A Comprehensive Survey on Video Saliency Detection with Auditory Information: the Audio-visual Consistency Perceptual is the Key!

Chenglizhao Chen[1,2] Mengke Song[1,2] Wenfeng Song[5] Li Guo[3†] Muwei Jian[4]
[1]College of Computer Science and Technology, China University of Petroleum (East China)
[2]Qingdao Institute of Software, China University of Petroleum (East China)
[3]College of Computer Science and Technology, Qingdao University
[4]School of Computer Science and Technology, Shandong University of Finance and Economics
[5]Computer School, Beijing Information Science and Technology University
Code & Data: https://github.com/MengkeSong/SCDL

***Abstract*—Video saliency detection (VSD) aims at fast locating the most attractive objects/things/patterns in a given video clip. Existing VSD-related works have mainly relied on the visual system but paid less attention to the audio aspect. In contrast, our audio system is the most vital complementary part of our visual system. Also, audio-visual saliency detection (AVSD), one of the most representative research topics for mimicking human perceptual mechanisms, is currently in its infancy, and none of the existing survey papers have touched on it, especially from the perspective of saliency detection. Thus, the ultimate goal of this paper is to provide an extensive review to bridge the gap between audio-visual fusion and saliency detection. In addition, as another highlight of this review, we have provided a deep insight into key factors that could directly determine AVSD deep models' performances. We claim that the audio-visual consistency degree (AVC) — a long-overlooked issue, can directly influence the effectiveness of using audio to benefit its visual counterpart when performing saliency detection. Moreover, to make the AVC issue more practical and valuable for future followers, we have newly equipped almost all existing publicly available AVSD datasets with additional frame-wise AVC labels. Based on these upgraded datasets, we have conducted extensive quantitative evaluations to ground our claim on the importance of AVC in the AVSD task. In a word, our ideas and new sets serve as a convenient platform with preliminaries and guidelines, all of which can potentially facilitate future works in further promoting state-of-the-art (SOTA) performance.**

***Index Terms*—audio-visual Fusion; Video Salieny Detection; semantical consistency.**

I. Introduction

Humans tend to be attracted by specific things, and this mechanism has its basic principle in general. But, outwardly, it could vary from different people and scenes, and, directly or indirectly, such differences are usually caused by either personality and individual differences or the exact environment [1], [2], [3]. For example, in an open wild, we may get attracted by a fantastic nature scene view and pay less attention to artificial subjects. However, things go differently in a downtown area, where magnificent artificial buildings could keep drawing our attention. Also, our attention could get shifted to "rare" elements — patterns that are anomalies for their nearby surroundings, and we have an academic name for all these objects/things/patterns attracting our attention — saliency.

†Corresponding author: Li Guo (ally_kwok@163.com)
The first two authors contribute equally to this paper.

a. Video Frame b. VSOD c. VFP

	Training Labels	Loss Functions	Network Designs
VSOD	Manual Annotations (Object-level, subfig-**b**)	Cross-Entropy loss (*focus on intact detection with sharp object boundary*)	Usually use the **bi-stream** network with **late** fusion
VFP	Human-eye Fixations (Scatter Points, subfig-**c**)	KL divergence, CC, NSS, and SIM loss (*focus the output's distribution aspect, paying no attention to the boundary issue*)	Frequently follow the **single-stream** structure with **early** or **mid** fusion

Fig. 1: The differences between **v**ideo **s**alient **o**bject **d**etection (VSOD) and **v**ideo **f**ixation **p**rediction (VFP) regarding "Training Labels", "Loss Functions", and "Network Designs". The training labels (**left** of the below table) of the VSOD task are object-level manual annotations (subfig-b), while the VFP are human-eye fixations (subfig-c). The loss function (**middle** of the below table) usually adopted by VSOD is cross-entropy loss which mainly focuses on intact detection with sharp object boundary. In contrast, the VFP focuses on the output's distribution aspect. Further, regarding network designs (**right** of the below table), the VSOD task mainly adopts bi-stream architectures considering both temporal and spatial information to implement a late fusion of two types of saliency, while the VFP task tends to employ single-stream structure with early or mid fusion.

In general, the saliency-related research activities [4] should come with a specific venue, *e.g.*, the visual saliency, which aims at segmenting the most eye-attracting objects or regions in a given scene. And the scenes are usually "expressed" in images or videos. Since video data is the main course of this survey, we shall omit image-based saliency works.

The current visual saliency detection research field can be roughly divided into two groups, *i.e.*, **v**ideo **s**alient **o**bject **d**etection (VSOD) and **v**ideo **f**ixation **p**rediction (VFP). The primary methodologies of VSOD and VFP are almost the same, whereas the existing hand-crafted methods [5], [6], [7], [8] mainly follow either top-down or bottom-up rationale.

TABLE I: Illustration of the main differences between the existing reviews and ours.

Reviews	Year	Publication	Contents
Katsaggelos *et.al* [13]	2015	P-IEEE	Audio-visual Fusion
Baltrusaitis *et.al* [14]	2018	T-PAMI	Multi-modality Machine Learning
Cong *et.al* [15]	2018	T-CSVT	RGBD/Video/Co-saliency Detection
Wang *et.al* [16]	2019	T-PAMI	Video Saliency Detection
Zhu *et.al* [17]	2021	IJAC	Audio-visual Localization/Correspondence
Chen *et.al* (Ours)	2022	T-CSVT	Audio-visual Saliency Detection

After entering the deep learning era, most of the existing works [9], [10], [11], [12] have adopted the end-to-end encoder-decoder network architecture, which, generally, belongs to the typical top-down category. Hence the difference between VSOD and VFP is the exact training ground truth data, training loss functions, and network architectures. For a better understanding, Fig. 1 has demonstrated such a difference.

Though our visual system is one of the most important venues for us to perceive the environment that we're in, our auditory system also plays an important role. For example, our attention could fast shift to a sounding object, showing that our auditory system can complement our visual system. Despite being complementary in general, these two venues have completely different perceptual mechanisms.

The visual venue is very informative yet with rather limited sensing scope (because of the limited **f**ield **o**f **v**ision, FOV). In contrast, the auditory venue is less informative, yet its sensing scope is dead-angle-free. Besides, different from the visual saliency research field, a pretty mature topic, audio-related saliency is in its infancy. Moreover, different from the visual saliency, a single modality task with abundant accessible training data, the available training data for the **a**udio-**v**isual **s**aliency **d**etection (AVSD) is in a critical shortage[1], which has resulted in a clear performance bottleneck, especially in this deep learning era.

Meanwhile, we have noticed that there exist massive researches [18], [19], [20] regarding the visual and auditory fusion. Their main interests usually focus on multimedia applications, *e.g.*, multi-modality information processing, filtering, and understanding, and these works are rarely intercrossed with the saliency detection research field. Though some of the existing fusion methods proposed in previous literatures [21], [22] can inspire and help the network design toward the saliency detection task, none of them has covered both saliency detection and audio-visual fusion. Thus, as shown in Fig. 2, this review mainly focuses on two topics, and we choose three concrete research fields as the main courses, *i.e.*, **v**ideo **s**aliency **d**etection (VSD), **a**udio-**v**isual **c**orrespondence (AVC), and **a**udio-**v**isual **s**aliency **d**etection (AVSD). Also, the differences between several existing reviews on audio-visual representation learning and ours have been illustrated in Table I.

Despite providing an extensive review, we have noticed that the **a**udio-**v**isual **c**onsistency (AVC) between audio and visual, a representative task considered in the multimedia research field [23], [24], [25], is the key factor to determining the overall performance of AVSD, while its importance has long been overlooked by our AVSD research field. To verify our claim, we have newly labeled all publicly available AVSD clips frame-by-frame and conducted massive quantitative experiments with them. This new finding can potentially benefit our audio-visual saliency detection research field shortly.

[1]widely-used VSD and VFP training datasets comprise totally 1.6K video clips, while the available data for AVSD is only 0.2K clips, not to mention the fact that the AVSD task is more challenge than VSD and VFP, and thus is more data-hungry.

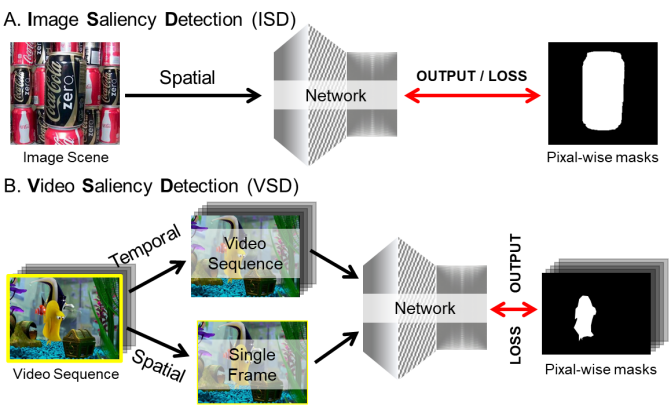


Fig. 3: Comparison between Image Saliency Detection and Video Saliency Detection. The former aims at detecting visually significant areas in the static image, while the latter contains not only the feature information of static images, but also the dynamic information between adjacent video sequences.

In a summary, significant highlights and contributions of this review include the following aspects:

- This review is the first attempt to bridge the gap between saliency detection and audio-visual fusion;
- We have extensively included the most recent deep learning-based works, making this review fresh and capable of helping new hands to join this new research topic;
- We have noticed one critical factor — the semantical consistency degree, which has been well studied by the multimedia research field while being completely omitted by our AVSD research field, could significantly influence the AVSD performance;
- For all widely-used existing AVSD datasets, we have newly equipped them with frame-wise semantical consistency degree labels, which could potentially benefit our research community.

II. Video Saliency Detection

A. *Image Saliency Detection* v.s. *Video Saliency Detection*

Image saliency detection (see in Fig. 3-A) aims at detecting the most eye-attracting-areas in the image, *e.g.*, region of interest, or regions with distinct patterns/textures/appearances. Thus, the primary problem scope of image saliency detection is usually localized in the spatial domain, where only a single image is considered each time when performing image saliency detection. Compared with images, videos additionally contain temporal information. Since the human visual system tends to pay more attention to dynamic changes, we shall simultaneously consider both static image and dynamic temporal information when performing video saliency detection (see in Fig. 3-B) — aiming at mimicking the human visual system.

