Space-Time Memory Network for Sounding Object Localization in Videos

Sizhe Li*
sli96@u.rochester.edu
Yapeng Tian*
yapengtian@rochester.edu
Chenliang Xu
chenliang.xu@rochester.edu

Department of Computer Science
University of Rochester
Rochester, USA

Abstract

Leveraging temporal synchronization and association within sight and sound is an essential step towards robust localization of sounding objects. To this end, we propose a space-time memory network for sounding object localization in videos. It can simultaneously learn spatio-temporal attention over both uni-modal and cross-modal representations from audio and visual modalities. We show and analyze both quantitatively and qualitatively the effectiveness of incorporating spatio-temporal learning in localizing audio-visual objects. We demonstrate that our approach generalizes over various complex audio-visual scenes and outperforms recent state-of-the-art methods. Code and data can be found at https://sites.google.com/view/bmvc2021stm.

1 Introduction

Neurological evidence suggests that human understandings of scenes predominantly rely on the integration of visual and auditory cues [3]. As humans, we form attention mechanisms to sounding sources by leveraging the temporal, cross-modal alignments between vision and sound, where understandings of the past tell us where and what to attend to next. For computational models, although there have been several developed sound source spatial localization frameworks [21, 22, 27], how much we gain from explicitly leveraging temporal correspondence that exists naturally in both videos and audios is yet to be explored.

However, considerations of temporal coherence are required to facilitate consistent understandings in complex scenes. Imagine a person playing a guitar in front of a wall of not-in-use guitars. In order to figure out which guitar is sounding and obtain stable localization results, we must take multiple timesteps into account. Hence, it is worthwhile to explore if learning temporal cues could benefit the localization of sounding objects in videos.

To localize visual objects associated with specific sound sources, most of the previous works capture audio-visual spatial correspondence using similarities between audio and visual modalities [2, 15, 21], cross-modal attention mechanisms [25, 27], and sounding class activation mapping [22]. Nevertheless, these methods often identify sounding objects for static images, and audio-visual temporal coherence in videos is commonly ignored.

*Equal contribution.

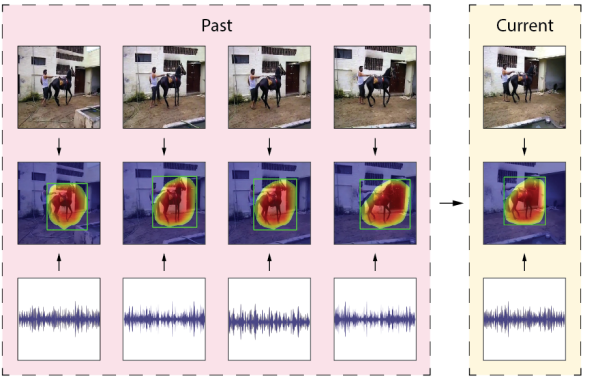


Figure 1: Our space-time memory network learns to attend to objects that currently sound by leveraging the temporal, cross-modal correspondence within sight and sound from the past. Here, given the frames of the two objects in motion and the sound of the horse walking, our model outputs stable and accurate localization results.

Consequently, the essential question we must answer is how to design an efficient multi-modal deep neural network architecture that exploits the temporal coherence in visual frames and the corresponding audio segments. In this paper, we propose a spatio-temporal attention-based memory module, that can learn rich reference information from uni-modal as well as cross-modal (audio-visual) representations. With temporal memory updates, our approach is more robust against appearance and acoustic changes than the previous methods. It yields more temporally consistent localization results and can handle the absence of audio-visual events. In particular, to demonstrate the values of multi-modal temporal learning in sounding object localization, we resort to an easily affordable, weakly-supervised task in classifying the audio-visual event category of a given video segment.

Herein, our main contributions are: (1) we propose a novel space-time memory network that learns representations of sounding objects to promote robust localization performance, as illustrated in Figure 1; (2) we validate the effectiveness of temporal learning in localizing sounding objects both quantitatively and qualitatively based on numerical benchmarks and visual interpretations; (3) we demonstrate that our approach generalizes over various complex audio-visual events and outperforms recent state-of-the-art methods.

