Hear The Flow: Optical Flow-Based Self-Supervised Visual Sound Source Localization (Supplementary Material)

Dennis Fedorishin* Deen Dayal Mohan* Bhavin Jawade Srirangaraj Setlur
Venu Govindaraju
University at Buffalo, Buffalo, New York, USA
{dcfedori,dmohan,bhavinja,setlur,govind}@buffalo.edu

A. Training And Evaluation Procedure

A.1. Dataset Construction

As mentioned in the main paper, we train our method on the VGG Sound [3] and Flickr SoundNet [1] datasets. Both of these datasets are originally aggregated from online video streaming services, YouTube and Flickr, respectively. The authors of both datasets release only video IDs hosted on the platform, meaning that users of these datasets need to individually download and preprocess videos. A portion of these videos have become unrecoverable, as they have either been removed or blocked on the platform.

To construct the training sets of both datasets, we download a random subset of videos in each dataset until enough available videos are collected to construct the 10k and 144k training sets. We then preprocess the videos, extract audio, and construct optical flow maps as described in the main paper.

For the VGG Sound Source [2] testing set, 5,158 YouTube IDs are provided as the official testing set. At the time of dataset construction, 488 videos are unrecoverable, resulting in a testing set of 4,670 samples. We use these available samples as the VGG Sound Source testing set to construct image-flow-audio pairs for evaluating our method. For the Flickr SoundNet test set, [4] provides 250 preprocessed image-audio testing pairs that are directly available from the author's official project page[1]. However, [4] does not provide the original videos of these 250 testing samples, which are required for constructing optical flow fields. We recover these original videos from the Flickr platform, find the corresponding video frame that the original test sample of that video belongs to, and construct an optical flow field using the subsequent frame. In this process, we are able to recover 178 videos, making our Flickr SoundNet consist of 178 image-flow-audio samples.

*Equal contribution authors in alphabetic order

[1] https://github.com/ardasnck/learning_to_localize_sound_source

Method	Training Set	cIoU$_{0.5}$	AUC$_{cIoU}$
LVS* [2]	Flickr 10k	0.659	0.529
HTF (Ours)		**0.718**	**0.558**
LVS* [2]	Flickr 144k	0.684	0.535
HTF (Ours)		**0.759**	**0.575**
LVS* [2]	VGGSound 144k	0.665	0.529
HTF (Ours)		**0.734**	**0.564**

Table 1. Quantitative results on the novel, expanded Flickr SoundNet testing dataset where models are trained on the two training subsets of Flickr SoundNet and VGG Sound 144k. "*" Denotes our faithful reproduction of the method.

A.2. Expanded Flickr SoundNet Test Set

The Flickr SoundNet test set, created by [4], contains 250 annotated samples randomly selected out of 2,786 total annotated samples. [4] originally used these annotated samples to explore supervised and semi-supervised learning methods for visual sound source localization. With the current research landscape of self-supervised sound source localization, the 250 testing samples are used for evaluation and the remaining annotated samples provided by [4] are disregarded. Since our method optimizes a self-supervised objective with no explicit annotations for training, one such alternative is to use these remaining annotated samples for expanding the Flickr SoundNet test set, as they are otherwise unused.

We collect and preprocess these remaining annotated samples to construct a novel Flickr SoundNet test set consisting of 1,769 samples. As shown in Table 1, we show results on the expanded Flickr SoundNet test set in a similar fashion to the quantitative results in the main paper on the official Flickr SoundNet test set. Specifically, we compare our method against LVS [2], where models are trained on both subsets of the Flickr SoundNet training sets, in addition to VGG Sound 144k.

As shown, our method significantly outperforms LVS [2] in the expanded testing scenario, showing our method is still

robust to a much larger-scale testing set that spans more sounding categories than the official test set. In addition, when comparing against testing on the official 250 samples, we see a reduction in performance across all methods, showing that expanding the testing set leads to a more challenging sound source localization scenario. For example, for our method trained on Flickr 144k, we achieve 0.865 cIoU and 0.639 AUC on the official test set, compared to 0.759 cIoU and 0.575 AUC on the expanded testing set we introduce. We believe this evaluation on 1,769 annotated samples instead of 250 samples offers a more robust and representative testing set, which can be used for future self-supervised visual sound localization works for improved evaluations.

