From Semantic Categories to Fixations: A Novel Weakly-supervised Visual-auditory Saliency Detection Approach

Guotao Wang[1] Chenglizhao Chen[2*] Dengping Fan[4] Aimin Hao[1,3,6] Hong Qin[5]

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
[2]College of Computer Science and Technology, Qingdao University
[3]Research Unit of Virtual Human and Virtual Surgery, Chinese Academy of Medical Sciences
[4]Inception Institute of Artificial Intelligence [5]Stony Brook University [6]Pengcheng Laboratory

Abstract

Thanks to the rapid advances in the deep learning techniques and the wide availability of large-scale training sets, the performances of video saliency detection models have been improving steadily and significantly. However, the 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 the real visual-audio environment. Hence, it would be neither efficiency nor necessary to re-collect real fixations under the same visual-audio circumstance. To address the problem, this paper advocate a novel approach in a weakly-supervised manner to alleviating the demand of large-scale training sets for visual-audio model training. By using the video category tags only, we propose the selective class activation mapping (SCAM), which follows a coarse-to-fine strategy to select the most discriminative regions in the spatial-temporal-audio circumstance. Moreover, these regions exhibit high consistency with the real human-eye fixations, which could subsequently be employed as the pseudo GTs to train a new spatial-temporal-audio (STA) network. Without resorting to any real fixation, the performance of our STA network is comparable to that of the fully supervised ones. Our code and results are publicly available at https://github.com/guotaowang/STANet.

1. Introduction and Motivation

In the deep learning era, we have witnessed a growing development in video saliency detection techniques [53, 34, 29, 14], where the primary task is to locate the most distinctive regions in a series of video sequences. At present, this field consists of two parallel research directions, i.e., the video salient object detection and the video fixation prediction. In practice, the former [19, 49, 41, 32, 13, 4, 5, 8] aims to segment the most salient objects with clear object boundaries (Fig. 1-A). The latter [35, 54, 12, 44, 18], as the main topic of this paper, predicts the real human-eye fixations in the form of scattered coordinates spreading over the entire scene without any clear boundaries (Fig. 1-B). In fact, this topic has long been investigated extensively in the past decades. Different from the previous works [39, 29, 51], this paper is interested in exploiting the deep learning techniques to predict fixations under the visual and audio circumstance, also known as visual-audio fixation prediction, and this topic is still in its early exploration stage.

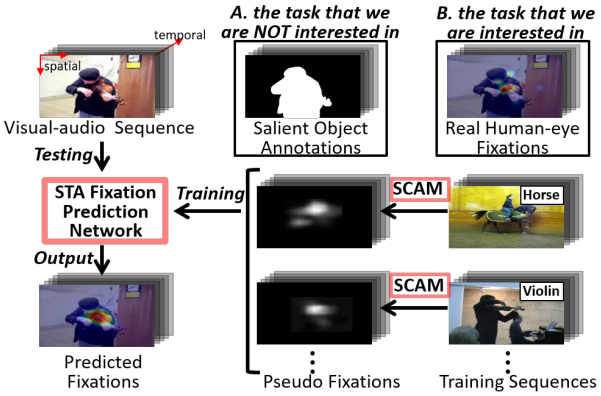

Figure 1. This paper mainly focuses on using a weakly-supervised approach to predicting **s**patial-**t**emporal-**a**udio (STA) fixations, where the key innovation is that, as the first attempt, we automatically convert semantic category tags to pseudo-fixations via the newly-proposed ***s**elective* **c**lass **a**ctivation **m**apping (SCAM).

At present, almost all **s**tate-**o**f-**t**he-**a**rt (SOTA) visual-audio fixation prediction approaches [47, 45] are developed with the help of the deep learning techniques, using the vanilla encoder-decoder structure, facilitated with various attention mechanisms, and trained in a fully-supervised manner. Albeit making progress, these fully-supervised approaches are plagued by one critical limitation (see below).

It is well known that a deep model's performance is heavily dependent on the adopted training set, and large-

*Corresponding Author

scale training sets equipped with real visual fixations are already accessible in our research community. However, it is time-consuming and laborious to re-collect real human-eye fixations in the visual-audio circumstance, thus, to our best knowledge, only a few visual-audio sequences are available for the visual-audio fixation prediction task, where only a small part of them are recommended for the network training, making the data shortage dilemma even worse. As a result, according to the extensive quantitative evaluation that we have done, almost all existing deep learning based visual-audio saliency prediction models [47, 45], though reluctant to admit, might be overfitted in essence.

To solve this problem, we seek to realize the visual-audio fixation prediction using a weakly-supervised strategy. Instead of using the labor-intensive frame-wise visual-audio **g**round **t**ruths (GTs), we devise a novel scheme to produce the GT-like visual-audio pseudo fixations by using the video category tags only. Actually, there already exist plenty of visual-audio sequences with well labeled semantic category tags (e.g., AVE set [46]), where most of them are originally collected for the visual-audio classification task.

Our approach is also inspired by the **c**lass **a**ctivation **m**apping (CAM, [64]) that has been used in the image object localization [57, 50, 43] and video object location [2, 3, 33]. The key rationale of CAM 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 relatively larger classification confidences than others.

Considering that we aim at the fixation prediction in the visual-audio circumstance, we propose the novel ***s**elective* **c**lass **a**ctivation **m**apping (SCAM), which relies on a coarse-to-fine strategy to select the most discriminative regions from multiple sources, where these regions exhibit high consistency with the real human-eye fixations. This coarse-to-fine methodology ensures the aforementioned less-discriminative scattered regions to be filtered completely, and the selection operation between different sources helps reveal the most discriminative regions, enabling the pseudo-fixations to be closer to the real ones. Once the pseudo-fixations have been obtained, a **s**patial-**t**emporal-**a**udio (STA) fixation prediction network will be trained, and it learns the common consistency of all pseudo-fixations. Consequently, it can predict fixations accurately for videos without being assigned to any semantic category tag in advance.

It is worth mentioning that this paper is one of the first attempts to explore the deep learning based visual-audio fixation prediction in a weakly-supervised manner, which is expected to contribute to visual-audio information integration and relevant applications in computer vision.

