JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Learning Audio-Visual Source Localization via False Negative Aware Contrastive Learning

Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, Nick Barnes

Australian National University
北京航空航天大学 BEIHANG UNIVERSITY 1952
上海人工智能实验室 Shanghai Artificial Intelligence Laboratory
UNIVERSITY OF OXFORD
oppo

Task definition: Audio-visual source localization

Problem: False negative issue

Optimizing objective of audio-visual contrastive learning in the existing methods:

$$\mathcal{L}_{\text{contrast\_i}} = -\log \frac{\exp\left[\frac{1}{\tau}\text{sim}(Z_i^a, Z_i^v)\right]}{\sum_j^b \exp\left[\frac{1}{\tau}\text{sim}(Z_i^a, Z_j^v)\right]} - \log \frac{\exp\left[\frac{1}{\tau}\text{sim}(Z_i^v, Z_i^a)\right]}{\sum_j^b \exp\left[\frac{1}{\tau}\text{sim}(Z_i^v, Z_j^a)\right]},$$

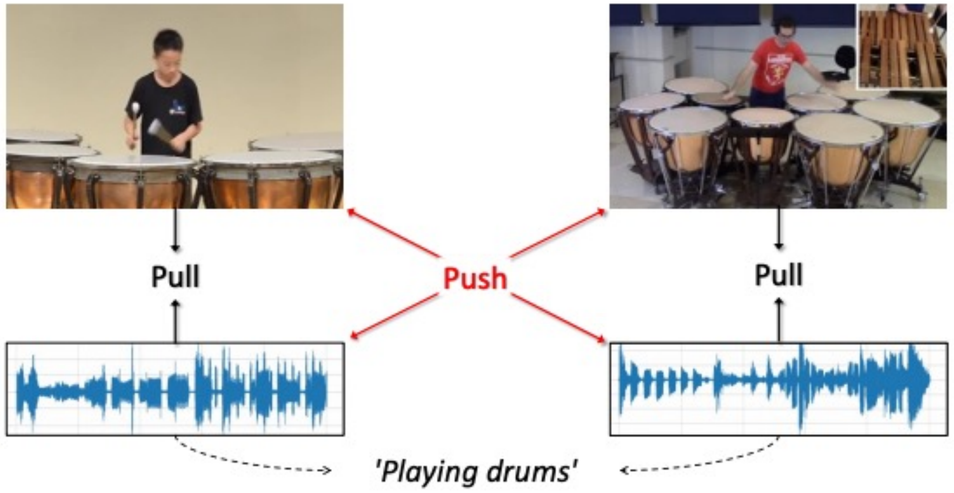


Figure 1. **False negative in audio-visual contrastive learning.** Audio-visual pairs with similar contents are falsely considered as negative samples to each other and pushed apart in the shared latent space, which we find would affect the model performance.