B. Additional Implementation Details

As mentioned in the main paper, we use ResNet18 feature extractors for the visual, audio, and flow portions of our method. For a given sample, the output features of the visual encoder, f_v, is a 7×7 spatial feature map where each spatial location has a feature vector of 512 units. These features, once attended over with the optical flow localization network, are used with the audio representation of the sample to construct S^{enh}, the sound source localization map. During inference, S^{enh} is upscaled to the size of the original image, which represents the visual sound source localization of that image. Furthermore, the attended visual and audio representations are both L_2 normalized before constructing S^{enh}.

For data augmentations, we randomly crop each image and optical flow map, in addition to a 50% chance of applying a horizontal flip to both. We normalize the images using the standardized ImageNet normalization statistics and the optical flow maps using a mean of 0 and standard deviation of 1.

C. Additional Qualitative Results

In Figure 1, we visualize examples comparing our method against LVS [2] on our expanded Flickr SoundNet test set, described in A.2. As shown, our method is able to reliably localize towards the visual sound source, both in the presence and absence of meaningful optical flow information. Further, we show that these otherwise unused labeled samples are of high quality and are a useful addition for better evaluating self-supervised visual sound source localization methods. These annotated samples, previously reserved for training, are not needed for the self-supervised learning objective.

D. Prior Work Reproduction

As mentioned in A.1, since many of the videos in the dataset are missing, generating perfromance numbers for prior works on the available test videos becomes important to fairly assess the contribution of our work. We present a comparison of the performance of [2] against our method, since it is the most relevant to our proposed approach. Reproducing other prior methods like [5] is a challenging task as the authors did not release pretrained models associated with their methods. Further, the public project repositories are missing relevant prepossessing or configuration files, which is required for a proper and fair reproduction of this method. To address this issue and further spur research, we open-source our code and other necessary resources for proper reproductions, available at `https://github.com/denfed/heartheflow`.

Method	Training Set	$\text{cIoU}_{0.5}$	AUC_{cIoU}
LVS [2]	Flickr 10k	0.582	0.525
LVS* [2]		0.730	0.578
LVS [2]	Flickr 144k	0.699	0.573
LVS† [2]		0.697	0.560
LVS* [2]		0.702	0.588
LVS [2]	VGGSound 144k	0.719	0.582
LVS* [2]		0.719	0.587

