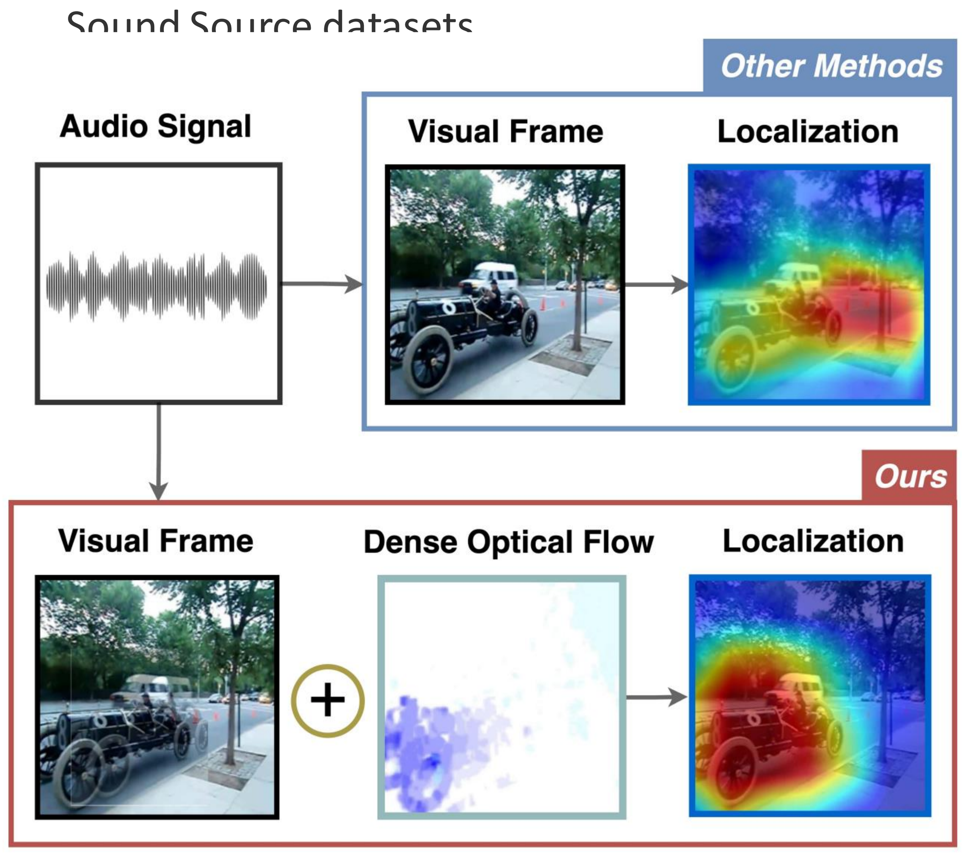
Hear The Flow: Optical Flow-Based Self-Supervised Visual Sound Source Localization

Dennis Fedorishin*, *Deen Dayal Mohan**, *Bhavin Jawade*, *Srirangaraj Setlur*, *Venu Govindaraju*

University at Buffalo, Buffalo, New York, USA

Abstract

Learning to localize the sound source in videos without explicit annotations is a novel area of audio-visual research. Existing work in this area focuses on creating attention maps to capture the correlation between the two modalities to localize the source of the sound. In a video, oftentimes, the objects exhibiting movement are the ones generating the sound. In this work, we capture this characteristic by modeling the optical flow in a video as a prior to better aid in localizing the sound source. We further demonstrate that the addition of flow-based attention substantially improves visual sound source localization. Finally, we benchmark our method on standard sound source localization datasets and achieve state-of-the-art performance on the SoundNet Flickr and VGG Sound Source datasets.



Background & Motivation

VISUAL SOUND SOURCE LOCALIZATION

- Given a video or image frame, localize the dominant sounding object(s)

