2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) | 978-1-6654-9346-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/WACV56688.2023.00231

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
{dcfedori,dmohan,bhavinja,setlur,govind}@buffalo.edu

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. Code: `https://github.com/denfed/heartheflow`.

1. Introduction

In recent years, the field of audio-visual understanding has become a very active area of research. This can be attributed to the large amount of video data being produced as part of user-generated content on social media and other platforms. Recent methods in audio-visual understanding have leveraged popular deep learning techniques to solve challenging problems such as action recognition [13], deepfake detection [34], and other tasks. Given a video, one such task in audio-visual understanding is to locate the object in the visual space that is generating the prominent audio content. When observing a natural scene, it is often trivial for a human to localize the region/object from which the sound originates. One of the main reasons for this is the binaural nature of the human hearing sense. However, the majority of audio-visual data in digital media is monaural, which complicates audio localization tasks. Furthermore, naturally occurring videos do not have explicit annotations of the location of the source of the audio in the image. This makes the task of training deep neural networks to understand audio-visual associations for localization a challenging task.

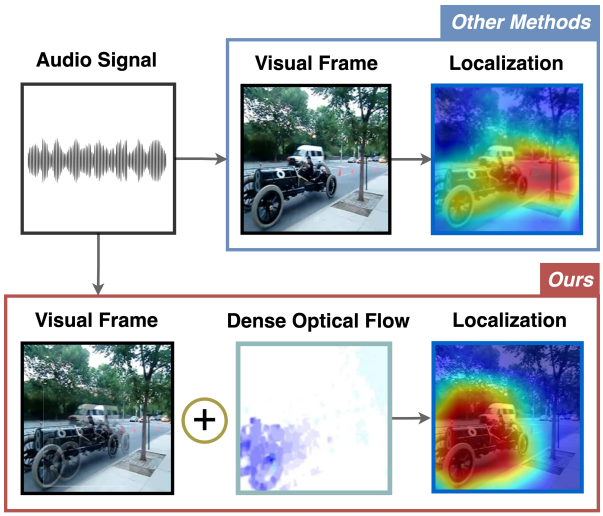


Figure 1. Given a video with audio, the goal of sound source localization is to localize the object/region producing the sound in a video frame. Our method introduces *optical flow* as an informative prior to improve visual sound source localization performance.

Owing to the success of self-supervised learning (SSL) in vision [8, 16], language [9, 26] and other multi-modal applications [2, 22], recent methods in sound source localization [6, 30] have adopted SSL based methods to overcome the need for annotations. One such method [6], finds the cosine similarity between the audio and visual representations extracted convolutionally at different spatial locations in the images. They rely on self-supervised training by creating positive and negative associations from these predicted similarity matrices. This bootstrapping approach has been shown to improve sound source localization.

Following this research finding, the majority of recent approaches in visual sound source localization have focused on creating robust optimization objectives for better audio-visual associations. However, one interesting aspect of the

*Equal contribution authors in alphabetic order

2642-9381/23/$31.00 ©2023 IEEE
DOI 10.1109/WACV56688.2023.00231

problem that has received relatively little attention is the creation of *informative priors* to improve the association of the audio to the correct "sounding object" (or the object producing the sound). Priors can be viewed as potential regions in the image from where the sound may originate. We can draw parallels to work in two-stage object detection methods, in which region proposal networks are used to identify regions in the image space that could potentially be objects. However, generating potential candidate regions for sound source localization is more challenging because the generated priors should be relevant from a multi-modal perspective. In order to generate these informative priors for where sounds possibly originate from, we leverage optical flow.

The intuition behind using optical flow to create an enhanced prior is the fact that optical flow can model patterns of apparent motion of objects. This is important as most often, an object moving in a video tends to be the sound source. Enforcing a constraint to prioritize the objects that tend to be in relative motion might lend itself to creating better sound source localizations. This paper proposes an optical flow-based localization network that can create informative priors for performing superior sound source localization. The contributions in this paper are as follows:

1. We explore the need for creating informative priors for visual sound source localization, which is a complementary research direction to prior methods.
2. We propose the use of optical flow as an additional source of information to create informative priors.
3. We design an optical flow-based localization network that uses cross-attention to form stronger audio-visual associations for visual sound source localization.
4. We run extensive experiments on two benchmark datasets: VGG Sound and Flickr SoundNet and demonstrate the effectiveness of our method. Our method consistently achieves superior results over the state-of-the-art. We perform rigorous ablation studies and provide quantitative and qualitative results, showing the superiority of our novel localization network.

