**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 Flicker 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
  \[
  A_{avg} = GAP(f_a) \\
  S = \sum_{i \in [1, m * n]} \frac{f_{visual} \cdot A_{avg}}{||f_{visual}||_1 \cdot ||A_{avg}||}
  \]

**OPTICAL FLOW CROSS-ATTENTION**

- Construct similarity matrix of visual and optical flow feature representations
  \[
  \beta = \text{softmax}(K_v \otimes Q_f / \sqrt{d})
  \]
- Create cross-attended visual and optical flow features
  \[
  E = \sum_{i \in [1, m]; j \in [1, n]} V_i^* \beta_{ij} f_j \\
  f_{atten} = f_{visual} \otimes E
  \]
- Add attended optical flow features to visual feature map and construct enhanced similarity map of audio features at each visual-flow spatial location
  \[
  f_{atten} = f_{visual} \otimes E
  \]

**SELF-SUPERVISED TRAINING**

- Threshold the similarity matrix into positive and negative pseudo masks
  \[
  PM^+ = \sigma(S_{k \rightarrow k} - \epsilon_p) / \tau \\
  PM^- = \sigma(S_{k \rightarrow k} + \epsilon_n) / \tau
  \]
- Construct positive and negative regions across samples in a batch and train with contrastive loss, like InfoNCE
  \[
  \text{Pos}_k = \frac{1}{1 - \text{PM}_k^+ (\text{PM}_k^+, S_{k \rightarrow k})} \\
  \text{Neg}_k = \frac{1}{1 - \text{PM}_k^+ (1 - \text{PM}_k^+, S_{k \rightarrow k}) + 1} \\
  L = \sum_k \left\{ \log \left( \frac{\exp(\text{Pos}_k)}{\exp(\text{Pos}_k) + \exp(\text{Neg}_k)} \right) \right\}
  \]

**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

**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 (CIFeR) and the National Science Foundation (NSF) under grant 1822190.

**REFERENCES**