Listen to Look into the Future: Audio-Visual Egocentric Gaze Anticipation

Bolin Lai[1], Fiona Ryan[1], Wenqi Jia[1], Miao Liu[2,*], and James M. Rehg[3,*]

[1] Georgia Institute of Technology
[2] GenAI, Meta
[3] University of Illinois Urbana-Champaign
{bolin.lai,fkryan,wenqi.jia}@gatech.edu miaoliu@meta.com
jrehg@illinois.edu

Abstract. Egocentric gaze anticipation serves as a key building block for the emerging capability of Augmented Reality. Notably, gaze behavior is driven by both visual cues and audio signals during daily activities. Motivated by this observation, we introduce the first model that leverages both the video and audio modalities for egocentric gaze anticipation. Specifically, we propose a Contrastive Spatial-Temporal Separable (CSTS) fusion approach that adopts two modules to separately capture audio-visual correlations in spatial and temporal dimensions, and applies a contrastive loss on the re-weighted audio-visual features from fusion modules for representation learning. We conduct extensive ablation studies and thorough analysis using two egocentric video datasets: Ego4D and Aria, to validate our model design. We demonstrate that audio improves the performance by +2.5% and +2.4% on the two datasets. Our model also outperforms the prior state-of-the-art methods by at least +1.9% and +1.6%. Moreover, we provide visualizations to show the gaze anticipation results and share additional insights into audio-visual representation learning. The code and data split are available on our website (https://bolinlai.github.io/CSTS-EgoGazeAnticipation/).

Keywords: Egocentric Vision · Gaze Behavior · Audio-Visual Learning

1 Introduction

A person's eye movements during their daily activities are reflective of their intentions and goals (see [18] for a representative cognitive science study). The ability to predict the future gaze targets of the camera-wearer from egocentric videos, known as *egocentric gaze anticipation*, is therefore a key step towards understanding and modeling cognitive processes and decision making. Furthermore, this capability could enable new applications in Augmented Reality and Wearable Computing, especially in social scenarios – for example, providing memory aids for patients with cognitive impairments, or reducing the latency of content delivery in such AR systems. However, forecasting the gaze fixations

* Equal corresponding author.

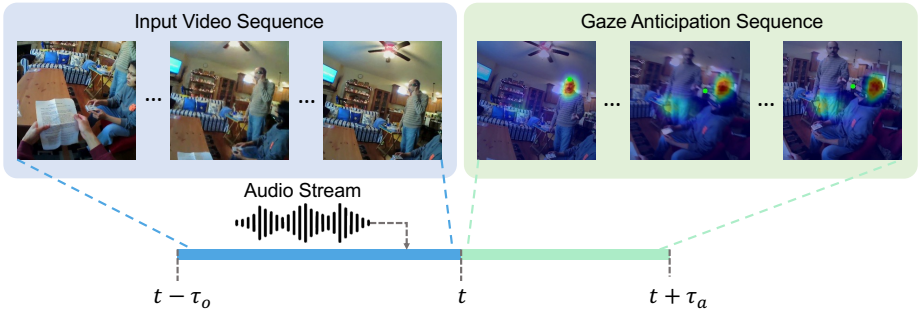

Fig. 1: The problem setting of egocentric gaze anticipation. τ_o denotes the observation time, and τ_a denotes the anticipation time. Given the video frames and audio signals of the Input Video Sequence, the model seeks to predict the gaze fixation distribution for the time steps in the Gaze Anticipation Sequence. Green dots indicate the gaze targets in future frames and the heatmap shows the gaze anticipation result from our model.

of a camera-wearer using only the egocentric view (*i.e.*, without eye tracking at testing time) is very challenging due to the complexity of egocentric scene content and the dynamic nature of gaze behaviors.

We argue that audio signals can serve as an important auxiliary cue for egocentric gaze forecasting. Consider the example in Fig. 1. In the input sequence, the camera view shifts from the paper held by the camera wearer to the standing speaker who asks a question. Then the sitting speaker on the far right answers the question, which is captured by the audio stream. In the anticipation sequence, the camera wearer's gaze shifts towards the sitting person's head after hearing her response. In this case, the audio stream (the sitting person's response) is an important stimulus that triggers this gaze movement. The influence of audio signals on eye movements is also evidenced by neuroscience research (e.g., [50]). Therefore, we address the problem of forecasting the gaze fixation of the camera-wearer in unseen future frames using a short egocentric video clip and corresponding audio. As shown in Fig. 1, fusion of the audio and video cues enables the model to correctly predict the future attention to the seated subject (i.e., the audio stream cues the model to anticipate a shift and the video stream makes it possible to identify which face is speaking).

Although many works have addressed egocentric gaze estimation [23–25, 29, 30, 33, 34, 55], the egocentric gaze *anticipation* task is largely understudied [68]. Moreover, how to leverage both the visual modality and the audio modality for egocentric gaze modeling has not been explored yet. Existing methods on audio-visual learning [5–7, 20, 46, 49, 59, 64] commonly fuse visual and audio embeddings simultaneously in time and space. However, such a fusion mechanism is not ideal under the egocentric setting, where the camera wearer's reaction to the audio stimuli causes a drastic change of camera viewpoint. In Fig. 1, as a reaction to the question and answer, the camera wearer shifts the attention from the paper to the standing person and then to the sitting person. The viewpoint

and scene also have changed because of head movement (see the first and last frame). Moreover, due to the natural delay of reaction time, the audio stimulus and gaze reaction will not occur at the same time. Therefore, predicting the future gaze behavior demands a model that can (1) learn possible viewpoint and scene change driven by the audio stream *over time* and (2) locate the potential future gaze target *in the visual space*. Fusing two modalities in time and space simultaneously may result in limited performance in the two targets because of spurious audio-visual correlations. Hence, a spatial-temporal separable fusion model is a better solution for egocentric gaze anticipation task.