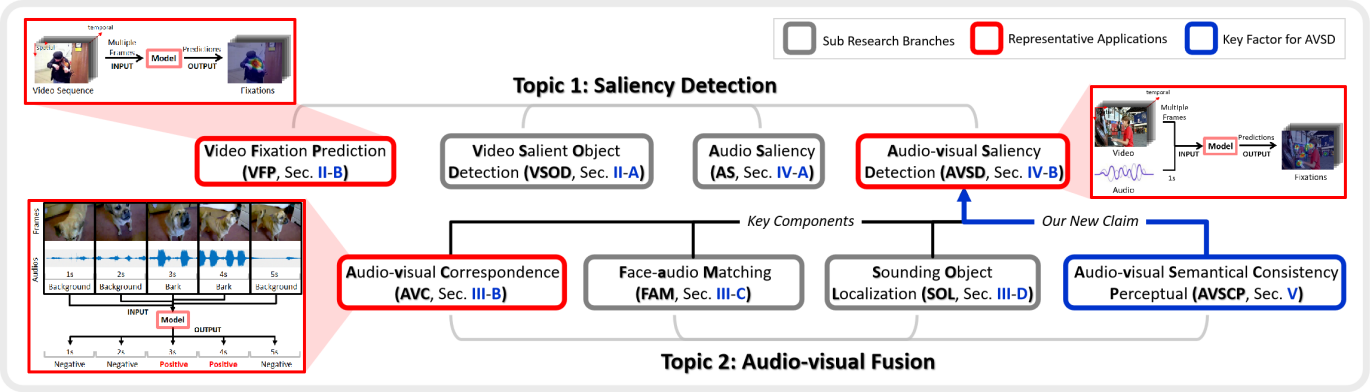


Fig. 2: Our review's structure covers two significant topics: 1) Video Saliency Detection and 2) Audio-Visual Multi-Modality Fusion. W.r.t., the most representative applications, we have highlighted them with red rectangular boxes. Also, we have newly argued that the audio-visual semantical consistency perceptual (highlighted by the blue box) is the key factor in determining the AVSD performance.

B. Video Salient Object Detection v.s. *Video Fixation Prediction*

In the video saliency detection research field, there exist two main research branches, including the **v**ideo **s**alient **o**bject **d**etection (VSOD) [7], [5], [26], [27], [28], [29] and the **v**ideo **f**ixation **p**rediction (VFP) [30], [31]. The major differences between VSOD and VFP lie in three aspects: training labels and loss functions and network designs.

As shown in Fig. 1 (left column in the table), training labels used for the VSOD task are binary masks, where all salient objects have been well annotated/segmented by humans. In contrast, the labels used in the VFP task are human-eye fixations (*i.e.*, individual pixel-wise coordinates) collected by eye-trackers directly, representing raw image regions that humans would pay attention to. In a word, training labels for the VSOD task are object-aware, while training labels for the VFP task are scattering locations.

Also, as shown in Fig. 1 (middle column in the table), the widely-used loss function in the VSOD task is the cross-entropy loss, while the VFP task usually prefers the **k**ullback-**l**eibler (KL) divergence, linear **c**orrelation **c**oefficient (CC) loss, **n**ormalized **s**canpath **s**aliency (NSS) loss, and **sim**ilarity (SIM) loss.

Further, the VSOD task mainly adopts bi-stream architectures considering both temporal and spatial information to implement late fusion. The VFP task tends to employ single-stream structure with early or mid fusion. See in Fig. 1 (right column in the table).

Despite using different training labels, loss functions, and network designs, there also exist multiple other distinguishing differences:

1) The VSOD task should additionally consider detections' integrity, *i.e.*, the detected salient regions should precisely comprise the entire salient object with all its subparts. However, the VFP task aims to simulate the human eye's fixation, and thus the detected results are not required to highlight the entire object.

2) The widely-used VSOD scenario could be fully automatic video segmentation. In this application, the saliency ranks of different objects tend to stay unchanged for a long time. However, the human eye's fixations are usually scattered locations, which are relatively weak in indicating those corresponding objects. In other words, fixations usually shift between objects.

In the following two subsections, we will review these two research branches respectively.

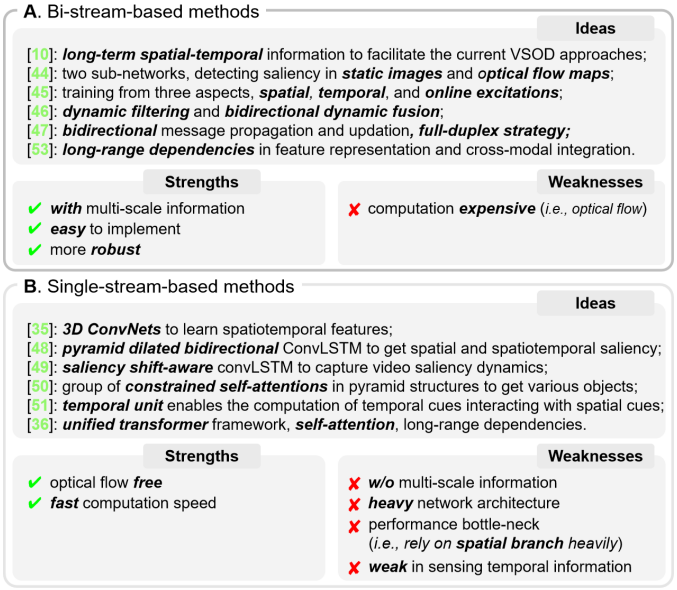


Fig. 4: The overall summary of VSOD including bi-stream-based methods (A) and single-stream-based methods (B).

C. Video Salient Object Detection (VSOD)

Task Definition. VSOD aims to locate and segment the most eye-attracting salient objects. Given a given video scene, VSOD can be regarded as a multi-task problem, where salient object localization and segmentation are performed simultaneously in an intensive, interactive manner between these two tasks. The major challenge of VSOD is how to appropriately fuse spatial and temporal information when formulating saliency decision rules, while these two independent information sources usually conflict with each other leading to learning ambiguity.

The existing **s**tate-**o**f-**t**he-**a**rt (SOTA) VSOD models [32], [33], [34], [35], [36], [37], [38], [39] can be divided into two

TABLE II: Strengths and weaknesses comparison between Optical Flow [48], LSTM [49], ConvLSTM [50], 3D Convolution [51] and Transformer [52] towards temporal sensing. {M. S.}: multi-scale, {O. F.}: optical flow, {P. B.}: performance bottleneck, {F. ST. I.}: full spatiotemporal interaction; ✗: *without*, ✓: *with*.

Methods	Implement	M. S.	Computation	Network	O. F.	P. B.	F. ST. I.
Optical Flow [48]	*easy*	✓	*expensive*	*light*	✓	✓	✗
LSTM [49]	*hard*	✗	*cheap*	*heavy*	✗	✓	✗
ConvLSTM [50]	*hard*	✗	*cheap*	*heavy*	✗	✓	✗
3D Convolution [51]	*easy*	✓	*cheap*	*light*	✗	✗	✓
Transformer [52]	*hard*	✓	*expensive*	*heavy*	✗	✓	✓

groups according to their network designs: 1) the bi-stream-based methods [40], [10], [41], [42], [43], and 2) the single-stream-based ones [44], [45], [46], [47].

The bi-stream-based models usually consist of two sub-branches, one for the motion saliency clues, whose input focuses on the temporal information (*e.g.*, optical flow data); another is the conventional color branch, which could be any off-the-shelf image salient object detection deep model. Note that the network architectures of these two branches could be the same, and the only difference is their training input, *i.e.*, optical flow result *vs.* color image.

The single-stream-based methods have abandoned the individual temporal computation, *e.g.*, the time-consuming optical flow [48]. Instead, it takes multiple frames as input each time, and then uses either LSTM [49], ConvLSTM [50], 3D convolution [51] or Transformer [52], [53] to sense temporal information. Detailed comparison results of these methods are shown in Table II. Compared with the bi-stream-based methods, this type of work has a significant advantage, *i.e.*, it could be 10 times faster in computation because the individual temporal information computation is the major efficiency bottleneck for the bi-stream-based approaches. More details regarding this issue can be found in [47].

In-depth Summary. As shown in Fig. 4, the two sub-branches of VSOD have been briefly summarized into subfig-A and subfig-B. The bi-stream-based methods Fig. 4-A utilizes optical flow to offer temporal information, but the computation of optical flow is time-consuming, slowing down the inference speed. Also, the bi-stream-based VSOD methods can make full use of multi-scale information via dense inter-stream short-connections in both encoder and decoder stages. Thus their results can retain good boundary sharpness. Further, the bi-stream network fashion is implementation friendly, requiring no complex architectures.

In sharp contrast to bi-stream-based methods, the single-stream-based methods Fig. 4-B have a distinct advantage, *i.e.*, because of being optical free, their computation speed is extremely fast. However, the single-stream-based methods also have several limitations. First, due to the attribute of free optical flow, the single-stream-based methods are usually weak in sensing temporal information. Second, they usually utilize early spatiotemporal fusion while omitting the spatial and temporal interaction in their decoding stage. Third, their network architectures are usually very complex. Last, full multi-scale interaction is beyond the reach of single-stream-based methods, and they rely on spatial branch heavily, resulting in unstable.

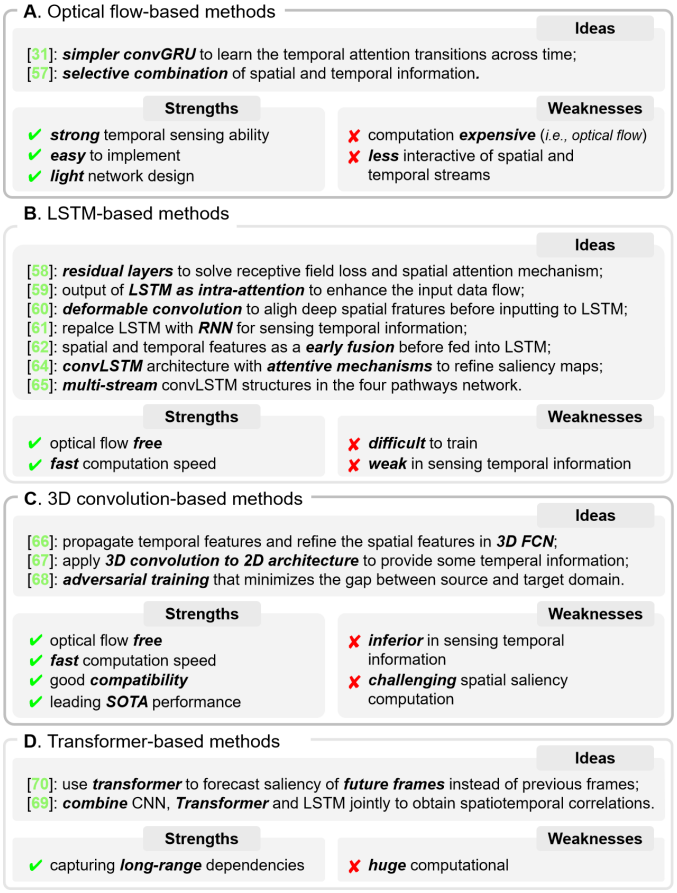


Fig. 5: The overall summary of VFP, where optical flow-based methods (A) are computation expensive, LSTM-based methods (B) are inferior to sense temporal information compared with optical flow, 3D-convolution-based methods (C) are the most inferior to the other two regarding the ability of temporal information sensing, and Transformer-based methods (D) are good at capturing long-range context dependencies.

