SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

February 2026, Vol. 69, Iss. 2, 122101:1–122101:19
https://doi.org/10.1007/s11432-024-4744-5

Weakly supervised visual-auditory fixation prediction with multigranularity perception

Guotao WANG[1], Chenglizhao CHEN[2,3,4*], Deng-Ping FAN[5,6,7], Aimin HAO[1] & Qinping ZHAO[1]

[1]*State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China*
[2]*College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China*
[3]*Qingdao Institute of Software, China University of Petroleum (East China), Qingdao 266580, China*
[4]*Shandong Provincial Key Laboratory of Intelligent Oil and Gas Industrial Software, Qingdao 266580, China*
[5]*Nankai International Advanced Research Institute (Shenzhen Futian), Shenzhen 518045, China*
[6]*Tianjin Visual Computing and Intelligent Perception Laboratory, Nankai University, Tianjin 300071, China*
[7]*College of Computer Science, Nankai University, Tianjin 300071, China*

Received 28 July 2024/Revised 19 February 2025/Accepted 8 April 2025/Published online 16 January 2026

Abstract Video saliency detection models have been achieving steady, significant improvements thanks to rapid advances in deep learning and the wide availability of large-scale training sets. However, deep learning-based visual-audio fixation prediction is still in its infancy. At present, only a few visual-audio sequences have been furnished, with real fixations being recorded in real visual-audio environments. Hence, it would neither be efficient nor necessary to recollect real fixations under the same visual-audio circumstances. To address this problem, this paper promotes a novel weakly supervised approach that alleviates the demand for large-scale training sets for visual-audio model training. By using only the video category tags, we propose the selective class activation mapping (SCAM) and its upgrade (SCAM+). In the spatial-temporal-audio circumstance, the former follows a coarse-to-fine strategy to select the most discriminative regions, which are usually capable of exhibiting high consistency with real human-eye fixations. The latter equips the SCAM with an additional multigranularity perception mechanism, making the whole process more consistent with that of the real human visual-audio system. Moreover, we distill knowledge from these regions to obtain completely new spatial-temporal-audio (STA) fixation prediction (FP) networks, enabling broad applications when video tags are unavailable. Without resorting to any real human-eye fixation, the performances of these STAFP networks are comparable to those of fully supervised networks. The code and results are publicly available at https://github.com/guotaowang/STANet.

Keywords weakly supervised learning, visual-audio fixation prediction, multigranularity perception

Citation Wang G T, Chen C L Z, Fan D-P, et al. Weakly supervised visual-auditory fixation prediction with multigranularity perception. Sci China Inf Sci, 2026, 69(2): 122101, https://doi.org/10.1007/s11432-024-4744-5

1 Introduction

In the deep learning era, we have witnessed a growing development in video saliency detection techniques [1, 2], where the primary task is to locate the most distinctive regions in a series of video sequences. At present, this research field consists of two parallel research directions, i.e., the video salient object detection (VSOD) and the video fixation prediction (VFP). In practice, the former [3–5] aims to segment the most salient objects with clear object boundaries (e.g., Figure 1(b)). The latter [6–10], the main topic of this paper, attempts to predict human-eye-like fixations—scattered coordinates spreading over the entire scene without any clear boundaries (e.g., Figure 1(a)).

Different from previous work on video saliency detection [1, 11], this paper focuses on two main aspects: i.e., (1) exploiting weakly-supervised techniques to generate pseudo-fixations under the visual-audio circumstances, and (2) using these pseudo-fixations to generate a saliency prediction model. The main focus of this paper is on the generation of pseudo-fixations. It is worth noting that this topic is still in its early exploration stage.

Visual-audio fixation prediction is a critical task for various video-related applications to help alleviate computation costs. For example, in the typical video compression application [12], if the compression method has been chosen, we can use the predicted visual-audio fixations to further beat the trade-off between the compression rate and the video quality after compression. That is, as shown in Figure 2(a), we shall compress those meaningful video contents—which are more likely to be highlighted by the predicted fixations—at a relatively low compression rate

* Corresponding author (email: cclz123@163.com)

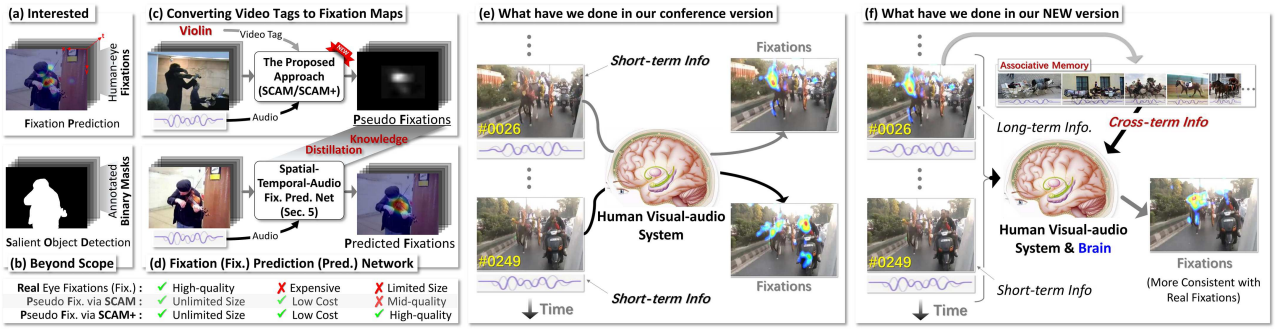

Figure 1 (Color online) Overview of our weakly supervised framework for spatial-temporal-audio (STA) fixation prediction. We convert semantic category tags into pseudo-fixations using the proposed selective class activation mapping (SCAM; Section 3) and its enhanced variant SCAM+ (Section 4), which is additionally equipped with a multigranularity perception module. The resulting pseudo-fixations serve as supervision for knowledge distillation, training two fixation prediction networks (STA and STA+; Section 5) to enable generic video fixation prediction without requiring video tags.

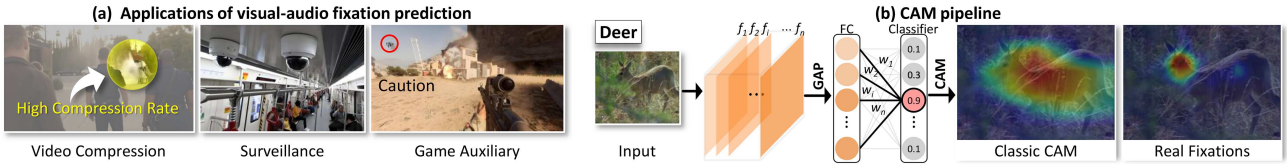

Figure 2 (Color online) (a) Applications of visual-audio fixation prediction: video compression [12], surveillance [13], and game auxiliary [14]. (b) Details of class activation mapping (CAM). FC: fully connected layer; GAP: global average pooling. The numbers shown in the classifier indicate class confidence scores.

to retain their overall visual quality. And, for those less important video contents, e.g., with fewer fixations, we can compress them with a high compression rate. In this way, we can ensure good overall video quality with less storage demand. By the way, there are also a bunch of other applications where the visual-audio fixation prediction is critical, e.g., surveillance [13] and game auxiliary [14].

At present, almost all state-of-the-art (SOTA) visual-audio fixation prediction approaches [1, 2, 15–17] are developed with the help of deep learning techniques, using the vanilla encoder-decoder structure, facilitated with various attention mechanisms, and trained in a fully supervised manner. Despite achieving progress, these fully supervised approaches are plagued by one critical limitation.

It is well known that a deep model's performance is heavily dependent on the adopted training set, and currently, large-scale visual-related training sets are accessible in our research community. However, the visual-audio circumstance-related training data are rather short since collecting real human-eye fixations in such multimodality circumstances is a time-consuming and laborious process. To the best of our knowledge, only a few visual-audio sequences equipped with real fixation data are available for the visual-audio fixation prediction task, where only a small part of them is recommended for network training, making the data shortage dilemma even worse.

To solve this problem, we seek to realize visual-audio fixation prediction using a weakly supervised strategy. Instead of using the labor-intensive framewise visual-audio ground truths (GTs), we devise a novel scheme to produce GT-like visual-audio pseudo fixations by using only video category tags as supervision. Actually, a plethora of visual-audio sequences with well-labeled semantic category tags already exist (e.g., the AVE set [18]), where most of them were originally collected for the visual-audio classification task. Note that, from a cost perspective, manually assigning semantic tags to videos is clearly more favorable than collecting real human-eye fixations. Thus, the key is how to convert video tags to real-fixation-like data.

Our approach is inspired by the class activation mapping (CAM [19]) that has been used in various object localization and object detection tasks [13, 20–23]. Our rationale relies on the fact that image regions with the strongest discriminative power regarding the classification task should be the most salient ones, where these regions usually tend to have larger classification confidences than others.

As seen in Figure 1(a), considering that we aim at fixation prediction in visual-audio circumstances, we present two practical innovations in formulating high-quality fixation maps: (1) selective class activation mapping (SCAM) and (2) SCAM+—an upgraded version of the SCAM. The key rationale of SCAM relies on the "data source" aspect, which performs a coarse-to-fine strategy to reveal the most discriminative regions from multiple sources (i.e., spatial, temporal, audio, and their combinations), where these regions exhibit high consistency with the real human-eye fixations. This coarse-to-fine methodology ensures that the aforementioned less discriminative scattered regions

are filtered completely, and the selection operation between different sources helps reveal the most discriminative regions, enabling the pseudo-fixations to be closer to real fixations.

Furthermore, since the human visual-audio system is clearly not independent of our brain, real human eye fixations are usually influenced by multiple physiological activities, e.g., short-term memory, long-term memory, associative memory, and semantic reasoning, making our attention mechanism a 'global' process in essence [5,24,25]. However, the proposed SCAM follows a typical 'local' mechanism that only considers 3 frames and a 1-second audio signal at most each time. Thus, we devise the upgraded version of SCAM, named SCAM+, which equips SCAM with an additional multigranularity perception mechanism to further boost the consistency degree between the derived pseudo-fixations and real fixations. By using both SCAM and SCAM+, we can automatically convert video tags to high-quality pseudofixation maps, and thus, theoretically, we can easily expand the existing video fixation training sets to unlimited sizes. To better highlight the novelty, we pictorially clarify the rationale of our approach in Figure 1(c).

In addition, to ensure a broad application, we shall use the pseudo-fixations derived by both SCAM and SCAM+ to train end-to-end fixation prediction networks, and this process could be regarded as a variant of knowledge distillation, where the end-to-end fixation prediction networks are clearly the students. As seen in Figure 1(d), the whole testing process does not require assigning semantic video tags to the input testing sequence, which enables the proposed approach to have wide applicability in real work.

This paper builds upon our conference version [26], which has been extended in two distinct aspects. The major innovations include "a completely new paradigm" to fully mimic real human visual fixations in the visual-audio environment and a series of novel technical ingredients to achieve this new paradigm. First, a new paradigm for fully mimicking real human visual fixations in the visual-audio environment. As shown in Figures 1(e) and (f), we have provided a demonstration regarding the concept difference between our conference version and the "new version". In short, our conference version has mainly focused on converting video tags to fixations that could be somewhat consistent with our visual-audio system, but our new version has additionally combined several brain-related activities that could influence human-eye fixations in the visual-audio environment, i.e., visual inertia and associative memory [27]. Our conference version has overlooked the brain-related activities when generating fixations. This is, of course, not an easy task to be mimicked by computers. Second, from the technical perspective, our new version has also presented multiple critical innovations that are indispensable for mimicking the real human visual-audio system. These technical innovations (e.g., C-GCN, a series of reasoning subnetworks, and the multi-modality decouple and recombine solution) are the plainest ones for achieving our new paradigm mentioned above. Compared with the previous conference version, our new version can also bring significant performance gains.

We summarize the key contributions as follows.

(1) From the source-wise perspective, we have presented a novel way (SCAM) to convert video tags to fixation maps, which could be consistent with real human-eye fixations.

(2) Inspired by the physiological mechanism of human attention, we devised an upgraded SCAM version, named SCAM+, which equips the SCAM with multigranularity perception ability, boosting the derived fixation maps' quality further.

(3) We provide one of the first attempts to predict visual-audio fixation in a weakly supervised manner, which is expected to contribute to visual-audio information integration and relevant CV-related applications, paving a new way to mimic human attention in a multimodality environment.

2 Related work

Video object localization. In video object location [19,22,28–30], each training sequence is tagged to a specific category, representing key objects or events, similar to how human fixations focus on significant areas in videos. Hence, generating pseudo-fixations from these category tags is theoretically plausible.