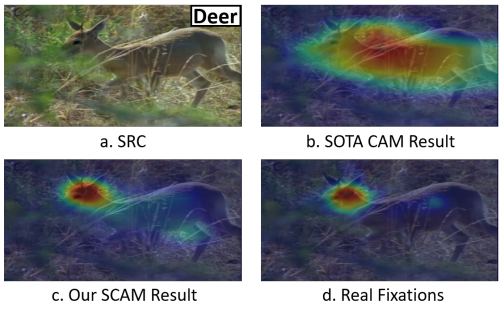

Figure 2. Existing SOTA approaches (e.g., Zeng et al. [57]) are mainly designed for locating salient objects rather than simulating human fixations; thus their results tend to be large scatter regions (b), which are quite different from the real fixations (d).

2. Related Work

Unsupervised Visual Fixation. Almost all conventional hand-crafted approaches should be categorized into the unsupervised class, and we will document several most representative ones. Fang et al. [15] detected visual saliency combining both spatial and temporal information founded upon uncertainty measures. Hossein et al. [22] proposed a model of visual saliency based on reconstruction error and cruder measurements of self information. Leboran et al. [30] implemented an explicit short term visual adaptation of the spatial-temporal scale decomposition feature to determine dynamic saliency. Let us now move to the deep learning based ones. Zhang et al. [62] learned saliency maps from multiple noisy unsupervised saliency methods and formulated the problem as the joint optimization of a latent saliency prediction module and a noise modeling module. Li et al. [31] adopted a super-pixel-wise variational autoencoder to better preserve object boundaries and maintain the spatial consistency (and also refer to Kim et al. [27])

Weakly-supervised Visual Fixation. Based on the pregiven image-level labels [50], points [40], scribbles [61], and bounding boxes [11], it can usually outperform the unsupervised approaches. Zeng et al. [58]proposed to combine bottom-up object evidences with top-down class confidence scores in the weakly-supervised object detection task. Zhang et al. [59] harnessed the image-level labels to produce reliable pixel-level annotations and design a fully end-to-end network to learn the segmentation maps.

Supervised Visual-audio Fixation. In recent years, the visual-audio saliency detection has received more attention than before, including STAVIS [47], DAVE [45], and AVC [37]. Since the audio source may correlate to some specific semantic categories, these models assume that the human-eye fixations may easily be affected by the audio source when the visual and audio sources are semantically synchronized, where the research foci of these models rely on designing better visual-audio fusion schemes. At present, there only exist totally 241 visual-audio sequences

with real fixations collected in the visual-audio circumstance, where these sequences are provided by [36, 9, 10] respectively. Motivated by these, this paper proposes to fully mine audio-visual pseudo-fixations in a weakly-supervised manner for the video fixation prediction task.

3. The Proposed Algorithm

3.1. Relationship between Video Tags and Fixations

In the video classification field, each training sequence is usually assigned with a semantic tag which associates this sequence with a specific video category. In general these semantic tags are assigned by performing the majority voting between multiple persons, aiming to represent the most meaningful objects or events in the given video. Similar to the process of tag assignment, the real human visual fixations tend to focus on those most meaningful and representative regions when watching a video sequence. Thus, formulating pseudo-fixations from video category tags is theoretically feasible.

3.2. Preliminary: Class Activation Mapping (CAM)

The fundamental idea of **c**lass **a**ctivation **m**apping (CAM) is to rely on the weighted summation of feature maps in the last convolutional layer to coarsely locate the most representative image regions regarding the current classification task. In practice, as can be seen in Fig. 3, those weights (w_i) correlated with the highest classification confidence in the last fully connected layer are selected to weigh the feature maps (f_i). The CAM — a 2-dimensional matrix, can be obtained by:

$$\mathrm{CAM} = \xi\Big[\sum_{i}^{d} w_i \times f_i\Big], \tag{1}$$

where d represents the feature channel number and $\xi[\cdot]$ is the normalization function. From the qualitative perspective, the CAM, which has been visualized in the bottom-right of Fig. 3, usually shows large feature response to frame regions (i.e., the '*motorbike*') that has contributed most regarding the classification task, and frequently, these regions usually correlate to the most salient object in the scene.

3.3. Limitation of the Conventional CAM

In fact, the CAM is quite different from the real human-eye fixations in essence. For example, as can be seen in Fig. 2-b, when performing the video classification task, the image regions that have contributed to the '*deer*' category most are capable of highlighting the salient object (the deer). Following this rationale, several previous works [63, 16, 48, 1, 7] have resorted to the CAM for locating salient objects. However, the CAMs obtained by these methods are quite different from the real human-eye fixations, and the main reasons for causing this difference mainly comprise the following two aspects.

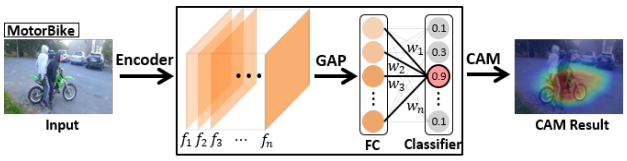

Figure 3. Illustration of the **c**lass **a**ctivation **m**apping (CAM) details. FC: fully connected layer; GAP: global average pooling; numbers in the classifier represent the classification confidences.

First, since both local and non-local deep features would contribute to the classification task, the CAMs tend to be large scatter regions. For example, as shown in Fig. 2, the main body of the deer can help the classifier to separate this image from other non-animal 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 system tends to pay more attention on the most discriminative image regions (e.g., the '*deer head*', see Fig. 2-d).

Second, most of the existing works [57, 50, 56, 55, 23] have only considered the spatial information when computing CAM. However, the real human-eye fixations are usually affected by multiple sources, including spatial, temporal, and audio ones. In fact, this multi-source nature has long been omitted by our research community, because, compared with the spatial information — a stable source, the other two sources (temporal and audio) are still considered to be rather unstable ones thus far, and this unstable attribute makes them difficult to be used for computing CAM. However, in many practical scenarios, it is exactly these two sources that could most benefit the classification task.

3.4. Computing SCAM on Multiple Sources

Compared with the single image case, the problem domain of our visual-audio case is much more complicated, where we need to consider multiple sources simultaneously, including spatial, temporal, and audio sources. As mentioned above, the conventional CAMs derived from using spatial information solely tend to be large scatter regions, which might be quite different from the real fixations; even worse, it cannot take full advantage of the complementary status between different sources in the spatial-temporal-audio circumstance. The main reason is that, in the spatial-temporal-audio circumstance, the feature maps tend to be multi-scale, multi-level, and multi-source, where all of them will jointly contribute to the classification task, thus there would be more false-alarms and redundant responses, making the CAMs far away from being the most discriminative regions.

