Audio-Visual Object Localization in Egocentric Videos

Chao Huang[1], Yapeng Tian[1], Anurag Kumar[2], Chenliang Xu[1]
[1]University of Rochester [2]Meta Reality Labs Research

1. Introduction

In the past decade, substantial efforts have been made to build benchmarks, establish new tasks, and create learning frameworks for egocentric video understanding. However, the majority of works focus on visual scene analysis. It presents significant challenges for designing powerful and robust egocentric video understanding systems. First, people with wearable devices usually record videos in naturalistic surroundings, where a variety of illumination conditions, object appearance, and motion patterns are shown. The dynamic visual variations introduce difficulties in achieving accurate visual perception. Second, egocentric scenes are often perceived within a limited field of view (FoV). The common body and head movements of users cause frequent view changes, which brings object deformation and creates dynamic out-of-scene content. Therefore, it is difficult to fully decode the surrounding information and perceive dynamic scenes in egocentric videos.

Audio, acting as an essential but less focused modality, often provides synchronized and complementary information with the video stream. In contrast to the variability of first-person visual footage, sound provides stable and persistent signals associated with the depicted events. Therefore, audio is an essential ingredient for "cooking" egocentric video understanding.

To capture fine-grained audio-visual association and tackle the challenges in egocentric vision, in this paper, we propose to solve an egocentric audio-visual object localization task, which aims to associate audio with dynamic visual scenes and localize sounding objects in egocentric videos. We develop a new framework to explicitly model the distinct characteristics of egocentric videos by integrating audio. Concretely, the egomotion in videos leads to various object deformations, making it difficult to consistently localize the sounding objects. Despite the downside, we found that the egomotion also provides rich geometry information about the underlying scene and hints at the relative geometric transformation between frames. Motivated by this, we use the predicted geometric transformation to mitigate the object deformation in the embedding space and align the visual features. Based on the aligned features, we further leverage the temporal contexts across frames to learn discriminative cues for localization. Our localization model aims to find precise audio-visual associations. However, the results might be affected by: (i) noisy audio with background sound components; (ii) unrelated visual contents, *e.g.*, other silent objects. To this end, we propose a cascaded feature enhancement module to mitigate the audio noise and improve cross-modal localization robustness. Our framework can be trained by taking nature audio-visual temporal synchronization as the "free" supervision. We also annotate an *Epic Sounding* dataset to facilitate quantitative comparison. In all, our contributions are: (i) an effective geometry-aware temporal aggregation approach to deal with unique egomotion in first-person videos; (ii) a novel cascaded feature enhancement module to progressively inject the audio and visual features with localization cues; and (iii) an *Epic Sounding Object* dataset with annotations on sounding objects to benchmark the localization performance in egocentric videos.

2. Problem and Proposed Framework

Given an egocentric video clip V containing T frames $I = \{I_i\}_{i=1}^{T}$ and its synchronized sound stream $s = \sum_{n=1}^{N} s_n$, $\mathcal{O} = \{O_i\}_{i=1}^{T}$ are sounding objectness maps that indicate locations of audible objects in the video frames. Here, s_n is the n-th sound source in the audio track. Note that there could be multiple sound sources mixed together ($N \geq 1$) and not all of them associate with visual objects. The task aims to predict each map O_i from the input frames I, and the audio s. The motivation is to capture audio associated visual objects and mitigate potential audio noises from audio-visual irrelevant sound sources. Since egocentric videos have a limited FoV, out-of-screen sound sources and egomotion originated from dramatic view changes are ubiquitous. These characteristics of egocentric video data make the problem very challenging.

