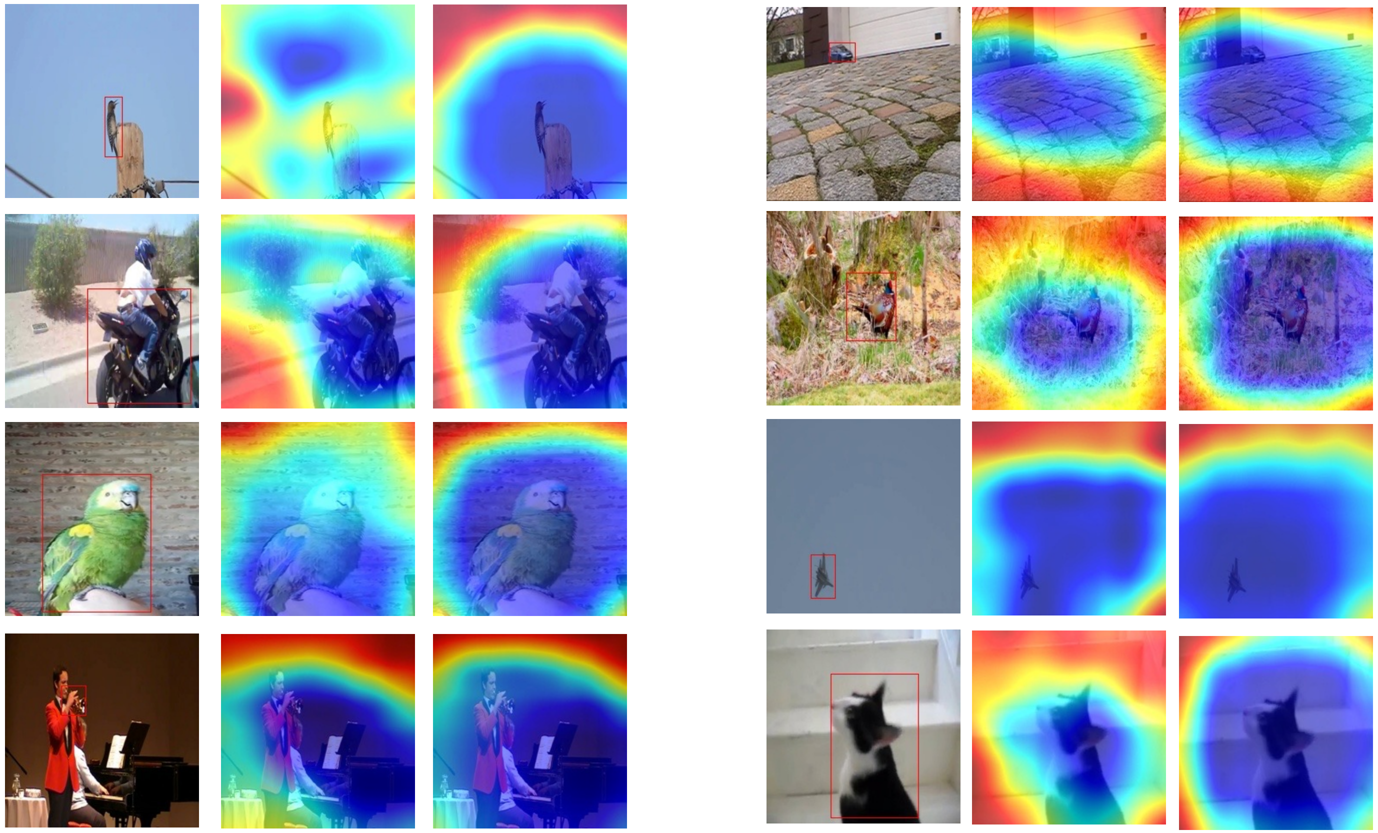
NEURAL INFORMATION PROCESSING SYSTEMS

✓ Audio-Visual Source Localization (AVSL) aims to locate sounding objects within video frames given the paired audio clips.