To overcome this problem, we propose to decouple the spatial-temporal-audio circumstance to three independent sources, i.e., spatial (S), temporal (T), and audio (A); in this way, we recombine them using three distinct fusion networks, i.e., S, ST, and SA classification nets (Fig. 4 and

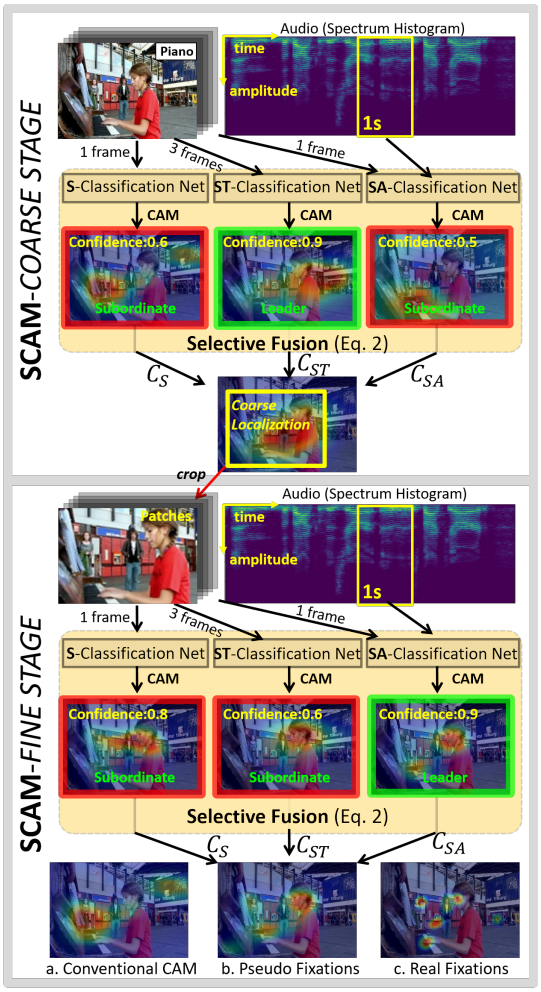

Figure 4. The proposed **s**elective **c**lass **a**ctivation **m**apping (S-CAM) follows the coarse-to-fine methodology, where the coarse stage localizes the region of interest and then the fine stage reveals those image regions with the strongest local responses. S: spatial; ST: spatiotemporal; SA: Spatial-audio.

Fig. 6). Then, the most discriminative regions that are more closer to the real fixations can be easily determined by selectively fusing CAMs derived from these classification nets. We name this process as *selective* **c**lass **a**ctivation **m**apping (SCAM), which can be detailed by Eq. 2.

$$\text{SCAM} = \xi\left[\frac{\oint\left(C_S^v\right)\cdot\Phi_S + \oint\left(C_{ST}^v\right)\cdot\Phi_{ST} + \oint\left(C_{SA}^v\right)\cdot\Phi_{SA} + \lambda}{\oint\left(C_S^v\right) + \oint\left(C_{ST}^v\right) + \oint\left(C_{SA}^v\right) + \lambda}\right], \tag{2}$$

where λ is a small constant for avoiding any division by zero; Φ_S, Φ_{ST}, and Φ_{SA} respectively represent the CAM derived from either S, ST, or SA classification nets; $C \in (0,1)^{1\times c}$ represents the classification confidences regarding c classes; $\xi[\cdot]$ is the normalization function; suppose the pre-given category tag of the S classification net is the v-th category in C, we use C_S^v to represent this confidence; $\oint(\cdot)$ is a soft filter (Eq. 3), which aims to compress those features of low classification confidences to be considered when computing the SCAM.

$$\oint\left(C_S^v\right) = \begin{cases} C_S^v & if \quad C_S^v > C_S^u|_{v\neq u, 1\leq u\leq c} \\ 0 & otherwise \end{cases}. \tag{3}$$

3.5. SCAM Rationale

Generally speaking, either spatial, temporal, or audio source could influence our visual attention, while, compared with the last two, the spatial source is usually more important and stable in practice. For example, a given frame may remain static for a long period of time, where the temporal source becomes completely absent; similar situation may also take place for the audio source. Thus, when we perform the classification task in computing CAM, the spatial information should be treated as the main source, while the other two can only be its complementary sources. This is the reason why we recombine S, T and A sources to S (no change), ST, and SA, respectively.

Considering all S, ST, and SA classification nets have already been trained on training instances labeled with video category tags only, most of these training instances usually perform very well if we feed them into these nets for testing. However, the CAMs derived from these nets are still rather different in essence, because their inputs are different, and we have demonstrated some most representative qualitative results in Fig. 5. Normally, the consistency level between CAM and real fixations is often positively related to the classification confidence level. By using the classification confidences as the fusion weights to compress those less trustworthy CAMs, the pseudo-fixations obtained by selectively fusing all these multi-source CAMs using Eq. 2 can be very closer to the real ones.

3.6. Multi-stage SCAM

Benefiting from the selective fusion over multiple sources, the proposed SCAM is able to outperform the conventional CAM in revealing pseudo-fixations. However, the pseudo-fixations produced by SCAM may still differ from the real ones occasionally, especially for scenes with complex background, where the pseudo-fixations tend to mess-up. The main reasons are two-fold: 1) complex video scenes usually contain more contents, yet it has been assigned with only one category tag, thus more contents belonging to out-of-scope categories may contribute to the classification task; 2) the aforementioned SCAM has followed the single-scale procedure, while, in sharp contrast, the real human visual system is a multi-scale one, where we tend to fast locate the region-of-interest unconsciously before assigning our real fixations to the local regions inside it. To further improve, we follow the coarse-to-fine methodology to sequentially perform SCAM twice. The