From a qualitative perspective, the CAM [19], which has been visualized in the part of Figure 2(b), usually shows a large feature response to frame regions (i.e., the '*Deer*') that has contributed most regarding the classification task, and these regions usually correlate to the most salient regions. More commonly, the CAM could be quite different from the real human-eye fixations. Actually, when performing the video classification task, the image regions with the most substantial contribution to the given category are capable of highlighting the salient object (i.e., the deer). Following this rationale, several previous studies [28,31] resorted to the CAM for the salient object localization task. However, the CAM results derived by these methods are different from real human-eye fixations, and the main reasons comprise the following 3 aspects.

First, since both local and nonlocal deep features contribute to the classification task, the CAMs tend to be

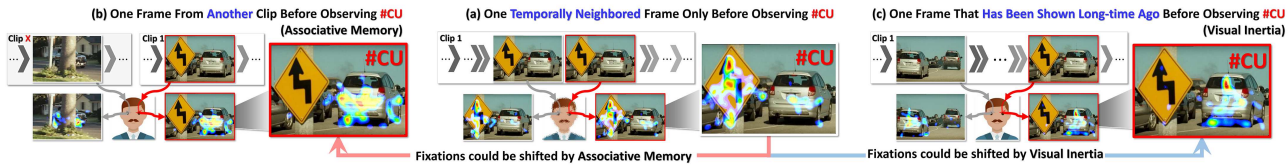

Figure 3 (Color online) Illustration of the importance of multigranularity perception. We have conducted additional experiments to support our claim, which can be found in Appendixes C and D.

large scattered regions. For example, as shown in Figure 2(b), the main body of the deer can help the classifier to separate this image from other nonanimal cases, while only the "*deer head*" can tell the classifier that the animal in this scene is a "*deer*". Instead of gazing at the "*main body*", our human visual-audio system tends to focus more attention on the most discriminative image regions (e.g., the "*deer head*").

Second, most of the existing studies [7, 21, 32] have only considered the spatial information when computing CAM. However, real human eye fixations are affected by multiple sources, including spatial, temporal, and audio sources. In fact, this multisource nature has long been omitted by our research community because compared with the spatial information—a stable source, the other 2 sources (temporal and audio) are still considered to be rather unstable sources thus far, and this unstable attribute makes them difficult to use for computing CAM. However, in many practical scenarios, these two sources are often the most beneficial for the classification task.

Third, real human fixation is not completely derived via the physiological reflex, but it is also subtly influenced by brain, which includes multiple physiological activities, implying that all instant visual stimuli, visual inertia, associative memory, and semantic reasoning [27] should be considered when performing CAM. For example, as seen in Figure 3(a), the "*road sign*" attracts massive eye fixations if only Clip 1 had been shown. However, the "*silver car*" could replace the "*road sign*" as the most salient region if either associative information (Clip X, Figure 3(b)) or long-term information (the early frames in Clip 1, Figure 3(c)) has been given in advance, where the associative information can be obtained by aligning a video frame from other video sequences sharing an identical video tag with the target sequence (Clip 1). Therefore, multigranularity perception is called for when resorting to the CAM to produce high-quality fixations.

3 Selective class activation mapping

In spatial-temporal-audio contexts, encoder-generated feature maps are multiscale, multilevel, and multisource, contributing jointly to classification. However, naively combining these sources without considering their complementary interplay can lead to false alarms and redundant responses.

To solve this problem, we propose to decouple the spatial-temporal-audio circumstance into 3 independent sources, i.e., spatial (S), temporal (T), and audio (A), and these sources will be individually fed into 3 classification nets (i.e., S, ST, and SA). By performing CAM over these classification nets, we can eventually obtain the source-wise CAMs. Clearly, this divide-and-conquer strategy can alleviate the redundancy problem effectively; however, to mimic the real human attention mechanism in spatiotemporal-audio circumstances, we shall "selectively fuse" them to localize the most discriminative regions. Thus, the key issue here is how to simultaneously achieve a source-wise complementary status and avoid accumulating redundant feature responses. Therefore, we present the selective class activation mapping (SCAM) fusion mechanism, whose technical details can be formulated by

$$\begin{aligned}&\mathrm{SCAM}=\mathcal{Z}\left(\frac{\left\|\mathrm{UC}\odot\mathrm{UR}\right\|_1+\lambda}{\left\|\mathrm{UC}\right\|_1+\lambda}\right),\\&\mathrm{UR}:\left[\Phi_{\mathrm{S}}\{i\},\Phi_{\mathrm{ST}}\{i\},\Phi_{\mathrm{SA}}\{i\}\right],\mathrm{UC}:\left[\oint\left(\mathrm{C_S}\{i\}\right),\oint\left(\mathrm{C_{ST}}\{i\}\right),\oint\left(\mathrm{C_{SA}}\{i\}\right)\right],\end{aligned}\tag{1}$$

where $\odot$ is the elementwise multiplicative operation; $\|\cdot\|_1$ denotes the L1-norm; λ is a small constant to avoid any division by zero; Φ_{S}, Φ_{ST}, and Φ_{SA} represent the CAM results derived from S, ST, and SA classification nets, respectively; $\mathcal{Z}(\cdot)$ is the min-max normalization operation. In addition, suppose the pre-given category tag for the classification network S is the i-th category among c classes, where $c=28$. Then we use $\mathrm{C_S}\{i\}$ to represent this confidence, where $\mathrm{C_S}\in(0,1)^{1\times c}$. $\oint(\cdot)$ is a soft filter (Eq. (2)) aiming to compress those features of low classification confidences to be considered when computing the SCAM. The detailed definition of $\oint(\cdot)$ is as follows:

$$\oint\left(\mathrm{C_S}\{i\}\right)=\begin{cases}\mathrm{C_S}\{i\}, & \text{if } \mathrm{C_S}\{i\}>\mathrm{C_S}\{u\}|_{i\neq u,1\leqslant u\leqslant c},\\ 0, & \text{otherwise.}\end{cases}\tag{2}$$

Figure 4 (Color online) Qualitative illustration of CAMs derived from different sources. 'Darbuka': a type of drum that is widely used in Middle Eastern and North African music. 'STA (S/SA/ST)-CAM': CAM obtained from the spatial-temporal-audio (spatial-/spatial-audio-/spatiotemporal) circumstances. 'SCAM' represents pseudo-fixations obtained by (1), where we can easily observe that the results in this column are highly consistent with the GTs.

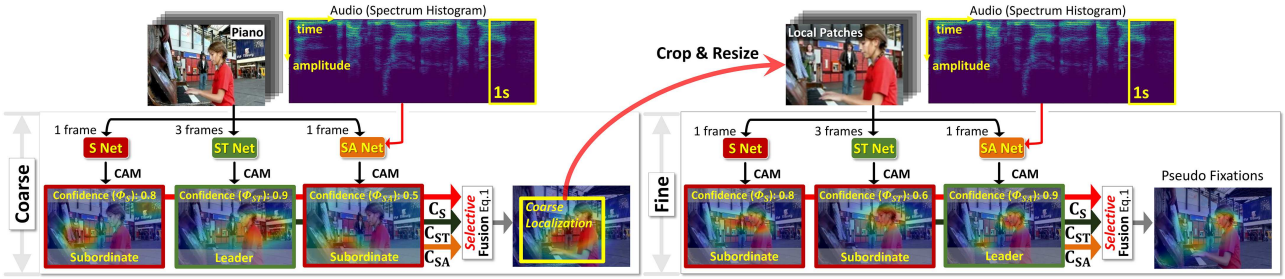

Figure 5 (Color online) Our selective class activation mapping (SCAM) follows the coarse-to-fine methodology, where COARSE stage localizes the regions of interest, and then FINE stage reveals those image regions with the strongest local responses. S: spatial; ST: spatiotemporal; SA: spatial-audio. The structures of S/ST/SA nets can be found in Figures 7(a)–(c).

3.1 SCAM rationales

Generally, either spatial, temporal, or audio sources could influence our visual attention; however, compared with the last two, the spatial source is usually more important and stable in practice.

Thus, in our classification nets, the spatial information should be treated as the main force, while the other two can only be its subordinates. This is the reason why we recombine S, T, and A sources to S (no change), ST, and SA, respectively.

However, the CAMs derived from these nets are still rather different in essence because their inputs are different, and we have demonstrated some of the most representative qualitative results in Figure 4.

3.2 Multistage SCAM

To enhance performance, SCAM is applied twice in a coarse-to-fine manner, where the coarse stage narrows the problem domain, allowing the fine stage to generate pseudo-fixations that better capture discriminative regions. As shown in Figure 5, the coarse stage localizes pseudo-fixations using a thresholded rectangular box (coarse location) and crops video sequences into patches. In the fine stage, these patches replace the original sequences as input for classification networks, where SCAM is reapplied to generate source-wise pseudo-fixations.

3.3 Fusion modules in SCAM

Following the previous studies [16,33], we have converted audio signals to 2D spectrum histograms U_i in advance, which are then fed into an off-the-shelf VGGSound $\mathcal{F}_{\text{VGGSound}}$ to obtain the corresponding deep features ($a_i = \mathcal{F}_{\text{VGGSound}}(U_i)$), and similarly, the spatial deep features ($s_i = \mathcal{F}_{\text{VGGNet}}(I_i)$) can be obtained by feeding each video frame (I_i) to the off-the-shelf VGGNet $\mathcal{F}_{\text{VGGNet}}$.

SA fuse module. The primary target of the SA fuse module is to integrate spatial information with audio signals. We use a series of deconvolutions to convert a_i's size to be identical to that of s_i. The detailed SA fusion can be detailed as $sa_i = Relu(\sigma(DeConv(\phi(a_i))) \odot s_i + s_i) \otimes s_i$, where $\otimes$ is the feature concatenation operation; $DeConv(\cdot)$ is the deconvolution operation; $\odot$ represents the elementwise multiplicative operation; $\sigma(\cdot)$ is the sigmoid function; and $Relu$ is the ReLU function. $\phi(\cdot)$ is a "binary switch" (i.e., the proposed audio switch, which will be detailed later) to eliminate those irrelevant audio signals.

ST fuse module. The methodology of the ST fuse module is quite similar to that of the SA fuse module, and the major difference is that the ST fuse module omits the audio switch. The main reason is that temporal information could be more likely to benefit its spatial counterpart, yet the audio signal could completely conflict

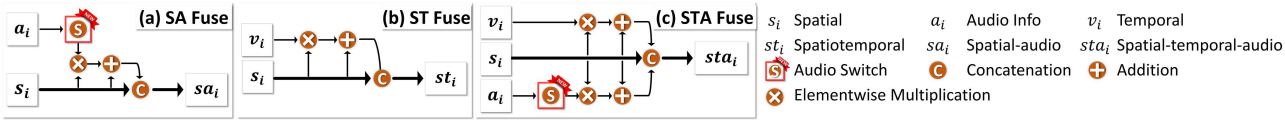

Figure 6 (Color online) Architectures of fusion modules adopted in Figure 7.

with the spatial clue, resulting in learning ambiguity. The technical details of ST fusion (see Figure 6(b)) can be formulated as $st_i = Relu(\sigma(v_i) \odot s_i + s_i) \otimes s_i$, where v_i represents the temporal information after using the 3D convolution and other symbols and operations are completely identical to those of SA fusion (Subsection 3.3).

Audio switch. The main function of this module is to alleviate the potential side effects from the audio signal when performing visual-audio fusion. In fact, the nature of the proposed "audio switch" is a plug-in tool, which can be achieved via an individual network with an identical structure to the 'SA' classification net. For a visual-audio fragment (1 frame and 1 second audio), we assign its binary label to '1' if the audio category predicted by the audio classification tool is identical to the pre-given video category; otherwise, we assign its binary label to '0'.

Classifiers. All classifiers adopted in the abovementioned networks are plain classifiers, which consist of only 2 steps. First, we use a 1×1 convolution to refine the channel size of the input feature tensor (i.e., s_i, sa_i, and st_i) to c. Next, we perform the global average pooling (GAP) operation to transform the tensor to a vector, whose dimension is identical to the video category number c.

Classification loss. All classification networks in SCAM have adopted an identical loss function, i.e., the standard multilabel soft margin (MSM) loss [34]. For better clarification, we take the "ST net" (e.g., Figure 7(c)) as an example here, and its classification loss can be seen as follows:

$$\mathcal{L}_{\mathrm{MSM}}(st_i, \mathrm{VC}) = -\frac{1}{c}\sum_{i=1}^{c} \mathrm{VC}_i \times \log\left(\frac{1}{1+e^{-st_i}}\right) + (1-\mathrm{VC}_i) \times \log\left(\frac{e^{-st_i}}{1+e^{-st_i}}\right), \tag{3}$$

where 'st_i' is the input of the 'classifier' which can be obtained via ST fusion, VC represents the exact class towards the given input, and c is the total video category number.