Pilot experiments: Impact of false negatives

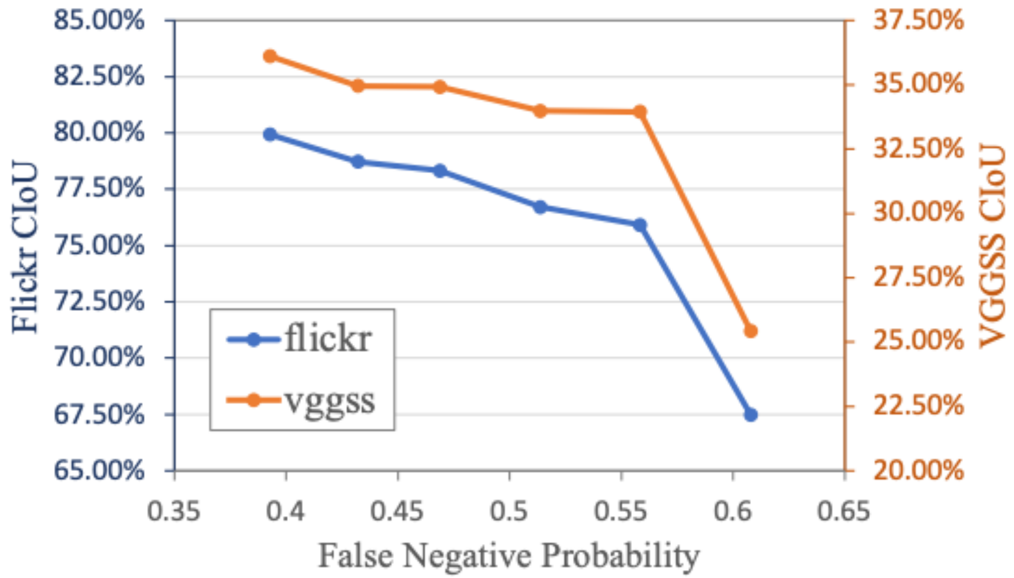
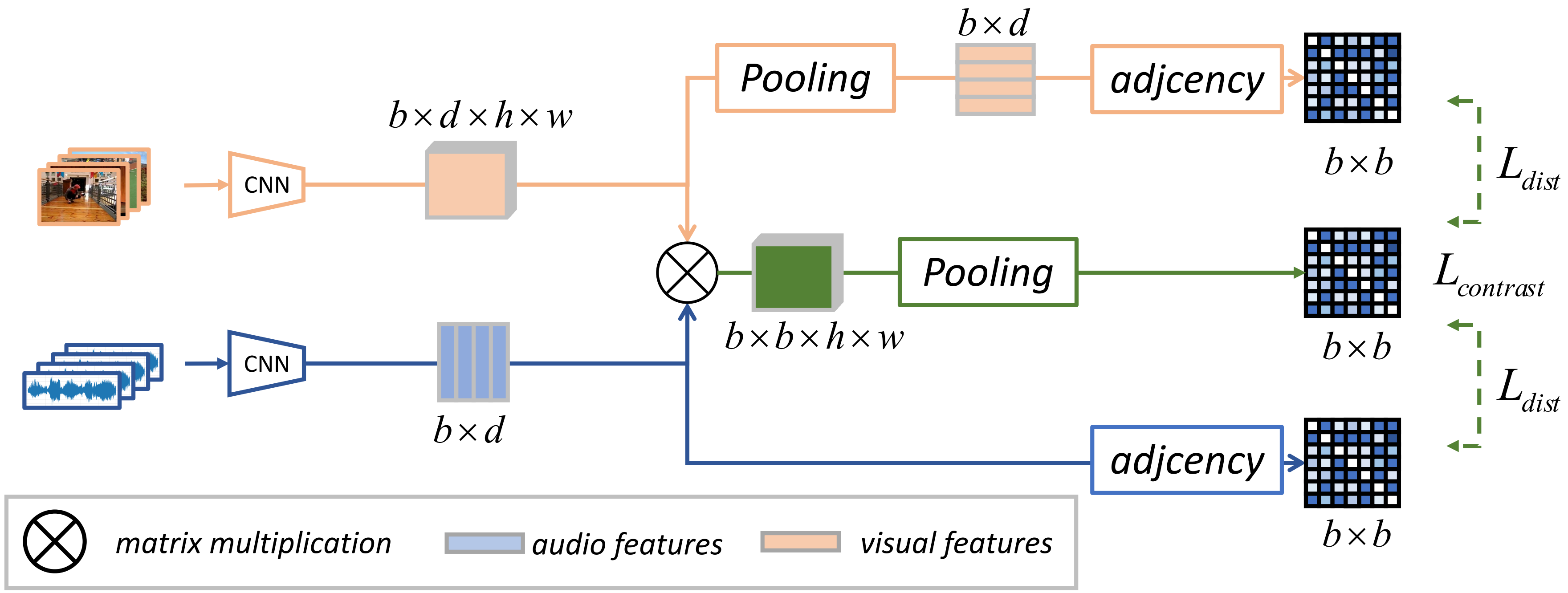


Figure 2. **Impact of false negatives on audio-visual representation learning**. We adopt Consensus Intersection over Union (CIoU) [28] as an evaluation metric (higher is better) and report results on FlickrSoundNet [6] and VGG-SS [9] test sets, depicted by blue and brown, respectively. An obvious performance decline is observed as the proportion of false negative samples increases.