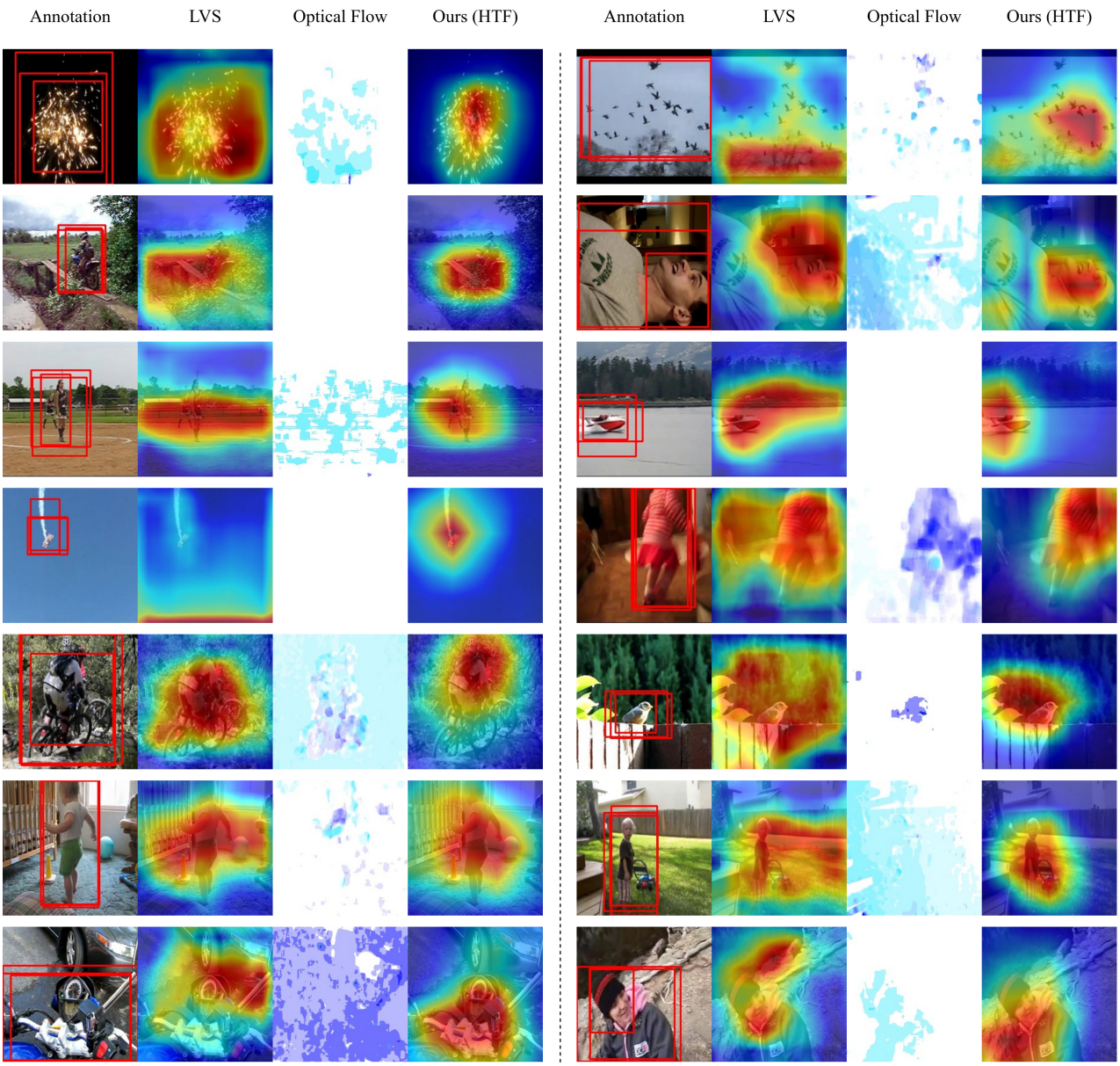
Table 2. Reproduction results of LVS [2] on the Flickr SoundNet testing dataset. "*" Denotes our faithful reproduction of the method, and "†" denotes our evaluation reproduction using officially provided model weights.

Method	Training Set	$\text{cIoU}_{0.5}$	AUC_{cIoU}
LVS [2]	VGGSound 10k	-	-
LVS* [2]		0.297	0.358
LVS [2]	VGGSound 144k	0.344	0.382
LVS† [2]		0.288	0.359
LVS* [2]		0.301	0.361

Table 3. Reproduction results of LVS [2] on the VGG Sound Source testing dataset. "*" Denotes our faithful reproduction of the method, and "†" denotes our evaluation reproduction using officially provided model weights.

As shown in Tables 2 and 3, we compare the reproduced results of LVS [2] against the author's original results described in [2]. For the Flickr SoundNet test set, our reproduced results are comparable with the original work. In certain cases, like training on Flickr 10k and Flickr 144k, we outperform the original results described in [2]. For VGG Sound, the authors highlight that some VGG Sound Source annotations are updated and result in a 2-3% difference in sound source localization performance on their official project page[2]. This difference is consistent with our reproduced results. Based on these results, we believe that our reproduction is faithful and hence we are able to provide a fair comparison to our method.

[2] `https://github.com/hche11/Localizing-Visual-Sounds-the-Hard-Way`



Visualization on expanded Flickr SoundNet test set

Figure 1. Qualitative results of our method on the expanded Flickr SoundNet test set described in A.2. Examples are from models trained on the Flickr 144k set. Figure is best viewed in color.

References

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, 2016.

[2] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021.

[3] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020.

[4] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018.

[5] Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, and Zhaoxiang Zhang. Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3222–3231, 2022.