Existing Problems:

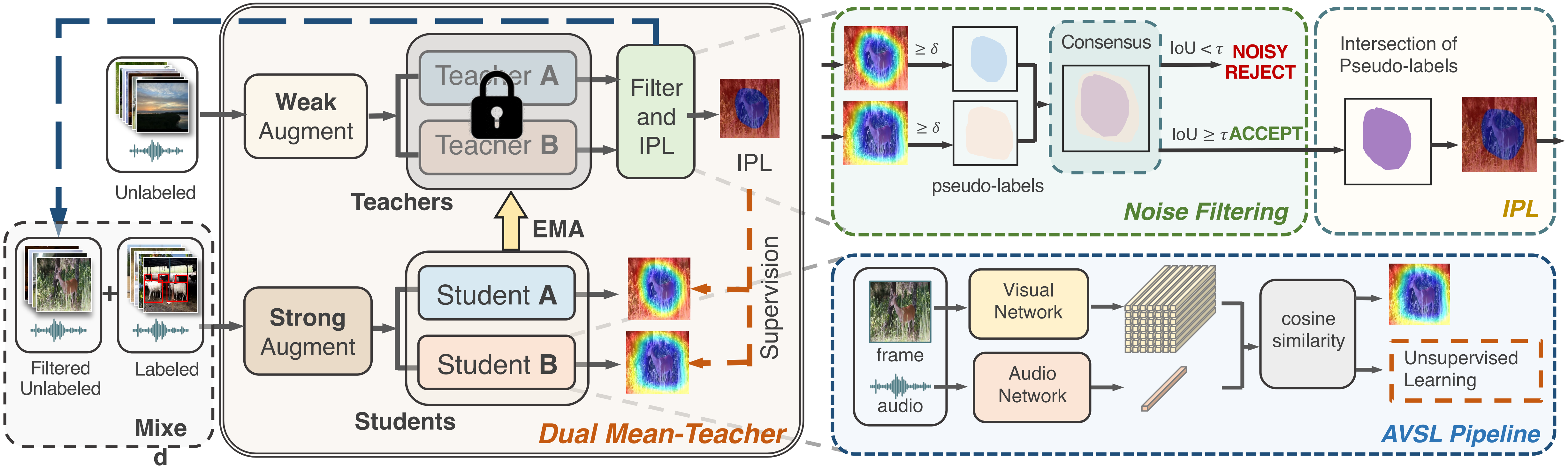
- small object
- blurry boundaries
- too much foreground
- inablity of specific site

Ground Truth | EZVSL | SLAVC | Ground Truth | EZVSL | SLAVC



Motivation

- How to eliminate the influence of confirmation bias?
- How to effectively utilize the abundance of unlabeled audio-visual pairs alongside the limited yet valuable annotated data.



DEMO VIDEO

Please turn on the speaker.