2 Related Work

2.1 Sounding Object Localization

Sounding object localization refers to the task of localizing visual objects/scenes associated with specific sounds in videos. Early works resorted to mutual information [13] and canonical correlation analysis [18] to perform localization and segmentation on sounding pixels. Recent efforts have learned deep audio-visual models to localize sounding pixels, using audio-visual embedding similarities [2, 15, 21], cross-modal attention mechanisms [25, 27], vision-to-sound knowledge transfer [6], sounding class activation mapping [16, 22], and sounding object visual grounding [30]. While these methods work well on a single sound source in the simple audio-visual scenes, they lack temporal knowledge and predict audible regions solely based on the association of the current video frame with the corresponding

audio segment. Most recently, Afouras et al. [1] compute audio-visual cross-modal attention to spatially localize sounding regions. Moreover, they incorporate temporal learning into the visual modality, where they propose to use optical flow that is separately learned to aggregate information over time and group sound sources into audio-visual objects. Their model also relies on speech-oriented tasks and scenes, assuming objects (speakers) of fixed size. By contrast, not only does our spatio-temporal attention mechanism consider both uni-modal and cross-modal representations, but it is also learned in an end-to-end manner. Hence, different from the previous methods, with learnable space-time memory modules, our model can effectively leverage multi-modal contexts for localizing sounding objects and thus is capable of handling diverse and complex audio-visual objects.

2.2 Audio-Visual Video Understanding

The community has attracted an increasing amount of interest in recent years since synchronized audio-visual scenes are widely available in videos. In addition to localizing sound sources, a wide range of tasks have been proposed, including audio-visual sound separation [7, 9, 26, 34, 35], audio-visual action recognition [10, 17, 19, 30], audio-visual event localization [27, 33], audio-visual video captioning [23, 28, 32], embodied audio-visual navigation [4, 8], audio-visual sound recognition [5], and audio-visual video parsing [29]. Our framework demonstrates that temporal learning facilitates better audio-visual understanding, which explicitly and subsequently benefits the localization performance.

3 Proposed Method

Our framework (see Figure 2) consists of three modules: audio and visual feature extraction, memory construction and propagation, and sounding object localization. Given a video, it learns to attend to objects that sound in video frames. The spatio-temporal attention mechanisms are designed to leverage both uni-modal and cross-modal representations. It is supported by the idea of accumulating the past evidence into memory, which is then aggregated and propagated onto the current timestep. Through the task of audio-visual event classification, our model facilitates audio-visual understanding by learning spatio-temporal attention mechanisms that locate sounding objects.

3.1 Audio-Visual Feature Extraction

Consider T audio-visual pairs $\{X_v^t, X_a^t\}_{t=1}^T$ as inputs, where $X_v^t \in \mathbb{R}^{H\times W\times 3}, X_a^t \in \mathbb{R}^{M\times N\times 1}$ denote the frame and its corresponding audio log-mel spectrogram at timestep t respectively. Let H, W, M, N denote the height, width, frequency, and time. For a given pair, we employ two convolutional encoders to project uni-modal features into C-dimensional joint audio-visual subspace for spatial-temporal attention in memory. Let h, w and m, n denote the corresponding spatial dimensions of the feature maps. We obtain visual feature map $x_v^t \in \mathbb{R}^{h\times w\times C}$ and audio feature map $x_a^t \in \mathbb{R}^{m\times n\times C}$. This is reflected by steps 1 to 2 in Figure 2.

3.2 Audio-Visual Memory Accumulation and Propagation

To efficiently accumulate and aggregate the temporal audio-visual evidence, we propose to build memory modules for audio and visual representations both separately and jointly,