3.4 SCAM vs. real fixation

The proposed SCAM is mainly based on performing selective fusion over discriminative regions revealed by the source-wise classification nets, i.e., S, ST, and SA nets. Although the pseudofixation maps generated by SCAM basically correlate to the most discriminative regions in the given visual-audio fragment, they might occasionally be different from real human-eye fixations (GTs), as demonstrated in Figure 8, i.e., SCAM vs. GTs.

From the data perspective, the total input of SCAM computation only comprises 3 neighboring video frames and a 1-second audio signal at most, which makes the proposed SCAM a typical local method. However, the real human visual-audio system follows a global manner, where the real fixations are jointly influenced by multiple factors (e.g., visual inertia and associative memory [27]). Thus, the challenge for using SCAM to produce high-quality pseudo-fixations relies on including the source-wise selection process (Eq. (1)) with global information, i.e., multigranularity information. We shall provide an in-depth explanation here.

First, the human visual-audio system tends to omit those image regions that are less salient from the source-wise perspective [31,35], e.g., that occurred repeatedly in the given video sequence, even the most discriminative regions in the current visual-audio fragment. For example, as seen in the 1st column of Figure 8, the real fixations have avoided the '*dog*' since this '*dog*' has occurred multiple times before, making it less salient than the '*hawk*'. However, the SCAM has only considered the current local visual-audio fragment, and thus its pseudofixation map still treats the '*dog*' as the salient one because, from the local perspective, the '*dog*' could also contribute greatly to the current classification task. Therefore, we incorporate both short-term and long-term information into the SCAM.

Second, the real fixation is influenced by the high-level semantic information that has been embedded in our brain. In fact, the human visual-audio system never works alone [36, 37], where the human brain plays a critical role; i.e., the real human fixation tends to vary between subjects with different ages, agendas, careers, etc. The main reason is that the high-level semantic knowledge that we have learned in our daily lives can directly determine where we really look at [38]. Despite the existence of this phenomenon, the principle regards the positive relationship between being discriminative and being salient still holds, and to further refine the pseudo-fixations derived by SCAM, we shall introduce the high-level semantic information into the SCAM. Therefore, we propose to consider the cross-term information, which could be obtained by introducing multiple visual-audio fragments cropped from other video sequences sharing an identical tag with the given sequence.

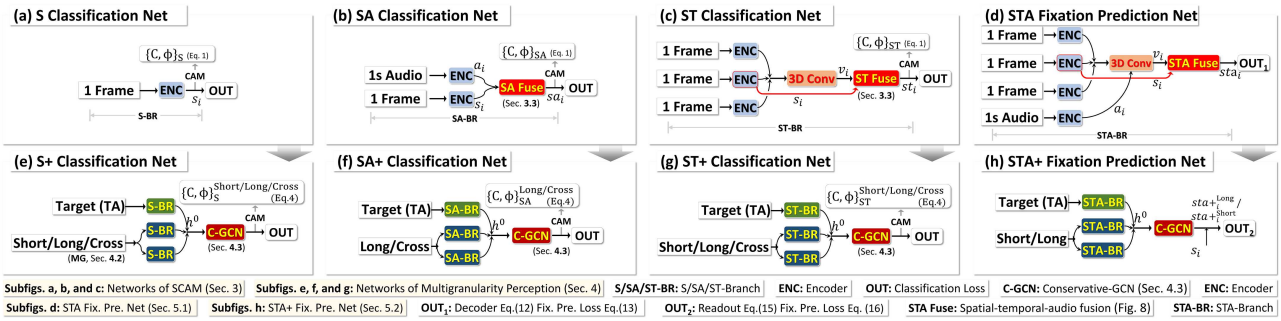

Figure 7 (Color online) Detailed network architectures used in this paper. (a)–(c) show the classification networks used in the proposed SCAM. (e)–(g) show the upgraded classification networks used in SCAM+, which couple the raw variants to form the SCAM+ model. Both SCAM and SCAM+ convert video tags into pseudo-fixations (pGTs), and SCAM+ consistently outperforms SCAM. C-GCN denotes the proposed m-step reasoning module; see Figure 10 and Subsection 4.3 for details. (d) illustrates the visual-audio fixation prediction network, which is trained using the pseudo-fixations as supervision, enabling generic fixation prediction without using video tags. (h) augments (d) with the proposed multigranularity perception module. Details of the "SA/ST Fuse" module are provided in Figure 6.

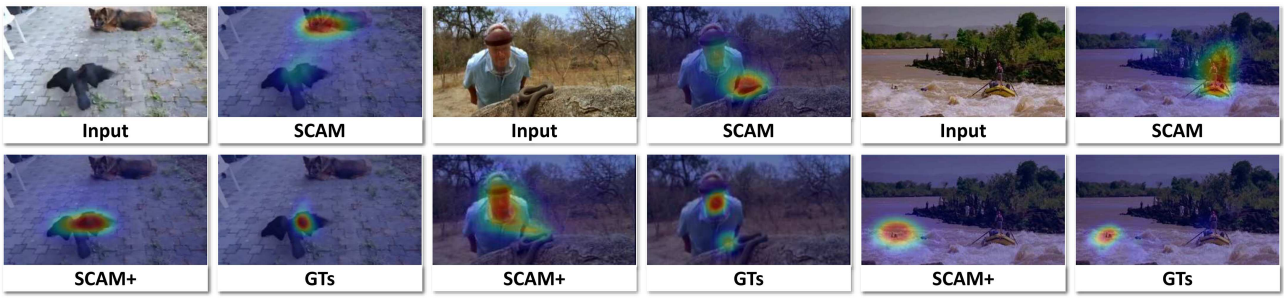

Figure 8 (Color online) Visual comparison of local and global information-based fusion. It can be observed that benefiting from the relational constraints between frames of global information, the global information-based method is capable of handling various challenging factors, such as location bias, color bias, and multiple objects.

4 Selective class activation mapping+

4.1 SCAM+ overview and technical pipeline

As seen in Figure 7, the major difference between SCAM and SCAM+ relies on the adopted classification nets, where only 3 classification nets, i.e., S, ST, and SA (Figures 7(a)–(c)), have been adopted by SCAM, while based on SCAM, the SCAM+ has additionally adopted 3 'types' of classification nets, i.e., S+, SA+, and ST+ (Figures 7(e)–(g)). Note that the combinations of short-/long-/cross-term information are used as the input of these new classification nets, leading to 8 additional classification nets with 8 classification confidences (C) and CAM maps (Φ), including $\{\mathrm{C},\Phi\}^{\mathrm{Short}}_{\mathrm{S+}}$, $\{\mathrm{C},\Phi\}^{\mathrm{Long}}_{\mathrm{S+}}$, $\{\mathrm{C},\Phi\}^{\mathrm{Cross}}_{\mathrm{S+}}$, $\{\mathrm{C},\Phi\}^{\mathrm{Short}}_{\mathrm{ST+}}$, $\{\mathrm{C},\Phi\}^{\mathrm{Long}}_{\mathrm{ST+}}$, $\{\mathrm{C},\Phi\}^{\mathrm{Cross}}_{\mathrm{ST+}}$, $\{\mathrm{C},\Phi\}^{\mathrm{Long}}_{\mathrm{SA+}}$, and $\{\mathrm{C},\Phi\}^{\mathrm{Cross}}_{\mathrm{SA+}}$. The first 3 of them can be derived from the S+ net, the middle 3 can be obtained from the ST+ net, and the last 2 are formulated via the SA+ net. We have omitted $\{\mathrm{C},\Phi\}^{\mathrm{Short}}_{\mathrm{SA+}}$ because the audio signal itself has already covered the short-term information. The SCAM+-based pseudo-fixations can be obtained via the following equation:

$$
\begin{aligned}
\mathrm{SCAM+} =& \frac{\left\|\mathrm{UC}\odot\mathrm{UR}+\mathrm{SC}\odot\mathrm{SR}+\mathrm{LC}\odot\mathrm{LR}+\mathrm{CC}\odot\mathrm{CR}\right\|_1+\lambda}{\left\|\mathrm{UC}+\mathrm{SC}+\mathrm{LC}+\mathrm{CC}\right\|_1+\lambda},\\
\mathrm{SR}:&\left[\Phi^{\mathrm{Short}}_{\mathrm{S+}}\{i\},\Phi^{\mathrm{Short}}_{\mathrm{ST+}}\{i\}\right],\mathrm{LR}:\left[\Phi^{\mathrm{Long}}_{\mathrm{S+}}\{i\},\Phi^{\mathrm{Long}}_{\mathrm{ST+}}\{i\},\Phi^{\mathrm{Long}}_{\mathrm{SA+}}\{i\}\right],\\
\mathrm{CR}:&\left[\Phi^{\mathrm{Cross}}_{\mathrm{S+}}\{i\},\Phi^{\mathrm{Cross}}_{\mathrm{ST+}}\{i\},\Phi^{\mathrm{Cross}}_{\mathrm{SA+}}\{i\}\right],\\
\mathrm{SC}:&\left[\oint\left(\mathrm{C}^{\mathrm{Short}}_{\mathrm{S+}}\{i\}\right),\oint\left(\mathrm{C}^{\mathrm{Short}}_{\mathrm{ST+}}\{i\}\right)\right],\ \mathrm{LC}:\left[\oint\left(\mathrm{C}^{\mathrm{Long}}_{\mathrm{S+}}\{i\}\right),\oint\left(\mathrm{C}^{\mathrm{Long}}_{\mathrm{ST+}}\{i\}\right),\oint\left(\mathrm{C}^{\mathrm{Long}}_{\mathrm{SA+}}\{i\}\right)\right],\\
\mathrm{CC}:&\left[\oint\left(\mathrm{C}^{\mathrm{Cross}}_{\mathrm{S+}}\{i\}\right),\oint\left(\mathrm{C}^{\mathrm{Cross}}_{\mathrm{ST+}}\{i\}\right),\oint\left(\mathrm{C}^{\mathrm{Cross}}_{\mathrm{SA+}}\{i\}\right)\right],
\end{aligned}
\tag{4}
$$

where the meanings of UC, UR, $\|\cdot\|_1$, λ, Φ, $\oint$, and $\odot$ can be found in (1).

Clearly, the proposed SCAM+ has included more granularity when producing pseudo-fixations. To facilitate

a better reading, 2 technical details should still be provided: (1) the exact definition and implementation of the abovementioned multigranularity information (Subsection 4.2) and (2) the exact network architectures, especially the fusion parts, of adopted classification nets (Subsection 4.3).

4.2 Multigranularity information definition

SCAM+ extends beyond the local information used in SCAM, which considers only three consecutive frames and a one-second audio signal (target data). Instead, it integrates multigranularity perception, incorporating spatial, temporal, and auditory cues to enhance fixation prediction. Inspired by human attention mechanisms, SCAM+ simulates short-term memory to process immediate stimuli, long-term memory to maintain fixation consistency over time, and cross-modal reasoning to dynamically integrate multimodal cues. This enables SCAM+ to adapt fixation predictions to dynamic environments, overcoming the limitations of conventional CAM-based approaches that focus only on static salient regions. We shall define these multigranularity components in the following.

Short-term information. In our approach, the short-term information acts as the instant visual stimuli to automatically suppress those less important scene contents and focus our attention on the remaining conspicuous ones. As seen in Figure 9, suppose the t-th video frame is the target frame marked by ①; its short-term spatial-temporal information ST+Short can be detailed as the neighboring 8 video frames marked by ⑨, and all 9 frames together can be divided into 3 groups to be fed into a three-stream classification net, i.e., the ST+ net detailed in Figure 7(g). Specifically, from the sole spatial source perspective (i.e., the S+ net, Figure 7(e)), the short-term information S+Short is the 2 neighboring frames of the target frame, i.e., the mark ②, which could be completely identical to the input of the ST net (Figure 7(c)) adopted in the SCAM, yet the underlying rationale is different in essence. The ST net resorts to a 3D convolution to acquire spatial-temporal information, yet in sharp contrast, the S+ net has aimed to learn the frame-level similarity relationship measurement to mimic the perceptual mechanism of the human association mechanism. Thus, both S+ net and ST net are complementary to each other.

Long-term information. As a complement part to the short-term information, the long-term information also serves the human visual-audio system as visual inertia to focus our attention on the most conspicuous scene regions in the entire video sequence. Long-term information refers to video frames and audio signals that are a long distance away from the current target frame (t). Suppose the t-th frame is the current target frame; the scope of long-term information could be any frame in the same video sequence with the target frame's neighbor frames excluded (i.e., 8 frames). From the spatial perspective, the long-term information S+Long of the t-th frame is 2 other frames that are temporally long distance away from the t-th frame and can be randomly selected (e.g., the mark ③ in Figure 9(a)), and these 3 frames (1 target frame and 2 long-term frames) will be fed into the S+ net, as shown in Figure 7(e). Similarly, from the spatial-temporal perspective, the long-term information ST+Long becomes 2 groups of 3 consecutive frames (e.g., the mark ⑩ in Figure 9(a)), which are also randomly selected. Thus, a total of 9 frames (3 of them are the target consecutive frames, and the other 6 frames are the long-term frames) will be fed into the ST+ net, as shown in Figure 7(g). With regard to the visual-audio perspective SA+Long, the only difference is that the audio signals have been additionally considered (e.g., the mark ⑥ in Figure 9(a)); thus, a total of 3 frames and 3 seconds audio signals are fed into the SA+ net, as demonstrated in Figure 7(f).