Our solutions: False Negative Aware Contrastive learning

1. FNS: False Negatives Suppression

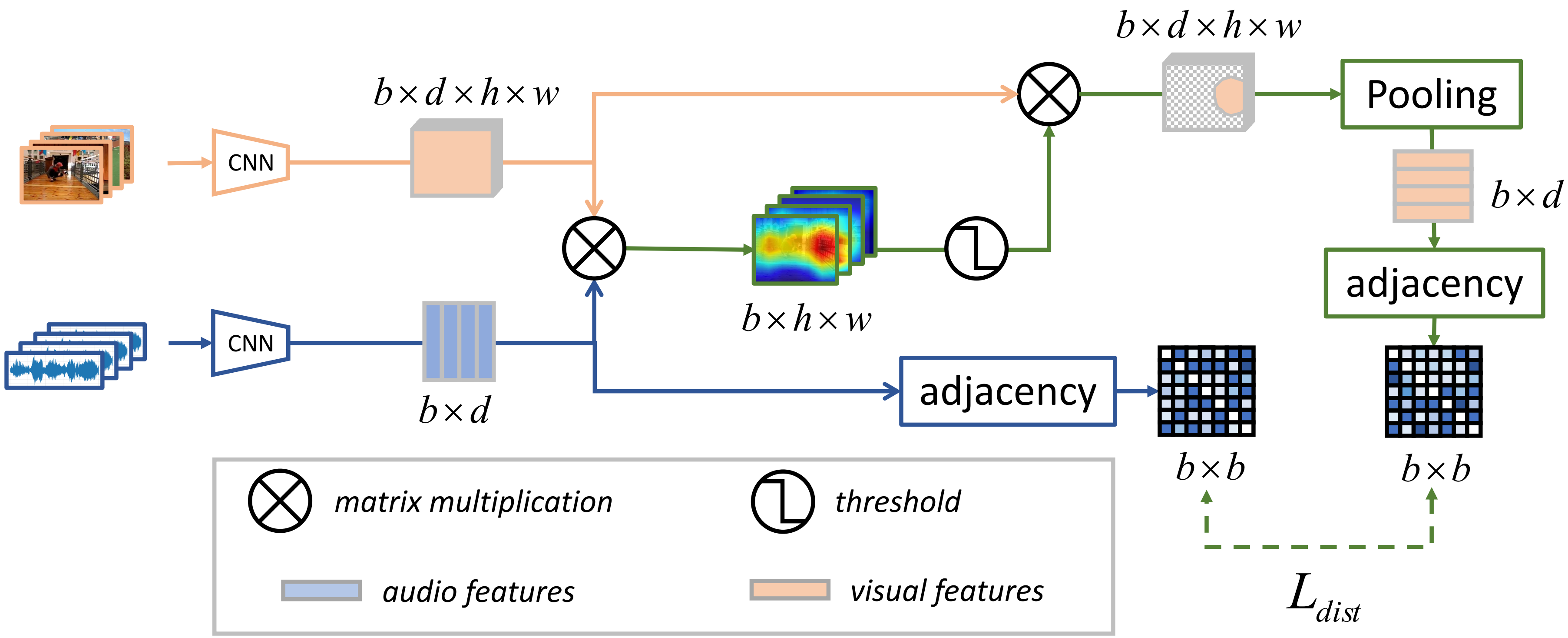


$$\mathcal{L}_{\text{FNS\_1}} = \frac{1}{b}\sum_{j}^{b} \mathcal{L}_{dist}(\text{sim}(Z_i^a, Z_j^v), \text{sim}(Z_i^a, Z_j^a)) \quad (2)$$

$$\mathcal{L}_{\text{FNS\_2}} = \frac{1}{b}\sum_{j}^{b} \mathcal{L}_{dist}(\text{sim}(Z_i^a, Z_j^v), \text{sim}(Z_i^v, Z_j^v)) \quad (3)$$