CHALLENGES

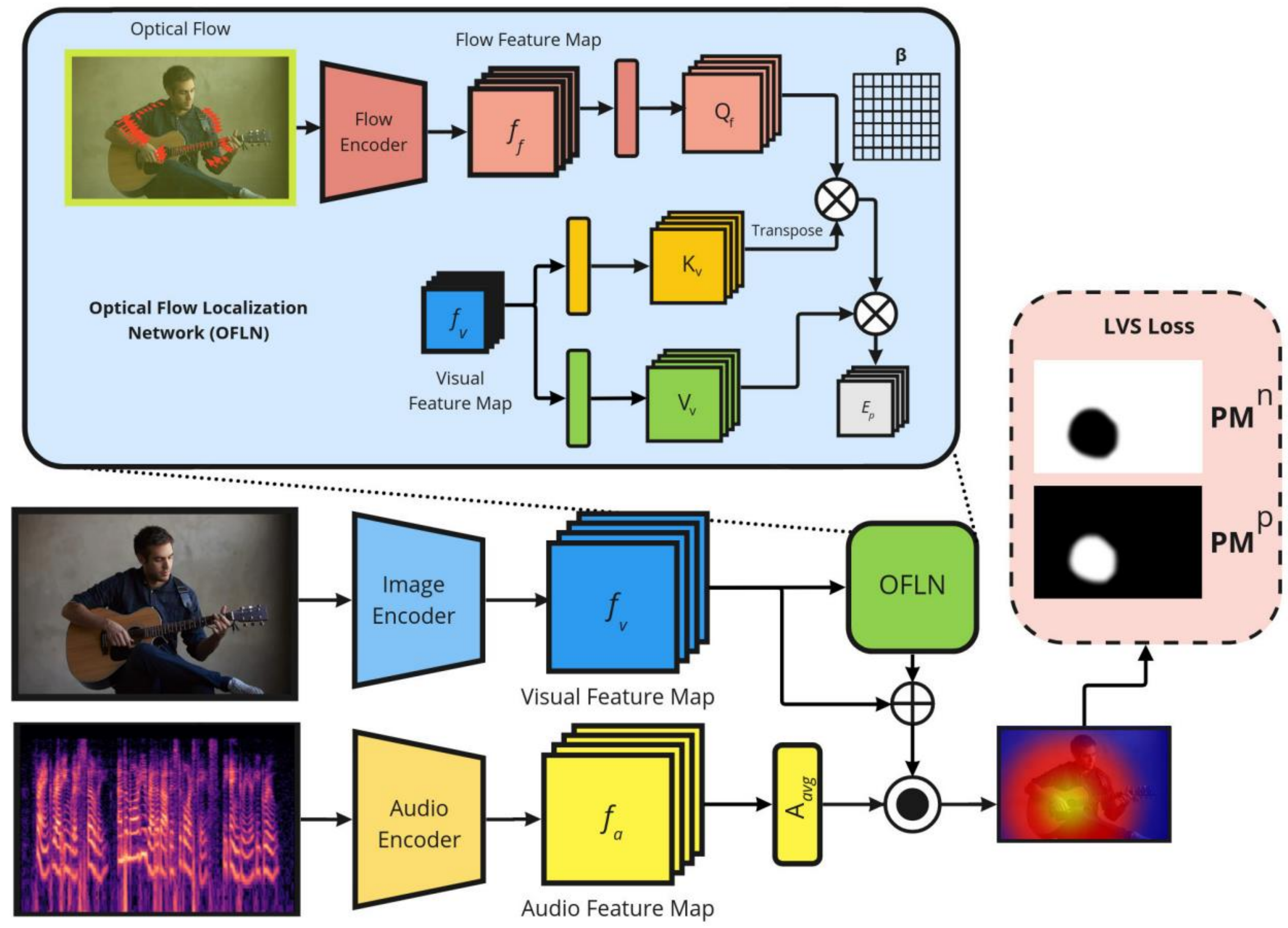
- Supervised methods require costly manually-labeled bounding boxes of sounding objects
- Creating audio-visual associations for localizing in a self-supervised settings is challenging

MOTIVATION

- Recent works focus on new loss formulations for improving contrastive representation learning
- We show the importance of *informative priors*, like optical flow, to improve VSSL with existing losses

Method

- **In a video, oftentimes the objects that are moving are making sounds**.
- We model this characteristic with our *Optical Flow Localization Network (OFLN)*



LOCALIZATION

- Localization using similarity of audio features at each visual spatial location

$$A_{avg} = GAP(f_a)$$

$$S = \frac{f_v^i . A_{avg}}{||f_v^i||.||A_{avg}||}, \forall i \in [1, m*n]$$

OPTICAL FLOW CROSS-ATTENTION

- Construct similarity matrix of visual and optical flow feature representations

$$\beta = softmax(\frac{K_v \odot Q_f}{\sqrt{d}})$$

- Create cross-attended visual and optical flow features

$$E = V_v^{ij}\beta^{ij}; \forall i \in [1, m]; \forall j \in [1, n]$$

- Add attended flow features to visual feature map and construct *enhanced* similarity map of audio features at each visual-flow spatial location

$$f_{enh} = f_v \oplus E_p$$

$$S^{enh} = \frac{f_{enh}^i . A_{avg}}{||f_{enh}^i||.||A_{avg}||}, \forall i \in m*n$$