Cross-term information. The human visual-audio system is influenced by associative memory, which accumulates basic impressions of different object categories. For example, when presented with multiple cars in a video sequence, our attention may be drawn to the one with the most distinctive appearance, relying on additional information about the current video category (cf. Figure 3). This phenomenon motivates us to consider cross-term information, where frames from other video sequences sharing the same video tag as the current sequence can serve as cross-term information. As seen in Figure 9(a), from the sole spatial domain perspective S+Cross, the cross-term information of the t-th frame in sequence 1 (with video category 2) can be any 2 frames randomly selected from other video sequences k that are also assigned with tag 2, e.g., the mark ④. Similarly, from either the spatial-temporal ST+Cross or visual-audio SA+Cross perspective, the cross-term information can be formulated accordingly as either mark ⑪ or mark ⑦.

4.3 Multigranularity perception

Compared with the SCAM, the upgraded version SCAM+ has adopted several new classification nets, i.e., S+, SA+, and ST+. As seen in Figure 7(g), all these classification networks have directly adopted part of the SCAM's classification nets (i.e., S, SA, and ST) as backbones, i.e., S-/SA-/ST-Branches. Since these feature backbones are almost unaltered, we focus on the subsequent fusion parts. In view of the human attention mechanism, the most valuable short-/long-/cross-term information could be the high-level semantic part, which motivates us to model the semantic relationship between target data and its short-/long-/cross-term companion. To achieve this, we devise

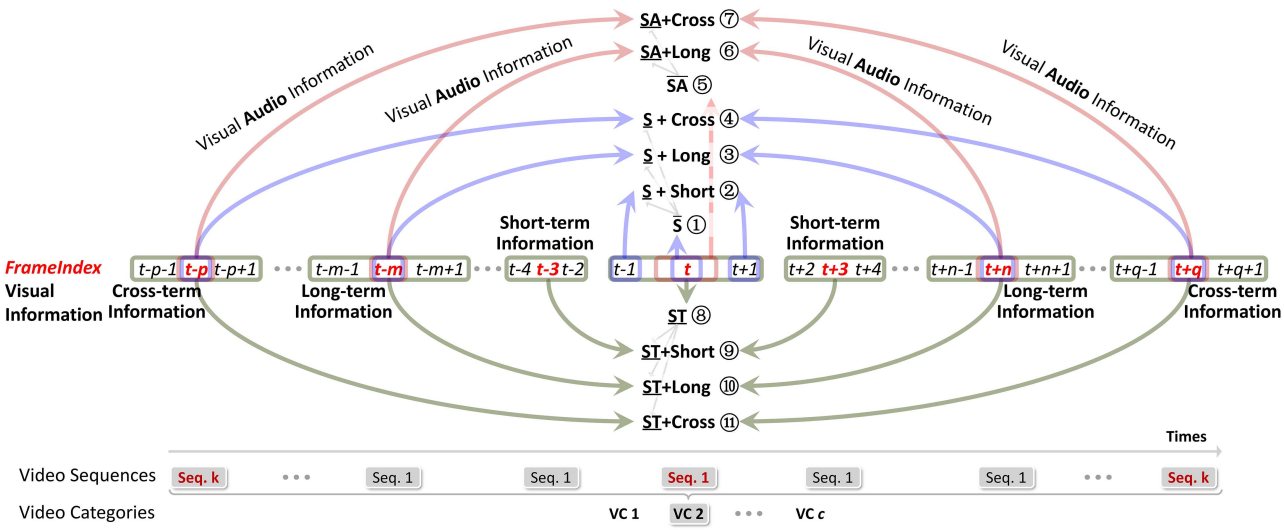

Figure 9 (Color online) Input data formulations and definitions of SCAM+. The numbers of k, m, n, p, and q are randomly selected.

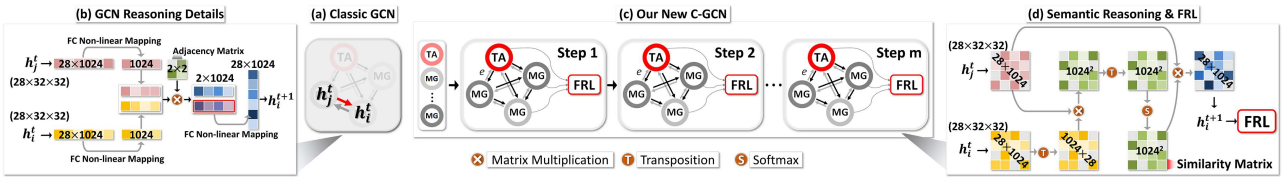

Figure 10 (Color online) Network architectures of classic GCN and the proposed C-GCN. C-GCN mainly consists of 2 parts, i.e., (1) semantic reasoning (Subsection 4.3.1), and (2) fixation refine layer (FRL, Subsection 4.3.2).

a variant of graph convolution network (GCN) [39] to combine the multigranularity information with source-wise information, named conservative-GCN (C-GCN), whose major highlights include (1) a completely new methodology for performing m-step semantic reasoning over multisource and multigranularity nodes and (2) a completely new fixation refine layer to ensure that the whole process is conservative enough to eliminate redundant feature responses.

4.3.1 *Semantic reasoning*

The proposed C-GCN effectively infers complex dependencies and contextual representations in high-level visual-audio semantic information. It achieves this by associating low-level visual-audio cues with high-level semantic information. By abstracting and propagating these cues, C-GCN aligns semantic representations with human interests, fully mimicking the intricate reasoning process of the human brain.

The C-GCN (Figure 10) consists of multiple nodes and edges $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, in which the $\mathcal{V}$ represents the node-set $\{\mathrm{TA}, \mathrm{MG}_1, \ldots, \mathrm{MG}_n\}$, where TA denotes the target node (i.e., the current input visual-audio fragment), and MGs include n nodes correlating to short/long/cross information. We use $\mathcal{E}$ to represent the edge set $\{e_{i,j}\}$, $i \neq j$; and $e_{\mathrm{TA},\mathrm{MG}_i} \in \mathcal{E}$ is the edge between nodes TA and MG_i.

As seen in Figures 7(e)–(g), the input of the C-GCN consists of 3 branches, i.e., the output of the S/SA/ST branch (s_i, sa_i, st_i) containing rich semantic information, and for simplicity, we uniformly define them as h_i^0, where the subscript i denotes that this feature belongs to the i-th node, the superscript 0 denotes the beginning of semantic reasoning. In the t-th reasoning stage, we use the interattention (F_{iatt}, Eq. (6)) to model the semantic relation of each 2 of the input nodes, i.e., the edge $e_{i,j}$, which can be formulated as follows:

$$\boldsymbol{e}^t_{\mathrm{TA},\mathrm{MG}_i} = Softmax\big(F_{iatt}(h^t_{\mathrm{TA}}, h^t_{\mathrm{MG}_i})^\top\big), \boldsymbol{e}^t_{\mathrm{MG}_i,\mathrm{TA}} = Softmax\big(F_{iatt}(h^t_{\mathrm{MG}}, h^t_{\mathrm{TA}_i})^\top\big), \tag{5}$$

where $\boldsymbol{e}^t_{\mathrm{TA},\mathrm{MG}_i} \in \mathbb{R}^{1024\times1024}$, $\top$ is the matrix transpose operation, $F_{iatt}(\cdot,\cdot)$ measures the consistency between its input and can be detailed as follows:

$$F_{iatt}(h^t_{\mathrm{TA}}, h^t_{\mathrm{MG}_i}) = \Big[\mathcal{R}_{28\times1024}\big(Conv_{1\times1}(h^t_{\mathrm{TA}})\big)\Big]^\top \circledast \Big[\mathcal{R}_{28\times1024}\big(Conv_{1\times1}(h^t_{\mathrm{MG}_i})\big)\Big]. \tag{6}$$

Here $\circledast$ denotes matrix multiplication, $h \in \mathbb{R}^{28\times32\times32}$, 32 is the size of the feature map, 28 is the total video

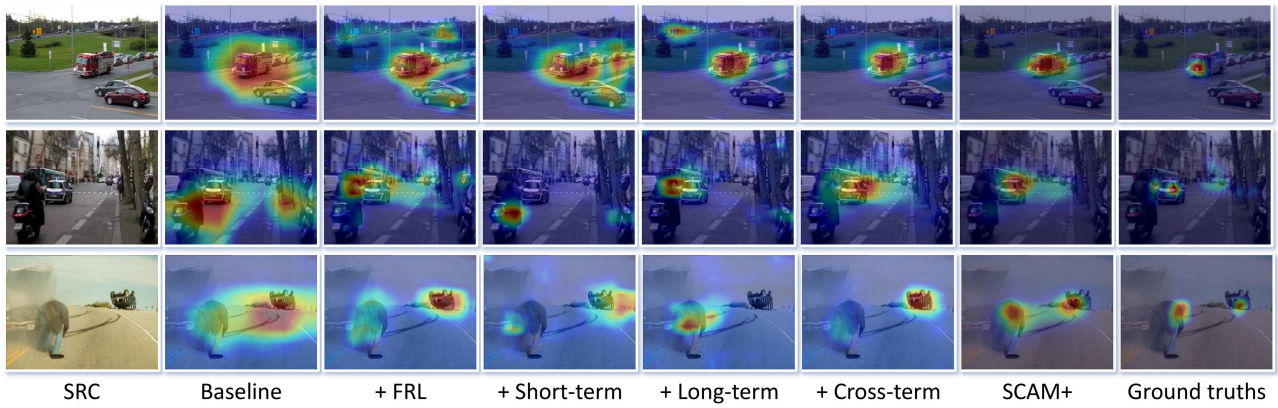

Figure 11 (Color online) Visualization of fixation predictions with different model variations, comparing fixation heatmaps generated by different models. The results show progressive improvements in attention localization as short-term, long-term, and cross-term information are incorporated, with SCAM+ producing the most human-like fixation maps.

category number (c), $\mathcal{R}_{28\times1024}(\cdot)$ reshapes its input to a matrix with size 28×1024, and $Conv_{1\times1}$ is a typical 1×1 convolution.

Actually, the rationale of edge $\boldsymbol{e}^t_{\mathrm{TA,MG}_i}$ is the feature similarity between two neighboring nodes. Since the primary function of the reasoning process is to exchange/share information between nodes and formulate a series of implicit principles toward the given learning objective, we update all nodes' statuses after each reasoning stage. We formulate the updating process towards the target node (TA) using the following equation:

$$h^{t+1}_{\mathrm{TA}} \leftarrow \mathcal{R}_{28\times32^2}\left(h^t_{\mathrm{TA}} \odot \sigma\left(\sum_{i\in\mathcal{N}_{\mathrm{TA}}} \mathcal{R}_{28\times1024}(h^t_{\mathrm{MG}_i}) \circledast \boldsymbol{e}^t_{\mathrm{TA,MG}_i}\right)\right), \tag{7}$$

where $\sigma(\cdot)$ is a typical sigmoid function; $\circledast$ denotes matrix multiplication; $\odot$ represents elementwise multiplicative operation; $\times$ is the standard multiplication; $\mathcal{N}_{\mathrm{TA}}$ includes all nodes neighboring the TA; and $\mathcal{R}_{28\times32^2}(\cdot)$ reshapes its input to a tensor with size 28×32^2. Similarly, the updating process towards the multigranularity nodes (e.g., MG_i) can be formulated as follows:

$$h^{t+1}_{\mathrm{MG}_i} \leftarrow \mathcal{R}_{28\times32^2}\left(h^t_{\mathrm{MG}_i} \odot \sigma\left(\sum_{j\in\mathcal{N}_{\mathrm{MG}_i}} \mathcal{R}_{28\times1024}(h^t_{\mathrm{MG}_j}) \circledast \boldsymbol{e}^t_{\mathrm{MG}_i,\mathrm{MG}_j} + \mathcal{R}_{28\times1024}(h^t_{\mathrm{TA}}) \circledast \boldsymbol{e}^t_{\mathrm{MG}_i,\mathrm{TA}}\right)\right). \tag{8}$$

By using the node-wise reasoning process, the high-level semantic information between the target input and its multigranularity information can be obtained. The final status of the target node h^m_{TA} (m is reasoning steps) is fed into a classifier. The reasoning steps, the number of MG nodes, and the replacement of C-GCN with Transformer are discussed in Appendixes C–E.