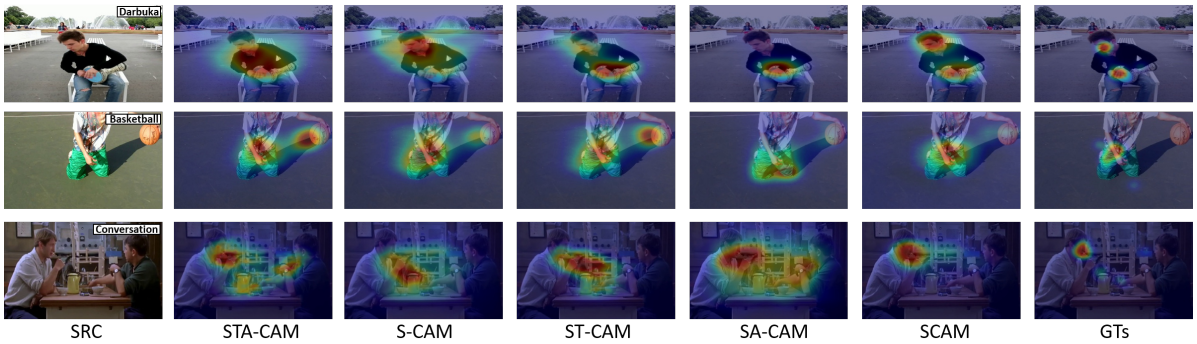

Figure 5. Qualitative illustration of CAMs derived from different sources. 'STA(S/SA/ST)-CAM': CAM obtained from the spatial-temporal-audio (spatial/spatial-audio/spatioteporal) circumstance; the 'SCAM' represents the pseudo-fixations obtained by Eq. 2, where we can easily observe that the results in this column can be very consistent with the GTs.

coarse stage decreases the given problem domain, thus the pseudo-fixations revealed in the fine stage are more likely to be those real discriminative regions, improving the overall performance significantly.

In the coarse stage, we use a rectangular box to tightly warp the pseudo -fixations that have been binarized by a hard threshold ($2 \times average$), and the video sequences will be cropped into video patches via these rectangular boxes. In the fine stage, the video sequences are replaced by the these video patches to be the classification nets' input, and we perform SCAM again to obtain the final pseudo-fixations. Compared with the conventional CAM (i.e., the CAM derived from the S classification net, Fig. 4-a), the pseudo-fixations (Fig. 4-b) obtained in this stage are clearly more consistent with the real fixations (Fig. 4-c), where the quantitative evidences can be seen in Sec. 4.

3.7. The Detail of Classification Nets

All networks adopted in this paper have followed the simplest encoder-decoder architecture. Following the previous work [45], we have converted audio signals to 2D spectrum histograms in advance. We use plain 3D convolution to sense temporal information. We believe all these implementations are quite simple and straightforward, and almost all network details have been clearly represented in Fig. 6. Enhanced alternatives, of course, could result in additional performance gain.

Audio Switch (ϕ). Different from the conventional implementation, we proposed the 'audio switch' module in both SA Fuse and STA Fuse (Fig. 6). The main function of this module is to alleviate the potential side-effects from the audio signal when performing the SA fusion and the STA fusion, and we will explain this issue as follows.

Compared with the temporal source, the audio source is usually associated with strong semantic information, making it more easily to influence its spatial counterpart. However, the audio source itself has a critical drawback, where video sequences may usually couple with meaningless background music or noise. In such case, fusing audio source with spatial source may make the classification task more difficult. In fact, the nature of the proposed 'audio switch' is a plug-in, and we implement it as an individual network with an identical structure as the SA classification net. Instead of aiming at the video classification task, this plug-in is trained on visual-audio data with binary labels considering that the current audio signal is really benefiting the spatial source. To obtain these binary labels automatically, we resort to an off-the-shelf audio classification tool (VggSound [6]), which was trained on a large-scale audio classification set including almost 300 categories. Our rationale is that the audio source would be able to benefit the spatial source only if it has been synchronized with its spatial counterpart, sharing an identical semantical information. Therefore, for a visual-audio fragment (1 frame and 1s 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'. Here we take the 'SA Fuse' for instance, where the SA fusion data flow can be represented as Eq. 4.

$$\mathrm{SA} \leftarrow Relu\Big(\sigma\big(\phi(\mathrm{A})\big) \odot \mathrm{S} + \mathrm{S}\Big), \tag{4}$$

where S denotes spatial flow; A denotes audio flow; $\odot$ is the typical element-wise multiplicative operation; $Relu(\cdot)$ denotes the widely-used **re**ctified **l**inear **u**nit (ReLU) activation operation; $\sigma(\cdot)$ is the sigmoid function; $\phi(\cdot)$ is the proposed audio switch, which returns 1 if the given audio can be classified (via VggSound [6]) to the category that is identical to the the pre-given tag. Our quantitative results suggest that the 'audio switch' can persistently improve the overall performance for about 1.5% averagely.

3.8. STA Fixation Prediction Network

The implementation of STA fixation prediction network is also very intuitive, where the spatial features are respectively fused with either temporal features or audio features in advance and later are combined via the simplest feature

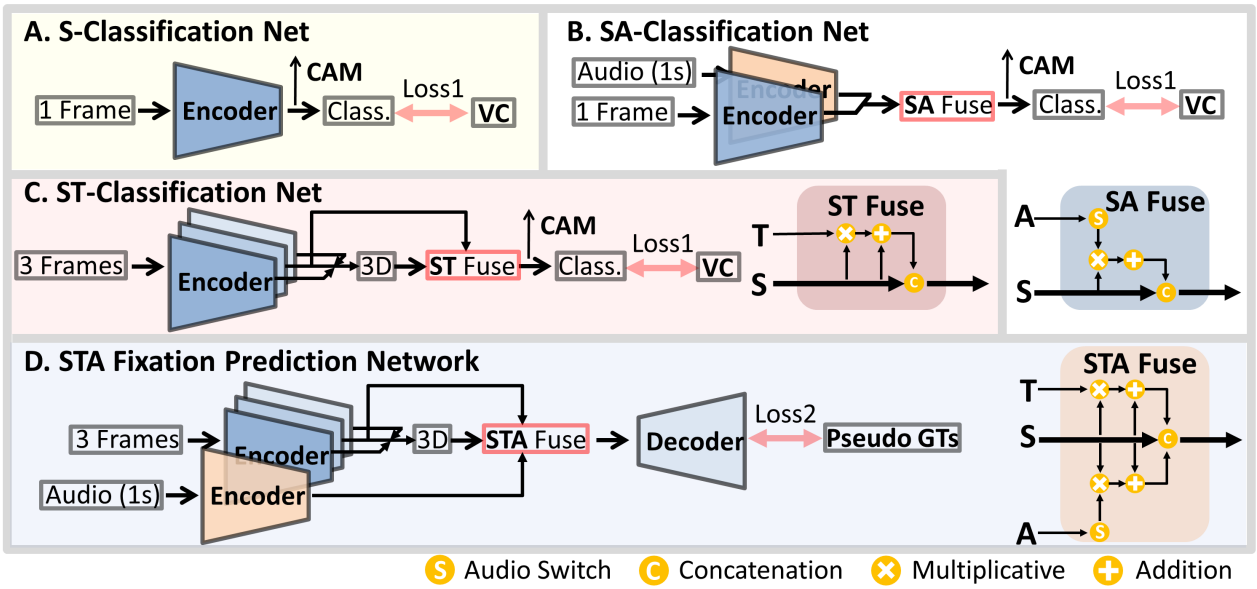

Figure 6. Network Details. Loss1: cross entropy loss; Loss2: binary cross entropy loss; 'Class.': Classification; **VC**: **V**ideo **C**ategories; Decoder: VGG16; '3D': 3D convolutional; A-C respectively show the network architectures of the adopted classification nets (Fig. 4); D is the STA fixation prediction network.