To address the challenges in our task, we propose a novel **C**ontrastive **S**patial-**T**emporal **S**eparable (**CSTS**) audio-visual fusion method for egocentric gaze anticipation. Specifically, we input the egocentric video frames and the corresponding audio spectrograms into a video encoder and an audio encoder respectively. Then we develop a spatial fusion module and a temporal fusion module in parallel based on self-attention mechanism for modeling the spatial and temporal audio-visual correlation *separately*, exactly addressing the aforementioned demands. The output representations from the two branches are merged by channel-wise reweighting and fed into a visual decoder to predict the future gaze target. We also propose a novel strategy that uses a multi-modal contrastive loss [2] on the *reweighted* representations (referred to as *post-fusion* contrastive loss) from the fusion modules to facilitate audio-visual correspondence learning. We demonstrate the benefits of our approach on two egocentric video datasets that capture social scenarios and everyday activities: Ego4D [16] and Aria [37]. The proposed model achieves state-of-the-art gaze anticipation performance on both datasets. Our contributions are summarized as follows:

- We introduce the first approach that utilizes video and audio signals for egocentric gaze anticipation.
- We propose a novel CSTS model that leverages a spatio-temporal separable fusion module and a post-fusion contrastive learning scheme to facilitate audio-visual representation learning for egocentric gaze anticipation.
- We present comprehensive experiment results on the Ego4D [16] and Aria [37] datasets. Our ablation studies show audio modality can improve the performance by +2.5% and +2.4% respectively in F1 score on Ego4D and Aria. The experiments also demonstrate our model outperforms prior state-of-the-art method by +1.9% and +1.6% in F1 score on the two datasets.

2 Related Work

Egocentric Gaze Modeling. Modeling human gaze behavior in egocentric videos is an important topic in egocentric vision. Most prior efforts target at egocentric gaze estimation [23–25, 29, 33, 34]. Huang *et al*. [24] propose learning temporal attention transitions from video features that reflect drastic gaze movements. Li *et al*. [34] and Huang *et al*. [23] utilize the correlation of gaze behaviors and actions, modeling them jointly with a convolutional network. Lai *et al*. [29] encode global scene context into a single global token and explicitly

model the global-local correlations in the visual embedding for gaze estimation. In contrast, egocentric gaze anticipation, which seeks to predict future gaze targets from past video frames, addresses an understudied dimension of modeling gaze. Zhang *et al*. [68] introduce this task and utilize a convolutional network and a discriminator to generate future video frames, which are further used to anticipate future gaze targets. They enhance their model by adding an additional branch for gaze forecasting [67]. All previous efforts on both egocentric gaze estimation and anticipation model gaze behavior from only the visual properties of the video stream, and do not consider the relationship between audio signals and gaze behavior. In this work, we introduce the first model that leverages both visual and audio signals for egocentric gaze anticipation task.

Audio-Visual Saliency Prediction. Audio-visual saliency prediction is a well-studied problem in computer vision [9, 40, 47, 48, 51, 53]. Another related research topic is sound source localization [5, 19–22, 52] which localizes sound source in the image/video corresponding to a given audio stream. Here, we mainly discuss previous approaches for fusing audio and visual representations in saliency prediction problem. Early CNN-based approaches adopt a late-fusion strategy [41,58–61] for saliency prediction. Recently, new findings suggest audio-visual fusion at the intermediate features is a more effective way to leverage advantages of both modalities [1,8,56,65] for saliency prediction. Jain *et al*. [26] investigate two fusion methods at the middle level which achieve new state of the art on multiple datasets. Yao *et al*. [66] propose to incorporate the audio signal at multiple decoder layers by using an inner-product operation. Similarly, Chang *et al*. [6] and Xiong *et al*. [64] merge audio features into visual features at multiple levels of the visual encoder. Notably, our problem differs from the audio-visual saliency prediction in two aspects: First, the goal of our task is forecasting gaze behavior in the *future*, while saliency prediction focuses more on studying human's attention mechanism in the *current* video frame. Second, our problem focuses egocentric videos that capture the *changing viewpoint* when people respond to audio and visual stimuli, while saliency prediction uses videos captured from a *fixed* viewpoint, and fail to reflect gaze reaction to real-time events. Apart from the difference on problem settings, we also want to emphasize that the transformer-based fusion methods have not been applied in the audio-visual saliency prediction problem. Moreover, we propose a well-motivated spatio-temporal separable fusion module to address this challenging problem

Contrastive Audio-Visual Representation Learning. Our work draws from a rich literature on leveraging contrastive learning to learn audiovisual feature representations [2–4, 15, 17, 28, 38, 39, 42–45]. These works learn correspondences between audio and visual signals in an self-supervised manner, constructing positive pairs from matching video frames and audio segments, and negative pairs from all other pairwise combinations. We employ a similar contrastive loss to learn correspondences between co-occurring audio and visual features. However, while prior methods calculate contrastive loss on the raw embedding from each modality, we propose to apply contrastive loss on re-weighted audio and visual representations from our proposed spatial and temporal fusion mechanism.

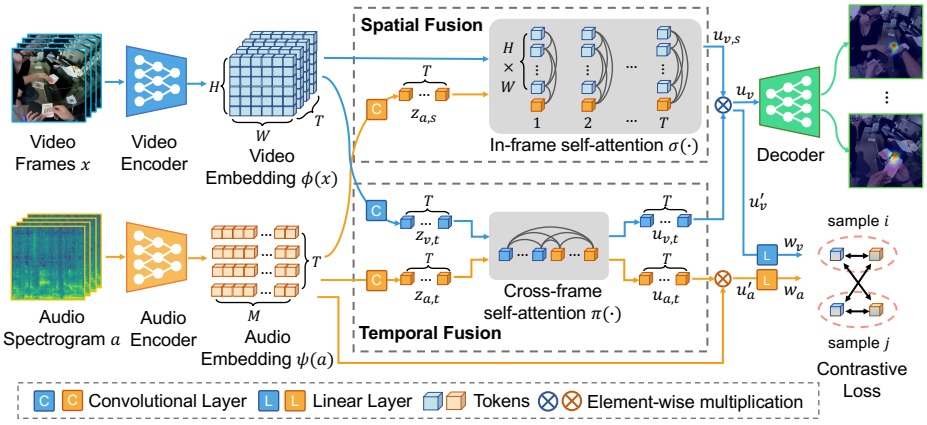

Fig. 2: Overview of the proposed model. The video embeddings $\phi(x)$ and audio embeddings $\psi(a)$ are obtained by two transformer-based encoders. We then model the correlations of visual and audio embeddings using two separate branches – (1) spatial fusion, which learns the spatial co-occurence of audio signals and visual objects in each frame, and (2) temporal fusion, which captures the temporal correlations and possible gaze movement. A contrastive loss is adopted to facilitate audio-visual representation learning. We input fused embeddings into a decoder for final gaze anticipation results.

3 Method

The egocentric gaze anticipation problem is illustrated in Fig. 1. Given egocentric video and audio samples from time $t - \tau_o$ to t, the goal is to predict the future gaze in each subsequent video frame from t to $t + \tau_a$ seconds. We denote the input video and audio as x and a, respectively, and model the gaze fixation as a probabilistic distribution on a 2D image plane (following [29, 34]).