2. Related Work

Generating robust multi-modal representations through joint audio-visual learning is an active area of research that has found application in multiple audio-visual tasks. Initial works in the area of joint audio-visual learning focus on probabilistic approaches. In [17], the audio-visual signals were modeled as samples from a multivariate Gaussian process, and audio-visual synchrony was defined as the mutual information between the modalities. [12] focused on first learning a lower-dimensional subspace that maximized mutual information between the two modalities. Furthermore, they explored the relationship between these audio-visual signals using non-parametric density estimators. [20] proposed a spatio-temporal segmentation mechanism that relies on the velocity and acceleration of moving objects as visual features and used canonical correlation analysis to associate the audio with relevant visual features. In recent years, deep learning-based methods have been used to explore the creation of better bimodal representations. They mostly employ two-stream networks to encode each modality individually and employ a contrastive loss-based supervision to align the two representations [19]. Methods like [1, 32] used source separation to localize audio via motion trajectory-based fusion and synchronization. Furthermore, [25] addressed the problem of separating multiple sound sources from unconstrained videos by creating coarse to fine-grained alignment of audio-visual representations. Additionally, methods like [24, 25] use class-specific saliency maps. [33] uses class attention maps to help generate saliency maps that are used for better sound source localization. More recently, methods have focused on creating objective functions specific to sound localization. [6] introduced the concept of tri-map, which incorporates background mining techniques into the self-supervised learning setting. The tri-map contains an area of positive correlation, no correlation (background), and an ignore zone to avoid uncertain areas in the visual space. [30] introduced a negative-free method for sound-localization by mining explicit positives. Further, this method uses a predictive coding technique to create better a feature alignment between the audio and visual modalities. These recent methods mainly focus on creating stronger optimization objectives for visual sound source localization. A complementary direction in the research landscape is to explore creating more informative priors for audio-visual association. In this paper, we explore one such idea, which leverages optical flow. The authors of [3] have explored the use of optical flow in the context of certain audio-visual tasks, like retrieval. In this work, we explore the use of optical flow as an informative prior for visual sound source localization.

Optical flow provides a means to estimate pixel-wise motion between consecutive frames. Early works [5, 18, 31] presented optical flow prediction as an energy minimization problem with several objective terms utilizing continuous optimization. Optical flow maps can be broadly divided into two types: Sparse and Dense. Sparse optical flow represents the motion of salient features in a frame, whereas Dense optical flow represents the motion flow vectors for the whole frame. Earlier methods for sparse optical flow estimation include the Lucas-Kanade algorithm [21], which utilizes brightness constancy equations to optimize a least squares approximation under the assumption that flow remains locally smooth and the relative displacement of neighboring pixels is constant. Farneback [10] proposed a dense optical flow estimation technique where quadratic polynomials were utilized to approximate pixel neighborhood for two

frames, and these polynomials were then used to compute global displacement. FlowNet [11] proposed the first CNN-based approach towards estimating optical flow maps where they computed static cross-correlation between intermediate convolutional feature maps for two consecutive frames and up-scale them to extract optical flow maps.

3. Method

In this section, we will first present the formulation of the sound source localization problem under a supervised setting. Following this, we will describe the current self-supervised approach, motivate the need for better localization proposals for sound source localization, and subsequently elaborate on the design and implementation of our novel optical flow-based sound source localization network.

3.1. Problem Statement

Given a video consisting of both the audio and visual modality, visual sound source localization aims to find the spatial region in the visual modality that generated the audio. Consider a video consisting of N frames. Let the image corresponding to a video frame be I, where $I \in \mathbb{R}^{W_i \text{x} H_i \text{x} 3}$ and A be the spectrogram representation generated out of the audio around the frame, where $A \in \mathbb{R}^{W_a \text{x} H_a \text{x} 1}$. The problem of audio localization can be thought of as finding the region in I that has a high association/correlation with A. More formally, this can be written as:

$$\begin{aligned} &f_v = \Phi(I;\theta_i); f_a = \Psi(A;\theta_j) \\ &P(I,A) = \omega(f_v, f_a) \end{aligned} \tag{1}$$

where $\Phi(I;\theta_i)$ and $\Psi(A;\theta_j)$ correspond to convolution neural network-based feature extractors associated with visual and audio modalities, and $f_v \in \mathbb{R}^{m \text{x} n \text{x} c}$ and $f_a \in \mathbb{R}^{m \text{x} n \text{x} c}$ are the corresponding lower dimensional feature maps, respectively. ω is the function that finds the association between the two modalities, and $P(I,A)$ is the region in the original image space with the source that generated the audio. It is important to note that extrapolating the association in the feature space to the corresponding region of the original image space (i.e $P(I,A)$) is trivial. Given the above-mentioned feature maps, one way of finding an association between the feature representations is:

$$\begin{aligned} &A_{avg} = GAP(f_a) \\ &S = \frac{f_v^i . A_{avg}}{||f_v^i||.||A_{avg}||}, \forall i \in [1, m*n] \end{aligned} \tag{2}$$

where $GAP(f_a)$ is the global-average-pooled representation of the audio feature map. S represents the cosine similarity of this audio representation to each spatial location in the visual feature map. Here m and n are the width and height of the feature map. If a binary mask $M \in \mathbb{R}^{m \text{x} n \text{x} 1}$ generated from a ground truth indicating positive and negative regions of audio-visual correspondence is available, we can formulate the learning objective in a supervised setting:

For a given sample k (with an image frame I_k and audio A_k) in the dataset, the positive and negative response can be defined as

$$\begin{aligned} \text{Pos}_k &= \frac{1}{|M_k|}\langle M_k, S_{k\to k}\rangle \\ \text{Neg}_k &= \frac{1}{|1-M_k|}\langle 1-M_k, S_{k\to k}\rangle + \frac{1}{m*n}\sum_{k\neq j}\langle 1, S_{k\to j}\rangle \end{aligned} \tag{3}$$

Here $S_{k\to k}$ refers to the cosine similarity S from Eq 2 when using I_k and A_k. Similarly, $S_{k\to j}$ is the cosine similarity when the image and audio are not from the same video. $\langle\cdot,\cdot\rangle$ denotes the inner product. The final learning objective has a similar formulation to [23]:

$$L = -\sum_{\mathcal{X}}\left\{\log\left(\frac{\exp(Pos_k)}{\exp(Pos_k)+\exp(Neg_k)}\right)\right\} \tag{4}$$

3.2. Self-Supervised Localization

In most real-world scenarios, the ground truth necessary to generate the binary mask M would be missing. Hence there is a need for a training objective that does not rely on explicit ground truth annotations. One way to achieve this objective is to replace the ground truth mask with a generated pseudo mask as proposed in [6]. The pseudo mask can be generated by binarizing the similarity matrix S based on a threshold. More specifically, given $S_{k\to k}$ from Eq 2, the pseudo mask can be written as:

$$PM = \sigma(S_{k\to k} - \epsilon)/\tau \tag{5}$$

where ϵ is a scalar threshold. σ denotes the sigmoid function that maps a similarity value in $S_{k\to k}$, that is below the threshold to value 0 and above the threshold to 1. τ is the temperature controlling the sharpness. Additionally, [6] further refines the pseudo mask by eliminating potentially noisy associations. This is done by considering separate positive and negative thresholds above and below the similarity value that is considered reliable. If a value is between these thresholds, it's considered a noisy association and is subsequently ignored. More formally:

$$\begin{aligned} \text{PM}_k^p &= \sigma(S_{k\to k} - \epsilon_p)/\tau \\ \text{PM}_k^n &= \sigma(S_{k\to k} - \epsilon_n)/\tau \\ \text{Pos}_k &= \frac{1}{|PM_k^p|}\langle PM_k^p, S_{k\to k}\rangle \\ \text{Neg}_k &= \frac{1}{|1-PM_k^n|}\langle 1-PM_k^n, S_{k\to k}\rangle \\ &\quad + \frac{1}{m*n}\sum_{k\neq j}\langle 1, S_{k\to j}\rangle \end{aligned} \tag{6}$$