concatenation operation. Then a typical decoder with three de-convolutional layers is used to convert the feature maps derived from the STA fusion module to the final fixations. The data flow in the 'STA Fuse' module can be formulated as Eq. 5.

$$\mathrm{STA} \leftarrow Relu\left[Cov\left(Con\left(\sigma\left(\phi(\mathrm{A})\right)\odot \mathrm{S}+\mathrm{S}, \sigma\left(\mathrm{T}\right)\odot \mathrm{S}+\mathrm{S}\right)\right)\right], \tag{5}$$

where $Con(\cdot)$ is the typical concatenation operation; $Cov(\cdot)$ denotes 1×1 convolution; all other symbols are identical to that in Eq. 4.

The binary cross entropy loss (L_B) adopted for the STA fixation prediction network training is detailed in Eq. 6, where 'N' denotes training instances number; 'PseudoGT' represents the pseudo-fixations obtained via the two-stage SCAM; $Dec(\cdot)$ denotes the decoder layers.

$$\begin{aligned} L_B = -\frac{1}{N}\sum_{i}^{N}\Big[&\mathrm{PseudoGT}_i \cdot log\big(Dec(\mathrm{STA}_i)\big) \\ &+ (1-\mathrm{PseudoGT}_i)\cdot log\big(1-Dec(\mathrm{STA}_i)\big)\Big]. \end{aligned} \tag{6}$$

A more powerful decoder equipped with multi-scale connections and channel-wise attentions may give rise to some additional performance gain, but full justification w.r.t. this issue is beyond the main topic of this paper, and we shall leave it for future research investigation. The training process of our STA fixation prediction network relies on the pseudo-fixations only, thus it is able to predict fixations for unseen visual-audio sequences without category tags.

4. Experiments and Validations

4.1. Datasets and Evaluation Metrics

Testing Sets. We have tested the proposed approach and other competitors on 6 datasets, including AVAD [36], Coutrot1 [9], Coutrot2 [10], DIEM [26], SumMe [20] and ETMD [28]. All these sets (241 sequences) are furnished with pixel-wise real fixations that are collected in the visual-audio circumstance.

Quantitative Metrics. Following the previous work [52], we have adopted 5 commonly-used evaluation metrics to measure the agreement between model predictions and the real human eye fixations, including AUC-Judd (AUC-J), **sim**ilarity metric (SIM), **s**huffled **AUC** (s-AUC), **n**ormalized **s**canpath **s**aliency (NSS), and linear **c**orrelation **c**oefficient (CC). Higher scores on each metric indicate better performance.

4.2. Implementation Details

Training Set. Recently, Google has released the Audioset [17], the largest visual-audio set thus far, and we use its subset **a**udio **v**isual **e**vent (AVE) [46][1] location dataset, which contains 4,143 sequences covering 28 semantic categories, as the classification training set for the S, ST and SA classification nets (Sec. 3.4).

Training Details. We follow the widely-used multi-stage training scheme. In the coarse-stage, all classification nets are trained on AVE set, where the batch size equals 20 and all video frames are resized to 256×256. Taking

[1] https://sites.google.com/view/audiovisualresearch

Table 1. Quantitative evidence towards the effectiveness of the proposed selective fusion scheme. This experiment is conducted on the AVAD set [36]. 'CO.': the coarse stage; 'FI.': the fine stage.

	Module	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
CO.	SCAM$_{avg}$	0.786	0.219	0.538	0.297	1.312
	SCAM$_{sel}$	**0.801**	**0.256**	**0.554**	**0.345**	**1.364**
FI.	SCAM$_{avg}$	0.845	0.303	0.573	0.399	1.797
	SCAM$_{sel}$	**0.873**	**0.334**	**0.580**	**0.438**	**2.018**

the cropped video patches as input, three completely new classification nets in the fine-stage will be trained, where the batch size equals 3 and all video patches are resized to 356×356. The STA fixation prediction network takes the pseudo-fixations as GTs, where video frames are resized to 356×356, thus the batch size is assigned to 3. All these training processes have adopted the **s**tochastic **g**radient **d**escent (SGD) optimizer with learning rate 0.001.

4.3. Component Evaluation

Effectiveness of the Selective Fusion. In order to selectively fuse the multi-source CAMs, we have adopted the classification confidences as the fusion weights (e.g., the C_S in Eq. 2). Actually, the effectiveness of this implementation is based on the pre-condition that the classification confidences are really positively related to the consistency level between CAMs and real fixations. To verify this issue, we have compared the proposed selective fusion and the conventional scheme (i.e., averaging all CAMs derived from different sources, SCAM$_{avg}$).

As shown in Table 1, the overall performance can be improved significantly via the proposed selective fusion SCAM$_{sel}$, where the proposed SCAM$_{sel}$ outperforms the SCAM$_{avg}$ by about 3%.

Effectiveness of the Multi-stage Methodology. To verity this aspect, we have respectively tested all CAMs derived from different sources (i.e., S, SA, ST, and STA) in different stages (i.e., COARSE and FINE, Table 2). We take the first row of Table 2 for instance, where the 'CAM$_S$' denotes the result of CAM obtained from the spatial source in the coarse-stage. Notice that the 'CAM$_S$' can also be used to represent the performance of the classic CAM that has been widely-adopted by previous works. The 'CAM$_{STA}$' denotes the performance of CAM derived using all spatial, temporal, and audio sources at the same time, where we simply adopt a network sharing similar fusion process to the proposed STA fixation prediction network (Fig. 6-D).

As documented in Table 2, all CAMs obtained in the fine-stage can significantly outperform those obtained in the coarse-stage. Meanwhile, we can easily observe that considering all sources simultaneously cannot take full advantage of the complementary nature between them, thus the performance improvement achieved by the CAM$_{STA}$ is really marginal. As for the proposed SCAM, it can persistently outperform the conventional CAM scheme in both stages. Also, by comparing the first and the last row, we can easily observe that the proposed multi-stage SCAM outperforms the classic CAM significantly, e.g., the CC metric has been improved from 0.261→0.438 and similar trends take place in other metrics.