D. *Video Fixation Prediction (VFP)*

Task Definition. Unlike the VSOD task using manual well-annotated object-wise binary masks as training objectives, the training GTs for the VFP task are scattered human-eye fixations collected by the eye trackers (*e.g.*, Tobbi, EyeLink, Smart Eye, and GazeTech). The earliest deep learning-based VFP approaches [54] followed the bi-stream structure, which belongs to the multi-task rationale, where one stream handles the fixation predictions in the spatial domain, and another stream focuses on the fixation predictions over the temporal scale. Thus, the key problem of the bi-stream-based VFP models is how to achieve the fusion balance between its sub-streams.

Optical Flow-based Approaches. The primary way for the bi-stream VFP models to sense temporal information is to take the optical flow results as the models' input. As shown in Fig. 6, almost all existing bi-stream VFP models have adopted the optical flow (*e.g.*, the most representative conventional one [48] and the deep learning-based ones,

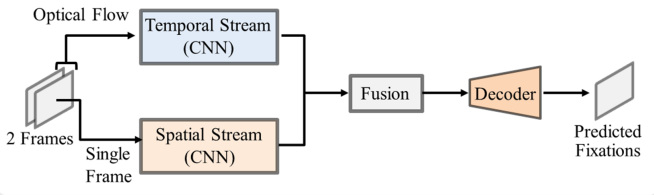


Fig. 6: Method pipeline of the optical flow-based bi-stream approaches mainly contains a temporal stream and a spatial stream followed by a fusion module and a decoder.

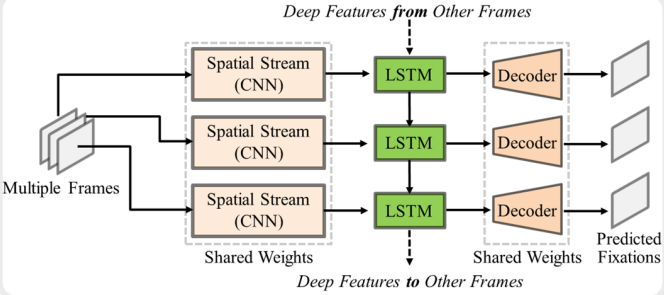


Fig. 7: Method pipeline of the **l**ong **s**hort-**t**erm **m**emory-(LSTM-) based approaches which usually follow the single-stream methodology.

such as FlowNet [55], [56]) as the temporal sub-stream to sensing temporal information. Here we just name a few most representative ones.

In [31], Lai *et al.* have made two key innovations: 1) a novel way for performing early fusion between spatial and temporal feature backbones, and 2) the **conv**olutional **G**ated **R**ecurrent **U**nit (convGRU) has been firstly applied for learning the temporal attention transitions across time, which can make the predicted video fixation maps temporally smooth. The major highlight of the fusion scheme is that the deep features obtained by the spatial and temporal feature backbones are connected densely via residual attention mechanism in a multi-scale way. Specifically, the exact fusion operation has biased toward the spatial information, where the deep features from the temporal backbone are only served as auxiliary stimuli. As a variant of the classic LSTM, the proposed convGRU has two advantages: 1) simpler network design and 2) slight performance improvement (less than 0.5%).

Following the bi-stream structure [30] also, Zhang *et al.* [57] have devised a novel fusion scheme. The key idea of the proposed fusion is to perform a selective combination of spatial and temporal information. The channel-wise attention has been used as the indicator to guide the selection process, and the rationale is that only those deep features with strong feature responses would be able to benefit from the detection task. In addition, the authors have devised a novel strategy that takes the spatial position of the salient objects in previous consecutive frames as the additional input, aiming at facilitating the estimation of temporal saliency by shrinking the problem domain. Consequently, the network's output could stay consistent (smooth) over the temporal direction.

Summary of Optical Flow-based VFP. As shown in Fig. 5-A, the major advantage of the optical flow-based approach is its strong temporal sensing ability (due to the usage of optical flow). The disadvantages are also clear: time-consuming and less interaction between spatial and temporal sources. The comprehensive summary of Optical Flow-based approaches can be seen in Fig. 5-A.

Long Short-Term Memory-based Approaches. Actually, most of the existing **s**tate-**o**f-**t**he-**a**rt (SOTA) VFP approaches have adopted the **l**ong **s**hort-**t**erm **m**emory (LSTM) to sense temporal information. The LSTM-based approaches usually follow the single-stream methodology compared to the optical flow-based ones. As can be seen in Fig. 7, this type of approaches usually adopts the **c**onvolutional **n**eural **n**etworks (CNN) to compute spatial deep features for each single frame. Then, to sense temporal information, all deep features computed individually via CNN are fed into the input gate of LSTM. Finally, a decoder is applied to produce the pixel-wise fixation prediction.

In [58], Wang *et al.* have completely followed the structure demonstrated in Fig. 7. However, some modifications have been made in the spatial stream, including 1) several residual layers were used to compensate for the loss of receptive field caused by removing the last two pooling layers of the VGG16 feature backbone; 2) the spatial attention mechanism was applied to the spatial-stream for facilitating the network training, where the static fixation GTs could be used as the attentions helping the network's training (*i.e.*, the dynamic saliency), which could be able to alleviate the demand of large scale of costly video fixation GTs.

Similar to [58], Linardos *et al.* in [59] have placed the LSTM in the middle stage of a typical encoder-decoder CNN. The LSTM collects the output of the encoder, and then its output, representing the spatiotemporal information, is fed to the decoder to formulate the fixation prediction. The major highlight of this work is the proposed recurrent mechanism, where the LSTM's output is used as an intra-attention to enhance the input data flow. Consequently, the network's ability to sense temporal information gets improved significantly.

To further enhance the sensing ability of temporal information, Chen *et al.* in [60] have taken 3 frames as the network's input each time. Then, the deep features computed from these frames are combined as the input of LSTM. Compared with the conventional LSTM-based approaches, which take only 1 frame as input each time, this method has considered 3 frames, and thus its temporal sensing ability could, of course, get enhanced. Since the spatial displacement occurs along the time scale, deep features computed from consecutive video frames are usually misaligned, which could confuse the subsequent learning process, blurring the final prediction results. To solve this problem, the authors have resorted to deformable convolution — an off-the-shelf tool that could dynamically learn the spatial positions of the convolutional kernels. By using the deformable convolution, the deep features before inputting into the LSTM are aligned.

It is worth mentioning that the LSTM can also be replaced by other networks which can sense temporal information. For example, Droste *et al.* in [61] have adopted the **r**ecurrent **n**eural **n**etwork (RNN), the early prototype of the LSTM,

for sensing temporal information, where the RNN is placed between the encoder and decoder, sharing a similar overall network structure to that of the [60].

Apart from the single-stream LSTM-based approaches mentioned above, there also exist several works [62] following the bi-stream methodology, where the spatial and temporal information interact with each other as an early fusion. Jiang *et al.* in [62] followed the typical bi-stream structure, in which a pruned form of YOLO is applied as the subnet for sensing the spatial information, and the temporal stream is a pruned FlowNet [55]. The multi-scale deep features provided by the spatial stream are collected via the concatenation and batch normalization operations, formulating a coarse localization mask to compress those non-salient backgrounds in the deep spatial features. Meanwhile, the deep features of the temporal stream are also assembled in a way identical to that used in the spatial stream. Finally, the multi-scale deep features assembled individually from the spatial and temporal streams are concatenated to be fed to an LSTM.

Similar to the early fusion adopted in [62], Wu *et al.* in [63] have committed one modification to enhance the temporal sensing ability: the inter-frame correlations are explored by performing the simple dot-product operator along the channel dimension. Besides, the authors have adopted the spatial attention-based shuffle operation to enhance the spatial stream, where the deep multi-level features are combined and later shuffled. Both these strong spatial deep features and cross-frame correlation features will be fed into a variant version of the LSTM, named the correlation-based ConvLSTM, where the input gate has been modified to an addition operation-based feature fusion; thus, it could be able to simultaneously take two different sources as its input.

Also, Cornia *et al.* in [64] have employed the ConvLSTM architecture with attentive mechanisms to refine the predicted saliency maps iteratively. Those predictions are combined with priors to model the tendency of humans to fix the center region of the image. In [65], Gorji *et al.* have deployed multi-stream ConvLSTM structures in the four pathways network, followed by an augmenting convnet that learns to combine the complementary and time-varying outputs of the ConvLSTMs by minimizing the relative entropy between the augmented saliency and viewers' fixation patterns on videos.

Summary of LSTM-based VFP. As shown in Fig. 5-B, the major advantage of the LSTM-based VFP is its faster computational speed. However, some of the most recent works [47] have argued that the nature of the LSTM might not be powerful to sense temporal information. The main reason is that such methods tend to focus solely on adjacent pixels, causing a loss in long-range dependency. Also, the LSTM-based VFP methods are usually difficult to train because the adopted LSTMs need to perform two tasks simultaneously, *i.e.*, 1) sense temporal information, and 2) fuse spatial and temporal information.

3D Convolution-based Approaches. Compared with the widely-used 2D convolution that can only sense spatial information, 3D convolution can sense both spatial and temporal information in a cubic way. As been discussed in [47], 3D convolution is generally inferior to its competitors (*e.g.*, LSTM [49] and optical flow [56]) in sensing temporal information, but it still has several unique advantages, *i.e.*, fast computation and good compatibility[2]. Also, to the best of our knowledge, 3D convolution-based VFP models [66] are generally leading the SOTA performance (simultaneously considering accuracy and efficiency), and the overall method pipeline of this type of approach has been provided in Fig. 8. We shall review several representatives here.

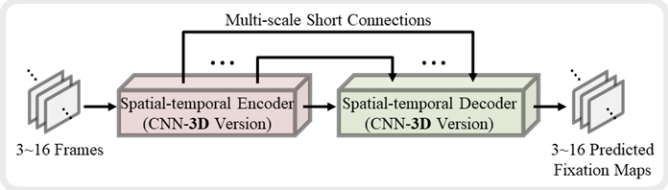


Fig. 8: Method pipeline of the 3D convolution-based approaches and the major highlight of these approaches is their capability of sensing both spatial and temporal information in a cubic way.