Notably, visual and audio signals have correlations in both spatial and temporal dimensions for gaze modeling. Spatially, the visual region that has a stronger correlation with the audio content (*e.g.*, faces correlated with speech) is more likely to be the potential future gaze target. Temporally, events in the audio signal may drive both egocentric viewpoint change (via head movement) and gaze movements as the camera wearer responds to new sounds. Our key insight is that separate spatial and temporal fusion channels can be a more effective way to model audio-video correlations in gaze anticipation problem.

Fig. 2 demonstrates the overview of our model. We exploit the transformer-based encoders $\phi(x)$ and $\psi(a)$ to extract the representations of the video frames x and audio signals a. We then employ a **C**ontrastive **S**patial-**T**emporal **S**eparable (CSTS) audio-visual fusion approach. Specifically, a spatial fusion module captures the correlation between audio embeddings and spatial appearance-based features; a temporal fusion module captures the temporal correlation between the visual and audio embeddings; and a contrastive loss is applied on fused audio-visual embeddings to facilitate the representation learning. Finally, spatially and

temporally fused audio-visual features are merged and fed into a decoder for future gaze anticipation.

3.1 Audio and Visual Feature Embedding

Visual Feature Embedding. We adopt the multi-scale vision transformer (MViT) architecture [12] as the video encoder $\phi(x)$. $\phi(x)$ splits the 3D video tensor input into multiple non-overlapping patches, and thereby extracts $T \times H \times W$ visual tokens with feature dimension D from x.

Audio Feature Embedding. We follow [27] to adopt a sliding window approach for audio signal preprocessing. Specifically, for a video frame at time step t_i, the corresponding audio segment has a range of $[t_i - \frac{1}{2}\Delta t_w, t_i + \frac{1}{2}\Delta t_w]$. We then use STFT to convert all audio segments into log-spectrograms and feed the processed audio segments into a transformer-based audio encoder $\psi(a)$. Since the audio stream has more sparse information than video stream, we adopt a light-weighted transformer architecture (inspired by [11, 14]) for the audio encoder $\psi(a)$. In this way, $\psi(a)$ extracts $T \times M$ tokens with feature dimension D from the audio inputs a.

3.2 Spatial-Temporal Separable Fusion

Spatial Audio-Visual Fusion. The spatial fusion branch identifies correlations between the audio signal corresponding to a video frame and its visual content in space. We first use convolutional operations to generate the audio representation $z_{a,s}$ for spatial fusion with dimensions $T \times 1 \times D$ from the audio embedding $\psi(a)$. This allows the model to extract a holistic audio embedding within each sliding window. We then input the visual embedding $\phi(x)$ and pooled audio embedding $z_{a,s}$ into an in-frame self-attention layer σ. In this layer, we masked out all cross-frame connections and only calculate the correlations among visual tokens within each frame and the corresponding single audio token. Therefore, the input to the spatial fusion consists of T groups of visual tokens, and T single audio embeddings. Formally, we have:

$$\phi(x) = \left[\phi(x)^{(1)}, ..., \phi(x)^{(T)}\right], \tag{1}$$

$$z_{a,s} = \left[z_{a,s}^{(1)}, ..., z_{a,s}^{(T)}\right], \tag{2}$$

where $\phi(x)^{(i)} \in \mathbb{R}^{1\times N\times D}$, $z_{a,s}^{(i)} \in \mathbb{R}^{1\times 1\times D}$ with $i \in \{1, ..., T\}$, and $N = H \times W$. Hence, the input from each time step is denoted as:

$$z_s^{(i)} = \left[\phi^{(i)}(x), z_{a,s}^{(i)}\right] \in \mathbb{R}^{1\times(N+1)\times D} \tag{3}$$

The in-frame self-attention operation for time step i can be written as:

$$\sigma(z_s^{(i)}) = Softmax\left(\boldsymbol{Q}_s^{(i)} {\boldsymbol{K}_s^{(i)}}^T / \sqrt{D}\right) \boldsymbol{V}_s^{(i)} \in \mathbb{R}^{1\times(N+1)\times D}, \tag{4}$$

where $\boldsymbol{Q_s^{(i)}}, \boldsymbol{K}_s^{(i)}, \boldsymbol{V}_s^{(i)}$ refer to query, key, and value of the spatial self-attention at time step i, respectively. We apply Eq. (4) independently for each time step i and have the following overall in-frame self-attention:

$$\sigma(z_s) = \left[\sigma(z_s^{(i)}), ..., \sigma(z_s^{(T)})\right] \in \mathbb{R}^{T\times(N+1)\times D}. \tag{5}$$

In practice, we input all tokens to the in-frame self-attention layer simultaneously, mask out cross-frame correlations and calculate Eq. (4) in one shot to speed up training. We further add two linear layers after the self-attention outputs $\sigma(z_s)$, following the standard self-attention layer design. The output of the spatial module is finally denoted as $u_s \in \mathbb{R}^{T\times(N+1)\times D}$.

Temporal Audio-Visual Fusion. The temporal fusion branch models relationships between audio and visual content across time. We apply two convolutional layers to integrate the embedding from each modality at each time step into a single token. The resulting visual and audio tokens are denoted as $z_{v,t} \in \mathbb{R}^{T\times 1\times D}$ and $z_{a,t} \in \mathbb{R}^{T\times 1\times D}$, respectively. Then we feed $z_t = [z_{v,t}, z_{a,t}] \in \mathbb{R}^{2T\times 1\times D}$ into a cross-frame self-attention layer π that can be formulated as:

$$\pi(z_t) = Softmax\left(\boldsymbol{Q}_t\boldsymbol{K}_t^T/\sqrt{D}\right)\boldsymbol{V}_t \in \mathbb{R}^{2T\times 1\times D}, \tag{6}$$

where $\boldsymbol{Q}_t, \boldsymbol{K}_t, \boldsymbol{V}_t$ are query, key and value matrices with dimension $2T \times 1 \times D$. Similar to the spatial fusion, two additional linear layers are added after $\pi(z_t)$ and result in the final temporal fusion output $u_t \in \mathbb{R}^{2T\times 1\times D}$.