Table 2. Quantitative evidence towards the effectiveness of the proposed multi-stage SCAM. This experiment is conducted on the AVAD set [36].

	Module	AUC-J ↑	SIM↑	s-AUC↑	CC↑	NSS↑
COARSE	CAM$_S$	0.774	0.202	0.545	0.261	1.269
	CAM$_{ST}$	0.785	0.223	0.536	0.269	1.292
	CAM$_{SA}$	0.780	0.214	0.542	0.277	1.276
	CAM$_{STA}$	0.793	0.227	0.551	0.293	1.273
	SCAM	0.801	0.256	0.554	0.345	1.364
FINE	CAM$_S$	0.834	0.291	0.574	0.376	1.528
	CAM$_{ST}$	0.843	0.289	0.571	0.372	1.581
	CAM$_{SA}$	0.845	0.304	0.564	0.384	1.622
	CAM$_{STA}$	0.856	0.296	0.579	0.415	1.803
	SCAM	**0.873**	**0.334**	**0.580**	**0.438**	**2.018**

Table 3. Quantitative evidence towards the effectiveness of the proposed **a**udio **s**witch (AS, Eq. 5). This experiment is conducted on the AVAD set [36].

Module	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
w/o. AS	0.864	0.330	0.571	0.421	1.833
w. AS	**0.873**	**0.334**	**0.580**	**0.438**	**2.018**

Effectiveness of the Proposed Audio Switch. In Table 3, we have reported the performance of the proposed model without ('w/o') using the **a**udio **s**witch (AS). We can observe that the audio switch is capable of improving the overall performance by about 2%. The main reason is that it can filter those meaningless background audio, alleviating the learning ambiguity when fusing unsynchronized spatial and audio information.

4.4. Quantitative Comparisons

We have compared our model (i.e., the STANet, which is trained using pseudo-fixations only) with other 14 SOTA methods, including 5 unsupervised methods, 5 weakly-supervised methods, and 4 fully-supervised methods on all 6 testing sets. Shown in Table 4, our method outperforms all unsupervised methods significantly, and it also outperforms the most recent weakly-supervised competitors (e.g., MWS [57] and WSSA [61]). In addition, our method achieves comparable result to the fully-supervised methods, especially, our method outperforms the fully-supervised DeepNet [39] for all testing sets except the Coutrot2. The main reason is that the semantic contents of the Coutrot2 set is quite different from that of the AVE set, while our model is weakly-supervised by the category tags of the AVE set. Notice that the performance of our approach can be boosted further by including more tagged sequences.

Different to the conventional video based CAM ap-

Table 4. Quantitative comparisons between our method with other fully-/weakly-/un-supervised methods. **Bold** means the best result. Due to the space limitation, we have included the corresponding qualitative comparisons in the uploaded '*supplementary material*'.

Means	DataSet	AVAD [36]					DIEM [26]					SumMe [20]				
	Methods	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
Un-supervised	ITTI [24]	0.688	0.170	0.533	0.131	0.611	0.663	0.217	0.583	0.137	0.555	0.666	0.151	0.559	0.097	0.436
	GBVS [21]	**0.854**	0.247	0.572	**0.337**	**1.556**	**0.830**	**0.318**	0.605	**0.356**	**1.277**	**0.808**	0.221	0.567	**0.272**	**1.134**
	SCLI [42]	0.747	0.210	0.535	0.170	0.792	0.739	0.267	0.590	0.207	0.779	0.746	0.209	0.577	0.184	0.796
	SBF [60]	0.833	**0.272**	0.576	0.308	1.489	0.759	0.292	0.608	0.301	1.081	0.783	**0.228**	0.590	0.230	1.023
	AWS-D [30]	0.825	0.221	**0.589**	0.304	1.378	0.733	0.250	**0.612**	0.301	1.128	0.747	0.192	**0.603**	0.186	0.853
Weakly-superwised	GradCAM++ [3]	0.777	0.273	0.559	0.255	1.217	0.732	0.216	0.583	0.271	0.778	0.774	0.217	0.593	0.225	0.924
	VUNP [33]	0.574	0.067	0.500	0.142	0.292	0.558	0.047	0.515	0.172	0.186	0.555	0.013	0.507	0.114	0.048
	WSS [50]	0.858	0.292	**0.592**	0.347	1.655	0.803	0.333	0.620	0.344	1.293	0.812	0.245	0.589	0.279	1.098
	MWS [57]	0.834	0.272	0.573	0.309	1.477	0.806	0.336	0.628	0.350	1.308	0.808	0.237	0.607	0.258	1.155
	WSSA [61]	0.807	0.261	0.574	0.285	1.339	0.767	0.305	0.608	0.311	1.178	0.755	0.225	0.585	0.231	1.058
	OUR(STANet)	**0.873**	**0.334**	0.580	**0.438**	**2.018**	**0.861**	**0.391**	**0.658**	**0.469**	**1.716**	**0.854**	**0.294**	**0.627**	**0.368**	**1.647**
Fully-supervised	DeepNet [39]	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
	SalGAN [38]	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**
	DeepVS [25]	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
	ACLNet [52]	**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
Means	DataSet	ETMD [28]					Coutrotl [9]					Coutrot2 [10]				
	Methods	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
Un-supervised	ITTI [24]	0.661	0.127	0.582	0.083	0.425	0.616	0.178	0.529	0.082	0.319	0.694	0.142	0.530	0.040	0.331
	GBVS [21]	**0.856**	0.226	0.613	**0.299**	**1.398**	**0.798**	**0.253**	0.526	**0.272**	**1.055**	0.819	**0.189**	0.577	**0.183**	1.071
	SCLI [42]	0.761	0.165	0.570	0.129	0.617	0.754	0.216	0.536	0.239	0.883	0.669	0.137	0.510	0.014	0.093
	SBF [60]	0.805	**0.232**	0.641	0.262	1.298	0.726	0.187	0.530	0.215	0.789	**0.827**	0.152	0.583	0.131	**1.101**
	AWS-D [30]	0.754	0.161	**0.664**	0.181	0.907	0.729	0.214	**0.581**	0.207	0.872	0.783	0.170	**0.590**	0.146	0.842
Weakly-superwised	GradCAM++ [3]	0.575	0.124	0.157	0.576	0.736	0.704	0.137	0.537	0.210	0.511	0.733	0.114	0.567	0.168	0.625
	VUNP [33]	0.505	0.030	0.103	0.132	0.593	0.589	0.063	0.514	0.152	0.304	0.661	0.101	0.536	0.162	0.491
	WSS [50]	0.854	0.277	0.661	0.334	1.650	0.772	0.247	**0.547**	0.233	0.975	0.835	0.208	0.578	0.192	1.178
	MWS [57]	0.833	0.237	0.649	0.293	1.425	0.743	0.231	0.528	0.201	0.798	0.839	0.188	0.581	0.168	1.197
	WSSA [61]	0.793	0.201	0.622	0.222	1.075	0.701	0.180	0.535	0.169	0.780	0.797	0.185	0.571	0.180	1.263
	OUR(STANet)	**0.908**	**0.318**	**0.682**	**0.448**	**2.176**	**0.829**	**0.306**	0.542	**0.339**	**1.376**	**0.850**	**0.247**	**0.597**	**0.273**	**1.475**
Fully-supervised	DeepNet [39]	0.889	0.225	0.699	0.387	1.900	0.824	0.273	0.559	0.340	1.410	0.896	0.201	0.600	0.301	1.820
	SalGAN [38]	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 [25]	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 [52]	**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