Min *et al.* in [67] have directly applied the 3D convolution to the conventional 2D encoder-decoder architecture, where the exact implementation is straightforward, *i.e.*, all 2D convolutions are replaced by 3D versions. Though the newly applied 3D convolution can provide some temporal information, one critical problem exists in the decoder. The widely-used unpooling operation cannot provide the exact spatial locations over the temporal scale, limiting its decoder's performance. To alleviate it, the authors have devised an auxiliary pooling scheme, whose key rationale is to record all spatial, temporal, and channel locations when performing pooling operations. Therefore, the unpooling operations in the decoder layers can re-use the reserved locations eventually.

Recently, Bellitto *et al.* in [68] have followed the 3D encoder-decoder network structure for the VFP task. The highlight of this approach is the newly proposed decoder, where two new concepts have been considered. To handle the domain shift problem, each side output of the encoder is assigned to an unsupervised binary classifier, whose primary objective is to follow the adversarial training that minimizes the gap between features learned from the source and target domain. Besides, for each layer in the decoder, multiple domain-specific priors are dynamically learned and incorporated to make the network domain-specific. This strategy could significantly improve quantitative scores further.

Summary of 3D Convolution-based VFP. As shown in Fig. 5-C, compared with the LSTM-based VFP, the 3D Convolution-based VFP methods usually have faster computation speed since 3D convolutions are more lightweight than LSTM. Also, as we have mentioned before, 3D convolution can serve any SOTA VFP as a plug-in. Thus any 2D convolution-based methods can be easily adapted to handle video data by replacing 2D convolutions with 3D convolutions.

Transformer-based Approaches. Comparing with the existing CNNs-based methods which have mainly adopted con-

[2]the powerful **generic nature** of the 3D convolution, *i.e.*, by replacing 2D convolutions to 3D convolutions, any 2D convolution-based image saliency detection models can be easily converted to adapt the saliency detection over video data without much network modifications

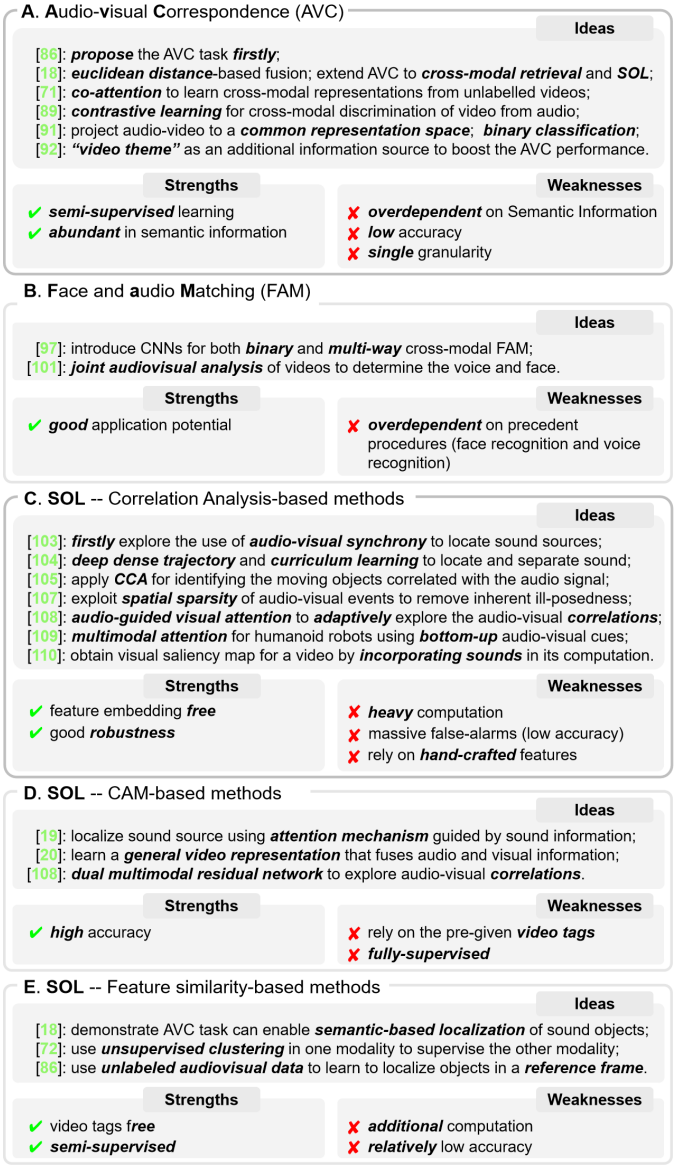


Fig. 9: The overall summary of Audio-Visual Multi-Modality Fusion, mainly consisting of three tasks: **a**udio-**v**isual **c**orrespondence (AVC), **f**ace and **a**udio **m**atching (FAM) and **s**ound-**o**bject **l**ocalization (SOL).

volutions to obtain features, Transformer-based approaches utilize self-attention machanism to perform feature extraction. Since CNN can only conduct locally, while self-attention machanisms can extract features globally, which is more powerful to model the long-range correlations between video frames in temporal sequences [69].

Ma *et al.* [70] have firstly introduced a pure-transformer framework for video saliency prediction. In the work, the authors have instructively forecasted visual saliency of future frames rather than merely focusing on previous frames. Since only performing VFP methods will make focusing regions lag behind the ongoing scenes, which may induce the disability to keep salient objects in the center of captured videos or keep following objects with high speeds. Thus, they have proposed a video saliency forecasting transformer to explore temporal and spatial semantic information from input videos for video saliency forecasting. And a cross-attention decoder is employed to eliminate the time dimension of the decoder feature by the time embedding layer.

Also, Wang *et al.* in [69] have introduced self-attention machanism to obtain spatiotemporal correlations between features and saliency regions. The key difference compared with [70] is that this work has combined CNN, Transformer and LSTM jointly to make fixation predictions. Specifically, a CNN-based multi-scale feature-fusion network aims at effectively extracting features in multi-category space and a CNN-based DConvLSTM, as the decoder, is used for dynamic information learning. The Transformer encoder serves to learn global correlation between pixels and human visual attention in both time and space domains.

Summary of Transformer-based VFP. As shown in Fig. 5-D, compared with the above-mentioned CNN-based VFP methods, the most strength of Transformer-based VFP is the aptitude for capturing long-range dependencies. Thus, it is good at sensing long-range temporal information. However, Transformer is a modeling approach based on pixel-to-pixel points, so the computational is undoubtedly huge.

III. Audio-Visual Multi-Modality Fusion

Unlike the visual signal, which determines human attention directly, the audio signal is usually the auxiliaries, influencing human attention subtly. For example, our attention can be easily attracted by a sounding object, *e.g.*, the sound of a dropping box hitting the floor. However, some audio signals are also completely helpless in drawing our attention, *e.g.*, background music. Thus, since the human visual field has blind spots, the audio signal, whose perception scope is almost 360°, should be appropriately used to complement visual in practice. With the development of deep learning techniques, more and more research attention has been paid to how to combine/fuse audio and visual for vision-related tasks, *e.g.*, sounding object localization [18], audio-visual synchronization [73], object tracking [74], and saliency detection [75]. Though the primary focus of this review is on saliency detection, we shall still review several most representative audio-visual-related tasks [76], [77], [78] in advance because these fusion-related arts can be directly referred to and get a deep insight into our audio-visual saliency detection. For a better reading, we propose to introduce three most representative tasks here, including **a**udio-**v**isual **c**orrespondence (AVC), **f**ace and **a**udio

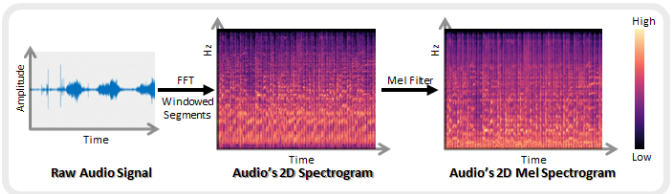
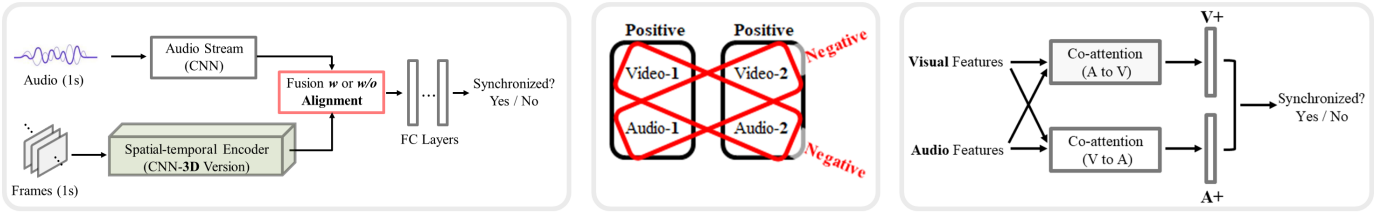


Fig. 10: Audio feature computation details. First, the raw 1D audio signal is transformed to a 2D spectrogram by **f**ast **F**ourier **t**ransform (FFT), thus the existing popular backbones (*e.g.*, VggNet [71] or ResNet [72]) can be used. Next, the Mel filter makes the 2D spectrogram more discriminative and sensitive to our human auditory system.



(a) (b) (c)

Fig. 11: (a) The widely-used network architecture for the **a**udio-**v**isual **c**orrespondence (AVC) task, which is a typical binary classification, where the key is how to align and fuse the audio and visual streams. (b) The widely-used **a**udio-**v**isual **c**orrespondence (AVC) training data formulation. The synchronized video and audio pairs are set to positive, whereas the unsynchronized video and audio pairs are set to negative. (c) Co-attention-based audio and visual feature fusion, where the outputs of the co-attention operation can be regarded as the upgraded versions, *i.e.*, A+ means upgraded audio features and V+ denotes upgraded visual features, where all those clearly unsynchronized information can be effectively excluded.

matching (FAM) and **s**ound-**o**bject **l**ocalization (SOL). The overall summary can be see in Fig. 9.

A. Preliminaries on Audio Feature Representation and Audio-Visual Feature Embedding

Given any 1-dimensional raw audio data, we can directly input it to an off-the-shelf feature backbone (*e.g.*, SoundNet [79] or VggSound [80]), where the raw audio is sequentially convoluted by a seises of 1D kernels. Also, the 1D audio signal can also be transformed to a 2D spectrogram, thus we can adopt the existing popular backbones (*e.g.*, VggNet [71] or ResNet [72]) instead, where the audio signal's 2D spectrogram can be visualized in the middle of Fig. 10. To make the audio's 2D spectrogram more discriminative and sensitive to our human auditory system, we can use the Mel filter — a predefined linear transformation [81], to convert the 2D spectrogram to Mel spectrogram (see the right part of Fig. 10).