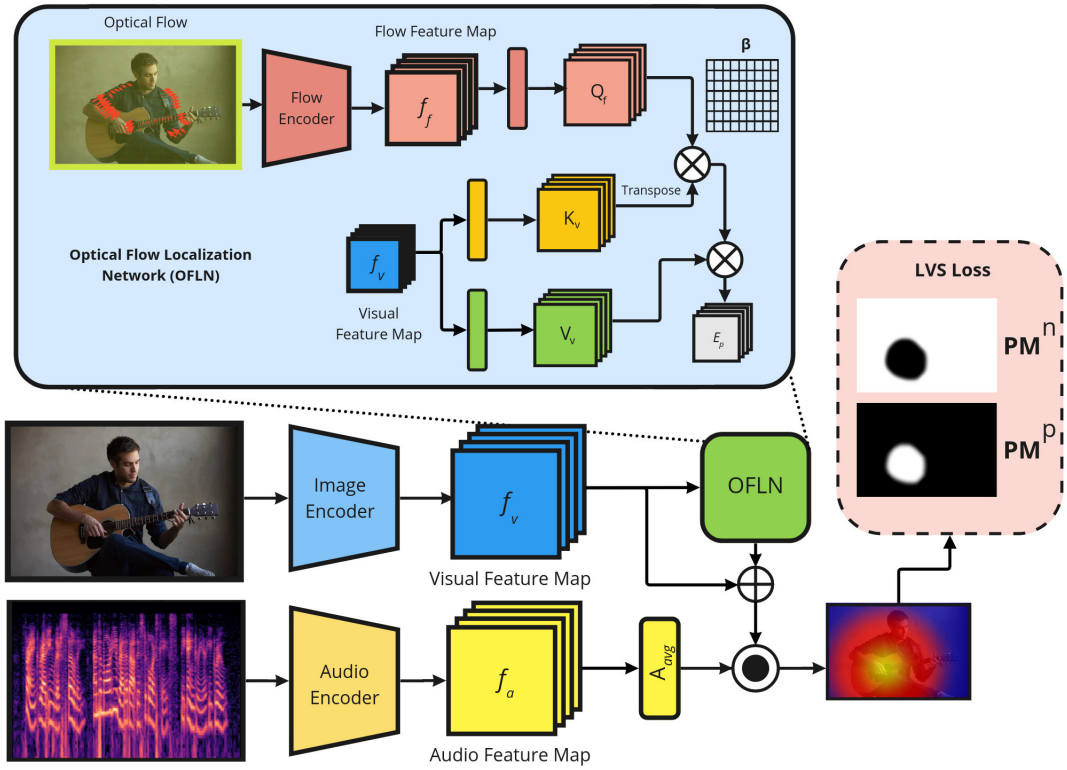


Figure 2. Overview of our optical flow-based sound localization method. Given a chosen frame of a video and the audio surrounding that frame, we extract features from both modalities, which are then used to attend towards sounding objects in the frame. We further compute the dense optical flow field from the chosen and subsequent frame and use flow features to attend towards moving objects in the frame.

Here ϵ_p and ϵ_n are the positive and negative thresholds, respectively. Once the positive and negative responses are computed, the overall training objective is similar to Eq 4.

In the above approach, it is logical to bootstrap the prediction and perform self-supervised training if the pseudo masks in Eq 5 generated at the initial training iterations resemble that of the ground truth. However, this is not guaranteed since the feature extractors associated with the individual modalities (in Eq 1) are randomly initialized. Therefore, a high or low value in the similarity matrix $S_{k\to k}$ during the initial iterations of the self-supervised training may not correspond to informative positive or negative regions since the feature extractors are not trained. If a feature extractor is initialized with pretrained weights from a classification task, for example, the visual extractor on ImageNet, the network will often activate towards objects in the image. Considering this characteristic as an *object-centric* prior, it may be useful for self-supervised sound localization as the most salient objects in a frame are often the ones emitting the sound. However, situations may arise where the source of the audio is not the most salient object in the frame. This would produce sub-optimal associations $S_{k\to k}$ in the initial iterations, which, when used for self-supervised training as mentioned in Eq 6 would lead to sub-optimal performance. As a result, there is a need to construct more meaningful priors when computing $S_{k\to k}$ to improve audio-visual associations, subsequently improving self-supervised learning.

3.3. Optical-Flow Based Localization Network

Having motivated the need for some meaningful priors that enable better audio-visual associations, we approach the problem from an object detection viewpoint. In earlier object detection methods such as R-CNN [15] and Fast R-CNN [14], a selective search was used as a method to generate region proposals. Selective search provided a set of probable locations where an object of interest may be present. An alternative to selective search-based approaches is a two-stage approach like in [27], using region proposal networks. Most of these region proposal networks have auxiliary training objectives in order to produce regions containing potential objects. Using these objectives to generate potential regions of interest in a self-supervised setting becomes challenging. Furthermore, generating candidate regions using selective search or regular region proposal networks, only based on visual modality, might not be well suited for enforcing priors for a cross-modal task such as visual sound source localization.

As a better alternative, we use optical flow to generate informative localization proposals. Optical flow using frames of a video can efficiently capture the objects that are mov-

ing. Most often, these objects are the source of the sound. Capturing optical flow in the pixel space can often be a good prior to improve audio-visual association. Furthermore, since the optical flow tends to focus on the relative motion of objects rather than the salient objects, it can complement the priors of the pre-trained vision model, which tends to focus on the latter. We design a network as shown in Figure 2, which takes in optical flow computed between two adjacent video frames and generates regions in the feature map f_v that act as priors to create better audio-visual associations. The localization network is comprised of a cross-attention between the feature representation extracted from image and flow modalities. Given the flow feature representation f_f and visual feature representation f_v, we project these feature representations using separate projection layers to create two tensors K_v and Q_f. β is computed as an outer product of the tensors K_v and Q_f along the channel dimensions. That is, if K_v and $Q_f \in \mathbb{R}^{mxnxd}$, then the resulting $\beta \in \mathbb{R}^{mxnxdxd}$ is computed as below:

$$\beta = softmax(\frac{K_v \odot Q_f}{\sqrt{d}}) \tag{7}$$

The softmax function is applied to the final dimension to normalize the attention matrix. The goal is to compute the attention to be applied to each spacial location, thus yielding a cross attention matrix of size dxd for each spatial location. We compute another tensor from the visual modality $V_v \in \mathbb{R}^{mxnxd}$. For each spatial location in V_v, we have a d dimensional representation which we multiply to the corresponding dxd attention matrix in β. That is :

$$E = V_v^{ij}\beta^{ij}; \forall i \in [1, m]; \forall j \in [1, n] \tag{8}$$

Finally E is projected back to produce the final cross-attended proposal prior $E_p \in \mathbb{R}^{mxnxc}$. In order to impose this prior for performing the audio-visual association, we add E_p to the visual feature map f_v as shown in Figure 2. The enhanced audio-visual association can be written as:

$$\begin{aligned} f_{enh} &= f_v \oplus E_p \\ S^{enh} &= \frac{f^i_{enh}.A_{avg}}{||f^i_{enh}||.||A_{avg}||}, \forall i \in m * n \end{aligned} \tag{9}$$

where $\oplus$ denotes element-wise addition. Once the enhanced audio-visual association is obtained, we use Eq 6 to compute the positive and negative responses. We train the entire network (feature extractors and localization network) end-to-end using the optimization objective mentioned in Eq 4.

4. Experiments

4.1. Datasets

For training and evaluating our proposed model, we follow prior work in this area and use two large-scale audio-visual datasets:

4.1.1 Flickr SoundNet

Flickr SoundNet [4] is a collection of over 2 million unconstrained videos collected from the Flickr platform. To directly compare against prior works, we construct two subsets of 10k and 144k videos that are preprocessed into extracted image-audio pairs, described further in Section 4.3. The Flickr SoundNet evaluation dataset consists of 250 image-audio pairs with labeled bounding boxes localized on the sound source in the image, manually annotated by [28].

4.1.2 VGG Sound

VGG Sound [7] is a dataset of 200k video clips spread across 309 sound categories. Similar to Flickr SoundNet, we construct subsets of 10k and 144k image-audio pairs to train our proposed model. For evaluation, we utilize the VGG Sound Source [6] dataset, which contains 5000 annotated image-audio pairs that span 220 sound categories. Compared to the Flickr SoundNet test set, which has about 50 sound categories, VGG Sound Source has significantly more sounding categories, making it a more challenging scenario for sound localization.

4.2. Evaluation Metrics

For proper comparisons against prior works, we use two metrics to quantify audio localization performance: Consensus Intersection Over Union (cIoU) and Area Under Curve of cIoU scores (AUC) [28]. cIoU quantifies localization performance by measuring the intersection over union of a ground-truth annotation and a localization map, where the ground-truth is an aggregation of multiple annotations, providing a single consensus. AUC is calculated by the area under the curve of cIoU created from thresholds varying from 0 to 1. In our experiments, we show results for cIoU at a threshold of 0.5, denoted by $cIoU_{0.5}$, and AUC scores, denoted by AUC_{cIoU}.

4.3. Implementation Details

In this paper, sound source localization is defined as localizing an excerpt of audio to its origin location in an image frame, both extracted from its respective video clip. For both Flickr SoundNet and VGG Sound, we extract the middle frame of a video along with 3 seconds of audio centered around the middle frame and a calculated dense optical flow field to construct an image-flow-audio pair. For the image frames, we resize images to 224×224 and perform random cropping and horizontal flipping data augmentations. To calculate an optical flow field corresponding to the middle frame, we take the middle frame and subsequent frame of a video V, denoted by V_t and V_{t+1} respectively, and use the Gunnar Farneback [10] algorithm to generate a 2-channel flow field corresponding to horizontal and vertical flow vectors denoting movement magnitude. We sim-

Method	Training Set	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**

Table 1. Quantitative results on the 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, and "†" denotes our evaluation reproduction using officially provided model weights.

Method	Training Set	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**

Table 2. Quantitative results on the VGG Sound Source testing dataset where models are trained on the two training subsets of VGG Sound.

Method	Testing Set	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**

Table 3. Quantitative results on the VGG Sound Source testing dataset on heard and unheard class subsets. Each model is trained on 50k samples belonging to 110 (heard) classes.

ilarly perform random cropping and horizontal flipping of the flow fields, which are performed consistently with image augmentations. For audio, we sample 3 seconds of the video at 16kHz and construct a log-scaled spectrogram using a bin size of 256, FFT window of 512 samples, and stride of 274 samples, resulting in a shape of 257×300.