SELF-SUPERVISED TRAINING

- Threshold the similarity matrix into positive and negative pseudo masks

$$\mathrm{PM}_k^p = \sigma(S_{k\to k} - \epsilon_p)/\tau$$
$$\mathrm{PM}_k^n = \sigma(S_{k\to k} - \epsilon_n)/\tau$$

- Construct positive and negative regions across samples in a batch and train with contrastive loss, like InfoNCE

$$\mathrm{Pos}_k = \frac{1}{|PM_k^p|}\langle PM_k^p, S_{k\to k}\rangle$$

$$\mathrm{Neg}_k = \frac{1}{|1 - PM_k^n|}\langle 1 - PM_k^n, S_{k\to k}\rangle + \frac{1}{m*n}\sum_{k\neq j}\langle 1, S_{k\to j}\rangle$$

$$L = -\sum_{\mathcal{X}}\left\{\log\left(\frac{\exp(Pos_k)}{\exp(Pos_k) + \exp(Neg_k)}\right)\right\}$$

Results

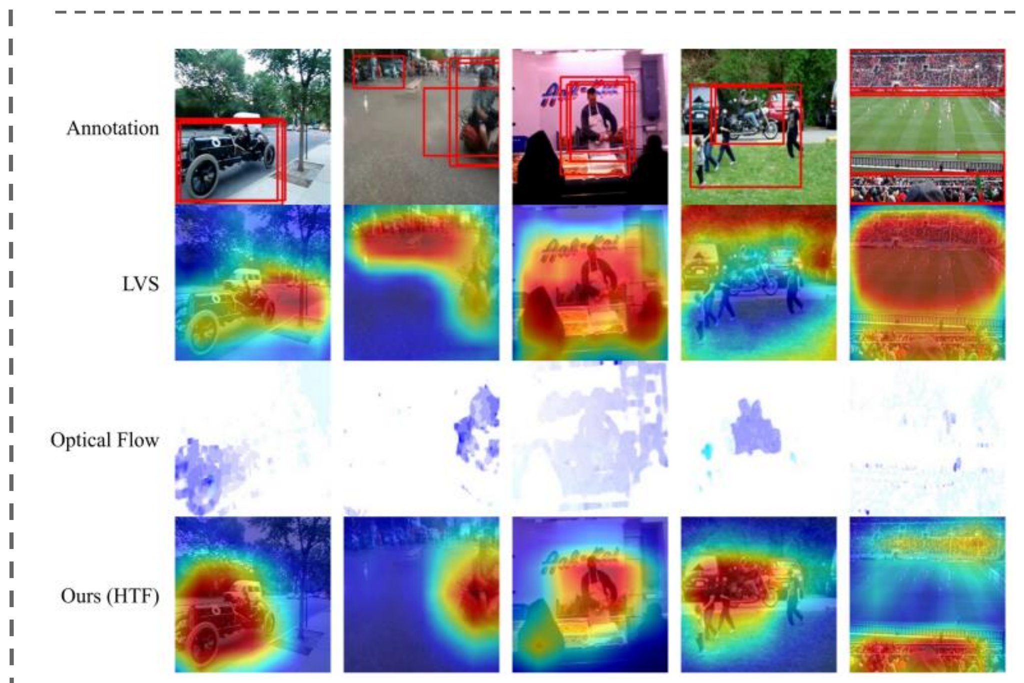
- State-of-the-art performance on Flickr SoundNet and VGG Sound Source testing datasets
- Using loss functions from previous works, **we show incorporating optical flow significantly improves VSSL**.