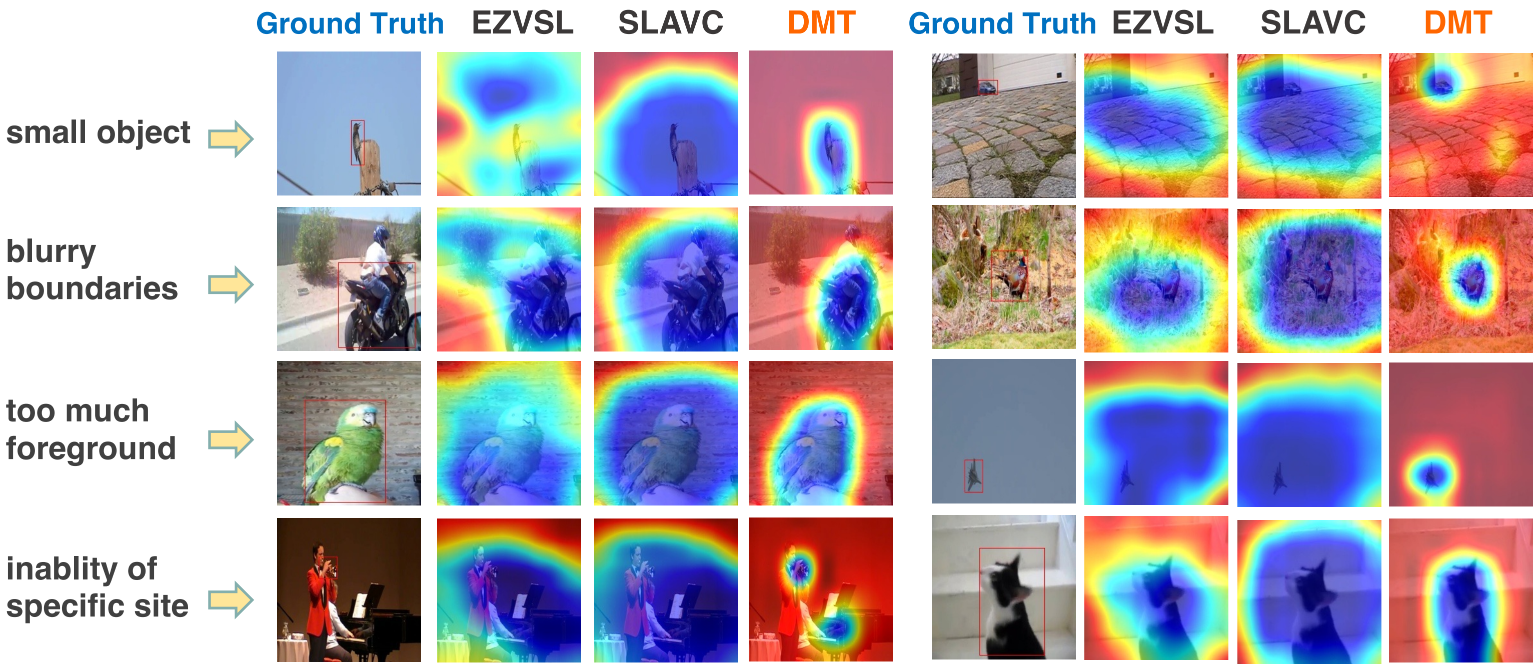
11428

Dual Mean-Teacher: An Unbiased Semi-Supervised Framework for Audio-Visual Source Localization

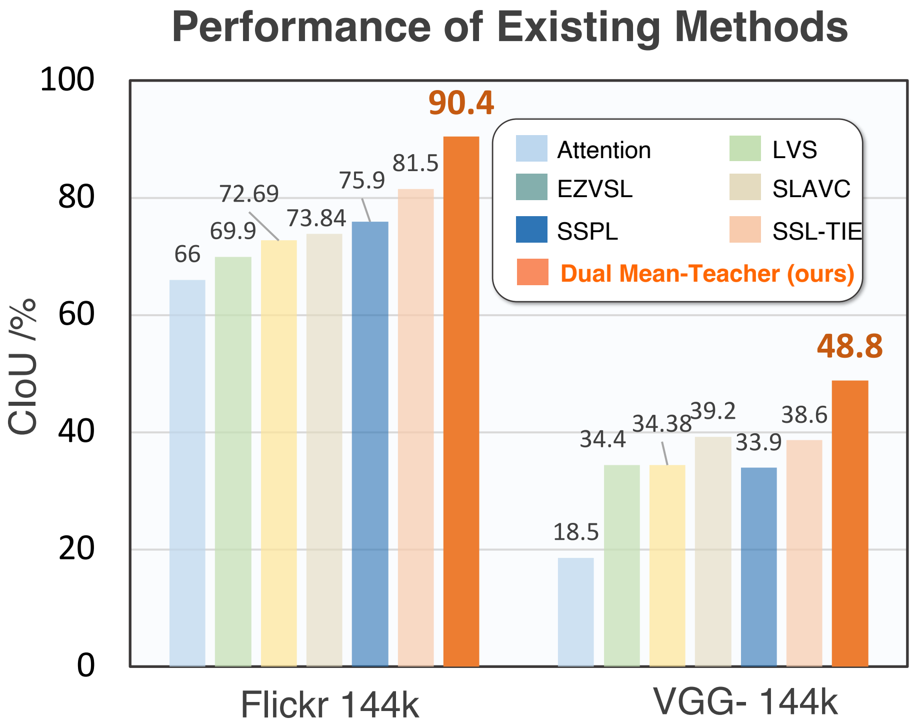
Performance

NEURAL INFORMATION PROCESSING SYSTEMS

Existing Problems can be solved well:



Performance:



Performance

NEURAL INFORMATION PROCESSING SYSTEMS

Table 1: Comparison results on Flickr-SoundNet. Models are trained on Flickr 10k and 144k. † indicates our reproduced results, others are borrowed from original papers. Attention10k-SSL is of 2k labeled data supervision. We report the proposed DMT results from both stages as `stage-2(stage-1)`. $|\mathcal{D}_l|$ denotes the number of labeled data.