Following [6], we use ResNet18 backbones as the visual and audio feature extractors. Similarly, we use ResNet18 as the optical flow feature extractor. We pretrain the visual and flow feature extractors on ImageNet and leave the audio network randomly initialized. During training, we keep the visual feature extractor parameters frozen. For all experiments, we train the model using the Adam optimizer with a learning rate of 10^{-3} and a batch size of 128. We train the model for 100 epochs for both the 10k and 144k sample subsets on Flickr SoundNet and VGGSound. We set $\epsilon_p = 0.65$, $\epsilon_n = 0.4$, and $\tau = 0.03$, as described in Eq 6.

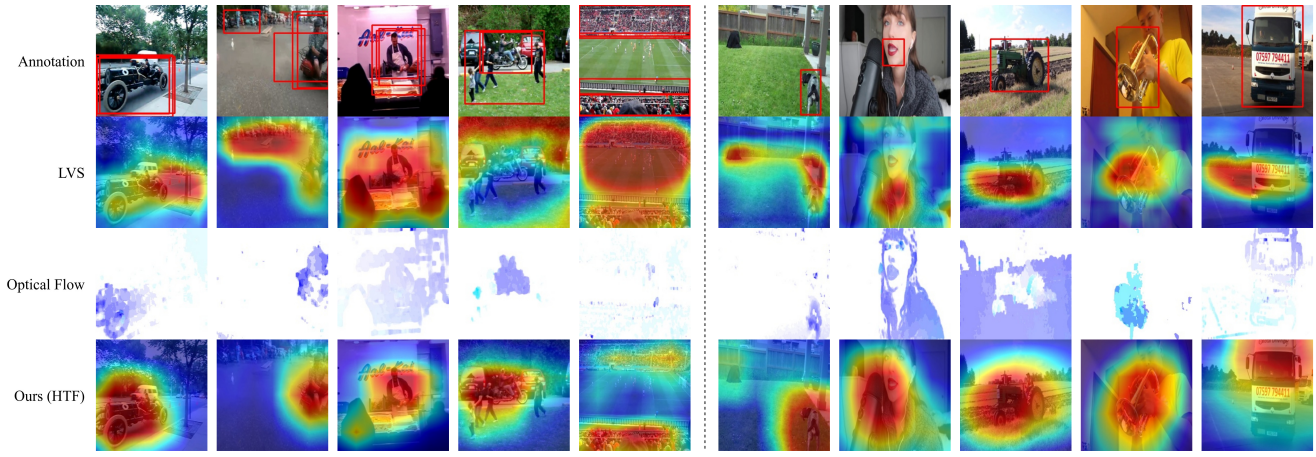
4.4. Quantitative Evaluation

In this section, we compare our method against prior works [1, 6, 19, 25, 28, 29, 30] on standardized experiments for self-supervised visual sound source localization. Results of various training configurations are reported in Tables 1 and 2 for the Flickr SoundNet and VGG Sound Source testing datasets, respectively.

As shown in Table 1 and 2, our method, HTF, significantly outperforms all prior methods, creating the new state-of-the-art on self-supervised sound source localization. On the Flickr testing set, we achieved an improved performance of 11.7% cIoU and 4.7% AUC when trained on 10k Flickr samples and 10.6% cIoU and 2.9% AUC when trained with 144k Flickr samples. Similarly, on the VGG Sound Source testing set, we improve by 7.9% cIoU and 2.9% AUC when trained on 10k VGG Sound samples and 5.5% cIoU and 2.0% AUC when trained on 144k samples.

Further, we investigate the robustness of our method by evaluating it across the VGG Sound and Flickr SoundNet datasets. Specifically, we train our model with 144k VGG Sound samples and test on the Flickr SoundNet test set. Comparing against [6, 29, 30], we significantly outperform all methods, as shown in Table 1, which shows our model is capable of generalizing well across datasets. We further investigate our method's robustness by testing on sound categories that are *disjoint* from what is seen during training. Following [6], we sample 110 sound categories from VGG Sound for training and test on the same 110 categories (heard) used during training and 110 other disjoint (unheard) sound categories. As shown in Table 3, we outperform [6] on both heard and unheard testing subsets. In addition, we highlight that the performance of the unheard subset slightly outperforms the heard subset, showing our model performs well against unheard sound categories.