Recently, there have been several works that have focused on the audio-visual feature embedding [21], [22], [82]. The main objective of audio-visual feature embedding is to obtain a generic feature representation. Thus in the spanned feature space, the embedded features can be informative and discriminative enough for specific applications, *e.g.*, the multi-modality image retrieval [83], audio-visual correspondence, face and audio matching, and sound object localization.

Tian *et al.* [84] have applied channel-wise attention to help selectively fuse audio and visual features. The motivation is very straightforward, which is based on an assumption, *i.e.*, that either visual or audio features might benefit the subsequent classification task, and thus the one with a higher feature response should be considered more during the fusion. Following this rationale, channel-wise attention has been applied simultaneously to both audio and visual streams. Then the exact modality-wise selection is achieved by performing a softmax. Note that this channel-wise attention-based multi-modality selective fusion has also been used in some existing VSOD approaches, *e.g.*, the classic MGA [40]. Recently, Gao *et al.* in [85] have adopted a distillation network to compute audio-visual features. A teacher network was initially trained in the visual domain, where the video tags were used as the classification supervision. Then, a student audio-visual network was trained by taking the predictions from the teacher network as its supervision. Thus the learned intermediate features can achieve automatic alignment between audio and visual and finally obtain a strong audio-visual feature embedding.

B. Audio-visual Correspondence (AVC)

Task Definition. The AVC task focuses on discovering the global semantic relation between audio and visual modalities, which takes both audio and visual information as input, then makes binary predictions on whether the given audio event is synchronized with the current visual event. For example, a barking dog might be out of the visual field, making the audio event unsynchronized with the visual event. In this case, the AVC task should make a negative prediction and vice versa. For a better understanding, we have provided a pictorial demonstration of the AVC task's overview in the bottom-left of Fig. 2. Also, Fig. 11 (a) demonstrates the AVC task more clearly from the network perspective. The nature of the AVC task is a typical binary classification, and the technical key is how to align and fuse audio and visual streams.

Arandjelovic *et al.* in [86] followed an identical network structure to that of Fig. 11 (a). Instead of focusing on the feature representation aspect, the primary interest of this work is to learn the relationship between single static frames and their audio counterparts. To fuse deep features derived from audio and visual sources, the authors have resorted to a series of feature reshape layers (*i.e.*, pooling layers). Hence, both features of audio and visual streams are reshaped to an identical size, which will be later fused via multiple fully-connected layers. The proposed training process requires no additional supervision data, where image and audio training instances pairs are automatically obtained by sampling two different videos, *i.e.*, picking a random frame from video-1 and a random 1-second audio clip from video-2, and please see Fig. 11 (b) for more details. Note that this strategy has been widely used in our AVC research community as the default training protocol.

Following the bi-stream structure, the same authors in [18] have made one significant modification regarding the audio-visual fusion part. In the early version [86], features derived from audio and visual streams are fused via the widely-used feature concatenation operation. However, the concatenation-based fusion tends to misalign both audio and visual signals, resulting in the fused audio-visual features being inadequate

for cross-modal retrieval. Thus, [18] has adopted the Euclidean distance-based fusion scheme to enforce the feature alignment process.

Also using the bi-stream framework, Cheng *et al.* in [73] have presented a fancy fusion scheme, where deep features respectively derived from either the audio steam or the visual stream are firstly combined by the newly designed "co-attention" operation, which has been shown in Fig. 11 (c). The primary objective of this co-attention operation is two-fold: 1) enhance audio-visual consistencies and 2) suppress those inconsistencies. As shown in Fig. 11 (c), the outputs of the co-attention operations can be regarded as the upgraded versions of the original input, *i.e.*, **A**+ and **V**+, where all those clearly unsynchronized information can be effectively excluded. In addition, the exact implementation of co-attention could be either the widely-used spatial attention [87] or the fancy transformer [88].

To further promote SOTA performance, the existing learning strategies (*e.g.*, contrastive learning [21]) can be used directly. Morgado *et al.* in [89] have applied contrastive learning to the AVC task, whose core idea can be briefly summarized as increasing the inter-class distance and decreasing the intra-class distance. In the implementation, training instances belonging to the intra-class are audio and visual pairs whose semantical feature distances are below the given hard threshold. And the rest of the audio-visual pairs are the inter-class cases. There also exist some other similar works (*e.g.*, [90]) which have adopted the existing learning strategies targeting better audio-visual feature embedding.

The AVC task can also be extended to tell if the current visual information is appropriate with the corresponding audio information. For example, it is inappropriate for a video frame to contain happy faces with a sad melody. To achieve this goal, Verma *et al.* in [91] have "weakly" divided the input audio-visual signals into three categories according to their intrinsic emotions, *i.e.*, positive, neutral, and negative, whose structure is almost identical to that of Fig. 11 (c). A similar solution can be found in [92], where the authors have adopted the *video theme* as an additional information source to boost the AVC performance. The "video theme" adopted in this paper is the manual video-level category tags. And the rationale of this work is to use the theme-based classification responses to eliminate instances whose audio-visual semantics are unsynchronized.

In-depth Summary. As shown in Fig. 9-A, most existing AVC methods have adopted semi-supervised learning, where *positive* and *negative* training instances are obtained by extracting visual and audio fragments from either the same sequence or different sequences. Since semantic consistency is the sole indicator to show if a given visual and audio pair corresponds, the existing AVC methods usually have very strong semantic information. Also, they are over-dependent on semantic information resulting in unstable and low-performance accuracy. Further, this type of approach can only conduct the AVC in single granularity — can only achieve batch-wise predictions rather than frame-wise predictions.

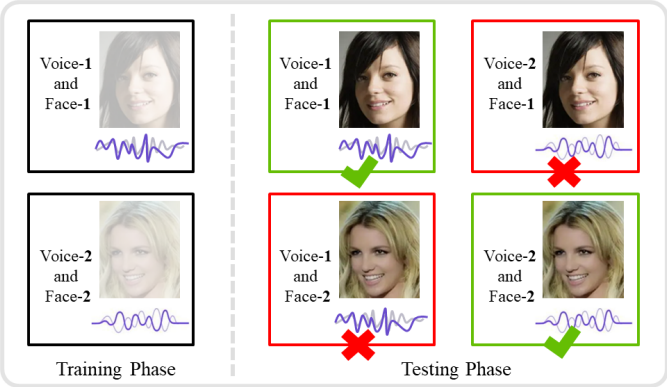


Fig. 12: Matching between faces and voices. The matched faces and voices pairs are set as positive (✓), where the unmatched faces and voices pairs are set as negative (✗).

C. *Face and Audio Matching (FAM)*

Task Definition. The primary task of FAM is that, given a face image/video and a voice sequence, FAM determines if the given face matches the voice. This task's overview can be seen in Fig. 12. The methodology of the FAM task is quite similar to that of the person **re-id**entification (ReID) [93], [94], [95], [96], while the major difference relies only on their feature modalities, where the ReID task only needs to consider the visual domain, while the FAM task needs to consider both audio and visual. Also, the key to succeeding in matching faces and voices heavily relies on the design of an appropriate audio-visual fusion.

Like the AVC task, Nagrani *et al.* in [97] directly treated the FAM task as a binary classification. In this work, the face and voice features are obtained by feeding the given face and voice to the existing feature backbone. The audio-visual fusion is implemented by the widely-used concatenation operation, and the final classification is fulfilled by the conventional fully-connected layers. Like the AVC task, the existing learning strategies could also be directly applied to the FAM task and bring solid performance gain, *e.g.*, triplet loss [98], [99] or contrastive loss [100]. Meanwhile, the FAM research field [101], [102] has also focused on feature embedding. Rather than performing the binary classification towards the matching problem, some other weakly-supervised classifications (*e.g.*, identity, gender, and nationality [98]) towards a single modality can also be used to implicitly obtain the aligned face-voice deep features.

In-depth Summary. As shown in Fig. 9-B, the FAM task is a representative application of AVC, but it is overdependent on precedent procedures, *i.e.*, face recognition, and voice recognition, making the existing FAM methods unstable.

D. *Sounding Object Localization (SOL)*

Task Definition. Given a pair of video and audio examples, the SOL task aims to locate the sounding object in visual space. and the task overview can be seen in Fig. 13. Recent works are mainly based on audio-visual synchronization, which jointly train visual and sound networks to obtain visual and audio features, respectively, then fuse the features, and

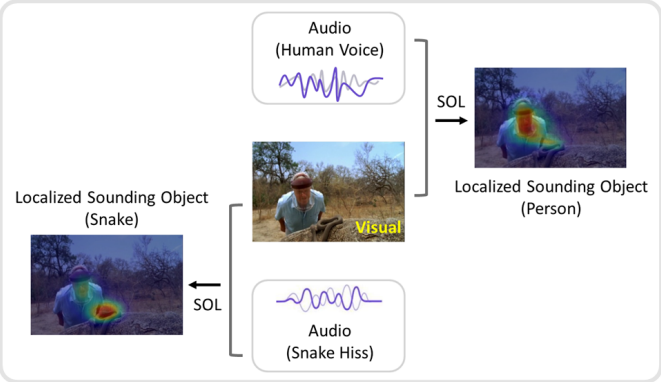


Fig. 13: The demonstration of the SOL tasks. The objective of SOL is to locate the sounding object in the visual space, *e.g.*, given a visual scene, the snake can be located with the snake hiss, while the person can be located with the human voice.

finally highlight the spatial regions sharing strong feature consistency with the audio counterpart.

Correlation Analysis-based SOL Approaches. The research of SOL has a long history, where the earliest work originates in 1999 [103]. In this work, Hershey *et al.* [103] have explored the correlation between audio and video signals. The idea itself is straightforward, whose rationale is that a spatial region containing a sounding object should have a large probability of exhibiting a strong correlation with the audio signal. Zhao *et al.* [104] have devised a Deep Dense Trajectory model and a curriculum learning scheme to locate and separate sound, which exploits the inherent coherence of audio-visual signals. The highlight of the work is learning the motion cues necessary for audio-visual sound separation. After that, several works have adopted various correlation analysis methods for the SOL task, and we shall briefly review them.

Izadinia *et al.* in [105] have applied the **c**anonical **c**orrelation **a**nalysis (CCA) [106] for identifying the moving objects which are heavily correlated with the audio signal. And similar attempts can be found in [107], [108]. Besides, several existing works [109], [110] have considered mutual information as the alternation, whose rationales are very similar to that of the CCA-based ones. In a word, the correlation analysis-based approaches are usually hand-crafted ones, which can only perform well when visual information has strong consistency with the audio counterpart.