Methods	Flickr 10k		Flickr 144k	
	CIoU	AUC	CIoU	AUC
Attention10k [13, 14]	43.60	44.90	66.00	55.80
CoarsetoFine [31]	52.20	49.60	–	–
DMC [2]	–	–	67.10	56.80
LVS [27]	58.20	52.50	69.90	57.30
EZVSL [8]	62.65	54.89	72.69	58.70
SLAVC† [9]	66.80	56.30	73.84	58.98
SSPL [28]	74.30	58.70	75.90	61.00
SSL-TIE† [29]	75.50	58.80	81.50	61.10
Attention10k-SSL [13, 14]	82.40	61.40	83.80	61.72
Ours ($\lvert\mathcal{D}_l\rvert = 256$)	87.20 (84.40)	65.77 (59.60)	87.60 (84.40)	66.28 (59.60)
Ours ($\lvert\mathcal{D}_l\rvert = 2k$)	87.80 (85.60)	66.20 (63.18)	88.20 (85.60)	66.63 (63.18)
Ours ($\lvert\mathcal{D}_l\rvert = 4k$)	**88.80** (86.20)	**67.81** (65.56)	**90.40** (86.20)	**69.36** (65.56)

Table 2: Comparison results on VGG-ss. Models are trained on VGG-Sound 10k and 144k.

Methods	VGG-Sound 10k		VGG-Sound 144k	
	CIoU	AUC	CIoU	AUC
Attention10k [13, 14]	16.00	28.30	18.50	30.20
LVS [27]	27.70	34.90	34.40	38.20
EZVSL [8]	32.30	33.68	34.38	37.70
SLAVC† [9]	37.80	39.48	39.20	39.46
SSPL [28]	31.40	36.90	33.90	38.00
SSL-TIE† [29]	36.80	37.21	38.60	39.60
Attention10k-SSL† [13, 14]	38.60	38.26	39.20	38.52
Ours ($\lvert\mathcal{D}_l\rvert = 256$)	41.20 (39.40)	40.68 (38.70)	43.60 (39.40)	41.88 (38.70)
Ours ($\lvert\mathcal{D}_l\rvert = 2k$)	43.20 (42.60)	40.82 (40.75)	45.60 (42.60)	43.24 (40.75)
Ours ($\lvert\mathcal{D}_l\rvert = 4k$)	**46.80** (43.80)	**43.18** (41.63)	**48.80** (43.80)	**45.76** (41.63)

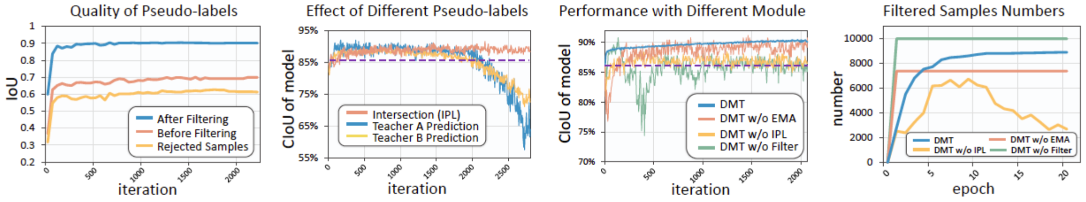
Table 4: Extension results of DMT with various audio backbones, with 'R', 'V' and 'S' denoting ResNet, VGGish and SoundNet.

Methods	Backbones	CIoU↑	AUC↑	MSE↓
EZVSL w/o DMT	R	62.65	54.89	0.428
EZVSL w/ DMT	R+V	85.30	65.80	0.312
EZVSL w/ DMT	R+S	85.95	66.12	0.298
EZVSL w/ DMT	V+S	**87.20**	**67.74**	**0.256**
SLAVC w/o DMT	R	66.80	56.30	0.386
SLAVC w/ DMT	R+V	86.10	66.24	0.288
SLAVC w/ DMT	R+S	86.30	66.58	0.283
SLAVC w/ DMT	V+S	**88.80**	**68.69**	**0.247**

Figure 3: Performance on music-domain.

Table 6: Performance on various labeled ratios % and multiple × on Flickr 144k.

Labeled ratio %	CIoU	AUC
0.5% (200/40k)	84.80	63.58
1% (400/40k)	86.20	65.16
2% (800/40k)	87.20	65.94
5% (2k/40k)	87.60	67.44
10% (4k/40k)	**88.40**	**68.12**
Multiple ×	**CIoU**	**AUC**
2.5× (4k/10k)	88.00	67.80
5× (4k/20k)	88.20	67.91
10× (4k/40k)	88.40	68.12
20× (4k/80k)	89.20	68.44
40× (4k/200k)	**91.20**	**71.36**



(a) Filter. (b) IPL. (c) Different modules. (d) Filtered number.

Figure 5: The effect of each component (Noise Filtering, IPL and EMA) in DMT to suppress confirmation bias, together with the number of filtered samples for pseudo labeling depicted in (d).