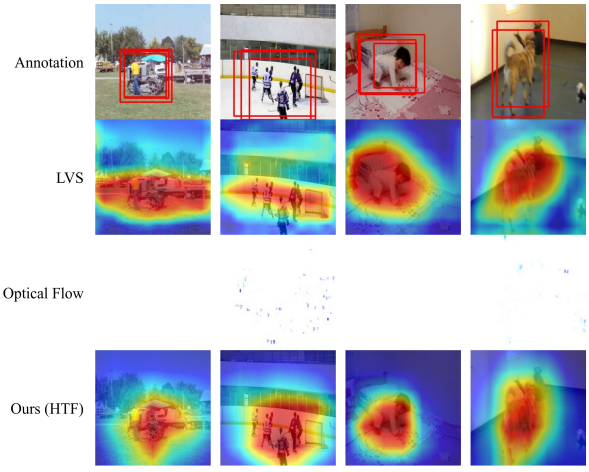
Utilizing a self-supervised loss formulation similar to [6], we see that our method significantly outperforms it on



(a) Visualization on Flickr SoundNet test set

(b) Visualization on VGG Sound Source test set

Figure 3. Qualitative results of our method on each testing set. Example visualizations are from models trained on each test set's respective 144k training set. Our method effectively localizes towards sounding objects exhibiting movement. Figure is best viewed in color.



Visualization on Flickr SoundNet test set

Figure 4. Qualitative results of our method on the Flickr SoundNet test set. Examples are from models trained on the Flickr 144k set. Even in the absence of meaningful optical flow, our method can still localize to the sound source. Figure is best viewed in color.

both testing datasets across all training setups and experiments. We highlight that these improvements are obtained from incorporating a more informative prior, based on optical flow, into the sound localization objective. In section 4.6, we further investigate the direct influence of incorporating optical flow along with our other design choices.

4.5. Qualitative Evaluation

In Figure 3, we visualize and compare sound localizations of LVS [6] and our method on the Flickr SoundNet and VGG Sound Source test sets. As shown, our method can accurately localize various types of sound sources. Comparing against LVS [6], we examine localization improvements across multiple samples, specifically where sounding objects exhibit a high flow magnitude through movement. For example, in the first column, LVS [6] localizes only a small portion of the sounding vehicle, while our method entirely localizes the vehicle, where a significant magnitude of flow is exhibited. In the fifth column, our method more accurately localizes to the two crowds in the stadium, both of which are sound sources exhibiting movement.

However, it is also important to investigate samples where little optical flow is present. It is possible that a frame in a video exhibits little movement, for example, a stationary car or person emitting noise. In these cases, there is no meaningful optical flow to localize towards. In Figure 4, we see that even in the absence of significant optical flow, our method still localizes on par or better compared to LVS [6]. This reinforces that optical flow is used as an *optional* prior, where areas of high movement when present, *may* be used to localize better, but is not required. In the following section, we further investigate the exact effects of introducing priors like optical flow into the self-supervised framework.

4.6. Ablation Studies

In this section, we explore the implications of our design choices with multiple ablation studies. As explained in section 3.2, we explore the need for informative priors to train a self-supervised audio localization network. We introduce optical flow as one of these priors, in addition to pretraining the vision network on ImageNet to provide an object-centric prior. In Table 4, we study the individual effects of each of these design choices, namely adding the flow attention mechanism, ImageNet weights for the vision encoder, and freezing the vision encoder during training.

When training the model without any priors (model 4.a), we see that performance suffers as there is little meaning-

Model	Flow	Pretrain Vision	Frozen Vision	cIoU$_{0.5}$	AUC$_{cIoU}$
a	×	×	×	0.129	0.275
b	×	✓	×	0.315	0.364
c	×	✓	✓	0.306	0.362
d	✓	×	×	0.271	0.343
e	✓	✓	×	0.382	0.394
f	✓	✓	✓	**0.393**	**0.398**

Table 4. Ablation study on incorporating optical flow and strategies on the vision encoder. All models are trained on VGG Sound 10k and tested on the VGG Sound Source test set.

ful information for the self-supervised objective. However, when simply adding the optical flow attention previously described (model 4.d), we see a large performance improvement as the network can now use optical flow to better localize, as moving objects are often the ones exhibiting sound. Similarly, when using ImageNet pretrained weights (Table 4.b), we see a significant performance improvement as the model now has an object-centric prior, where salient objects in an image are often exhibiting sound. When combining both priors (model 4.e), we see even further performance improvements, which show the importance of incorporating *multiple* informative priors for the self-supervised sound localization objective.