Different from the correlation analysis-based approaches, which are mainly interested in the SOL task and designed mainly for videos with plain audio signals, there also exist several works [111], [112] which have investigated the stereo cases, *i.e.*, videos with a stereo audio signal. The key idea of this branch of work is very simple — the sounding object's spatial location can be coarsely determined by analyzing the difference between the individual soundtracks. Let's take the dual soundtrack. For example, the audio signal of a sounding object should first arrive at the left microphone if the sounding object is located on the left. Theoretically, this type of approach can achieve the best SOL performance. However, the stereo audio signal requirement inevitably narrows the broad applications.

Summary of Correlation Analysis-based SOL. As shown in Fig. 9-C, the major advantage of correlation analysis-based SOL methods is that they do not need to perform specific visual-audio feature embedding. Instead, they usually adopt various off-the-shelf feature computation tools, most of which are hand-crafted ones, to span high-dimensional feature spaces for both visual and audio data. By using correlation analysis methods, they can reveal the potential visual-audio consistencies for localizing the sounding visual areas. However, due to the hand-crafted nature, such methods are generally time-consuming. And because of the limitations of hand-crafted features (*i.e.*, weak discriminative ability), their results usually exist massive false alarms.

Deep Learning-based SOL Approaches. Recently, SOL-related works are all based on deep learning [113], [114], [115], [116], whose key idea is to perform audio and visual feature embedding. And most of them can be roughly divided into two groups: 1) the **c**lass **a**ctivation **m**apping- (CAM-) based ones and 2) the feature similarity-based ones.

The CAM-based approaches [108], [117], [19], [20] usually adopt the conventional classification network, *e.g.*, the image scene classification. Outwardly, the primary objective of their network training is to achieve a good classification performance. The real purpose is to utilize the classification task to formulate the audio-visual feature embedding. Because the sounding objects' audio signal can significantly contribute to the classification task, we can infer that image regions with strong audio-visual feature responses tend to comprise the sounding object. Following this rationale, the CAM-based SOL methods can directly utilize the feature response map provided at the fusion module's last layer as the SOL result. Since the CAM's computation is fully automatic, the nature of the CAM-based SOL methods is implicit.

Summary of CAM-based SOL. As shown in Fig. 9-D, the greatest strength of CAM-based approaches is the high accuracy because they have additionally used video tags for network training. However, their fully-supervised manner is a clear shortcoming, limiting their broad application.

The feature similarity-based SOL methods [86], [18], [74]

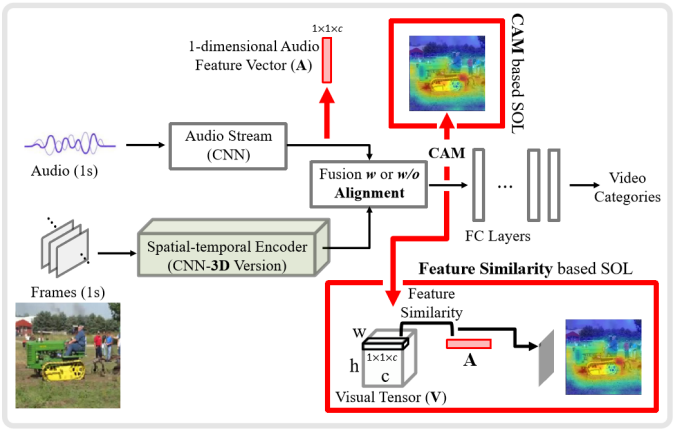


Fig. 14: Demonstration of the CAM-based SOL and the feature similarity-based SOL. The former searches for strong audio-visual feature response to localize the sounding object, while the latter uses the pixel-wise audio-visual feature similarity instead.

are slightly different from the CAM-based ones. Instead of using the implicit manner, this branch of work has adopted the explicit way. That is, after the classifier training, two separate deep feature representations can be derived from the feature backbones' (*e.g.*, Vgg and VggSound) bottom layers, *i.e.*, a deep visual feature (a 3-dimensional tensor) and a deep audio feature (a 1-dimensional vector). Then, because those pixels belonging to the sounding object tend to have a strong audio-visual correlation, the pixel-wise audio-visual feature similarity (*e.g.*, the widely-used Euclidean distance and Cosine similarity) can be applied to locate the sounding pixels. To facilitate a better understanding, we have provided a pictorial demonstration in Fig. 14.

Summary of feature similarity-based SOL. As shown in Fig. 9-E, compared with CAM-based SOL approaches, feature similarity-based methods are video tags free and trained in a semi-supervised way, avoiding disturbances of unfaithful video tags. The weakness is the additional computation of the feature similarity computation.

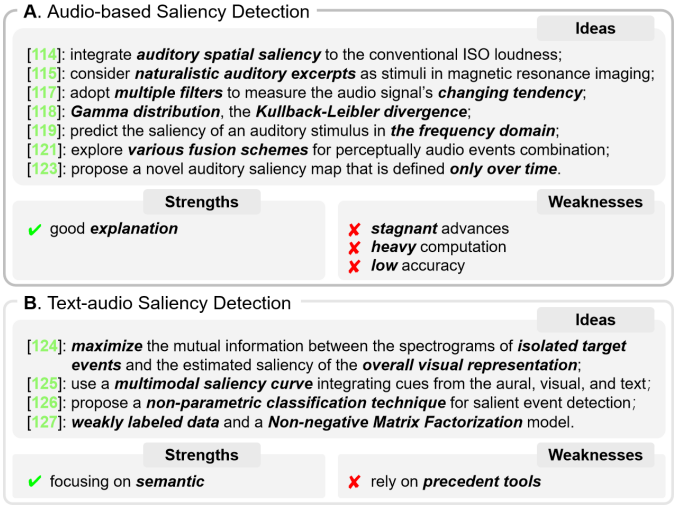


Fig. 15: The overall summary of **A**udio **S**aliency **D**etection (ASD). The former (A) considers audio signals solely, while the latter (B) takes both audio and text signals.

IV. Audio Related Saliency Detection

A. *Audio Saliency Detection (ASD)*

Task Definition. The ASD task is designed to detect drastic changes in audio signals which could attract human attention. Compared with visual saliency, ASD is a relatively easy task because the audio signal is less informative than the visual signal. By considering audio solely, saliency detection can still be performed, *a.k.a.*, audio saliency detection or salient event detection, and there exist multiple works [118], [119], most of which are non-deep learning-based ones, and we shall briefly review them here.

Following the rationale proposed in the earliest Itti's classic work [120] — salient regions should exhibit high contrast to their surroundings, Kayser *et al.* in [121] have investigated the audio saliency detection task. In this work, the authors have adopted multiple filters to measure the audio signal's changing tendency, *i.e.*, the first derivative of intensity and frequency over the time scale. Because, for a short time span, salient audio fragments usually come with a large difference from the rest, their temporal-scale changing tendency can be very effective in evaluating saliency. Following a similar idea, Schauerte *et al.* [122] have adopted the KL-divergence between two audio fragments' 2D spectral histograms. Compared with the previous work [121], which could be regarded as a "local" audio saliency approach, this new work is a non-local one. Also, based on the 2D spectral histogram, Tsuchida *et al.* in [123] have proposed a novel non-local signal feature representation method. For each cell in 2D spectral histogram, the authors have used **p**rincipal **c**omponent **a**nalysis (PCA) to extract the non-local feature. Audio saliency can be obtained based on these features by performing contrast computation over the newly devised feature subspace.

Also, the audio signal's amplitude and frequency are the widely-used computational unit for the salient event detection [124]. Zlatintsi *et al.* in [125] have converted both audio amplitude and frequency to 3D feature via the Teager energy [126]. This work has strongly assumed that people tend to be attracted by sudden loudness. Thus, the authors have directly considered the averaged audio's amplitude, frequency, and newly devised energy to measure the saliency degree. Beyond the amplitude and frequency-based representations, Merve *et al.* in [127] have further devised several novel feature representations (*e.g.*, envelope feature, bandwidth feature, rate feature, pitch feature) for audio signals over the time scale. Finally, this work follows the conventional common thread, *i.e.*, the contrast computation, for each of the newly devised features to obtain multiple bottom audio saliency. The final saliency is achieved by combining them via simple linear fusion.

In-depth Summary. We have summarized the ASD works in Fig. 15-A. Most of the existing ASD methods are explainable because the audio signal's changing tendency fits the human attention mechanism. Generally, the existing methods require heavy computation and achieve low accuracy because hand-crafted methods such as Gamma distribution calculations are less discriminative and time-consuming. Further, the advances toward audio saliency detection are relatively slow, the widely-used methodologies are still limited to the conventional hand-crafted ones, and the deep learning-related researches are quite rare. Considering the importance of audio saliency, this field deserves intense research attention in the near future.

B. *Text-audio Saliency Detection (TASD)*

Task Definition. The TASD task aims at detecting drastic changes over audio signals with the help of text modality, which focuses on the relationship between text information and audio saliency [128], [129] instead of merely audio signals, unlike the above-mentioned audio saliency methods. Existing TASD task methods mainly compute similarity matrices or clustering to measure the consistency between text and audio.

Zlatintsi *et al.* in [130] have fused both text information and audio signal before computing the audio saliency, where the key rationale of the adopted fusion is to calculate the feature similarity (*e.g.*, mutual information) between text and

audio. Another most representative work could be the [131], which has adopted a non-negative matrix factorization model to measure the consistency between text and audio.

In-depth Summary. As shown in Fig. 15-B, the text-audio saliency detection methods mainly focus on semantic information of text and audio, similar to the AVC task. But they rely more on precedent tools, which limits their development.

C. Audio-visual Saliency Detection (AVSD)

Task Definition. The AVSD task is designed to mimic our human attention mechanism in a visual-audio environment because both visual and audio stimuli could cause attention shifting. Thus, the ultimate goal of AVSD is to highlight those video regions which are simultaneously salient in both visual and audio sources.

As the main topic of this review, we shall give a more detailed introduction and discussion of the SOTA audio-visual saliency detection approaches. However, to our knowledge, this topic is definitely in its infancy, and only several deep learning-based works exist. Thus, we take the exact audio-visual fusion scheme as the starting point. Therefore, some related works mentioned in the previous sections might be referenced here for a better understanding.

Hand-crafted Fusions for AVSD. Most of the existing hand-crafted approaches [132], [133] follow the bi-stream structure, which is almost the same as the AVC task reviewed above. Given a video sequence, saliency detection over either the audio or visual channel is computed first. Then the audio-visual saliency can be derived by designing an appropriate fusion scheme. Any off-the-shelf audio/visual saliency detection methods can be used directly, making the exact fusion scheme the key to the overall performance.