4.3.2 *Fixation refine layer*

In most cases, the C-GCN-based reasoning process can appropriately fuse both multisource and multigranularity information. However, there still exists one issue regarding the reasoning process to be handled. Although the reasoning steps can implicitly incorporate the multigranularity information, the addition operation-based updating process (e.g., Eq. (7)) could lead to the target node's feature response map being redundant, resulting in pseudo-fixations that differ from the real fixations (e.g., Figure 11). Thus, for each reasoning stage, we have additionally assigned one fixation refine layer (FRL, Figure 10) to eliminate redundant feature responses. The technical details of the proposed FRL are as follows.

First, we establish a binary matrix $\mathrm{rMASK}\in\{0,1\}^{32\times32}$ to indicate which elements in the fused feature tensor have relatively large feature responses, and these qualified elements are more likely to belong to the most discriminative frame regions. The computation of rMASK is as follows:

$$\mathrm{rMASK} = \left[max\Big(\mathrm{cMean}\big(h^t_i\big)\Big) \times \mathcal{T}_d - \mathrm{cMean}\big(h^t_i\big)\right]_+, \tag{9}$$

where $\times$ is the standard multiplication, $\mathrm{cMean}(\cdot)$ is a channelwise average operation, which converts a tensor to a matrix, and $max(\cdot)$ is a typical max operation returning the largest value in its input matrix; h^t_i is a tensor feature,

which can be obtained by either (7) or (8); $[\cdot]_+$ converts all its negative elements to 0 and converts the remaining positive elements to 1; and $\mathcal{T}_d$ is a predefined hard threshold.

Next, we use rMASK to filter redundant information in h_i^t, and this process can be formulated as

$$h_i^{t'} = \frac{1}{2} \times \Big[h_i^t \odot \Big(\mathcal{T}_r \times \sigma \left(\mathrm{cMean}(h_i^t) \right) \odot \mathrm{rMASK} + \sigma \left(\mathrm{cMean}(h_i^t) \right) \odot (1 - \mathrm{rMASK}) \Big) + h_i^t \Big], \tag{10}$$

where $\odot$ is the elementwise multiplicative operation and $\mathcal{T}_r$ is the refinement rate assigned by a predefined value. The rationale of (10) can be explained as: since the rMASK can indicate those spatial regions with large feature responses, we use it as an attention filter ($\mathcal{T}_r$ controls the compression rate) for those less trustworthy (i.e., with relatively low responses) regions—those regions are more likely to be redundant. The exact choices of $\mathcal{T}_d$ and $\mathcal{T}_r$ will be discussed in Appendix F.2.

5 Generic fixation prediction networks

SCAM+ (Eq. (4)) generates extensive pseudo-visual-audio fixation maps, which generally align with real fixations. However, its reliance on human-provided video tags limits its applicability. To address this, we use SCAM+-based maps as learning objectives for knowledge distillation, enabling end-to-end fixation prediction without human-provided tags.

5.1 Source-wise fixation prediction network

As seen in Figure 7(d), the implementation of spatial-temporal-audio (STA) fixation prediction network is very intuitive, where the spatial features s_i are fused with either temporal features v_i or audio features a_i in advance and later combined via the simplest feature concatenation operation. Then, a typical decoder with 3 deconvolutional layers is used to convert the feature maps derived from the "STA Fuse" module to fixation maps. The data flow in the "STA Fuse" module can be formulated as

$$sta_i = Relu\Big[Cov\Big(\Big(\sigma\big(\phi(a_i)\big) \odot s_i + s_i \Big) \otimes \Big(\sigma\big(v_i\big) \odot s_i + s_i \Big) \Big) \Big], \tag{11}$$

where $\otimes$ is the typical concatenation operation; $\phi(\cdot)$ is the proposed audio switch (Subsection 3.3); $Cov(\cdot)$ denotes 1×1 convolution; all other symbols are identical to those in SA fusion (Subsection 3.3).

The sta_i will be fed into the decoder, whose output ($\widehat{sta}_i$) can be formulated as

$$\widehat{sta}_i = \uparrow (Ref(Ref(\mathrm{F3}) \otimes \uparrow (Ref(Ref(\mathrm{F4}) \otimes \uparrow (Ref(Ref(\mathrm{F5}) \otimes sta_i)))))), \tag{12}$$

where $\uparrow(\cdot)$ is the upsampling operation, $Ref(\cdot)$ refines all its input to a fixed channel number (32), $\otimes$ is the feature concatenation operation, and F3, F4, and F5 are the side outputs of the encoder (ENC) correlated to the target frame.

For the training process, we choose 2 typical loss functions, i.e., the binary cross-entropy loss $\mathcal{L}_{\mathrm{BCE}}$ [40] and the Kullback-Leibler divergence loss $\mathcal{L}_{\mathrm{KL}}$ [34], and thus the overall loss function $\mathcal{L}_{\mathrm{STA}}$ can be formulated simply as

$$\mathcal{L}_{\mathrm{STA}} = \mathcal{L}_{\mathrm{BCE}}(\widehat{sta}_i, \mathrm{pGT}) + \mathcal{L}_{\mathrm{KL}}(\widehat{sta}_i, \mathrm{pGT}). \tag{13}$$

Clearly, the training process of the proposed STA fixation prediction network relies on pseudo-fixations (pGTs) only; thus, it is able to perform end-to-end fixation prediction for unseen visual-audio sequences without any category tag.

5.2 Multigranularity fixation prediction network

As shown in Figure 7(h), the STA fixation prediction network can also be upgraded by combining it with the multigranularity perception mechanism. We name the upgraded version STA+, whose input also consists of 3 parts, which are the output of the STA net with different input data.

In the STA+ net, we shall only additionally consider both short-term and long-term information, where the cross-term information has been omitted to ensure its versatility. Similar to the classification nets mentioned before, we continue using the proposed C-GCN as the blender. Thus, we independently train 2 STA+ nets, where the input

data (3 parts) are {1 target + 2 short} or {1 target + 2 long}, and the dataflow of the fusion part in these 2 STA+ nets (e.g., Figure 7(h)) can be expressed uniformly as the following equation:

$$\begin{aligned} sta{+}_i^{\text{Short/Long}} = \Big\{h^m_{\text{TA}}, h^m_{\text{MG}_1}, h^m_{\text{MG}_2}\Big\} =& \Omega\Big[sta_i,\ sta_i^{\text{Short}}(1),\ sta_i^{\text{Short}}(2)\Big] \\ & \text{or } \Omega\Big[sta_i,\ sta_i^{\text{Long}}(1),\ sta_i^{\text{Long}}(2)\Big], \end{aligned} \tag{14}$$

where $sta{+}_i^{\text{Short/Long}}$ is the output of the m step reasoning, which includes 3 tensors (i.e., $h^m_{\text{TA}}, h^m_{\text{MG}_1}, h^{\text{m}}_{\text{MG}_2}$) with size $28 \times 32 \times 32$, and these tensors correlate to 3 different nodes (i.e., 1 TA node and 2 MG nodes, respectively, see (7) and (8)); $\Omega[\cdot]$ denotes the proposed C-GCN mentioned in Subsection 4.3; $sta_i^{\text{Short}}(1)$ and $sta_i^{\text{Short}}(2)$ denote 2 individual outputs of the 'STA-Branch' using different short-term fragments as input, and, $sta_i^{\text{Long}}(1)$ and $sta_i^{\text{Long}}(2)$ are 2 outputs correlated to 2 different long-term input fragments.

The 'ReadOut' module takes $sta{+}_i^{\text{Short/Long}}$ as input individually and then outputs 9 fixation maps. Taking the target node for instance, we abbreviate $sta{+}_i^{\text{Short/Long}}$ as h^m_{TA}, where m is the total reasoning steps, and the output of the 'ReadOut' can be formulated as

$$\hat{h}_{\text{TA}} = RO(h^m_{\text{TA}}) = Ref_1\Big(\uparrow\big(Ref_{32}(h^m_{\text{TA}} \otimes S)\big)\Big), \tag{15}$$

where S denotes the spatial feature correlated to the middle frame of the target fragment, e.g., the s_i in Figure 7(d); RO represents the 'ReadOut' in Figure 7(h); $Ref_p(\cdot)$ is a refinement operation which uses 1×1 convolution to reduce input's channel number to p; $\uparrow(\cdot)$ upsamples its input to 356×356. The output could become $\hat{h}_{\text{MG}_1}$ and $\hat{h}_{\text{MG}_2}$ by alternately feeding $h^m_{\text{MG}_1}$ and $h^m_{\text{MG}_2}$ to the 'ReadOut' module of STA+.

We assign an individual loss function for each fixation prediction net, and the loss function adopted by these 2 fixation prediction nets (STA+$^{\text{Short}}$, and STA+$^{\text{Long}}$, 14) can be uniformly expressed as

$$\begin{aligned} \mathcal{L}_{\text{STA+}} =& \mathcal{L}_{\text{BCE}}(\hat{h}_{\text{TA}}, \text{pGT}_{\text{TA}}) + \mathcal{L}_{\text{KL}}(\hat{h}_{\text{TA}}, \text{pGT}_{\text{TA}}) + \mathcal{L}_{\text{BCE}}(\hat{h}_{\text{MG}_1}, \text{pGT}_{\text{MG}_1}) \\ & + \mathcal{L}_{\text{KL}}(\hat{h}_{\text{MG}_1}, \text{pGT}_{\text{MG}_1}) + \mathcal{L}_{\text{BCE}}(\hat{h}_{\text{MG}_2}, \text{pGT}_{\text{MG}_2}) + \mathcal{L}_{\text{KL}}(\hat{h}_{\text{MG}_2}, \text{pGT}_{\text{MG}_2}), \end{aligned} \tag{16}$$

where $\text{pGT}_{\text{TA/MG}_1/\text{MG}_2}$ denotes the correlated pseudofixation maps of different nodes, the input $\hat{h}_{\text{TA}}$ can be obtained via (15) directly, and the computation of $\hat{h}_{\text{MG}_1/\text{MG}_2}$ shall replace Eq. (15)'s input to $h^m_{\text{MG}_1}$ and $h^m_{\text{MG}_2}$ accordingly.

5.3 Predicted visual-audio fixations

Thus far, we can obtain 3 individual fixation prediction networks mentioned in Subsections 5.1 and 5.2; i.e., from the source-wise perspective, we can obtain an STA net, and from the multigranularity aspect, we can additionally obtain 2 STA+ nets, which correlate to short-term or long-term versions, respectively (Eq. (14)). These 3 fixation prediction nets complement each other in essence, and we shall combine their predictions to derive the final fixation maps that could outperform each of them.

We define the predicted fixation map of the 3 prediction nets (STA, STA+$^{\text{Short}}$, and STA+$^{\text{Long}}$) as PF_{sta}, $\text{PF}^{\text{Short}}_{sta+}$, and $\text{PF}^{\text{Long}}_{sta+}$. The final predicted visual-audio fixation map (PF_{final}) can be formulated as follows:

$$\text{PF}_{final} = \frac{1}{2} \times \mathcal{Z}\Big(\mathcal{Z}(\text{PF}_{sta}) \odot \mathcal{Z}(\text{PF}^{\text{Short}}_{sta+}) \odot \mathcal{Z}(\text{PF}^{\text{Long}}_{sta+})\Big) + \frac{1}{2} \times \mathcal{Z}\Big(\text{PF}_{sta} + \text{PF}^{\text{Short}}_{sta+} + \text{PF}^{\text{Long}}_{sta+}\Big), \tag{17}$$

where $\mathcal{Z}(\cdot)$ is the typical min-max normalization function, $\times$ represents the standard multiplicative operation. and $\odot$ denotes the element multiplicative operation. Eq. (17) mainly consists of 2 parts, where the left part obtains the common consistency, while the right part serves as the complementary fixation map to ensure good robustness. We verify the advantages of this fusion scheme over conventional plain schemes from a quantitative perspective in Appendix F.6.

6 Experiments and validations

6.1 Implementation details

Training dataset. We use the Audio-Visual Event (AVE) dataset [18], a subset of Google's AudioSet, containing 4143 sequences across 28 categories, as the classification training set for S, ST, SA, S+, ST+, and SA+ networks (Sections 3 and 4).

Training processes. SCAM and SCAM+ are trained using a multistage scheme to generate pseudo-GTs for S, ST, SA, S+, ST+, and SA+ branches. In the coarse stage, SCAM processes video frames resized to 256×256 with a batch size of 20. In the fine stage, three classification networks are trained separately with 356×356 patches and a batch size of 3. SCAM+ follows a similar setup with a batch size of 16. STA and STA+ use pseudo-fixations as GTs, trained with 356×356 frames, a batch size of 3, and SGD at a 0.00005 learning rate.