Method	Training Set	$\mathrm{cIoU}_{0.5}$	AUC_{cIoU}
Attention [28]	Flickr 10k	0.436	0.449
CoarseToFine [25]		0.522	0.496
AVObject [1]		0.546	0.504
LVS* [6]		0.730	0.578
SSPL [30]		0.743	0.587
HTF (Ours)		**0.860**	**0.634**
Attention [28]	Flickr 144k	0.660	0.558
DMC [19]		0.671	0.568
LVS* [6]		0.702	0.588
LVS† [6]		0.697	0.560
HardPos [29]		0.762	0.597
SSPL [30]		0.759	0.610
HTF (Ours)		**0.865**	**0.639**
LVS* [6]	VGGSound 144k	0.719	0.587
HardPos [29]		0.768	0.592
SSPL [30]		0.767	0.605
HTF (Ours)		**0.848**	**0.640**

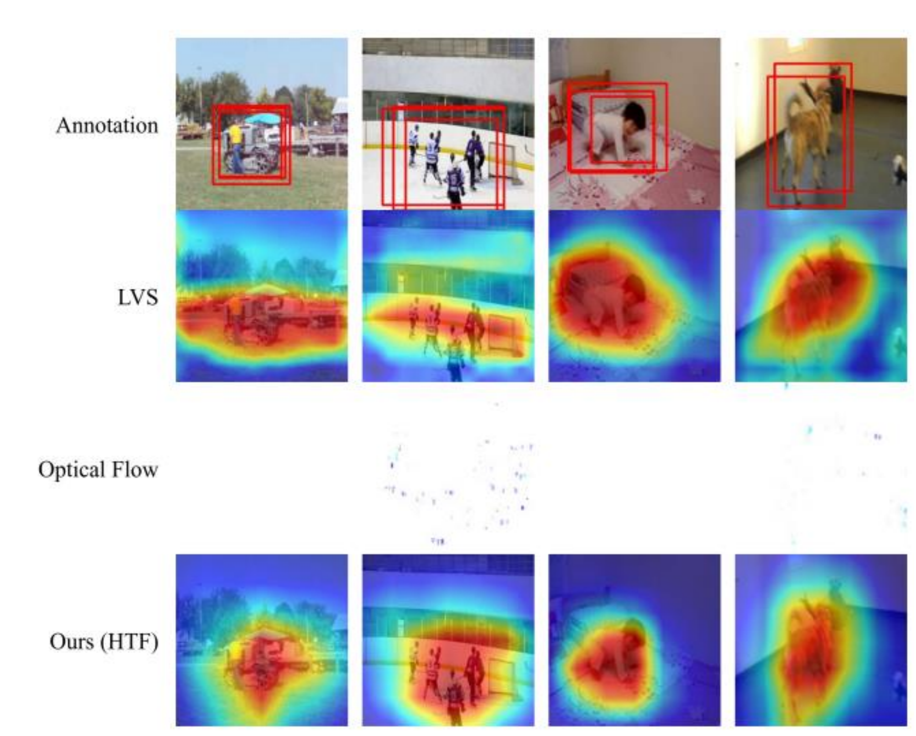
Method	Training Set	$\mathrm{cIoU}_{0.5}$	AUC_{cIoU}
Attention [28]	VGGSound 10k	0.160	0.283
LVS* [6]		0.297	0.358
SSPL [30]		0.314	0.369
HTF (Ours)		**0.393**	**0.398**
Attention [28]	VGGSound 144k	0.185	0.302
AVObject [1]		0.297	0.357
LVS* [6]		0.301	0.361
LVS† [6]		0.288	0.359
HardPos [29]		0.346	0.380
SSPL [30]		0.339	0.380
HTF (Ours)		**0.394**	**0.400**

- Strong ability to generalize across datasets and unheard sound classes

Method	Testing Set	$\mathrm{cIoU}_{0.5}$	AUC_{cIoU}
LVS* [6]	VGGSS Heard 110	0.251	0.336
HTF (Ours)		**0.373**	**0.386**
LVS* [6]	VGGSS Unheard 110	0.270	0.349
HTF (Ours)		**0.393**	**0.400**



- The proposed OFLN generalizes well even in the absence of meaningful optical flow



Conclusion

- We explore and usefulness of *informative priors* to train self-supervised visual sound source localization models
- We incorporate optical flow with our novel OFLN, achieving **state-of-the-art results across all VSSL benchmarks**

Acknowledgments

- *This work was supported by the Center for Identification Technology Research (CITeR) and the National Science Foundation (NSF) under grant 1822190.*

REFERENCES

- Chen, Honglie, et al. "Localizing visual sounds the hard way." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

dcfedori@buffalo.edu

Download our code:
https://github.com/denfed/heartheflow