Many works [134], [135], [136], [137] have adopted the multiplicative-based fusion because it can effectively enhance the consistency and compress the inconsistency between audio and visual saliency-related features. After all, those real salient regions tend to be salient in both the audio domain and visual domain simultaneously. The limitation of the multiplicative-based fusion is also quite clear — it tends to get confused if there exist multiple visual and saliency features. In cases with multiple audio and visual features, Coutrot *et al.* in [138] have adopted the linear fusion, where the fusion weights are computed via the classic **e**xpectation-**m**aximization (EM) algorithm, a statistical method using training samples to estimate the relative importance of each feature aiming to maximize the global likelihood of the mixture model. Further, Sidaty *et al.* in [139] have conducted an extensive evaluation regarding different fusion schemes, including maximum, addition, average, multiplication, and non-linear combination-based fusion schemes. As expected, all such simple fusions are inferior to the non-linear fusion because, in most cases, the audio and visual saliency could have different contributions to the final audio-visual saliency. The given video scene and content usually determine the exact contribution degree. Yet, these naive fusion schemes are not flexible enough, failing to achieve the optimal balance between audio and visual.

Also, some works [140], [141], [142] have adopted correlation analysis fusions for AVSD. From the experiment perspective, Min *et al.* in [140] have conducted extensive verifications of the human eye fixations in conditions with and without audio signals. Their results indicate that audio signals can significantly influence human attention only if the salient object is visually non-salient yet salient in the audio channel; otherwise, the audio information is completely helpless. This work also inspires us that an audio-visual saliency detection method should bias more towards visual signals in most cases. Following the same rationale, Min *et al.* in [141] have adopted the classic **c**anonical **c**orrelation **a**nalysis (CCA) to localize spatial regions which have demonstrated strong audio-visual consistency. Since the audio and visual saliency cues have been computed, the fusion process mainly targets highlighting the visual regions correlated well to the audio. More recently, Min *et al.* in [142] have further considered the deep learning-based saliency cues. And the CCA has been replaced by its upgraded variant — the **k**ernel **c**anonical **c**orrelation **a**nalysis (KCCA), to measure the audio-visual correlation. The main reason is that the CCA can only correlate linear relationships. At the same time, the KCCA can map features to higher-dimensional feature spaces and increase the nonlinearity, which could be more practical in the audio-visual saliency detection task.

To further explore the advantages and disadvantages of the existing fusion schemes, Tsiami *et al.* in [143] have compared three widely-used audio-visual fusion schemes, *e.g.*, direct fusion (*i.e.*, the multiplicative-based fusion), linear correlation coefficient [144], and mutual information [145]. The authors have combined the existing visual saliency models with the off-the-shelf audio saliency models by using one of these fusion schemes alternatively. The quantitative results have reached a clear conclusion, *i.e.*, the exact optimal fusion scheme is determined by multiple factors, including the quality of low-level saliency cues and the input video data. For "raw" hand-crafted saliency cues computed by models which are good at measuring saliency from the temporal scale, the correlation coefficient could be the best choice since it mainly considers the temporal consistency between audio and visual. As for the case where the raw saliency cues have been incorporated with spatial information, mutual information could be the optimal choice. However, things could be changed for those "refined" saliency cues — saliency cues obtained via deep learning-based top-down models, where the direct fusion usually exhibits the best fusion performance because the refined saliency cues are generally more trustworthy than those raw ones. Thus they could be directly used to complement their counterparts.

Summary of Hand-crafted Fusions. As shown in Fig. 16-A, the existing hand-crafted AVSD methods are very easy to implement. However, such methods have a critical limitation, *i.e.*, the adopted hand-crafted fusions cannot well handle the complex complementary relationships between visual and audio signals due to the limited flexibility. These methods also need huge computation to obtain hand-crafted features.

SOTA Deep Learning-based AVSD Methods. After en-

A. Hand-crafted Fusions

Ideas

[132]: provide an ***overview*** of audiovisual saliency map models;
[133]: predict gaze maps by fusing ***spatial***, ***temporal*** and ***auditory attentional maps***;
[134]: ***first*** devise a ***lightweight*** audio-visual saliency model by ***deep CCA***;
[135]: ***new*** AVSD ***dataset***, ***temporal attention mechanism***, and ***multi-scale AV fusion***;
[136]: insert ***360° spatial-temporal visual learning*** and ***spatial auditory*** in ODVs;
[137]: use ***event-based visual sensors*** and ***microphones on a robot*** for AVSD;
[138]: employ ***speaker diarization algorithm*** to predict the eye movements;
[139]: testify simple fusions of audio and video are ***inferior*** to non-linear fusions in AVSD;
[140]: illustrate in ***what situation*** audio will ***influence*** visual attention;
[141]: adopt ***CCA*** to localize ***spatial regions*** with strong audio-visual ***consistency***;
[142]: repalce CCA with ***KCCA*** to map features to ***higher-dimensional*** feature spaces.

Strengths

✔ ***easy*** to implement

Weaknesses

✘ relying on ***hand-crafted*** fusions
✘ ***less*** interactive of spatial and temporal streams

B. Deep learning-based -- Plain Fusion

Ideas

[146]: utilize ***plain concatenation*** operations to fuse audio and video features;
[149]: use a ***multisensory framework*** of audio-visual signals for AVSD.

Strengths

✔ ***simple*** to implement
✔ good ***compatibility***

Weaknesses

✘ ***inferior*** performance
✘ ***lost*** in spatial Information
✘ ***limited interaction*** between visual and audio

C. Deep learning-based -- Spatial Alignment

Ideas

[73]: use ***de-convolution*** for the audio-visual alignment in a ***coarse-to-fine*** strategy.

Strengths

✔ retaining ***spatial information***

Weaknesses

✘ ***weak interaction*** between visual and audio

D. Deep learning-based -- Bi-linear Fusion

Ideas

[147]: use ***bilinear fusion*** to handle the ***dimension mismatched*** problem;
[148]: adopt ***trilinear interpolation*** and ***3D convolutions*** as decoder modules.

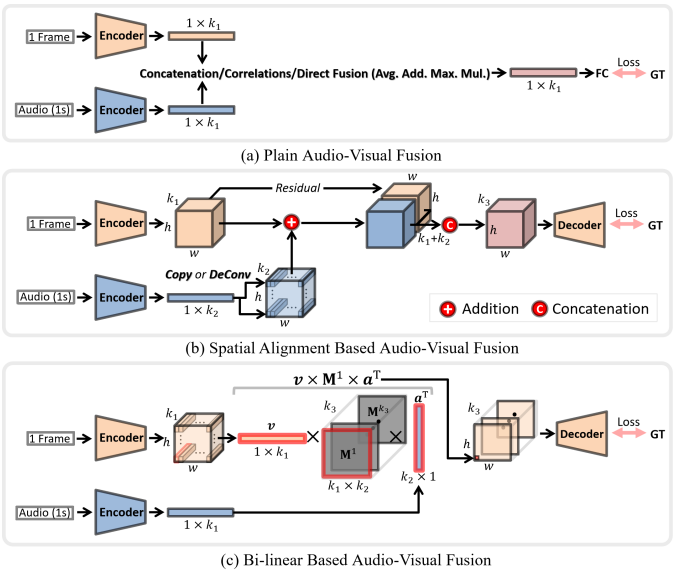
Strengths

✔ ***strong interaction*** between visual and audio
✔ retaining intact ***spatial information***

Weaknesses

✘ ***NULL***

Fig. 16: The overall summary of **A**udio-**v**isual **S**aliency **D**etection (AVSD), which is the main topic of this review. Details of deep-learning-based methods are demonstrated in Fig. 17.



(a) Plain Audio-Visual Fusion

(b) Spatial Alignment Based Audio-Visual Fusion

(c) Bi-linear Based Audio-Visual Fusion

Fig. 17: The most representation fusion schemes for audio-visual saliency detection. Among them, (a) merely utilizes the conventional plain concatenation operations to integrate audio and visual features; (b) treats the audio part as auxiliary information, and the embedded semantical consistency is used to highlight the corresponding spatial regions; (c) adopts a dimension transformation matrix to handle the dimension mismatched problem, which doesn't require the identical dimension size of the individual audio and visual saliency cues.

tering the deep learning era, massive deep learning-based visual saliency models have been proposed. However, to the best of our knowledge, there only exist five deep learning-based audio-visual saliency detection models [146], [147], [148], [75], [149]. Here we shall provide a detailed review of these works respectively. For a better understanding, we have provided multiple method pipelines to clarify the audio-visual fusion methodology regarding these SOTA deep learning-based audio-visual saliency detection works. As can be seen in Fig. 17, all three sub-figures respectively correlate to the SOTA models mentioned above: sub-figure (a) [146], [149], sub-figure (b) [75], and sub-figure (c) [148], [147].

We shall first introduce the [146], [149]. As shown in Fig. 17 (a), the audio-visual fusion adopts the conventional plain concatenation operations, which takes both audio and visual feature tensors as input, and the saliency predictions are obtained via a typical decoder after concatenating both audio and visual tensors. Specifically, because the audio modality has a completely different formation from the visual modality, it is required to ensure that the audio's tensor feature has the same size as its visual counterpart. The overall method rationale of this work is very straightforward, and other existing ones could replace the concatenation-based fusion, *e.g.*, direct fusions, and correlation analysis tools, where similar works have been widely adopted by the AVC task, which have been reviewed in Sec. III-B.

Summary of Deep Learning-based Plain Fusion. As shown in Fig. 16-B, the advantage of plain fusions in the deep learning era is easy to implement with good compatibility. However, the disadvantage is also clear, *i.e.*, inferior performance compared to other methods using more fancy fusion logics (*e.g.*, bi-linear fusion). That is, plain fusions (*e.g.*, vector-based feature concatenation) could lead to a loss in spatial information, resulting in limited interaction between visual and audio.

As illustrated in Fig. 17 (b), the spatial alignment-based audio-visual fusion can bias the fusion toward the visual part, where the deep audio feature, which usually is a 1-dimensional vector with the same size as the visual tensor's channel number, is either de-convolved or copied to correlate to each spatial location. This implementation has treated the audio as auxiliary information, where the embedded semantical consistency is the key factor in highlighting the corresponding spatial regions as the salient ones. The AVC task has widely used the "copy" scheme. However, to our knowledge, [75] is the first attempt to use de-convolution for the audio-visual alignment. Also, either the copy or the de-convolution-based alignment can be combined with the popular "residual" operation to focus the fusion process on the visual signal because, in most cases, the visual signal is stronger in determining human attention than the audio signal.