Our solutions: False Negative Aware Contrastive learning

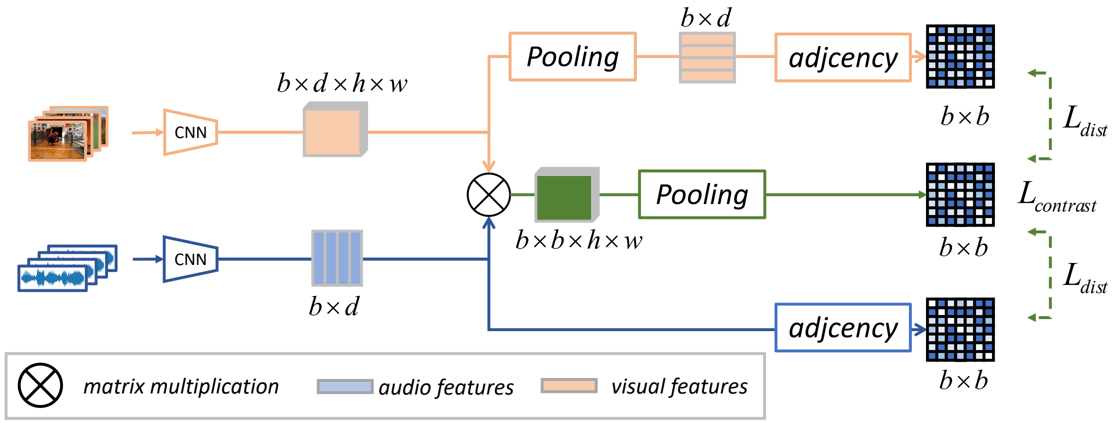
2. TNE: True Negatives Enhancement.

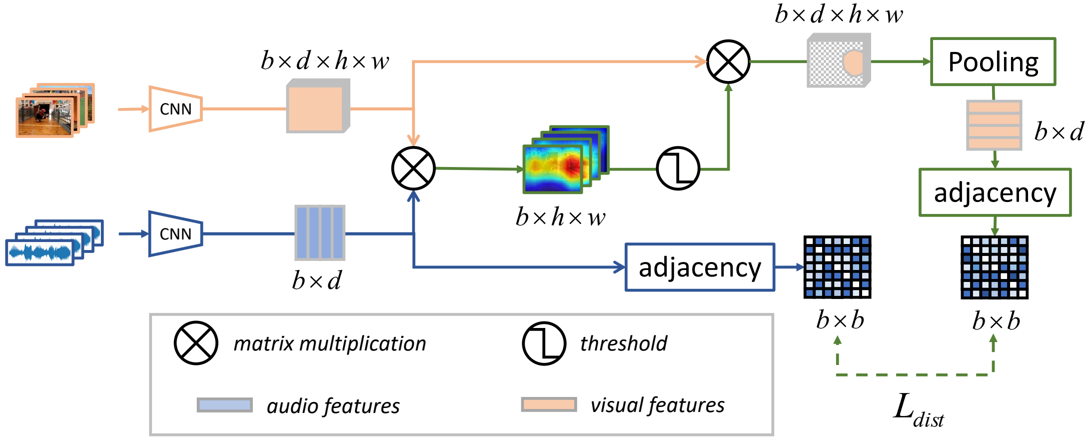


$$\mathcal{L}_{\text{TNE}} = \frac{1}{b}\sum_{j}^{b} \mathcal{L}_{dist}(\text{sim}(Z_i^a, Z_j^a), \text{sim}(Z_i^s, Z_j^s))$$

Our solutions: False Negative Aware Contrastive learning

1. FNS: False Negatives Suppression
2. TNE: True negative enhancement





$$\mathcal{L}_i = \mathcal{L}_{\text{contrast\_i}} + \alpha\mathcal{L}_{\text{FNS\_1}} + \beta\mathcal{L}_{\text{FNS\_2}} + \gamma\mathcal{L}_{\text{TNE}}$$

Performances on Flickr, VGG-SS, Heard 110, Unhear 110 and AVSbench:

Table 1. Quantitative results of the model trained with Flickr 10k and 144k. Note that 'EZ-VSL + OGL' corresponds to the main results reported in [21]. 'EZ-VSL' indicates our reproduced results without OGL, which are not reported in [21]. We reproduce the results with the trained weights and code provided by [21].

Train set	Method	Flickr CIoU(%)	Flickr AUC(%)	VGG-SS CIoU(%)	VGG-SS AUC(%)
Flickr 10k	Attention10k [28]	43.60	44.90	-	-
	CoursetoFine [26]	52.20	49.60	-	-
	AVObject [1]	54.60	50.40	-	-
	LVS [8]	58.20	52.50	-	-
	EZ-VSL* [21]	62.24	54.74	19.86	30.96
	Ours	**84.33**	**63.26**	**35.27**	**38.00**
	EZ-VSL + OGL [21]	81.93	62.58	37.61	39.21
	Ours + OGL	**84.73**	**64.34**	**40.97**	**40.38**
Flickr 144k	Attention10k [28]	66.00	55.80	-	-
	DMC [15]	67.10	56.80	-	-
	LVS [8]	69.90	57.30	-	-
	HardPos [29]	75.20	59.70	-	-
	EZ-VSL* [21]	72.69	58.70	30.27	35.92
	Ours	**78.71**	**59.33**	**33.93**	**37.29**
	EZ-VSL + OGL [21]	83.13	63.06	41.01	40.23
	Ours + OGL	**83.93**	**63.06**	**41.10**	**40.44**

Table 3. Quantitative results on Heard 110 and Unheard 110. For a fair comparison, the results of EZ-VSL [21] and ours are integrated with the OGL module.