To solve the egocentric audio-visual object localization task, we propose a new framework as shown in Figure 1. Our model first extracts representations from the audio s and sampled frames I, and then performs cascaded feature enhancement on both audio and visual branches. The training of audio enhancement network is driven by mix-and-separate strategy [4, 7, 8], targeting to disentangle visually

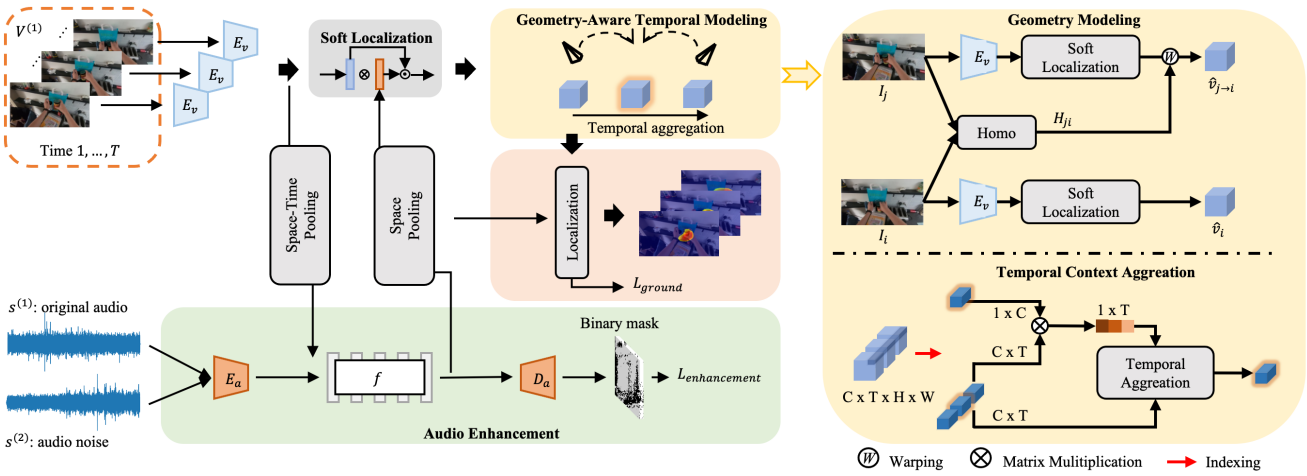

Figure 1. An overview of our egocentric audio-visual object localization framework. In the beginning, our model extracts deep features from the video and audio streams. Then the audio and visual features are fed into the cascaded feature enhancement module to inject localization cues for both branches. Such module is additionally trained with "mix-and-separation" strategy. Next, our geometric-aware temporal modeling block leverages the relative geometric information between visual frames and performs alignment based temporal context aggregation to get the final visual features for localization.

indicated sound sources. Benefiting from the enhanced audio features, the sounding-irrelevant visual features can be reweighted by means of audio-guided cross-modal attention. To deal with the egomotion in egocentric videos, our model estimates the homography transformation between the frames, and then apply it to align frame-level features and aggregate temporal contexts. During training, we take the audio-visual temporal synchronization as the supervision signal and use audio-guided cross-modal attention to learn the map O_i. The network can be optimized in terms of computed maps from sampled positive and negative audio-visual pairs with Contrastive Learning. To further leverage the temporal contexts in I, we solve the training in a Multiple Instance Learning (MIL) setting.

2.1. Feature extraction

Visual representations. As shown in Figure 1, we first use a shared visual encoder network E_v to extract visual feature maps for each input frame I_i. We adopt a pre-trained Dilated ResNet model and remove the final fully-connected layers. We can subsequently obtain a group of feature maps $v = \{v_i\}_{i=1}^{T}$, where $v_i \in \mathbb{R}^{c \quad h_v \quad w_v}$. Here c is the number of channels and $h_v \times w_v$ denotes the spatial content.

Audio representations. To extract audio representations from raw waveform, we first transform audio stream s into a magnitude spectrogram X with the short-time Fourier transform (STFT). Then, we can compute audio features by means of CNN encoder E_a in the Time-Frequency (T-F) space. The resultant audio features $a = E_a(s), a \in \mathbb{R}^{c \quad h_a \quad w_a}$. c is set as 512 in our experiment.