Merging of Two Fusion Modules. After obtaining audio-visual representations from the two fusion modules, we merge the two branches by reweighting the output from spatial fusion with the temporal weights from temporal fusion in each channel, which produces a new representation for each modality that has been refined by multimodal spatial and temporal correlation. Specifically, we break down the output from spatial fusion $u_s \in \mathbb{R}^{T\times(N+1)\times D}$ into $u_{v,s} \in \mathbb{R}^{T\times N\times D}$ and $u_{a,s} \in \mathbb{R}^{T\times 1\times D}$, and the output from temporal fusion $u_t \in \mathbb{R}^{2T\times 1\times D}$ into $u_{v,t} \in \mathbb{R}^{T\times 1\times D}$ and $u_{a,t} \in \mathbb{R}^{T\times 1\times D}$. The reweighted visual representation is formulated as

$$u_v = u_{v,s} \otimes u_{v,t} \in \mathbb{R}^{T\times N\times D}, \tag{7}$$

where $\otimes$ denotes element-wise multiplication with broadcast mechanism. u_v is then fed into a decoder to generate final prediction for future gaze target. We follow [29] to add skip connections from the video encoder to the decoder and optimize the network with a KL-Divergence loss $\mathcal{L}_{kld}$.

3.3 Contrastive Learning for Audio-Visual Fusion

In addition to using KL-Divergence loss to supervise gaze anticipation, we propose to leverage the intrinsic alignment of visual and audio modalities to learn a more robust audio-visual representation by using a contrastive learning scheme. Multi-modal contrastive loss has been proved to be effective in self-supervised

learning [2, 3]. Rather than calculating the contrastive loss directly on the raw embedded features, we innovatively propose to use the reweighted video and audio representations from the spatial and temporal fusion modules, which has not been studied in prior works. In our experiments, we show this is a more effective representation learning method for egocentric gaze anticipation.

To this end, we reweight the raw audio embedding $\psi(a) \in \mathbb{R}^{T\times M\times D}$ from the audio encoder by temporal weights $u_{a,t}$ from the temporal fusion module in a similar way to Eq. (7). We then get the reweighted audio feature as

$$u_a = \psi(a) \otimes u_{a,t} \in \mathbb{R}^{T\times M\times D} \tag{8}$$

We don't use an additional learnable token to aggregate information from other tokens as prior works did [2, 3, 35]. We instead average all tokens of u_v and u_a respectively to obtain the single-vector representations $u'_v, u'_a \in \mathbb{R}^{1\times D}$ and then map them to a low-dimensional common space using linear layers followed by L2 normalization. It can be formulated as $w_v = Norm\,(f_1(u'_v))$ and $w_a = Norm\,(f_2(u'_a))$, where $f_1(\cdot), f_2(\cdot)$ are linear layers. The resulting visual vector and audio vector are denoted as $w_v, w_a \in \mathbb{R}^{1\times D'}$, where D' is the new dimension of the common space. Within each mini-batch, corresponding audio and visual embeddings are considered as positive pairs, and all other pairwise combinations are considered as negative. Following [35], we calculate video-to-audio loss and audio-to-video loss separately. The video-to-audio contrastive loss is defined as

$$\mathcal{L}^{v2a}_{cntr} = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\log\frac{\exp({w_v^{(i)}}^T w_a^{(i)}/\mathcal{T})}{\sum_{j\in\mathcal{B}}\exp({w_v^{(i)}}^T w_a^{(j)}/\mathcal{T})}, \tag{9}$$

where $\mathcal{B}$ is the training batch $\mathcal{B} = \{1, 2, \ldots, n\}$ and $\mathcal{T}$ is the temperature factor. Superscripts (i) and (j) denote the i-th and j-th samples in the batch. The audio-to-video loss is defined in a symmetric way. Finally, the contrastive loss is defined as $\mathcal{L}_{cntr} = \mathcal{L}^{v2a}_{cntr} + \mathcal{L}^{a2v}_{cntr}$. $\mathcal{L}_{kld}$ and $\mathcal{L}_{cntr}$ are linearly combined with a parameter α for the final training loss, *i.e.*, $\mathcal{L} = \mathcal{L}_{kld} + \alpha\mathcal{L}_{cntr}$.

3.4 Implementation Details

In our experiments, we set the observation time τ_o as 3 seconds and the anticipation time τ_a as 2 seconds. For video input, we sample 8 frames from the observable segment and resize to a spatial size of 256×256. For audio input, following [27], we first resample the audio signal to 24kHz and set a time window with $\Delta t_w = 1.28s$ to crop the audio segment corresponding to each video frame. We then convert it to a log-spectrogram using a STFT with window size 10ms and hop length 5ms. The number of frequency bands is set as 256 resulting in a spectrogram matrix of size 256×256. The output of the decoder is the gaze distribution on 8 frames uniformly sampled from the 2-second anticipation time. More details about model architecture and training hyper-parameters can be found in supplementary.

4 Experiments

4.1 Experiment Setup

Datasets. We conduct experiments on two egocentric datasets that contain aligned video and audio streams and gaze tracking data – Ego4D[1] [16] and Aria [37]. Note that another widely used gaze estimation benchmark EGTEA Gaze+ [34] does not release audio data and thus is not usable for our study. Other popular egocentric video datasets, such as Epic-Kitchens [10] and Charades-Ego [54], are also not applicable to our task because they don't have eye-tracking data. Please refer to the supplementary for more details on the two datasets, our data preprocessing and train/test splits.

Evaluation Metrics. As suggested in recent work on egocentric gaze estimation [29], AUC score can easily get saturated due to the long-tailed distribution of gaze on 2D video frames. Therefore, we follow [29, 34] to adopt F1 score (*primary*), recall and precision as our evaluation metrics.

4.2 Ablation Study

We first quantify the performance contribution of each key module from our proposed method. Specifically, we denote the model only using our proposed spatial fusion module as *S-fusion*, the model only using our proposed temporal fusion module as *T-fusion*, the model using both modules and our spatial-temporal separable fusion strategy *without* the contrastive learning schema as *STS*. We finally present the performance of our full CSTS model (*i.e.*, *STS* + contrastive learning). As demonstrated in Tab. 1, compared with models trained solely on RGB frames (Vision only), S-fusion and T-fusion boost the F1 score by +1.4% and +1.5% on Ego4D, and +1.1% and +1.1% on Aria. Moreover, the STS model further achieves a F1 score of 39.2% on Ego4D and 59.3% on Aria. These results suggest that both the spatial and and the temporal correlation between video and audio signal play a vital role for egocentric gaze anticipation. Contrastive loss further improves F1 score by +0.5% and +0.6% suggesting its contributions to audio-visual representative learning. We also observe that the full model doesn't achieve the best in recall. This is because some incomplete baselines don't leverage audio modality as effectively as the full model and thus produce more uncertainty in output, resulting in higher recall and lower precision. Therefore, we consider F1 as the *primary* metric. Similar phenomenon is also observed in the following experiments.