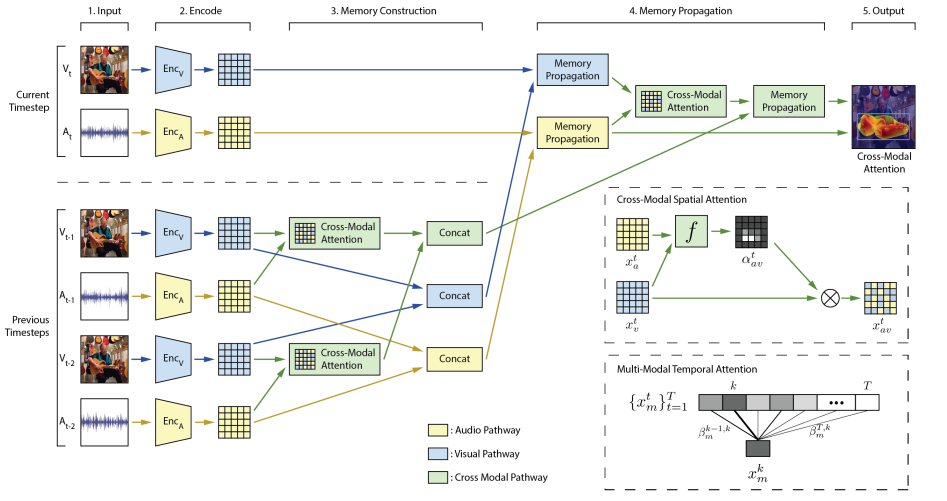


Figure 2: **Overview of the Space-Time Memory Network**: (1-2) extracts audio and visual features from inputs. (3) constructs one cross-modal memory module and two uni-modal memory modules. (4) first propagates the uni-modal memory, then computes cross-modal attention, and propagates the cross-modal memory. (5) computes spatial attention from the memory-propagated audio features to the memory-propagated audio-visual features. The outputs is an attention map that localizes sounding objects, which is used for downstream audio-visual event classification during training.

which we collectively refer to as the *Spatio-Temporal Memory Layer*. Here, we introduce the attention mechanisms in space and time, then describe how they form the memory layer.

Cross-Modal Spatial Attention f**.** We measure the similarity between the extracted uni-modal features x_v^t and x_a^t at the given timestep t. We denote the cross-modal spatial attention function as $f(x_a^t, x_v^t) = \alpha_{av}^t$, such that $f : \mathbb{R}^{m\times n\times C} \times \mathbb{R}^{h\times w\times C} \mapsto \mathbb{R}^{h\times w}$. For each position (i, j) in x_v^t, the attention weight can be computed by:

$$\alpha_{av}^t(i,j) = \frac{exp(x_v^t(i,j)\phi(x_a^t)^\intercal)}{\sum_{i,j} exp(x_v^t(i,j)\phi(x_a^t)^\intercal)}. \tag{1}$$

Here, we adopt dot product as a generic choice to compute the spatial similarity and $\phi(\cdot)$ is a global pooling operation over the spatial dimension of its input. We compute the cross-modal audio-visual features $x_{av}^t \in \mathbb{R}^{h\times w\times C}$ by multiplying the learned spatial attention map with every channel $c \in \{1,...,C\}$ in the visual feature map. Specifically, $x_{av}^t(i,j,c) = \alpha_{av}^t(i,j) * x_v^t(i,j,c)$. For every pixel in a given frame X_v^t, we have thus found its affinity with the aligned audio input X_a^t via cross-modal spatial attention f on the feature level.

Multi-Modal Temporal Attention g**.** Our memory formulation is generic and can be adapted to both uni-modal and cross-modal representations. Consider the modality $m \in \{a, v, av\}$ and the previous timestep $k \in \{1,...,K\}$. We seek to measure the importance of x_m^k with respect to x_m^t from the current timestep. We denote the multi-modal temporal attention function as $g(x_m^t, x_m^k) = \beta_m^{k,t}$, where h_m, w_m denote the spatial dimensions of the feature maps of modality m, such that $g : \mathbb{R}^{h_m\times w_m\times C} \times \mathbb{R}^{h_m\times w_m\times C} \mapsto \mathbb{R}$.

We adopt the generic multi-head scaled dot-product attention from Vaswani et al. [31],

where we view x_m^t as queries and x_m^k as keys and values. For each of the L attention heads, we globally pool the input feature maps into feature vectors, and compute the set of attention weights over K timesteps. Concretely, the multi-modal temporal attention function is formulated as:

$$\beta_m^{k,t} = \frac{1}{L}\sum_{l=1}^{L} A(\phi(x_m^t)W_l^{query}, \phi(x_m^k)W_l^{key}) \tag{2}$$

where $A(I_{query}, I_{key}) = softmax(\frac{I_{query}I_{key}^T}{\sqrt{C}})$ denotes the attention function, C corresponds to the number of dimensions for keys and queries, and W_l^{query}, W_l^{key} are the learnable projections.

Having computed the set of attention weights, we propagate the memory to the current timestep t to obtain $\dot{x}_m^t = \beta_m^{t,t} * x_m^t + \sum_{k=1}^{K} \beta_m^{k,t} * x_m^k$. Hence, the multi-modal temporal attention takes into account the importance of the past k to the present t.