2.2. Cascaded feature enhancement

In practice, there could be sound sources, s_n, which are out-of-screen in egocentric videos due to the limited FoV. For instance, *frying sound* and *human speech* may simultaneously be captured, while only the visual object of frying sound is presented in the scene. In this case, these additional sound sources are essentially noise and can corrupt subsequent audio representations. We propose to mitigate these acoustic noises by disentangling visually guided audio representations from the input audio.

Audio feature enhancement. Our goal here is to separate the noisy components from the audio features. However, the final localization objective cannot provide direct supervision to guide the disentanglement. Inspired by the audio-visual source separation works [4, 7, 8], we adopt the commonly used "mix-and-separate" idea to provide additional disentanglement supervision.

Given the current audio as $s^{(1)}$, we randomly sample another audio stream $s^{(2)}$ from a different video and mix them together to obtain an input audio signal $s = s^{(1)} + s^{(2)}$. We can then obtain a mixed spectrogram X and the other two original magnitude spectrograms $X^{(1)}$ and $X^{(2)}$, respectively. We define the audio feature enhancement as a function $f(\cdot, \cdot; \theta_1)$, which takes the mixed spectrogram X and visual feature vector as input. The output audio feature of f should be disentangled from audio noise. In the module, we apply spatial average pooling and temporal max pooling on visual feature maps v to obtain a visual feature vector $g_v \in \mathbb{R}^c$. Then we replicate this feature vector $h_a \times w_a$ times and tile them to match the shape of audio features. Lastly, the audio feature enhancement is presented as $\hat{a} = f(a, g_v; \theta_1)$. In practice, we implement the disentanglement network f as

a two-layer MLP. Since the additive audio signal is known, it is natural to supervise the training of f by solving a spectrogram mask generation task. Concretely, we add a task decoder D_a following the disentanglement network to output a binary mask M_{pred} [8] (as shown in Figure 1). Note that we use a U-NET style network [4] with skip connection for effective predictions. The value of ground truth mask M_{gt} is calculated by determining whether the original input sound is dominant at locations (u, v) in the T-F space. In this case, we compute the per-pixel $L2$ loss and therefore the enhancement learning objective can be written as $\mathcal{L}_{enhancement} = ||M_{pred} - M_{gt}||_2$.

Soft localization. While the audio feature is enhanced, the visual feature maps may contain sound-irrelevant regions which require further improvement. To this end, we propose to highlight the spatial regions that are more likely to be associated with the on-screen sounds. To achieve this, we apply max pooling on the spatial dimensions of $\hat{a}$, obtaining an audio feature vector $g_{\hat{a}}$. For every visual feature v_i, we compute the cosine similarity as : $S_i : S_i[x, y] = v_i[x, y] \times g_{\hat{a}}$, where both features are l_2 normalized and $\times$ denotes matrix multiplication. Softmax is used on S_i to generate a soft mask that represents the audio-visual correspondence. Hence, each v_i can be attended with the calculated weights $\hat{v}_i = Softmax(S_i) \cdot v_i$.

2.3. Geometry-Aware Temporal Modeling

The uniqueness in egocentric videos such as egomotion and distinct object appearance pose significant challenges on egocentric audio-visual object localization. However, such temporal variations in egocentric videos also reveal rich geometric cues to recover the scene from changing viewpoints. In our work, we leverage egomotion to estimate the relative geometric transformation between frames, and then apply the transformation at the feature-level to perform geometry-aware temporal aggregation (as shown in Figure 1). Given $\{I_i\}_{i=1}^T$ and their features $\{\hat{v}_i\}_{i=1}^T$, we take a $\hat{v}_i$ as query each time and use the others as supports to aggregate temporal contexts. For clarity, we decompose the geometry-aware temporal aggregation into two parts:

Geometry modeling. Our objective in this step is to compute the geometric transformation that represents the egomotion between frames. We found that homography estimation, which can align images taken from different perspectives, can be approximately served as the way to measure geometric transformation. Specifically, we adopt a commonly used idea that combines SIFT + RANSAC to solve homography. With the query frame I_i and a supporting frame I_j, we use $h(\cdot)$ to denote the entire process $\mathcal{H}_{ji} = h(I_j, I_i)_{j \to i}$, where $\mathcal{H}_{ji}$ represents the homography transformation from frame I_j to I_i. With the computed homography transformation, we can then transform visual features $\hat{v}_j$ to $\hat{v}_{ji}$, which should be under the viewpoint of I_i. Note that since the resolution of feature maps is scaled down compared to the raw frame size, the homography matrix $\mathcal{H}$ should also be downsampled using the same scaling factor. The feature transformation can be written as: $\hat{v}_{ji} = \mathcal{H}_{ji} \otimes \hat{v}_j$, here $\otimes$ represents the warping operation.

Temporal aggregation. For the query feature $\hat{v}_i$, we now have a set of aligned features $\{\hat{v}_{ji}\}_{j=1}^T$. To aggregate the temporal contexts, we propose to compute the correlation between features at the same locations but from different frames. Concretely, we implement the aggregation as a temporal attention network (see Figure 1):

$$z_i[x, y] = \hat{v}_i[x, y] + Softmax(\frac{\hat{v}_i[x, y]\hat{\boldsymbol{v}}[x, y]^T}{\sqrt{d}})\hat{\boldsymbol{v}}[x, y], \tag{1}$$

where $\hat{\boldsymbol{v}} = [\hat{v}_{1i}; ...; \hat{v}_{Ti}]$ is the concatenation of frame features. The scaling factor d equals the feature dimension and $(\cdot)^T$ represents the transpose operation. The same aggregation is run over all the spatial locations (x, y) to get the final visual feature z_i.

2.4. Training Objective

With the audio feature vector $g_{\hat{a}}$ and the visual features $\{z_i\}_{i=1}^T$, we compute the audio-visual attention map S_i for each frame I_i. We then adopt a differential thresholding on S_i to generate pseudo sounding objectness map $m_i = sigmoid((S_i - \epsilon)/\tau)$, which represents the location of sounding objects. Here ϵ is the threshold and τ denotes the temperature that controls the sharpness.

Since multiple audio-visual pairs are provided, we can solve the localization task in MIL setting to reduce the uncertainty. Concretely, we use a softmax MIL pooling function to aggregate the concatenated attention maps $\boldsymbol{S} = [S_1; ...; S_T]$ by assigning different weights W_t to S_t at different time steps: $\overline{S} = \sum_{t=1}^T (W_t \cdot \boldsymbol{S})[:, :, t]$. In this way, for each video clip V in the batch, we can define its positive signal as $P = \frac{1}{\overline{jm}j}\langle \overline{m}, \overline{S} \rangle$ and negative training signals as $N = \frac{1}{hw}\langle \mathbf{1}, S_{neg} \rangle$. We obtain S_{neg} by associating the current visual inputs I with audios from other video clips. $\mathbf{1}$ denotes an all ones tensor with shape $h \times w$. Therefore, the localization and the overall objectives are:

$$\mathcal{L}_{ground} = -\frac{1}{N}\sum_{k=1}^{N}[\log \frac{\exp(P_k)}{\exp(P_k) + \exp(N_k)}], \tag{2}$$

$$\mathcal{L} = \mathcal{L}_{ground} + \lambda \mathcal{L}_{enhancement}. \tag{3}$$

3. Experimental Results

In this work, we train our framework on the large-scale egocentric video dataset Epic-Kitchen [3], which contains synchronized video and audio recordings and covers diverse acoustic kitchen scenarios. To facilitate quantitative comparison, we annotate an *Epic-Sounding Object* dataset and