Testing data sets. To test the performance of the proposed approach, we adopted 6 testing datasets, including AVAD [41], Coutrot1 [42], Coutrot2 [43], DIEM [44], SumMe [45], and ETMD [46].

Quantitative metrics. Following prior research [34, 47], we adopted five standard metrics (AUC-J, s-AUC, NSS, SIM, and CC) to compare our model's saliency predictions with actual human eye movements.

6.2 Effectiveness evaluation on different components

6.2.1 *Effectiveness of the proposed multistage rationale*

In Subsection 3.2, we adopted a coarse-to-fine approach to perform SCAM twice, mimicking the human attention mechanism. To verify this strategy, we tested CAM results from different sources (S, SA, ST) at various stages (COARSE vs. FINE). The quantitative evaluation in Table 1 (marked by ①) shows the consistency between CAM results and real human-eye fixation data. The FINE stage significantly improves results, with a 40% boost when using a single source (S, SA, or ST; see lines 1–3 and 6–8) and notable improvements with multiple sources (see lines 4 and 10). Comparing lines 5 and 11 demonstrates a solid performance gain, highlighting the generic nature of the multistage strategy.

6.2.2 *Effectiveness of the proposed selective fusion*

To selectively fuse the multisource CAM results (i.e., Φ_{S}, Φ_{SA}, and Φ_{ST}, Eq. (1)), we have adopted the classification confidences as the fusion weights ($\mathrm{C_S}$, $\mathrm{C_{SA}}$, and $\mathrm{C_{ST}}$). This is based on the precondition that higher classification confidence indicates better alignment with real fixations. To verify this issue, we have compared the proposed selective fusion (i.e., SCAM, Eq. (1)) with the conventional fusion scheme that averages all CAM results: $\mathrm{AC} = \frac{1}{3}\big(\Phi_{\mathrm{S}} + \Phi_{\mathrm{SA}} + \Phi_{\mathrm{ST}}\big)$, where all symbols have the same definitions as those of Eq. (1), and we show the corresponding quantitative result of this scheme in the 'G' column of Table 1. The mark ② in Table 1 has highlighted the advantage of the proposed SCAM against the conventional scheme, i.e., SCAM vs. AC, where the overall quantitative performance can obtain persistent improvement for all considered metrics.

6.2.3 *The effectiveness of the proposed audio switch*

In Table 1, we have reported the model's performance without the audio switch (Subsection 3.3), highlighted by mark ③. Comparing lines 9 and 11, and lines 22 and 24, shows that the audio switch improves performance by about 1.5% on average by filtering meaningless background audio and reducing learning ambiguity from unsynchronized spatial and audio information.

6.2.4 *Effectiveness of multigranularity selective fusion*

To study the influence of the proposed multigranularity information on the proposed SCAM, we tested all CAMs derived from different sources (i.e., S, SA, and ST) coupled with multigranularity information, i.e., short-term, long-term and cross-term information (the 'I', 'G', and 'K' columns of Table 1).

As seen in Table 1, marked by ④, we found the performance gain when using each type of multigranularity information with the source-wise CAM results. Meanwhile, we have also tested using all types of multigranularity information simultaneously, where the 'M' and 'N' columns denote the averaged and selectively fused SC, LC, and CC, respectively.

Clearly, compared with the CAM result obtained via a single source (line 12), the multigranularity information (line 13) could bring solid performance gain. However, simply combining the single-source-based CAM result with one type of multigranularity information might still be inferior to the SCAM, which can be confirmed by comparing lines 11 and 13. The main reason is that, compared with the source-wise information, the multigranularity information could be less important to the classification task, motivating us to use it to serve the SCAM as the subordinates.

As expected, since the 'M' column (line 23) has adopted multiple multigranularity information, the overall performance could be improved compared with the single multigranularity information-based ones (e.g., lines 13–20). In addition, benefiting from selective fusion, the 'N' column (lines 21, 22, and 24) could further improve

Table 1 Quantitative evidence towards the component studies for STA and STA+. To facilitate the display, we use different colors to distinguish different components. This experiment is conducted on the AVAD set [41]. Here max J↑, S↑, A↑, C↑, and N↑ indicate maximal AUC-J, SIM, s-AUC, CC, and NSS, respectively. A: spatial (S); B: spatial + audio (SA); C: spatial + temporal (ST); D: coarse stage; E: fine stage (Subsection 3.2); F: audio switch (Subsection 3.3); G: average all CAM results (S, ST, SA); H: selective class activation mapping; I: multigranularity (short-term only); J: multigranularity (long-term only); K: multigranularity (cross-term only); L: fixation refine layer (FRL, Subsection 4.3.2); M: averaged short, long, and cross; N: selective combine short, long, and cross; O: combine SCAM with multigranularity; ① Subsection 6.2.1: multi-stage SCAM (Subsection 3.2); ② Subsection 6.2.2: SCAM; ③ Subsection 6.2.3: audio switch (Subsection 3.3); ④ Subsection 6.2.4: multigranularity (Subsection 4.2); ⑤ Subsection 6.2.4: SCAM+; ⑥ Subsection 6.2.5: FRL (Subsection 4.3.2).

		Major components															Metrics & results					
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	J↑	S↑	A↑	C↑	N↑	
①⑥	1	✓			✓												0.774	0.202	0.545	0.261	1.269	
①	2		✓		✓		✓										0.785	0.223	0.536	0.269	1.292	
①	3			✓	✓												0.780	0.214	0.542	0.277	1.276	
①②	4	✓	✓	✓	✓		✓	✓									0.786	0.219	0.538	0.297	1.312	
①②	5	✓	✓	✓	✓		✓		✓								0.801	0.256	0.554	0.345	1.364	SCAM
①②	6	✓				✓											0.834	0.291	0.574	0.376	1.528	
①②	7		✓			✓	✓										0.843	0.289	0.571	0.372	1.581	
①②	8			✓		✓											0.845	0.304	0.564	0.384	1.622	
②③	9	✓	✓	✓		✓			✓								0.864	0.33	0.571	0.421	1.833	
②	10	✓	✓	✓		✓	✓	✓									0.845	0.303	0.573	0.399	1.797	
②③	11	✓	✓	✓		✓	✓		✓								0.873	0.334	0.58	0.438	2.018	
⑥	12	✓			✓								✓				0.807	0.233	0.546	0.304	1.435	
④	13	✓								✓			✓				0.829	0.277	0.554	0.372	1.556	
④	14			✓						✓			✓				0.839	0.268	0.565	0.379	1.568	
④	15	✓									✓		✓				0.833	0.28	0.559	0.374	1.55	
④	16		✓				✓				✓		✓				0.835	0.271	0.556	0.377	1.574	
④	17			✓							✓		✓				0.836	0.272	0.55	0.373	1.569	+Multi-granularity
④	18	✓										✓	✓				0.841	0.269	0.561	0.368	1.553	
④	19		✓				✓					✓	✓				0.838	0.293	0.562	0.381	1.561	
④	20			✓								✓	✓				0.831	0.284	0.558	0.382	1.585	
④⑥	21	✓	✓	✓			✓			✓	✓	✓			✓		0.856	0.312	0.564	0.436	1.857	
③④⑥	22	✓	✓	✓						✓	✓	✓	✓		✓		0.875	0.329	0.587	0.461	1.893	
③④⑥	23	✓	✓	✓			✓			✓	✓	✓	✓	✓			0.864	0.319	0.585	0.432	1.903	
④⑥	24	✓	✓	✓			✓			✓	✓	✓	✓		✓		0.886	0.336	0.596	0.467	2.156	
⑤	25	✓	✓	✓		✓	✓			✓	✓	✓	✓			✓	**0.887**	**0.361**	**0.595**	**0.49**	**2.318**	SCAM+

the overall performance. After combining full source-wise information and full multigranularity information, the proposed SCAM+, marked by ⑤, achieves the best performance.

6.2.5 *Effectiveness of fixation refine layer*

To verify the effectiveness of the proposed fixation refine layer (FRL), we conducted a series of evaluations in Table 1, and the corresponding results are marked by ⑥.

First, as seen in line 12, we applied the FRL to the spatial source-based coarse stage feature, where the FRL is directly applied to S_i (Figure 7(a)), then generated the FRL-based CAM result. Comparing line 12 with line 1, a clear performance improvement can be observed easily. Second, we have also tried to remove the FRL from the multigranularity (the 'N' column) to verify the performance gap. By comparing line 21 with line 24 of the 'L' column, we can notice a clear performance decrease.

6.3 Quantitative comparisons with SOTA work

6.3.1 *Quantitative comparisons with weakly supervised methods*

As shown in Table 2, we compare our method with other fully- and weakly-supervised methods, e.g., SSCAM, Sco-CAM, ISCAM, EGCAM, ECAM, LCAM (related details in [48]); SPG [49], VUNP [22], WSS [29], MWS [28], WSSA [31], STANet [26]; DeepNet [11], SalGAN [40], DeepVS [50], ACLNet [47], HD2S [10], AVCRL [37], TMFINet [1], GSGNet [51], MSPI [36], WSVS [52], KDCI [38], CBA [53], EAVOL [54], and BCAM [55] on 6 datasets. STANet and STANet+ consistently outperform weakly supervised baselines across multiple datasets, with STANet+ achieving a +0.46 NSS increase and a +7.4% AUC-Judd improvement. These performance gains

Table 2 Quantitative comparison with SOTA methods on six datasets. Bold means the best result, STANet+^{b} denotes the proposed STANet+ trained on the VGGSound dataset [33]. In addition, we have also provided some representative "qualitative comparisons" between our approach and the SOTA work, which can be found in Appendix E. Here max J↑, S↑, A↑, C↑, and N↑ indicate maximal AUC-J, SIM, s-AUC, CC, and NSS, respectively. Compared with fully supervised methods, our method has adopted a different training protocol; i.e., SOTA fully supervised methods were trained on the widely-used training set (2857 videos [47]) with real human-eye fixations, yet our method was solely trained on the AVE set, a sub-set of the AudioSet, containing 4143 clips equipped with semantic tags.