proaches [2, 3, 33] which tend to highlight the single object persistently, the frame regions highlighted by our approach the most discriminative ones, might vary from frame to frame, because, in the visual-audio circumstance, either spatial, temporal, or audio could alternatively contribute most to the classification task. This attribute is very consistent with the real human fixation, because we human never pay our attention to a fixed location for a long period of time, especially in the visual-audio circumstance.

4.5. Limitation

We have only considered one single semantic tag for each visual-audio sequence, while, in practice, a video sequence could be assigned with multiple tags. Thus, our method might not be able to perform very well for sequences with massive out-of-scope semantic contents. This problem can be alleviated by including more data with multiple tags, which calls for new research in the near future.

5. Conclusion and Future Work

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, the proposed SCAM is able to produce pseudo-fixations that are more consistent with the real ones. The key technical innovations include the multi-source based selective fusion and its multi-stage methodology, where the effectiveness has been respectively validated by the component evaluation. We have also compared our model — the STA fixation prediction network trained using our pseudo-fixations, with other SOTA methods. The results favor our new method over unsupervised and weakly-supervised methods. They also show that our method is even better than some fully-supervised methods. In the near future, we are interested in exploring a full-automatical way for mining category tags which contribute most regarding the classification task. Thus, the pseudo fixations derived from this new small-group tags may be more consistent with the real ones.

Acknowledgments. This research was supported in part by the National Key R&D Program of China (2017YFF0106407), the National Natural Science Foundation of China (61802215, 61806106, 61672077, and 61532002), the Natural Science Foundation of Shandong Province (ZR2019BF011) and the National Science Foundation of the USA (IIS-1715985 and IIS-1812606).

References

[1] Aditya Arun, CV Jawahar, and MPawan Kumar. Dissimilarity coefficient based weakly supervised object detection. In *CVPR*, pages 9432–9441, 2019. 3

[2] SarahAdel Bargal, Andrea Zunino, Donghyun Kim, Jianming Zhang, Vittorio Murino, and Stan Sclaroff. Excitation backprop for rnns. In *CVPR*, pages 1440–1449, 2018. 2, 8

[3] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and VineethN Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *WACV*, pages 839–847, 2018. 2, 8

[4] Chenglizhao Chen, Shuai Li, Yongguang Wang, Aimin Hao, and Hong Qin. Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion. *IEEE TIP*, 26(7):3156–3170, 2017. 1

[5] Chenglizhao Chen, Guotao Wang, Chong Peng, Xiaowei Zhang, and Hong Qin. A video saliency detection model in compressed domain. *IEEE TIP*, 29(1):1090–1100, 2020. 1

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, pages 721–725, 2020. 5

[7] Ze Chen, Zhihang Fu, Rongxin Jiang, Yaowu Chen, and XianSheng Hua. Slv: Spatial likelihood voting for weakly supervised object detection. In *CVPR*, pages 12995–13004, 2020. 3

[8] Runmin Cong, Jianjun Lei, Huazhu Fu, Fatih Porikli, Qingming Huang, and Chunping Hou. Video saliency detection via sparsity-based reconstruction and propagation. *IEEE TIP*, 28(10):4819–4831, 2019. 1

[9] Antoine Coutrot and Nathalie Guyader. How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of Vision*, 14(8):5–5, 2014. 3, 6, 8

[10] Antoine Coutrot and Nathalie Guyader. Multimodal saliency models for videos. In *From Human Attention to Computational Attention*, pages 291–304. 2016. 3, 6, 8

[11] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, pages 1635–1643, 2015. 2

[12] Richard Droste, Jianbo Jiao, and JAlison Noble. Unified image and video saliency modeling. *arXiv preprint arXiv:2003.05477*, 2020. 1

[13] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *CVPR*, pages 8554–8564, 2019. 1

[14] Yuming Fang, Guanqun Ding, Jia Li, and Zhijun Fang. Deep3dsaliency: Deep stereoscopic video saliency detection model by 3d convolutional networks. *IEEE TIP*, 28(5):2305–2318, 2018. 1

[15] Yuming Fang, Zhou Wang, Weisi Lin, and Zhijun Fang. Video saliency incorporating spatiotemporal cues and uncertainty weighting. *IEEE TIP*, 23(9):3910–3921, 2014. 2

[16] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *ICCV*, pages 9834–9843, 2019. 3

[17] JortF Gemmeke, DanielPW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, RChanning Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, pages 776–780, 2017. 6

[18] Siavash Gorji and JamesJ Clark. Going from image to video saliency: Augmenting image salience with dynamic attentional push. In *CVPR*, pages 7501–7511, 2018. 1

[19] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Mingming Cheng, and Shaoping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *AAAI*, pages 10869–10876, 2020. 1

[20] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc VanGool. Creating summaries from user videos. In *ECCV*, pages 505–520, 2014. 6, 8

[21] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2007. 8

[22] Sayed HosseinKhatoonabadi, Nuno Vasconcelos, IvanV Bajic, and Yufeng Shan. How many bits does it take for a stimulus to be salient? In *CVPR*, pages 5501–5510, 2015. 2

[23] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, pages 5001–5009, 2018. 3

[24] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI*, 20(11):1254–1259, 1998. 8