We further explore the effects of freezing the vision encoder during training. As previously mentioned, a network pretrained on a classification task such as ImageNet will often have high activations around salient objects (an object-centric prior). When training in a self-supervised setting, the network may divert from its original weights and instead have a less object-centric focus, which may be sub-optimal for sound source localization. When freezing the network in the non-flow setting (model 4.c), we see performance slightly decrease compared to the unfrozen counterpart (model 4.b). However, when freezing the network in the optical flow setting (model 4.f), we see a slight improvement over the flow setting with an unfrozen vision encoder (model 4.e). We infer that enforcing the vision encoder to keep its object-centric characteristics while the flow encoder can reason and attend towards other parts of the image produces a more informative representation, improving localization performance.

Finally, we explore variations of the optical flow encoder to better understand how the optical flow information is being used. We replace the learnable ResNet18 encoder with a single max pooling layer to see if the simple presence of movement is still informative for localizing sounds. As shown in Table 5, when using a simple max pooling layer (model 5.a), we still notice a significant performance improvement over the network without optical flow (models 4.a-c). However, we see a further improvement over the

Model	Flow Network	Training Set	cIoU$_{0.5}$	AUC$_{cIoU}$
a	MaxPool	VGGS 10k	0.379	0.393
b	**ResNet18**		**0.393**	**0.398**
c	MaxPool	VGGS 144k	0.381	0.393
d	**ResNet18**		**0.394**	**0.400**

Table 5. Ablation study on configurations of the flow encoder network across training settings. All models are tested on the VGG Sound Source test set.

max pooling layer when using a learnable encoder, like a ResNet18 network. While the max pooling layer only captures the presence of movement at a particular location, a learnable encoder allows deeper reasoning of the flow information. For example, the eighth column in Figure 3 shows an optical flow field where the sounding object (tractor) is not moving but rather the environment around it is. In this case, with the max pooling encoder, the network is biased away from the sounding object, whereas a learnable encoder can better reason about the flow in the given frame, improving overall localization performance.

5. Conclusion

In this work, we introduce a novel self-supervised sound source localization method that uses optical flow to aid in the localization of a sounding object in a frame of a video. In a video, moving objects are often the ones making the sound. We take advantage of this observation by using optical flow as a prior for the self-supervised learning setting. We formulate the self-supervised objective and describe the cross-attention mechanism of optical flow over the corresponding video frame. We evaluate our approach on standardized datasets and compare against prior works and show state-of-the-art results across all experiments and evaluations. Further, we conduct extensive ablation studies to show the necessity and effect of including informative priors like optical flow, into the self-supervised sound localization objective to improve performance.

While we explore optical flow in this work, there are other priors that may be explored to improve sound source localization further. For example, pretraining the audio encoder can likely provide a better understanding of the *class* of the sound being emitted, which can then be used to help localize toward that specific object. Further, improving the optical flow generation, for example, using flow estimation methods or aggregating flow across multiple frames, can potentially improve the optical flow signal to ultimately improve overall localization performance. We leave exploration of these hypotheses for future work.

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

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision*, pages 208–224. Springer, 2020.

[2] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021.

[3] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018.

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

[5] M. J. Black and P. Anandan. A framework for the robust estimation of optical flow. In *Fourth International Conf. on Computer Vision, ICCV-93*, pages 231–236, Berlin, Germany, May 1993.

[6] 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.

[7] 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.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer, 2003.

[11] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. *CoRR*, abs/1504.06852, 2015.

[12] John W Fisher III, Trevor Darrell, William Freeman, and Paul Viola. Learning joint statistical models for audio-visual fusion and segregation. *Advances in neural information processing systems*, 13, 2000.

[13] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020.

[14] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[15] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[17] John Hershey and Javier Movellan. Audio vision: Using audio-visual synchrony to locate sounds. *Advances in neural information processing systems*, 12, 1999.

[18] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artif. Intell.*, 17(1–3):185–203, aug 1981.

[19] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9248–9257, 2019.

[20] Hamid Izadinia, Imran Saleemi, and Mubarak Shah. Multimodal analysis for identification and segmentation of moving-sounding objects. *IEEE Transactions on Multimedia*, 15(2):378–390, 2012.

[21] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, page 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc.

[22] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv:2112.12750*, 2021.

[23] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.

[24] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[25] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *European Conference on Computer Vision*, pages 292–308. Springer, 2020.

[26] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[28] 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.

[29] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Learning sound localization better from semantically similar samples. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4863–4867. IEEE, 2022.

[30] 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.

[31] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l1 optical flow. In *Proceedings 29th DAGM Symposium "Pattern Recognition"*, pages 214–223. Springer, 2007. 29th DAGM Symposium on Pattern Recognition : DAGM 2007 ; Conference date: 12-09-2007 Through 14-09-2007.

[32] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1735–1744, 2019.

[33] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[34] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14800–14809, 2021.