	Method	AVAD					DIEM					SumMe					ETMD					Coutrot 1					Coutrot 2				
		J↑	S↑	A↑	C↑	N↑	J↑	S↑	A↑	C↑	N↑	J↑	S↑	A↑	C↑	N↑	J↑	S↑	A↑	C↑	N↑	J↑	S↑	A↑	C↑	N↑	J↑	S↑	A↑	C↑	N↑
Fully-supervised	DeepNet	0.869	0.256	0.561	0.383	1.850	0.832	0.318	0.622	0.407	1.520	0.848	0.227	0.645	0.332	1.550	0.889	0.225	0.699	0.387	1.900	0.824	0.273	0.559	0.346	1.410	0.896	0.201	0.600	0.301	1.820
	SalGAN	0.886	0.360	0.579	0.491	2.550	0.857	0.393	0.660	0.486	1.890	0.875	0.289	0.688	0.397	1.970	0.903	0.311	0.746	0.476	2.460	0.853	0.332	0.579	0.416	1.850	0.933	0.290	0.618	0.439	2.960
	DeepVS	0.896	0.391	0.585	0.528	3.010	0.840	0.392	0.625	0.452	1.860	0.842	0.262	0.612	0.317	1.620	0.904	0.349	0.686	0.461	2.480	0.830	0.317	0.561	0.359	1.770	0.925	0.259	0.646	0.449	3.790
	ACLNet	0.905	0.446	0.560	0.580	3.170	0.869	0.427	0.622	0.522	2.020	0.868	0.296	0.609	0.379	1.790	0.915	0.329	0.675	0.477	2.360	0.850	0.361	0.542	0.425	1.920	0.926	0.322	0.594	0.448	3.160
	HD2S	0.905	0.462	0.607	0.600	3.100	0.880	0.502	0.701	0.604	2.200	0.871	0.349	0.628	0.360	1.880	0.913	0.412	0.700	0.502	2.540	0.852	0.396	0.582	0.467	2.070	0.890	0.331	0.645	0.451	2.830
	AVCRL	0.925	0.482	0.620	0.626	3.570	0.894	0.492	0.698	0.592	2.330	0.895	0.343	0.671	0.439	2.250	0.935	0.433	0.741	0.566	3.050	0.880	0.415	0.604	0.499	2.440	0.954	0.527	0.729	0.748	5.450
	TMFINet	0.900	0.515	0.618	0.675	3.640	0.895	0.529	0.701	0.648	2.440	0.870	0.378	0.646	0.468	2.140	0.902	0.411	0.702	0.524	2.680	0.870	0.418	0.605	0.522	2.420	0.920	0.425	0.689	0.591	4.360
	GSGNet	0.910	0.400	0.605	0.548	2.870	0.865	0.328	0.599	0.417	1.900	0.878	0.319	0.692	0.420	1.940	0.910	0.341	0.754	0.502	2.600	0.865	0.328	0.599	0.417	1.900	0.931	0.300	0.681	0.450	2.930
	MSPI	0.932	0.522	0.612	0.688	**3.820**	0.906	0.527	0.694	0.650	**2.590**	0.899	0.367	0.682	0.476	**2.450**	0.935	0.442	0.746	0.592	**3.150**	0.890	0.432	0.606	0.540	**2.580**	0.961	0.553	0.715	0.766	**6.210**
Weakly-supervised	ISCAM	0.774	0.195	0.545	0.240	1.030	0.774	0.256	0.619	0.282	0.990	0.761	0.183	0.582	0.208	0.900	0.738	0.147	0.585	0.157	0.710	0.704	0.201	0.520	0.171	0.650	0.480	0.135	0.409	0.018	0.120
	SSCAM	0.777	0.186	0.531	0.228	0.970	0.750	0.242	0.604	0.248	0.870	0.763	0.179	0.591	0.206	0.830	0.730	0.144	0.595	0.149	0.680	0.686	0.194	0.521	0.154	0.590	0.502	0.141	0.417	0.003	0.020
	ScoCAM	0.772	0.196	0.548	0.237	1.020	0.770	0.257	0.614	0.279	0.980	0.753	0.182	0.577	0.202	0.800	0.737	0.148	0.581	0.157	0.700	0.708	0.203	0.518	0.176	0.670	0.538	0.143	0.423	0.018	0.070
	EGCAM	0.737	0.222	0.533	0.212	0.910	0.758	0.310	0.618	0.308	1.100	0.741	0.220	0.583	0.215	0.870	0.687	0.156	0.570	0.124	0.580	0.640	0.193	0.509	0.114	0.410	0.575	0.107	0.425	0.031	0.180
	ECAM	0.725	0.219	0.526	0.205	0.880	0.740	0.294	0.605	0.273	0.970	0.727	0.211	0.570	0.198	0.790	0.683	0.151	0.555	0.116	0.540	0.647	0.200	0.508	0.130	0.460	0.610	0.104	0.415	0.037	0.220
	LCAM	0.776	0.199	0.542	0.241	1.030	0.773	0.259	0.616	0.285	1.010	0.778	0.189	0.593	0.228	0.910	0.749	0.151	0.581	0.168	0.750	0.699	0.201	0.516	0.168	0.620	0.511	0.141	0.400	0.003	0.030
	WSS	0.858	0.292	0.592	0.347	1.660	0.803	0.333	0.620	0.344	1.290	0.812	0.245	0.589	0.279	1.100	0.854	0.277	0.661	0.334	1.650	0.772	0.247	0.547	0.233	0.980	0.835	0.208	0.578	0.192	1.180
	SPG	0.662	0.176	0.506	0.165	0.730	0.713	0.238	0.579	0.233	0.860	0.714	0.182	0.561	0.209	0.910	0.695	0.138	0.550	0.144	0.690	0.650	0.187	0.505	0.142	0.530	0.511	0.123	0.464	0.017	0.070
	MWS	0.834	0.272	0.573	0.309	1.480	0.806	0.336	0.628	0.350	1.310	0.808	0.237	0.607	0.258	1.160	0.833	0.237	0.649	0.293	1.430	0.743	0.231	0.528	0.201	0.800	0.839	0.188	0.581	0.168	1.200
	WSSA	0.807	0.261	0.574	0.285	1.340	0.767	0.305	0.608	0.311	1.180	0.755	0.225	0.585	0.231	1.060	0.793	0.201	0.622	0.222	1.080	0.701	0.180	0.535	0.169	0.780	0.797	0.185	0.571	0.180	1.260
	VUNP	0.574	0.067	0.500	0.142	0.290	0.558	0.047	0.515	0.172	0.190	0.555	0.013	0.507	0.114	0.050	0.505	0.030	0.103	0.132	0.590	0.589	0.063	0.514	0.152	0.300	0.661	0.101	0.536	0.162	0.490
	WSVS	0.873	0.285	0.590	0.409	1.970	0.838	0.333	0.661	0.418	1.610	0.846	0.270	0.640	0.354	1.610	0.882	0.246	0.691	0.383	1.920	0.820	0.263	0.547	0.313	1.440	0.866	0.233	0.607	0.315	1.870
	BCAM	0.728	0.234	0.529	0.252	1.210	0.812	0.305	0.601	0.324	1.090	0.800	0.212	0.591	0.315	1.330	0.809	0.212	0.568	0.210	1.050	0.691	0.213	0.501	0.254	1.400	0.767	0.158	0.563	0.142	1.250
	EAVOL	0.792	0.266	0.519	0.295	1.320	0.807	0.340	0.613	0.349	1.040	0.801	0.227	0.616	0.307	1.260	0.815	0.231	0.586	0.256	1.230	0.745	0.219	0.517	0.258	1.030	0.766	0.180	0.586	0.182	1.510
	KDCI	0.708	0.166	0.517	0.159	0.709	0.693	0.226	0.530	0.196	0.666	0.701	0.162	0.520	0.152	0.589	0.745	0.144	0.553	0.175	0.783	0.672	0.189	0.493	0.149	0.543	0.601	0.139	0.561	0.052	0.249
	CBA	0.774	0.215	0.524	0.257	1.080	0.803	0.308	0.603	0.343	1.200	0.809	0.225	0.596	0.280	1.160	0.801	0.187	0.573	0.238	1.070	0.758	0.237	0.524	0.240	0.914	0.758	0.188	0.579	0.172	0.898
	STANet	0.873	0.334	0.580	0.438	2.020	0.861	0.391	0.658	0.469	1.720	0.854	0.294	0.627	0.368	1.650	0.908	0.318	0.682	0.448	2.180	0.829	0.306	0.542	0.339	1.380	0.850	0.247	0.597	0.273	1.480
	STANet+	0.887	0.361	0.595	0.490	2.320	0.884	0.436	0.679	0.544	2.040	0.866	0.323	0.634	0.402	1.760	0.910	0.328	0.683	0.452	2.300	0.840	0.315	0.552	0.354	1.430	0.862	0.267	0.612	0.349	1.950
	STANet+^{b}	**0.892**	**0.366**	**0.603**	**0.508**	**2.430**	**0.887**	0.433	**0.687**	**0.558**	**2.130**	**0.868**	**0.319**	**0.642**	**0.417**	**1.860**	**0.915**	**0.345**	**0.699**	**0.493**	**2.480**	0.843	0.315	0.555	**0.370**	**1.540**	**0.871**	**0.269**	**0.609**	**0.348**	**1.970**

Table 3 Impact of different "Thresholds" ($\mathcal{T}_h$) on the saliency detection results of weakly supervised methods on the AVAD set [41]. "0.0" means no threshold is applied on the saliency result, and "0.3" means that the saliency value below 0.3 is set to 0.

T_h	MWS			WSSA			WSS		
	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑
0.0	**0.807**	**0.285**	1.339	**0.834**	**0.309**	**1.477**	**0.858**	**0.347**	**1.655**
0.3	0.799	0.285	1.339	0.801	0.296	1.444	0.827	0.337	1.623
0.5	0.801	0.285	1.344	0.772	0.262	1.419	0.813	0.333	1.617
0.7	0.800	0.285	**1.347**	0.735	0.254	1.279	0.788	0.320	1.574
T_h	ScoCAM			LCAM			SGradCAMpp		
	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑
0.0	**0.772**	**0.237**	**1.016**	**0.776**	**0.241**	**1.033**	**0.809**	**0.275**	**1.181**
0.3	0.763	0.231	0.989	0.764	0.233	0.992	0.806	0.269	1.151
0.5	0.735	0.224	0.969	0.737	0.227	0.974	0.726	0.181	0.789
0.7	0.687	0.207	0.916	0.698	0.221	0.978	0.630	0.122	0.501

stem from STANet+'s ability to adapt fixation predictions based on context, rather than persistently highlighting a single object like video-based CAM methods [22, 30]. This dynamic adaptation aligns more closely with human attention behavior in visual-audio environments.

To make weakly supervised salient object detection models (e.g., MWS [28], WSSA [31], WSS [29], ScoCAM [56], LCAM [48], and SGradCAMpp [57]) more comparable to real fixation prediction, their outputs were thresholded to produce smaller, fixation-like regions. However, as shown in Table 3, this thresholding approach is ineffective because it does not address the fundamental difference between saliency maps and real human fixations. Unlike these methods, SCAM+ dynamically integrates spatial, temporal, and auditory cues, enabling it to generate fixation predictions that better align with human attention behavior.

6.3.2 *Quantitative comparisons with fully supervised methods*

As shown in Table 2, our method achieves comparable results to the fully supervised methods and even outperforms some of them. Note that the performance of our approach can be boosted further by including more tagged audiovisual sequences (e.g., STANet+^{b} in Table 2).

Furthermore, for fair comparison, we also tested another experiment (as shown in Table 4). Both our method and the other three representative fully supervised methods [11, 40, 47] are trained on an identical training set, i.e., the widely-used 168 video clips (70% of the total 241 sequences) obtained from the six visual-audio sets [41–46]. All these clips have been equipped with real fixations. In our training approach, we have omitted the real fixations and instead assigned each clip an appropriate video tag. The results reported in Table 4 suggest that our approach cannot perform well on such a small-scale set, and other SOTA methods also get degenerate. As shown in Table 2, we can achieve some performance gain by using the VGGSound dataset [33] (STANet+^{b}), but this gain is at a relatively large expense on computation. The main reason is that a large portion of videos in VGGSound is irrelevant to the adopted testing sets.

6.4 Computational cost analysis

To analyze the computational efficiency of our model, we report the FLOPs, MACs, and parameter count at different processing stages. Table 5 quantifies the computational cost of SCAM+. The introduction of multigranularity processing increases model complexity, which we address using knowledge distillation.

Table 5 presents the results after distillation. The distilled STANet+ model achieves a 52% reduction in FLOPs and an 89% reduction in parameters while maintaining high prediction accuracy. Furthermore, even on resource-constrained devices such as Snapdragon 8 Gen 3 and RTX 3060, the performance drop remains minimal (only about 1% precision loss), confirming the feasibility of our approach in diverse deployment settings. Overall, while SCAM+ is computationally intensive during training, the distilled STANet+ network reduces inference costs, making it more suitable for real-world applications.

6.5 Limitation

We evaluated SCAM+ on the real-world surveillance and video tagging dataset, MOTChallenge [58], to assess its performance across various real-world scenarios, including surveillance and video tagging. The results on the MOTChallenge dataset are presented in Figure 12(a). SCAM+ effectively predicts fixations in scenarios like parking

Table 4 Qualitative comparison between our method and the other three representative fully supervised SOTA models (ACLNet [47], SalGAN [40], DeepNet [11]). All methods were trained on an identical set (168 sequences), where the real fixations are used when training fully supervised SOTA, while, in our training, only the assigned video tags are used. The numeric results are obtained by testing these models on the AVAD set [41].

	Model	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑		Model	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
STANet+	Weakly supervised	0.887	0.361	0.595	0.490	2.318	SalGAN	Weakly supervised	0.886	0.360	0.579	0.491	2.550
STANet+	Re-Trained on 168	0.856	0.264	0.609	0.378	1.711	SalGAN	Re-Trained on 168	0.845	0.327	0.556	0.382	1.854
ACLNet	Fully supervised	0.905	0.446	0.560	0.580	3.170	DeepNet	Fully supervised	0.869	0.256	0.561	0.383	1.850
ACLNet	Re-trained on 168	0.838	0.329	0.579	0.393	1.843	DeepNet	Re-trained on 168	0.852	0.247	0.582	0.410	1.750

Table 5 Computational cost comparison before and after knowledge distillation. Distillation reduces FLOPs by 50%–80%, enabling STANet+ to be deployed on lower-end hardware while maintaining accuracy.