Test Set	Method	CIoU(%)	AUC(%)
Heard 110	LVS [8]	28.90	36.20
	EZ-VSL [21]	37.25	38.97
	Ours	**39.54**	**39.83**
Unheard 110	LVS [8]	26.30	34.70
	EZ-VSL [21]	39.57	39.60
	Ours	**42.91**	**41.17**

Table 2. Quantitative results of models trained with VGG-SS 10k and 144k.

Train set	Method	Flickr CIoU(%)	Flickr AUC(%)	VGG-SS CIoU(%)	VGG-SS AUC(%)
VGGSound 10k	LVS [8]	61.80	53.60	-	-
	EZ-VSL* [21]	63.85	54.44	25.84	33.68
	Ours	**85.74**	**63.66**	**37.29**	**38.99**
	EZ-VSL + OGL [21]	78.71	61.53	38.71	39.80
	Ours + OGL	**82.13**	**63.64**	**40.69**	**40.42**
VGGSound 144k	Attention10k [28]	-	-	18.50	30.20
	DMC [15]	-	-	29.10	34.80
	AVObject [1]	-	-	29.70	35.70
	LVS [8]	73.50	59.00	34.40	38.20
	HardPos [29]	76.80	59.20	34.60	38.00
	EZ-VSL [21]	79.51	61.17	34.38	37.70
	Ours	**84.73**	**63.76**	**39.50**	**39.66**
	EZ-VSL + OGL [21]	83.94	63.60	38.85	39.54
	Ours + OGL	**85.14**	**64.30**	**41.85**	**40.80**

Table 4. Zero-shot results on AVSBench S4 and MS3 [38]. All models are pretrained on VGGSound-144k dataset.

Test set	Method	mIoU	FScore
S4	LVS	23.69	.251
	EZ-VSL	26.43	.292
	Ours	**27.15**	**.314**
MS3	LVS	18.54	.174
	EZ-VSL	21.36	.216
	Ours	**21.98**	**.225**

Mining the potential False Negatives

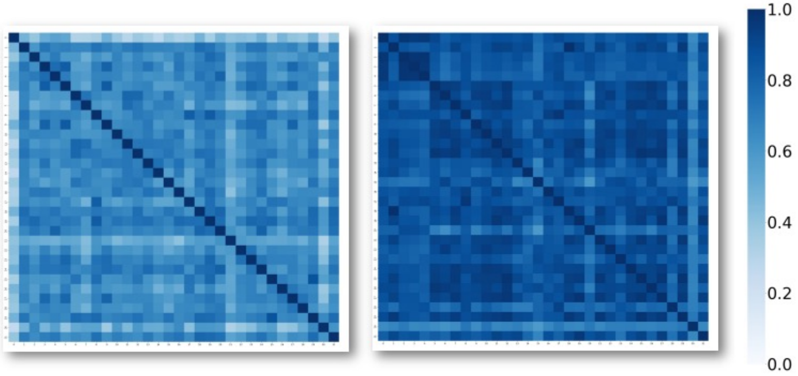


Figure 7. Cross-modal similarity matrix predicted by EZ-VSL (left) and ours (right) when all samples in the batch belong to the same category, *namely*, they are false negatives of each other. All values are normalized between 0 to 1.

Table 6. Audio-visual similarities with different data. TN: all samples in the batch belong to different categories. FN: all samples in the batch belong to the same category.

Method	TN ↓	FN ↑
LVS	0.4484	0.5102
EZ-VSl	0.5858	0.5938
Ours	**0.3812**	**0.6554**

Limitations and future works:

- We can’t recognize multiple object instances that belong to the same semantic class.
- More fine-grained future fusion method is anticipated.

Thank you!