Spatio-Temporal Memory Layer. We now describe the algorithm that combines the two proposed attention mechanisms in space and time. Our memory layer first leverages the uni-modal association in time by applying the temporal attention g to x_a^t, x_v^t, respectively. It then considers the cross-modal association in space by applying the spatial attention f to $\dot{x}_a^t, \dot{x}_v^t$. Finally, the memory layer derives the cross-modal association in time by applying the temporal attention g to $\dot{x}_{av}^t$. This is reflected by steps 3 to 4 in Figure 2 and Algorithm 1.

Algorithm 1 Spatio-Temporal Memory Layer Pseudocode

1: **procedure** MEMORYLAYERFORWARD($x_a^t, x_v^t, \{(x_a^k, x_v^k, x_{av}^k)\}_{k=1}^K$)
2: $\quad \dot{x}_a^t = TemporalAttention(x_a^t, \{x_a^k\}_{k=1}^K)$ ▷ Uni-Modal Temporal Attention
3: $\quad \dot{x}_v^t = TemporalAttention(x_v^t, \{x_v^k\}_{k=1}^K)$ ▷ Uni-Modal Temporal Attention
4: $\quad \dot{x}_{av}^t = SpatialAttention(\dot{x}_a^t, \dot{x}_v^t)$ ▷ Cross-Modal Spatial Attention
5: $\quad \ddot{x}_{av}^t = TemporalAttention(\dot{x}_{av}^t, \{x_{av}^k\}_{k=1}^K)$ ▷ Cross-Modal Temporal Attention
6: $\quad$ **return** $\ddot{x}_{av}^t, \dot{x}_a^t$
7: **end procedure**

3.3 Localizing sounding objects

To extract audio-visual objects from various audio-visual events, we cannot impose an assumption on the sizes of the objects. We propose two post-processing approaches, using contour detection and pre-trained object-proposal-networks, respectively.

Given outputs $\dot{x}_a^t, \ddot{x}_{av}^t$ from the memory layer, we compute the cross-modal spatial attention map $\dot{\alpha}_{av}^t = f(\dot{x}_a^t, \ddot{x}_{av}^t)$, which we view as the final sounding object localization map. This is reflected by step 5 in Figure 2. We generate bounding boxes by applying Otsu's threshold [20] and contour detection to the normalized spatial attention map. Alternatively, the attention map $\dot{\alpha}_{av}^t$ can also be incorporated into robust object instances generated by out-of-the-box object proposal methods. Given frame X_v^t, we extract N object instances using a region-proposal-network (RPN). We convert them into binary masks $\{m_n^t\}_{n=1}^N$, with 1 indicating the instance and 0 otherwise. We calculate the individual score of each box S_n^t as the weighted sum between m_n^t and $\dot{\alpha}_{av}^t$, or $S_n^t = \sum_{i,j} m_n^t(i,j) * \dot{\alpha}_{av}^t(i,j)$, and apply non-maximum suppression (NMS) to filter overlapping boxes.

3.4 Learning Spatio-Temporal Attention Mechanisms

We utilize the easily affordable, weakly-supervised classification task on audio-visual event categories to learn the proposed spatio-temporal attention mechanisms. We fuse the outputs from the memory layer, including the uni-modal audio feature vector $\dot{x}_a^t$ and the cross-modal audio-visual feature map $\dot{x}_{av}^t$, to obtain a joint representation. In particular, we sum $\dot{x}_{av}^t$ over its spatial dimensions, since it is already weighted by the spatial attention f, and concatenate the result with $\dot{x}_a^t$. This gives us the final output of our network at timestep t, denoted as $\mathcal{O}^t \in \mathbb{R}^{2C}$. Concretely, $\mathcal{O}^t = [\sum_{i,j} \dot{x}_{av}^t(i,j); \dot{x}_a^t]$, where $[...;...]$ denotes concatenation. This joint audio-visual representation is used to estimate the audio-visual event category for the given video segment using a multilayer-perceptron (MLP) and the cross-entropy loss.

4 Experiments

4.1 Datasets

To carefully and temporally evaluate the localization performance, we emphasize three aspects when tailoring dataset: the scope of the sounding object categories covered, the scale of the testing examples contained, and the number of frames per video densely annotated. Therefore, we use AVE dataset [27] for scope and AudioSet-Instrument dataset [11] for scale. For both datasets, we densely annotated frames in the test videos.