[25] Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, and Zulin Wang. Deepvs: A deep learning based video saliency prediction approach. In *ECCV*, pages 602–617, 2018. 8

[26] Parag K. Mital, Tim J. Smith, Robin L. Hill, and John M. Henderson. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, 3(1):5–24, 2011. 6, 8

[27] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *ICCV*, pages 6092–6101, 2019. 2

[28] Petros Koutras and Petros Maragos. A perceptually based spatio-temporal computational framework for visual saliency estimation. *Signal Processing: Image Communication*, 38:15–31, 2015. 6, 8

[29] Qiuxia Lai, Wenguan Wang, Hanqiu Sun, and Jianbing Shen. Video saliency prediction using spatiotemporal residual attentive networks. *IEEE TIP*, 29:1113–1126, 2019. 1

[30] Victor Leboran, Anton GarciaDiaz, XoseR FdezVidal, and XoseM Pardo. Dynamic whitening saliency. *IEEE PAMI*, 39(5):893–907, 2016. 2, 8

[31] Bo Li, Zhengxing Sun, and Yuqi Guo. Supervae: Superpixelwise variational autoencoder for salient object detection. In *AAAI*, volume 33, pages 8569–8576, 2019. 2

[32] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *ICCV*, pages 7274–7283, 2019. 1

[33] Zhenqiang Li, Weimin Wang, Zuoyue Li, Yifei Huang, and Yoichi Sato. Towards visually explaining video understand-

ing networks with perturbation. In *WACV*, pages 1120–1129. 2, 8

[34] Panagiotis Linardos, Eva Mohedano, JuanJose Nieto, NoelE OConnor, Xavier GiroiNieto, and Kevin McGuinness. Simple vs complex temporal recurrences for video saliency prediction. *arXiv preprint arXiv:1907.01869*, 2019. 1

[35] Kyle Min and JasonJ Corso. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *ICCV*, pages 2394–2403, 2019. 1

[36] Xiongkuo Min, Guangtao Zhai, Ke Gu, and Xiaokang Yang. Fixation prediction through multimodal analysis. *ACM TOMM*, 13(1):1–23, 2016. 3, 6, 7, 8

[37] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Xiao-Ping Zhang, Xiaokang Yang, and Xinping Guan. A multimodal saliency model for videos with high audio-visual correspondence. *IEEE TIP*, 29:3805–3819, 2020. 2

[38] Junting Pan, CristianCanton Ferrer, Kevin McGuinness, NoelE OConnor, Jordi Torres, Elisa Sayrol, and Xavier GiroiNieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017. 8

[39] Junting Pan, Elisa Sayrol, Xavier GiroiNieto, Kevin McGuinness, and NoelE OConnor. Shallow and deep convolutional networks for saliency prediction. In *CVPR*, pages 598–606, 2016. 1, 7, 8

[40] Rui Qian, Yunchao Wei, Honghui Shi, Jiachen Li, Jiaying Liu, and Thomas Huang. Weakly supervised scene parsing with point-based distance metric learning. In *AAAI*, volume 33, pages 8843–8850, 2019. 2

[41] Sucheng Ren, Chu Han, Xin Yang, Guoqiang Han, and Shengfeng He. Tenet: Triple excitation network for video salient object detection. In *ECCV*, pages 212–228, 2020. 1

[42] Dmitry Rudoy, DanB Goldman, Eli Shechtman, and Lihi ZelnikManor. Learning video saliency from human gaze using candidate selection. In *CVPR*, pages 1147–1154, 2013. 8

[43] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019. 2

[44] Meijun Sun, Ziqi Zhou, Qinghua Hu, Zheng Wang, and Jianmin Jiang. Sg-fcn: A motion and memory-based deep learning model for video saliency detection. *IEEE transactions on cybernetics*, 49(8):2900–2911, 2018. 1

[45] HamedR Tavakoli, Ali Borji, Esa Rahtu, and Juho Kannala. Dave: A deep audio-visual embedding for dynamic saliency prediction. *arXiv preprint arXiv:1905.10693*, 2019. 1, 2, 5

[46] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, pages 247–263, 2018. 2, 6

[47] Antigoni Tsiami, Petros Koutras, and Petros Maragos. S-tavis: Spatio-temporal audiovisual saliency network. In *CVPR*, pages 4766–4776, 2020. 1, 2

[48] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *CVPR*, pages 2199–2208, 2019. 3

[49] Bo Wang, Wenxi Liu, Guoqiang Han, and Shengfeng He. Learning long-term structural dependencies for video salient object detection. *IEEE TIP*, 29:9017–9031, 2020. 1

[50] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, pages 136–145, 2017. 2, 3, 8

[51] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *IEEE TIP*, 27(5):2368–2378, 2017. 1

[52] Wenguan Wang, Jianbing Shen, Fang Guo, MingMing Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *CVPR*, pages 4894–4903, 2018. 6, 8

[53] Wenguan Wang, Jianbing Shen, Jianwen Xie, MingMing Cheng, Haibin Ling, and Ali Borji. Revisiting video saliency prediction in the deep learning era. *IEEE PAMI*, 2019. 1

[54] Xinyi Wu, Zhenyao Wu, Jinglin Zhang, Lili Ju, and Song Wang. Salsac: A video saliency prediction model with shuffled attentions and correlation-based convlstm. pages 12410–12417, 2020. 1

[55] Zhenheng Yang, Dhruv Mahajan, Deepti Ghadiyaram, Ram Nevatia, and Vignesh Ramanathan. Activity driven weakly supervised object detection. In *CVPR*, pages 2917–2926, 2019. 3

[56] Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Cap2det: Learning to amplify weak caption supervision for object detection. In *ICCV*, pages 9686–9695, 2019. 3

[57] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, Lihe Zhang, Mingyang Qian, and Yizhou Yu. Multi-source weak supervision for saliency detection. In *CVPR*, pages 6074–6083, 2019. 2, 3, 7, 8

[58] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In *ICCV*, pages 8292–8300, 2019. 2

[59] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. 34:12756–12772, 2020. 2

[60] Dingwen Zhang, Junwei Han, and Yu Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *ICCV*, pages 4048–4056, 2017. 8

[61] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *CVPR*, pages 12546–12555, 2020. 2, 7, 8

[62] Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtash Harandi, and Richard Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *CVPR*, pages 9029–9038, 2018. 2

[63] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas Huang. Adversarial complementary learning for weakly supervised object localization. In *CVPR*, pages 1325–1334, 2018. 3

[64] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 2