	Computational cost of SCAM+													
	S					SA				ST				
	Coarse	Fine	+Short	+Long	+Cross	Coarse	Fine	+Long	+Cross	Coarse	Fine	+Short	+Long	+Cross
FLOPs	59.00	77.06	92.92	92.92	92.92	58.48	79.32	200.46	200.46	156.23	172.51	590.32	590.32	590.32
MACs	29.41	38.41	46.41	46.41	46.41	29.18	41.23	100.14	100.14	75.31	86.07	294.93	294.93	294.93
Params	58.03	58.03	138.36	138.36	138.36	371.12	371.12	39.47	39.47	200.09	200.09	21.86	21.86	21.86
SUM FLOPs: 3053.24; MACs: 1523.91; Params: 1818.08														

	Computational cost of STANet+					
	LOPs	MACs	Params	Device	Precision	AUC-J
STANet	264.72	131.99	116.54	Snapdragon 8 Gen 3		0.873
STA+Short	593.78	296.65	39.49	RTX 3060/3070/2080 Ti	**1%↓**	0.886
STA+Long	593.78	296.65	39.49			
STANet+	**1452.28**	**725.29**	**195.52**	RTX 3090	**0.1%↓**	**0.887**
Percent	**52%↓**	**52% ↓**	**89% ↓**			

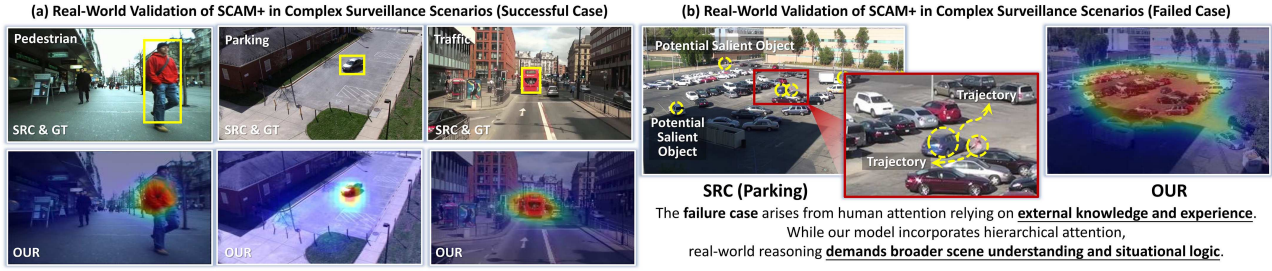

Figure 12 (Color online) Real-world validation of SCAM+ in surveillance settings, evaluated on the MOTChallenge benchmark [58]. (a) Successful cases where SCAM+ captures human-like attention in pedestrian tracking, parking, and traffic monitoring; (b) a challenging parking-lot scene where human attention relies on external knowledge and situational logic that SCAM+ does not explicitly model.

lots, urban traffic, and pedestrian tracking, with its predictions aligning closely with the bounding boxes. However, as shown in Figure 12(b), we observed that SCAM+ faces challenges in more complex scenes. For example, in a parking lot environment, even humans need to combine pedestrian and vehicle motion trajectories with logical reasoning to make accurate predictions. While SCAM+ incorporates semantic reasoning, it still struggles in scenarios that require a deeper understanding based on life experience and commonsense knowledge. Future improvements could focus on integrating external knowledge and enhancing reasoning capabilities to better handle complex real-world environments.

7 Conclusion

In this paper, we have detailed a novel scheme for converting video-audio semantic category tags to pseudo-fixations. Compared with the widely used CAM mechanism, the proposed SCAM and SCAM+ are able to produce pseudo-fixations that are more consistent with real human fixations. We have also compared our model, the STANet+ fixation prediction network trained using our pseudo-fixations, with other SOTA methods. The results favor our new method over unsupervised and weakly supervised methods. Besides, they also show that our method is even better than some fully supervised methods.

Acknowledgements This work was supported by Guangxi Science and Technology Major Program (Grant No. GuiKeAA24206017), Fundamental Research Funds for the Central Universities, National Key R&D Program of China (Grant No. 2023YFC3604500), National Natural Science Foundation of China (Grant Nos. 62172246, 62476143), Shandong Provincial Natural Science Foundation for Outstanding Young Scientists (Grant No. ZR2024YQ071), and Youth Innovation and Technology Support Program for Colleges and Universities in Shandong Province (Grant No. 2021KJ062).

Supporting information Appendixes A–F. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Zhou X, Wu S, Shi R, et al. Transformer-based multi-scale feature integration network for video saliency prediction. IEEE Trans Circuits Syst Video Technol, 2023, 33: 7696–7707
2 Xiong J, Wang G, Zhang P, et al. Casp-net: rethinking video saliency prediction from an audio-visual consistency perceptual perspective. In: Proceedings of CVPR, 2023
3 Li H, Chen G, Li G, et al. Motion guided attention for video salient object detection. In: Proceedings of ICCV, 2019
4 Chen C, Li S, Wang Y, et al. Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion. IEEE Trans Image Process, 2017, 26: 3156–3170
5 Chen C, Wang G, Peng C, et al. Improved robust video saliency detection based on long-term spatial-temporal information. IEEE Trans Image Process, 2020, 29: 1090–1100
6 Jian M, Lam K M, Dong J, et al. Visual-patch-attention-aware saliency detection. IEEE Trans Cybern, 2015, 45: 1575–1586
7 Deng S, Zhuo W, Xie J, et al. QA-CLIMS: question-answer cross language image matching for weakly supervised semantic segmentation. In: Proceedings of MM, 2023. 5572–5583
8 Qiao M, Xu M, Wen S, et al. Saliency prediction of sports videos: a large-scale database and a self-adaptive approach. In: Proceedings of ICASSP, 2024
9 Xiong J, Zhang P, You T, et al. Diffsal: joint audio and video learning for diffusion saliency prediction. In: Proceedings of CVPR, 2024
10 Bellitto G, Proietto Salanitri F, Palazzo S, et al. Hierarchical domain-adapted feature learning for video saliency prediction. Int J Comput Vis, 2021, 129: 3216–3232
11 Pan J, Sayrol E, GiroiNieto X, et al. Shallow and deep convolutional networks for saliency prediction. In: Proceedings of CVPR, 2016
12 Yang W, Huang H, Hu Y, et al. Video coding for machines: compact visual representation compression for intelligent collaborative analytics. IEEE Trans Pattern Anal Mach Intell, 2024, 46: 5174–5191
13 Jin Z, Xie J, Wu B, et al. Weakly supervised pedestrian segmentation for person re-identification. IEEE Trans Circuits Syst Video Technol, 2023, 33: 1349–1362
14 Jiang J, Chen B, Pan J, et al. Forkmerge: mitigating negative transfer in auxiliary-task learning. In: Proceedings of NIPS, 2024
15 Tsiami A, Koutras P, Maragos P. Stavis: spatio-temporal audiovisual saliency network. In: Proceedings of CVPR, 2020
16 Tavakoli H, Borji A, Rahtu E, et al. Dave: a deep audio-visual embedding for dynamic saliency prediction. 2019. ArXiv:1905.10693
17 Wen S, Yang L, Xu M, et al. Saliency prediction on mobile videos: a fixation mapping-based dataset and a transformer approach. IEEE Trans Circuits Syst Video Technol, 2024, 34: 5935–5950
18 Tian Y, Shi J, Li B, et al. Audio-visual event localization in unconstrained videos. In: Proceedings of ECCV, 2018
19 Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization. In: Proceedings of CVPR, 2016
20 Xie J, Xiang J, Chen J, et al. C2am: contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In: Proceedings of CVPR, 2022
21 Chen J, Lu W, Li Y, et al. Adversarial learning of object-aware activation map for weakly-supervised semantic segmentation. IEEE Trans Circuits Syst Video Technol, 2023, 33: 3935–3946
22 Li Z, Wang W, Li Z, et al. Towards visually explaining video understanding networks with perturbation. In: Proceedings of WACV, 2021
23 Xie J, Luo C, Zhu X, et al. Online refinement of low-level feature based activation map for weakly supervised object localization. In: Proceedings of ICCV, 2021. 132–141
24 Wang B, Liu W, Han G, et al. Learning long-term structural dependencies for video salient object detection. IEEE Trans Image Process, 2020, 29: 9017–9031
25 Huang Y, Wu Q, Zhang Z, et al. Meta clothing status calibration for long-term person re-identification. IEEE Trans Image Process, 2024, 33: 2334–2346
26 Wang G, Chen C, Fan D P, et al. From semantic categories to fixations: a novel weakly-supervised visual-auditory saliency detection approach. In: Proceedings of CVPR, 2021
27 Sun J, Han G, Zeng Z, et al. Memristor-based neural network circuit of full-function Pavlov associative memory with time delay and variable learning rate. IEEE Trans Cybern, 2019, 50: 2935–2945
28 Zeng Y, Zhuge Y, Lu H, et al. Multi-source weak supervision for saliency detection. In: Proceedings of CVPR, 2019
29 Wang L, Lu H, Wang Y, et al. Learning to detect salient objects with image-level supervision. In: Proceedings of CVPR, 2017
30 Chattopadhay A, Sarkar A, Howlader P, et al. Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks. In: Proceedings of WACV, 2018
31 Zhang J, Yu X, Li A, et al. Weakly-supervised salient object detection via scribble annotations. In: Proceedings of CVPR, 2020
32 Xie J, Hou X, Ye K, et al. Clims: cross language image matching for weakly supervised semantic segmentation. In: Proceedings of CVPR, 2022. 4483–4492
33 Chen H, Xie W, Vedaldi A, et al. Vggsound: a large-scale audio-visual dataset. In: Proceedings of ICASSP, 2020
34 Borji A. Saliency prediction in the deep learning era: successes and limitations. IEEE Trans Pattern Anal Mach Intell, 2021, 43: 679–700
35 Fan D P, Wang W, Cheng M M, et al. Shifting more attention to video salient object detection. In: Proceedings of CVPR, 2019
36 Xie J, Liu Z, Li G, et al. Audio-visual saliency prediction with multisensory perception and integration. Image Vision Computing, 2024, 143: 104955
37 Ning H, Zhao B, Hu Z, et al. Audio-visual collaborative representation learning for dynamic saliency prediction. Knowledge-Based Syst, 2022, 256: 109675
38 Shao F, Luo Y, Gao F, et al. Knowledge-guided causal intervention for weakly-supervised object localization. IEEE Trans Knowl Data Eng, 2024, 36: 6477–6489
39 Kipf T, Welling M. Semi-supervised classification with graph convolutional networks. In: Proceedings of ICLR, 2017
40 Pan J, Ferrer C, McGuinness K, et al. Salgan: visual saliency prediction with generative adversarial networks. In: Proceedings of CVPR SUNw, 2017
41 Min X, Zhai G, Hu C, et al. Fixation prediction through multimodal analysis. In: Proceedings of Visual Communications and Image Processing (VCIP), 2016
42 Coutrot A, Guyader N. How saliency, faces, and sound influence gaze in dynamic social scenes. J Vision, 2014, 14: 5
43 Coutrot A, Guyader N. Multimodal saliency models for videos. In: From Human Attention to Computational Attention. New York: Springer, 2016. 291–304

44 Mital P K, Smith T J, Hill R L, et al. Clustering of gaze during dynamic scene viewing is predicted by motion. Cogn Comput, 2011, 3: 5–24
45 Gygli M, Grabner H, Riemenschneider H, et al. Creating summaries from user videos. In: Proceedings of ECCV, 2014
46 Koutras P, Maragos P. A perceptually based spatio-temporal computational framework for visual saliency estimation. Signal Processing-Image Communication, 2015, 38: 15–31
47 Wang W, Shen J, Guo F, et al. Revisiting video saliency: a large-scale benchmark and a new model. In: Proceedings of CVPR, 2018
48 Jiang P T, Zhang C B, Hou Q, et al. LayerCAM: exploring hierarchical class activation maps for localization. IEEE Trans Image Process, 2021, 30: 5875–5888
49 Zhang X, Wei Y, Kang G, et al. Self-produced guidance for weakly-supervised object localization. In: Proceedings of ECCV, 2018
50 Jiang L, Xu M, Liu T, et al. Deepvs: a deep learning based video saliency prediction approach. In: Proceedings of ECCV, 2018
51 Xie J, Liu Z, Li G, et al. Global semantic-guided network for saliency prediction. Knowledge-Based Syst, 2024, 284: 111279
52 Zhou L, Zhou T, Khan S, et al. Weakly supervised visual saliency prediction. IEEE Trans Image Process, 2022, 31: 3111–3124
53 Yasuki S, Taki M. Cam back again: large kernel CNNs from a weakly supervised object localization perspective. In: Proceedings of CVPR, 2024
54 Huang C, Tian Y, Kumar A, et al. Egocentric audio-visual object localization. In: Proceedings of CVPR, 2023. 22910–22921
55 Zhu L, She Q, Chen Q, et al. Background-aware classification activation map for weakly supervised object localization. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 14175–14191
56 Wang H, Wang Z, Du M, et al. Score-cam: score-weighted visual explanations for convolutional neural networks. In: Proceedings of CVPRW, 2020
57 Omeiza D, Speakman S, Cintas C, et al. Smooth Grad-CAM++: an enhanced inference level visualization technique for deep convolutional neural network models. 2019. ArXiv:1908.01224
58 Dendorfer P, Osep A, Milan A, et al. MOTChallenge: a benchmark for single-camera multiple target tracking. Int J Comput Vis, 2021, 129: 845–881