4.3 Analysis on Fusion and Contrastive Learning Strategies

Directly feeding all visual and audio tokens into a fusion layer (*i.e.*, joint fusion) is a widely used approach for audio-visual saliency prediction [6, 59, 64] and action recognition [14, 27, 62]. To show the superiority of the proposed spatial-temporal separable (*STS*) fusion approach in handling the unique challenges of

[1]We only use the subset collected in social scenarios [31, 32].

Table 1: Ablations on each key component of our proposed model. *CSTS* (highlighted in green) refers to the complete model of our approach. The best results are highlighted with **boldface**. Please refer to Sec. 4.2 for more discussions.

Methods	Ego4D			Aria		
	F1 Score	Recall	Precision	F1 Score	Recall	Precision
Vision only	37.2	54.1	28.3	57.5	62.4	53.3
S-fusion	38.6	**54.1**	30.1	58.6	**67.1**	52.0
T-fusion	38.7	53.8	30.1	58.6	65.9	52.8
STS	39.2	53.7	30.8	59.3	66.8	53.3
CSTS	**39.7**	53.3	**31.6**	**59.9**	66.8	**54.3**

Table 2: Analysis on proposed fusion strategies. The best results are highlighted with **boldface**. *STS* (highlighted in green) refers to the proposed spatial-temporal separable fusion method (without contrastive learning). More discussions are in Sec. 4.3.

Methods	Ego4D			Aria		
	F1 Score	Recall	Precision	F1 Score	Recall	Precision
Vision only	37.2	**54.1**	28.3	57.5	62.4	53.3
Linear	38.2	53.0	29.9	58.1	65.9	51.9
Bilinear	37.6	52.8	29.2	57.7	66.8	50.8
Concat.	38.1	53.6	29.5	58.0	66.8	51.2
Vanilla SA	38.5	53.3	30.1	58.0	67.2	51.1
STS	**39.2**	53.7	**30.8**	**59.3**	**66.8**	**53.3**

our task, we provide additional comparison with four joint fusion strategies that are widely used in audio-visual saliency prediction and audio-visual action recognition. Specifically, the four strategies are (1) fusing two modalities with a few linear layers [14] (denoted as *Linear*); (2) feeding video and audio embeddings to a single bilinear layer [26,66] (denoted as *Bilinear*); (3) concatenating audio and visual embeddings along channel dimension (denoted as *Concat.*) as in [26, 27]; (4) feeding all embedded video and audio tokens together into a standard self-attention layer (denoted as *Vanilla SA*), inspired by [36, 64]. We replace our fusion modules with the four strategies in our framework for a fair comparison. We elaborate the implementation details of each baseline in supplementary.

As shown in Tab. 2, Linear, Bilinear, Concat. and Vanilla SA methods have limited improvement over the vision-only baseline, suggesting that previous fusion strategies for audio-visual saliency prediction and general action recognition are sub-optimal for our problem setting. In contrast, our proposed fusion strategy (STS) yields larger performance boost (+2.0% on Ego4D and +1.8% on Aria) even without using the contrastive loss, which shows the benefits of spatial-temporal separable fusion mechanism. The possible reason is that prior joint fusion methods are designed for third-person videos without a drastic viewpoint change. However, forecasting gaze in egocentric view has the unique challenges

Table 3: Analysis on the proposed contrastive learning schema. *Post Contr* denotes our post-fusion contrastive learning. *STS + Post Contr* refers to the complete CSTS model. The best results are highlighted with **boldface**. More discussions are in Sec. 4.3.

Methods	Ego4D			Aria		
	F1 Score	Recall	Precision	F1 Score	Recall	Precision
Vanilla SA	38.5	53.3	30.1	58.0	**67.2**	51.1
SA + Vanilla Contr	38.5	52.4	30.5	58.4	67.0	51.8
SA + Post Contr	38.9	**54.4**	30.3	58.8	66.4	52.8
STS	39.2	53.7	30.8	59.3	66.8	53.3
STS + Vanilla Contr	39.0	53.7	30.6	59.1	66.5	53.1
STS + Post Contr	**39.7**	53.3	**31.6**	**59.9**	66.8	**54.3**

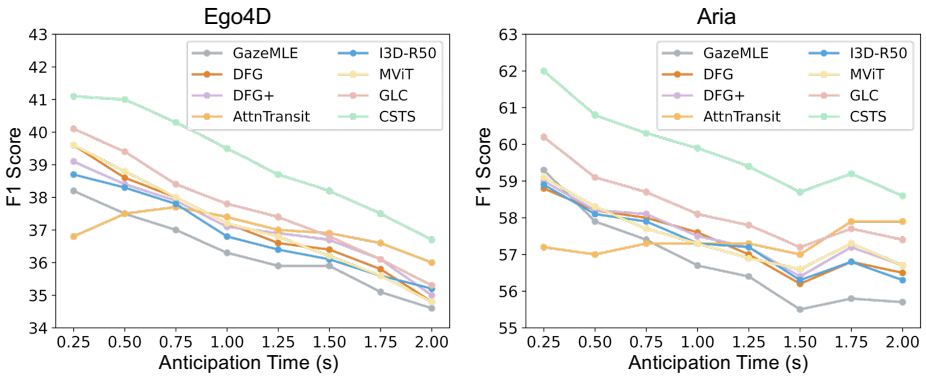

Fig. 3: The performance of gaze anticipation in each frame. Our model (CSTS) consistently outperforms all prior methods by a notable margin.

caused by camera movement and the latency of gaze response to audio stimuli. Our approach fuses two modalities in space and time separately and hence avoids spurious correlations that may happen in joint fusion baselines.

We also evaluate the benefits of our proposed post-fusion contrastive learning scheme in Tab. 3. Here, we consider another baseline (denoted as *Vanilla Contr*) that calculates the contrastive loss using raw video and audio embeddings (*i.e.*, $\phi(x)$ and $\psi(a)$ in Fig. 2), as is typical in prior work [15, 17, 38, 57]. Our novel strategy of adding contrastive loss on fused features is denoted as *Post Contr*. Vanilla Contr makes only minor differences on Vanilla SA model and even slightly reduces performance when accompanied by our proposed STS mechanism. In contrast, our proposed Post Contr scheme improves the performance of Vanilla SA by +0.4% and 0.8% and improves STS by +0.5% and +0.6% on the two datasets. These results further suggest that post-fusion contrastive learning is more robust for audio-visual learning in our task. More experiments of different contrastive learning strategies are provided in supplementary.

Table 4: Comparison with state-of-the-art models on egocentric gaze anticipation. We also adapt previous egocentric gaze estimation approaches to the anticipation setting for a more thorough comparison. The best results are highlighted with **boldface**. The green row shows our model performance. Please refer to Sec. 4.4 for more discussions.