AVE: Audio-Visual Event Dataset. AVE dataset [27] consists of 4143 10-second video clips that are labeled with the audio-visual event, covering 28 categories. Although the dataset is also temporally labeled with audio-visual event boundaries, to demonstrate the effectiveness of our weakly-supervised learning framework, we include the whole length of every video. We adopt the original data split, including 3339 videos for training, 402 for validation, and 402 for testing.

AudioSet-Instrument. AudioSet-Instrument dataset is a subset of AudioSet [11] that consists of 101076 10-second video clips, spanning across 15 instrument categories. It contains challenging video clips, where many are of poor quality with multiple sound sources. We use 100221/424/424 for training/validation/testing.

Annotations. We annotated test sets in AVE and AudioSet-Instrument, where we created bounding boxes for sounding objects in the duration of the audio-visual event. For reproducibility, we will release the annotations with source code.

4.2 Implementation Details

Data Preprocessing. During training, we randomly sample one second and extract its corresponding audio-visual pair $\{(X_v^t, X_a^t)\}_{t=1}^T$, where T denotes the number of frames, as well as the number of audio segments, we sample per second. In practice, we use $T = 4$. During the evaluation, T corresponds to the number of frames that are annotated in the given video. Final input frames are of spatial dimensions 256×256. During training, this is achieved by resizing the image by a scale of 1.1 and randomly crop it to the desired size. During the evaluation, every frame is directly resized to the target dimensions. Audio inputs are re-sampled to 16 kHz mono and its corresponding spectrogram is computed through Short-Time Fourier Transform with a Hann window size of 25ms and a hop length of 10ms. For every input frame, we use the output spectrogram of its temporally-aligned audio segment.

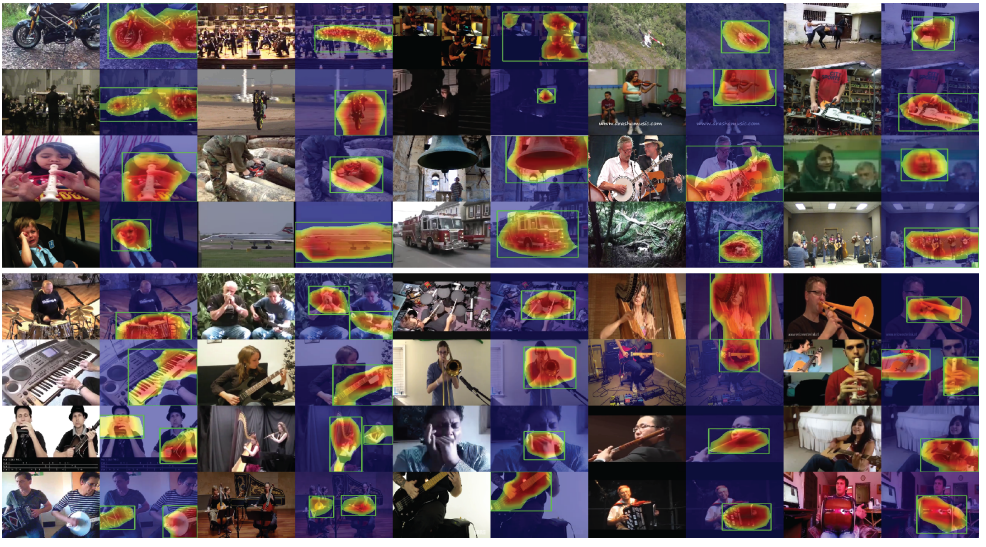


Figure 3: **Qualitative Visualizations**: We show localization results from the two datasets. Top four rows contain results from the AVE dataset. Bottom four rows contain results from the Audioset-Instrument dataset. Bounding boxes are extracted from the cross-modal spatial attention maps, using Otsu's threshold and contour detection.

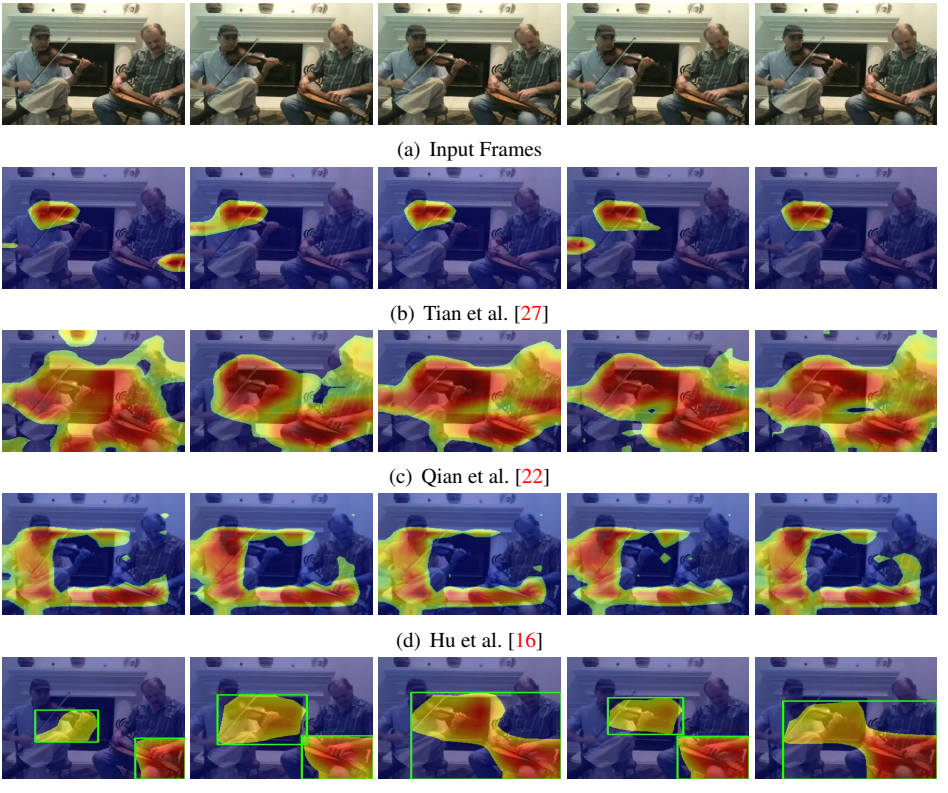
Encoder. We use ResNet-152 [12] to extract 2048-D per-frame visual features and VGGish [14] to extract 512-D per-segment audio features, they are then each projected to a 256-D joint feature space by a stack of 1x1 convolutions with ReLU non-linearity.

Memory Module. We use Multi-head scaled dot product attention with 256-D embedding dimension and 1 head.

Sounding Object Extraction. To localize objects that sound from the cross-modal spatial attention map, we use bilinear interpolation to resize to 256×256. We then normalize the attention map by subtracting the minimum and dividing by the maximum, resulting in a lower bound of 0 and an upper bound of 1. We use Otsu's threshold and contour detection to extract bounding boxes for experimental comparison and ablation study. For the RPN-based extraction method demonstrated in Figure 7, we used a Faster R-CNN model with a ResNet-50-FPN backbone [12, 24], pretrained on COCO train2017.

Hyperparameters. In our experiments, we use batch size 128 and 30 epochs. Our framework is trained with Adam optimizer with the initial learning rate of 10^{-4} on four NVIDIA 1080Ti GPU.

Evaluation Metrics. To evaluate the localization performance of our framework, we employ Intersection over Union (IoU) for the box-level localization performance and Consensus Intersection over Union (cIoU) proposed in [25] for the pixel-level localization performance. We now expand on details of cIoU calculations. Given an annotated frame, we convert the ground truth bounding boxes $\{b_j\}_{j=1}^{N}$ into a binary ground truth map $\mathbf{g}$, where 1 indicates that a pixel is sounding and 0 otherwise. Given predicted location map α, we define cIoU under threshold τ as: $cIoU(\tau) = \frac{\sum_{i \in \mathcal{A}(\tau)} \mathbf{g}_i}{\sum_i \mathbf{g}_i + \sum_{i \in \mathcal{A}(\tau) - \mathcal{G}} 1}$. Here, i is the pixel index of the map. $\mathcal{A}(\tau) = \{i | \alpha_i > \tau\}$ denotes the set of pixels with attention intensity higher than the threshold, and $\mathcal{G} = \{i | \mathbf{g}_i > 0\}$ represents the set of pixels annotated as positive. We use 0.5 as the cIoU



(a) Input Frames

(b) Tian et al. [27]

(c) Qian et al. [22]

(d) Hu et al. [16]

(e) Our model

Figure 4: **Experimental Comparison**: We qualitatively compare the localization performance of our framework with two recent methods by Tian et al. [27] and Qian et al. [22].

threshold in our experiments.