Methods	Ego4D			Aria		
	F1 Score	Recall	Precision	F1 Score	Recall	Precision
Center Prior	13.6	9.4	24.2	24.9	17.3	44.4
GazeMLE [34]	36.3	52.5	27.8	56.8	64.1	51.0
AttnTrans [24]	37.0	**55.0**	27.9	57.4	65.5	51.0
I3D-R50 [13]	36.9	52.1	28.6	57.4	63.6	52.2
MViT [12]	37.2	54.1	28.3	57.5	62.4	53.3
GLC [29]	37.8	52.9	29.4	58.3	65.4	52.6
DFG [68]	37.2	53.2	28.6	57.4	63.6	52.3
DFG+ [67]	37.3	52.3	29.0	57.6	65.5	51.3
CSTS	**39.7**	53.3	**31.6**	**59.9**	**66.8**	**54.3**

4.4 Comparison with State-of-the-art Methods

Most existing works on egocentric gaze modeling target at egocentric gaze estimation rather than anticipation. In order to provide a thorough comparison, in addition to comparing against SOTA egocentric gaze anticipation models (DFG [68], DFG+ [67]), we also adapt the recent SOTA egocentric gaze estimation model GLC [29] and all baselines from [29] (I3D-Res50 [63], MViT [12], GazeMLE [34] and AttnTrans [24]) to the anticipation task.

As presented in Tab. 4, our method outperforms its direct competitor DFG+, which is the previous SOTA model for egocentric gaze anticipation, by +2.4% F1 on Ego4D and +2.3% F1 on Aria. Note that the original DFG and DFG+ used a less powerful backbone encoder, so for fair comparison, we reimplement their method using the same MViT backbone as our method. We also observe that methods originally designed for egocentric gaze estimation still work as strong baselines for the egocentric gaze anticipation task. Our proposed CSTS model also outperforms these methods, surpassing the recent SOTA for egocentric gaze estimation – GLC by +1.9% F1 on Ego4D and +1.6% F1 on Aria. In addition, We also incorporate audio stream into the strongest baseline (GLC) by a straightforward concatenation whose F1 score is 38.1% on Eg4D and 58.5% on Aria. The marginal gain over GLC (+0.3%/+0.2%) suggests that simply using audio stream in a strong baseline without specific design leads to sub-optimal solution in egocentric gaze anticipation problem, which in turn validates the effectiveness and necessity of our approach.

In addition, we evaluate gaze anticipation on each anticipation time step independently and compare with previous methods in Fig. 3. Unsurprisingly, the anticipation problem becomes more challenging as the anticipation time step increases farther into the future. Our CSTS method consistently outperforms all baselines at all future time steps. Moreover, we note that our model also

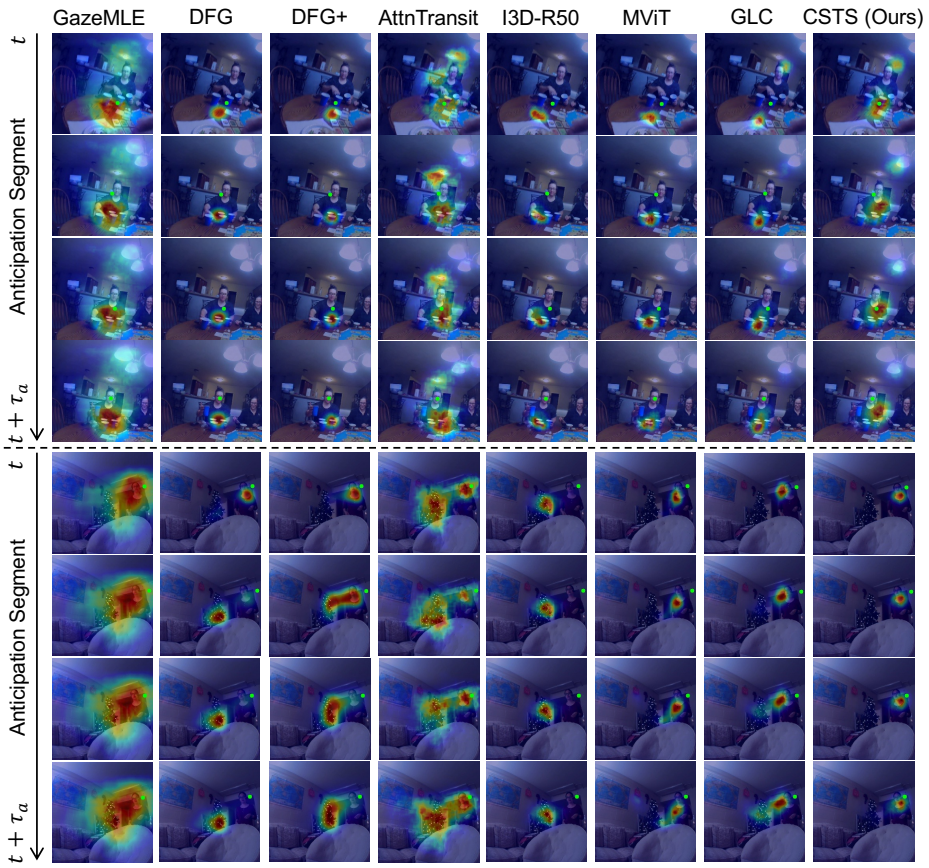

Fig. 4: Egocentric gaze anticipation results from our model and other baselines. We show the results of four future time steps uniformly sampled from the anticipation segments. Green dots indicate the ground truth gaze location.

produces new SOTA results on egocentric gaze *estimation*, demonstrating the generalizability and robustness of our approach across gaze modeling tasks. We include these results in supplementary.

4.5 Visualization of Predictions

We visually showcase the anticipation results of CSTS and the baselines in Fig. 4. We can see that GazeMLE [34] and AttnTransit [24] produce more uncertainty in prediction heatmaps. Other methods fail to anticipate the true gaze target, and are likely to be misled by other salient objects. Our CSTS approach produces the best gaze anticipation results among all methods. We attribute this improvement to our novel model design that effectively addresses the unique challenges of forecasting gaze targets in egocentric view.

Fig. 5: Visualization of the spatial correlation weights. All video frames are sorted in a chronological order indexed by the numbers on the top-right corner.