4.3 Results

In Figure 3, we illustrate sounding object localization results on AVE and Audioset-Instrument datasets. We observe that our model is capable of correctly discovering sounding regions for a wide range of sound sources in challenging unconstrained videos.

Experimental Comparison. To further validate the effectiveness of our space-time memory network, we compare it with three recent methods: Owens and Efros [21], Tian et al. [27], Qian et al. [22], and Hu et al. [16][1]. We demonstrate the quantitative results in Table 1 and the qualitative results in Figure 4[2]. We find that our framework outperforms the compared approaches on AVE and Audioset-Instrument datasets both quantitatively and qualitatively, which substantiates the benefits of the proposed space-time memory network in localizing dynamic audio-visual objects.

Ablation Study. To evaluate the effectiveness of our proposed memory module, we conduct

[1] Afouras et al. [1] was not compared since their framework is trained on speech-oriented downstream tasks and imposes an assumption on the size of the sounding object.

[2] More results can be found in our supplementary material.

Method	AVE		Audioset-Instrument	
	cIoU@0.5	IoU@0.5	cIoU@0.5	IoU@0.5
Tian et al. [27]	11.73	18.81	25.11	33.20
Owens et al. [21]	13.96	22.64	21.79	38.08
Qian et al. [22]	24.16	16.82	20.16	31.31
Hu et al. [16]	21.25	33.05	33.74	40.02
Ours	**37.78**	**37.50**	**51.06**	**56.49**

Table 1: Localization results of different methods on AVE and Audioset-Instrument datasets. All methods are evaluated by IoU@0.5 and cIoU@0.5. The top-1 result in each column is highlighted.

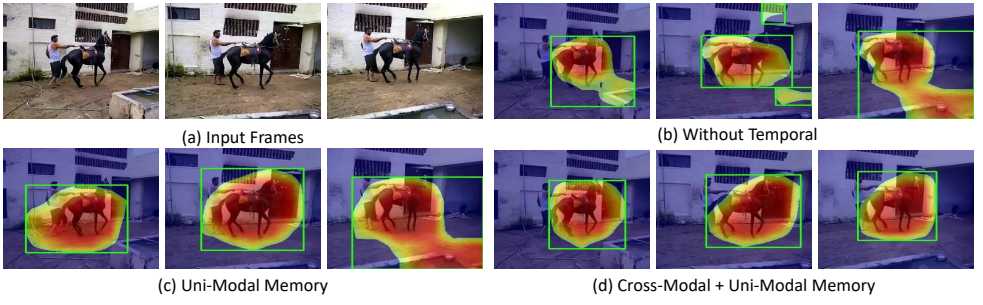


Figure 5: **Ablation Study**: Visualizations of sounding object localization from the three ablative groups. Here, only horse is making audible sounds.

ablation study for the following models: (1) *Cross-Modal Memory + Uni-Modal Memory* (C+U): we employ both uni-modal and cross-modal temporal learning modules. (2) *Uni-Modal Memory* (U): we remove the cross-modal memory module, propagating memory on a uni-modal level only. (3) *Without Temporal*: we remove both uni-modal and cross-modal memory modules. Without temporal learning, the baseline model associates frames and sounds on single timesteps. We show the quantitative results in Table 2 and the qualitative results in Figure 5. We find that (1) outperforms the other two ablative groups numerically and demonstrates significantly more robust visualization results.

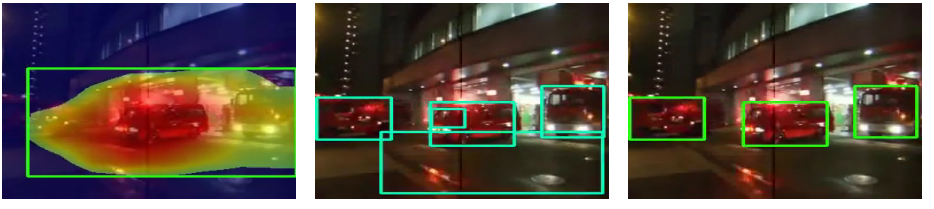
Method	AVE		Audioset-Instrument	
	cIoU@0.5	IoU@0.5	cIoU@0.5	IoU@0.5
Without Temporal	34.81	33.82	44.30	51.57
U	36.41	35.23	48.96	53.19
C+U	**37.78**	**37.50**	**51.06**	**56.49**

Table 2: **Ablation Study**: Localization results of the ablative groups on AVE and Audioset-Instrument datasets. All methods are evaluated by IoU@0.5 and cIoU@0.5. The top-1 result in each column is highlighted.

Handling the absence of audio-visual events. Given that not all audio-visual segments contain audio-visual events, we further demonstrate the robustness of our model in a challenging example in Figure 6. While the cello performer has lifted her bow up, it is impossible to tell whether the flute is sounding merely from sight. Our framework identifies the non-sounding frames by resorting to sound.

Figure 6: **Handling the absence of audio-visual events:** In this challenging example (top), both performers stop playing between the third and the fourth steps, and resume on the fifth step. We show our model performance (bottom), where background class is predicted following an absence of audio-visual events.

Audio-Visual Object Grounding. Following the RPN-based approach to extract sounding objects, we demonstrate how our cross-modal attention map can be incorporated to further refine localization performance in Figure 7.



(a) Attention Output (b) Box Proposals (c) AV Grounding

Figure 7: **Audio-Visual Object Grounding:** Using a pretrained RPN, our grounding approach can further refine the localization performance of our framework. Here, we show (a) the final model output followed by contour detection, (b) the extracted box proposals, and (c) the audio-visual object grounding results.

5 Conclusion

In this paper, we investigate the effectiveness of multi-modal temporal learning in localizing audio-visual objects. We propose a novel space-time memory framework to address the problem. Results from experimental comparison and ablation study support our claim both objectively and subjectively that multi-modal temporal learning is crucial for robust sounding object localization performance.

Acknowledgements: We would like to thank the anonymous reviewers for the constructive comments. This work was supported in part by NSF 1741472 and 1909912. The article solely reflects the opinions and conclusions of its authors but not the funding agents.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *ECCV*, 2020.

[2] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *ECCV*, 2018.

[3] Zelie Britton and Qadeer Arshad. Vestibular and multi-sensory influences upon self-motion perception and the consequences for human behavior. *Frontiers in Neurology*, 2019.

[4] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigaton in 3d environments. In *ECCV*, 2020.

[5] Haytham M. Fayek and Anurag Kumar. Large scale audiovisual learning of sounds with weakly labeled data. In *IJCAI*, 2020.

[6] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *ICCV*, 2019.

[7] Chuang Gan, Deng Huang, Hang Zhao, Joshua B. Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *CVPR*, 2020.

[8] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and J. Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. *ICRA*, 2020.

[9] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, 2018.

[10] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *CVPR*, 2020.

[11] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[13] John Hershey and Javier Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *NIPS*, 2000.

[14] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson. Cnn architectures for large-scale audio classification. In *ICASSP*, 2017.

[15] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *CVPR*, 2019.

[16] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. In *NeurIPS*, 2020.

[17] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *ICCV*, 2019.

[18] E. Kidron, Y. Y. Schechner, and M. Elad. Pixels that sound. In *CVPR*, 2005.

[19] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *ICCV*, 2019.

[20] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE SMC*, 1979.

[21] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018.

[22] Rui Qian, Heinrich Dinkel Di Hu, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *ECCV*, 2020.

[23] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *ICCV*, 2019.

[24] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*, 2017.

[25] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, 2018.

[26] Yapeng Tian and Chenliang Xu. Can audio-visual integration strengthen robustness under multimodal attacks? In *CVPR*, June 2021.

[27] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018.

[28] Yapeng Tian, Chenxiao Guan, Goodman Justin, Marc Moore, and Chenliang Xu. Audio-visual interpretable and controllable video captioning. In *CVPR Workshops*, 2019.

[29] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *ECCV*, 2020.

[30] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In *CVPR*, June 2021.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[32] Xin Wang, Yuan-Fang Wang, and William Yang Wang. Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning. In *NAACL-HLT*, 2018.

[33] Y. Wu, L. Zhu, Y. Yan, and Y. Yang. Dual attention matching for audio-visual event localization. In *ICCV*, 2019.

[34] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018.

[35] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *ICCV*, 2019.