4.6 Visualization of Learned Correlations

We provide further insight on our model by visualizing the audio-visual correlations from the spatial fusion module. For each time step t, we calculate the correlation of each visual token with the single audio token and map it back to the input frames. The correlation heatmaps are shown in Fig. 5. In the first example, the speaker in the middle speaks, then turns her head around to talk with a social partner in the background (frame 1-3). We observe that our model captures that the audio signal has the highest correlation with spatial region of the speaker while she is speaking. Then, when she stops talking and turns her head back, the correlation is highest in the background regions, indicating the potential location of her social partner. The second example illustrates a similar phenomenon: the model captures the speaker at the beginning when she is talking, then attends to background locations when she stops. These examples suggest our model has the capability to model the audio-visual correlations in spatial dimension to learn a robust audio-visual representation.

5 Conclusion

In this paper, we propose a novel contrastive spatial-temporal separable fusion approach (CSTS) for egocentric gaze anticipation. Our key contribution is breaking down the fusion of the audio and visual modalities into a separate spatial fusion module for learning the spatial co-occurrence of visual features and audio signals, and a temporal fusion module for modeling the changing viewpoint and scene driven by audio stimuli. We further adopt a contrastive loss on the reweighted audio-visual representations from the fusion modules to facilitate multimodal representation learning. We demonstrate the benefits of our proposed model design on two egocentric video datasets: Ego4D and Aria. Our work is a key step for probing into human cognitive process with computational models, and provides important insights into multimodal representation learning, visual forecasting and egocentric video understanding.

Acknowledgements

Portions of this work were supported in part by a gift from Meta and a grant from the Toyota Research Institute University 2.0 program. The second author is supported by an NSF Graduate Research Fellowship.

References

1. Agrawal, R., Jyoti, S., Girmaji, R., Sivaprasad, S., Gandhi, V.: Does audio help in deep audio-visual saliency prediction models? In: Proceedings of the 2022 International Conference on Multimodal Interaction. pp. 48–56 (2022)
2. Akbari, H., Yuan, L., Qian, R., Chuang, W.H., Chang, S.F., Cui, Y., Gong, B.: Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. Advances in Neural Information Processing Systems **34**, 24206–24221 (2021)
3. Alayrac, J.B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., Zisserman, A.: Self-supervised multimodal versatile networks. Advances in Neural Information Processing Systems **33**, 25–37 (2020)
4. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE international conference on computer vision. pp. 609–617 (2017)
5. Arandjelovic, R., Zisserman, A.: Objects that sound. In: Proceedings of the European conference on computer vision (ECCV). pp. 435–451 (2018)
6. Chang, Q., Zhu, S.: Temporal-spatial feature pyramid for video saliency detection. Cognitive Computation (2021)
7. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16867–16876 (2021)
8. Cheng, S., Gao, X., Song, L., Xiahou, J.: Audio-visual salieny network with audio attention module. In: 2021 2nd International Conference on Artificial Intelligence and Information Systems. pp. 1–5 (2021)
9. Coutrot, A., Guyader, N.: Multimodal saliency models for videos. From Human Attention to Computational Attention: A Multidisciplinary Approach pp. 291–304 (2016)
10. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European conference on computer vision (ECCV). pp. 720–736 (2018)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (2020)
12. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6824–6835 (2021)
13. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)

14. Gao, R., Oh, T.H., Grauman, K., Torresani, L.: Listen to look: Action recognition by previewing audio. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10457–10467 (2020)
15. Gong, Y., Rouditchenko, A., Liu, A.H., Harwath, D., Karlinsky, L., Kuehne, H., Glass, J.: Contrastive audio-visual masked autoencoder. International Conference on Learning Representations (2022)
16. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18995–19012 (2022)
17. Gurram, S., Fang, A., Chan, D., Canny, J.: Lava: Language audio vision alignment for contrastive video pre-training. arXiv preprint arXiv:2207.08024 (2022)
18. Hayhoe, M., Ballard, D.: Eye movements in natural behavior. Trends in cognitive sciences **9**(4), 188–194 (2005)
19. Hu, D., Nie, F., Li, X.: Deep multimodal clustering for unsupervised audiovisual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9248–9257 (2019)
20. Hu, D., Qian, R., Jiang, M., Tan, X., Wen, S., Ding, E., Lin, W., Dou, D.: Discriminative sounding objects localization via self-supervised audiovisual matching. Advances in Neural Information Processing Systems **33**, 10077–10087 (2020)
21. Hu, X., Chen, Z., Owens, A.: Mix and localize: Localizing sound sources in mixtures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10483–10492 (2022)
22. Huang, C., Tian, Y., Kumar, A., Xu, C.: Egocentric audio-visual object localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22910–22921 (2023)
23. Huang, Y., Cai, M., Li, Z., Lu, F., Sato, Y.: Mutual context network for jointly estimating egocentric gaze and action. IEEE Transactions on Image Processing **29**, 7795–7806 (2020)
24. Huang, Y., Cai, M., Li, Z., Sato, Y.: Predicting gaze in egocentric video by learning task-dependent attention transition. In: Proceedings of the European conference on computer vision (ECCV). pp. 754–769 (2018)
25. Huang, Y., Cai, M., Sato, Y.: An ego-vision system for discovering human joint attention. IEEE Transactions on Human-Machine Systems **50**(4), 306–316 (2020)
26. Jain, S., Yarlagadda, P., Jyoti, S., Karthik, S., Subramanian, R., Gandhi, V.: Vinet: Pushing the limits of visual modality for audio-visual saliency prediction. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3520–3527. IEEE (2021)
27. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5492–5501 (2019)
28. Korbar, B., Tran, D., Torresani, L.: Cooperative learning of audio and video models from self-supervised synchronization. Advances in Neural Information Processing Systems **31** (2018)
29. Lai, B., Liu, M., Ryan, F., Rehg, J.: In the eye of transformer: Global-local correlation for egocentric gaze estimation. British Machine Vision Conference (2022)
30. Lai, B., Liu, M., Ryan, F., Rehg, J.M.: In the eye of transformer: Global–local correlation for egocentric gaze estimation and beyond. International Journal of Computer Vision **132**(3), 854–871 (2024)

31. Lai, B., Zhang, H., Liu, M., Pariani, A., Ryan, F., Jia, W., Hayati, S.A., Rehg, J., Yang, D.: Werewolf among us: Multimodal resources for modeling persuasion behaviors in social deduction games. Association for Computational Linguistics: ACL 2023 (2023)
32. Lee, S., Lai, B., Ryan, F., Boote, B., Rehg, J.M.: Modeling multimodal social interactions: New challenges and baselines with densely aligned representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14585–14595 (2024)
33. Li, Y., Fathi, A., Rehg, J.M.: Learning to predict gaze in egocentric video. In: Proceedings of the IEEE international conference on computer vision. pp. 3216–3223 (2013)
34. Li, Y., Liu, M., Rehg, J.: In the eye of the beholder: Gaze and actions in first person video. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
35. Lin, K.Q., Wang, A.J., Soldan, M., Wray, M., Yan, R., Xu, E.Z., Gao, D., Tu, R., Zhao, W., Kong, W., et al.: Egocentric video-language pretraining. Advances in Neural Information Processing Systems (2022)
36. Lin, Y.B., Sung, Y.L., Lei, J., Bansal, M., Bertasius, G.: Vision transformers are parameter-efficient audio-visual learners. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2023)
37. Lv, Z., Miller, E., Meissner, J., Pesqueira, L., Sweeney, C., Dong, J., Ma, L., Patel, P., Moulon, P., Somasundaram, K., Parkhi, O., Zou, Y., Raina, N., Saarinen, S., Mansour, Y.M., Huang, P.K., Wang, Z., Troynikov, A., Artal, R.M., DeTone, D., Barnes, D., Argall, E., Lobanovskiy, A., Kim, D.J., Bouttefroy, P., Straub, J., Engel, J.J., Gupta, P., Yan, M., Nardi, R.D., Newcombe, R.: Aria pilot dataset. `https://about.facebook.com/realitylabs/projectaria/datasets` (2022)
38. Ma, S., Zeng, Z., McDuff, D., Song, Y.: Active contrastive learning of audio-visual video representations. International Conference on Learning Representations (2020)
39. Ma, S., Zeng, Z., McDuff, D., Song, Y.: Contrastive learning of global-local video representations. arXiv preprint arXiv:2104.05418 (2021)
40. Min, X., Zhai, G., Gu, K., Yang, X.: Fixation prediction through multimodal analysis. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **13**(1), 1–23 (2016)
41. Min, X., Zhai, G., Zhou, J., Zhang, X.P., Yang, X., Guan, X.: A multimodal saliency model for videos with high audio-visual correspondence. IEEE Transactions on Image Processing **29**, 3805–3819 (2020)
42. Morgado, P., Li, Y., Nvasconcelos, N.: Learning representations from audio-visual spatial alignment. Advances in Neural Information Processing Systems **33**, 4733–4744 (2020)
43. Morgado, P., Misra, I., Vasconcelos, N.: Robust audio-visual instance discrimination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12934–12945 (2021)
44. Morgado, P., Vasconcelos, N., Misra, I.: Audio-visual instance discrimination with cross-modal agreement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12475–12486 (2021)
45. Patrick, M., Asano, Y., Kuznetsova, P., Fong, R., Henriques, J.F., Zweig, G., Vedaldi, A.: Multi-modal self-supervision from generalized data transformations. arXiv preprint (2020)
46. Qian, R., Hu, D., Dinkel, H., Wu, M., Xu, N., Lin, W.: Multiple sound sources localization from coarse to fine. In: Computer Vision–ECCV 2020: 16th European

Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. pp. 292–308. Springer (2020)
47. Ratajczak, R., Pellerin, D., Labourey, Q., Garbay, C.: A fast audiovisual attention model for human detection and localization on a companion robot. In: VISUAL 2016-The First International Conference on Applications and Systems of Visual Paradigms (VISUAL 2016) (2016)
48. Ruesch, J., Lopes, M., Bernardino, A., Hornstein, J., Santos-Victor, J., Pfeifer, R.: Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub. In: 2008 IEEE International Conference on Robotics and Automation. pp. 962–967. IEEE (2008)
49. Ryan, F., Jiang, H., Shukla, A., Rehg, J.M., Ithapu, V.K.: Egocentric auditory attention localization in conversations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14663–14674 (2023)
50. Schaefer, K., Süss, K., Fiebig, E.: Acoustic-induced eye movements. Annals of the New York Academy of Sciences **374**, 674–688 (1981)
51. Schauerte, B., Kühn, B., Kroschel, K., Stiefelhagen, R.: Multimodal saliency-based attention for object-based scene analysis. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1173–1179. IEEE (2011)
52. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., Kweon, I.S.: Learning to localize sound source in visual scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4358–4366 (2018)
53. Sidaty, N., Larabi, M.C., Saadane, A.: Toward an audiovisual attention model for multimodal video content. Neurocomputing **259**, 94–111 (2017)
54. Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K.: Charades-ego: A large-scale dataset of paired third and first person videos. arXiv preprint arXiv:1804.09626 (2018)
55. Soo Park, H., Shi, J.: Social saliency prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4777–4785 (2015)
56. Tavakoli, H.R., Borji, A., Rahtu, E., Kannala, J.: Dave: A deep audio-visual embedding for dynamic saliency prediction. arXiv preprint arXiv:1905.10693 (2019)
57. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
58. Tsiami, A., Koutras, P., Katsamanis, A., Vatakis, A., Maragos, P.: A behaviorally inspired fusion approach for computational audiovisual saliency modeling. Signal Processing: Image Communication **76**, 186–200 (2019)
59. Tsiami, A., Koutras, P., Maragos, P.: Stavis: Spatio-temporal audiovisual saliency network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4766–4776 (2020)
60. Wang, G., Chen, C., Fan, D.P., Hao, A., Qin, H.: From semantic categories to fixations: A novel weakly-supervised visual-auditory saliency detection approach. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15119–15128 (2021)
61. Wang, G., Chen, C., Fan, D.P., Hao, A., Qin, H.: Weakly supervised visual-auditory fixation prediction with multigranularity perception. arXiv preprint arXiv:2112.13697 (2021)
62. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12695–12705 (2020)

63. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
64. Xiong, J., Wang, G., Zhang, P., Huang, W., Zha, Y., Zhai, G.: Casp-net: Rethinking video saliency prediction from an audio-visual consistency perceptual perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6441–6450 (2023)
65. Yang, Q., Li, Y., Li, C., Wang, H., Yan, S., Wei, L., Dai, W., Zou, J., Xiong, H., Frossard, P.: Svgc-ava: 360-degree video saliency prediction with spherical vector-based graph convolution and audio-visual attention. IEEE Transactions on Multimedia (2023)
66. Yao, S., Min, X., Zhai, G.: Deep audio-visual fusion neural network for saliency estimation. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 1604–1608. IEEE (2021)
67. Zhang, M., Ma, K.T., Lim, J.H., Zhao, Q., Feng, J.: Anticipating where people will look using adversarial networks. IEEE transactions on pattern analysis and machine intelligence **41**(8), 1783–1796 (2018)
68. Zhang, M., Teck Ma, K., Hwee Lim, J., Zhao, Q., Feng, J.: Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4372–4381 (2017)