Summary of Deep Learning-based Spatial Alignment. As shown in Fig. 16-C, compared with the plain fusion mentioned

above, the existing spatial alignment-based AVSD methods (see Fig. 17-b) can well retain overall spatial structure by spanning a dummy audio feature tensor with identical size to the visual counterpart. However, the interaction between visual and audio is still very weak in such methods.

Lastly, as demonstrated in Fig. 17 (c), we introduce the bi-linear audio-visual fusion, which has been adopted by [147], [148] and achieved the leading SOTA performance. Compared with either the plain or spatial alignment-based fusion, the bi-linear fusion has one significant advantage: it doesn't require the individual audio and visual saliency cues to have an identical dimension size, where a dimension transformation matrix, *i.e.*, see the **M** in the sub-figure C, is adopted to handle the dimension mismatched problem. The bi-linear fusion also has its own limitation, *i.e.*, the semantical correspondence between audio and visual channels has been destroyed, making modeling complex audio and visual interactivity very difficult. In sharp contrast, the spatial alignment-based fusion could make full use of the semantical information provided by either off-the-shelf visual (*e.g.*, ResNet50) or audio (*e.g.*, VggSound) feature backbone, where the learned semantical information could shrink the problem domain effectively. As a result, the audio-visual complementary fusion status could be easily reached even in a complex audio-visual environment.

Summary of Deep Learning-based Bi-linear Fusion. As shown in Fig. 16-D, the bi-linear fusion achieves the best performance since it can retain the overall spatial structure well and enable strong visual-audio interaction. In a word, the deep learning-based bi-linear fusion can simultaneously perform visual-audio embedding and aligning without much additional computation cost. And bi-linear fusion should be a common thread for multi-modality feature fusion in this deep learning era.

V. Audio-visual Semantical Consistency Perceptual

A. Preliminary

Existing **a**udio-**v**isual **s**aliency **d**etection (AVSD) works mainly adopt bi-stream network architecture, where audio saliency and visual saliency are computed individually and combined later as the final output. When the audio signal is inconsistent with the visual signal, the audio saliency is completely helpless to complement the visual saliency, which takes up about 60% of all cases. For example, in an image, two persons are talking. At the same time, the background music comes from the outside; in this case, the audio signal cannot benefit the visual in determining saliency.

Inspired by previous multimedia related works [23], [24], [25], we propose to introduce the "**a**udio-**v**isual **c**onsistency (AVC)" into our saliency detection research field. The major highlight of our approach is its generic usage, which can upgrade any SOTA bi-stream-based AVSD model from "AVC-unaware" to "AVC-aware". The optimal audio-visual fusion is very difficult to achieve if the adopted AVSD model is AVC-unaware because the model is completely blind and thus cannot completely omit the audio when the audio is not corresponding to the visual. Thus, in facing weak audio signals, an AVSD model taking both audio and visual is inferior to the model using the visual solely, yet this "binary switch" cannot be achieved if the model is AVC-unaware.

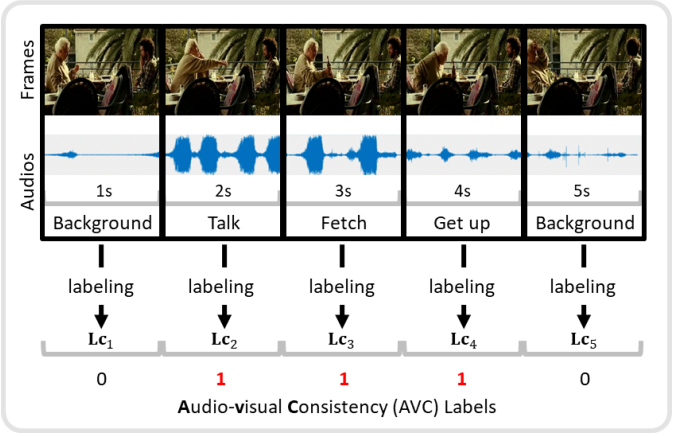


Fig. 18: Detailed demonstration of our AVC annotation. In the selected video clip, when two men are talking, the sounds of 1st and 5th seconds are the background music, whose AVC labels are set to 0, meaning that the audio and the visual are semantically mismatched. While, from the 2nd to the 4th seconds, the audio signals are talking sound, fetching sound, and getting up sound, respectively, and thus we label them as 1 because these audio-visual fragments are clearly matched.

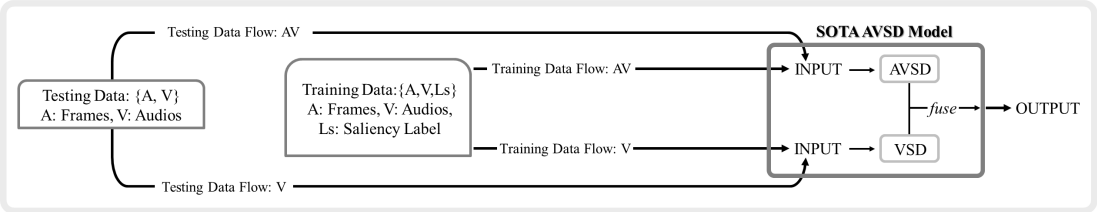
An intuitive way to convert an AVC-unaware AVSD model to AVC-aware is to resort to an additional module that can automatically predict whether the currently given audio is consistent with the visual. Therefore, to fully realize our idea — making any existing bi-stream AVSD model AVC-aware, two things should be prepared in advance: 1) train the aforementioned classifier, and 2) integrate the classifier into the AVSD model. Next, we shall respectively detail each of them in the following subsections.

B. Audio-visual Consistency Labeling

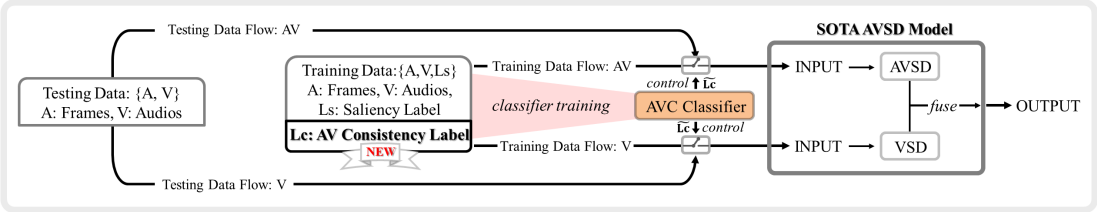
The AVC classifier can, of course, be trained via the above-mentioned weakly-supervised method, shown in Fig. 11 (b). However, the overall performance of this method is usually too limited to benefit the saliency detection task. Thus, we propose to utilize the fully-supervised method to train the audio-visual consistency classifier.

We shall manually equip each video frame with AVC labels to achieve this goal. We manually provide all the existing benchmark AVSD datasets with binary AVC labels, and a representative pictorial demonstration has been shown in Fig. 18. Thus, each audio-visual fragment will be assigned to 1 or 0 labels accordingly. Suppose all existing training instances (with N frames) can be represented as: $\{\mathrm{A_i}, \mathrm{V_i}, \mathrm{Ls_i}\}$, where $i \leq N$, A, and V respectively denote the audio and visual, and Ls is the corresponding fixation map. During the annotation process, if the audio sound is made by the salient object[3], we regard that the audio and visual are consistent. Thus we assign the AVC label as 1. Otherwise, if the audio is unseen background music or off-screen sound, the AVC label of this audio-visual fragment is set as 0. For a better understanding,

[3]We manually regard an object as salient if it has the highest fixation number in the scene.



(a) Conventional AVSD Training and Testing



(b) Our Audio-visual Consistency-aware AVSD Training and Testing

Fig. 19: Demonstrations of the differences between the conventional audio-visual saliency detection model training/testing pipeline (a) and the newly modified training/testing pipeline (b). The advocated AVC classifier can be trained by the newly annotated AVC labels and then dynamically control the data flow of the adopted bi-stream SOTA AVSD model. The $\tilde{\mathbf{Lc}}$ denotes the binary output of the AVC classifier. By equipping the existing SOTA AVSD model with the AVC classifier, we can make the original AVC-unaware AVSD model AVC-aware, achieving persistent performance improvement.

we have provided a pictorial demonstration regarding how to perform the proposed AVC labeling process, which can be found in Fig. 18.

After the annotation process[4], each training instance can be converted to:

$$\{\mathrm{A_i}, \mathrm{V_i}, \mathrm{Ls_i}\} \rightarrow \{\mathrm{A_i}, \mathrm{V_i}, \mathrm{Ls_i}, \mathbf{Lc_i}\}, \quad \mathbf{Lc_i} \in \{0, 1\}, \mathrm{i} \in [1, N], \quad (1)$$

where **Lc** denotes the newly annotated audio-visual consistency label, and N denotes the total frame number. We have newly annotated all publicly available AVSD benchmarks, totally 5 sets (or 6 if the Coutrot set is divided into Coutrot1 set and Coutrot2 set) consisting of 241 video clips involving 300,000 frames. These newly annotated datasets are now publicly available[5].

C. The Proposed AVC-aware AVSD Model

The conventional **a**udio-**v**isual **s**aliency **d**etection (AVSD) training and testing protocol has been shown in Fig. 19 (a), where the AVSD model is a typical bi-stream fusion net, which combines its AVSD and **v**isual **s**aliency **d**etection (VSD) to formulate the final result. The VSD stream is the mainstream, and the AVSD stream is the auxiliary stream, where audio and visual are fused early via fusion schemes mentioned in Fig. 17, to promote the VSD stream further. As we have mentioned, this typical AVSD training and testing protocol are completely AVC-unaware. The later fusion (*i.e.*, fuse VSD with AVSD) could even degenerate the overall performance when the given audio and visual are mismatched.

To handle the above-mentioned problem, we propose the AVC-aware training and testing protocol, which has been shown in Fig. 19 (b), whose major difference to (a) is the newly provided AVC classifier, and this classifier can be trained by using the newly equipped AVC labels. In our implementation, we use an identical classifier structure to AVID [89] to automatically predict the AVC degree of the current input audio-visual fragment, outputting 0 or 1. Notice that other classifier structures can also be used, and we have tested several others, where the quantitative result (Table V) suggests that the AVID is the best choice.

As shown in Fig. 19 (b), the newly proposed AVSD model can be trained in the typical end-to-end way, where the AVC classifier serves the existing SOTA bi-stream AVSD model, *i.e.*, Fig. 19 (b), as "binary switchers" to control the INPUT of the adopted SOTA AVSD model. In other words, the output of the AVC classifier determines whether or not the single V flow or both V and AV flows are to be used in the subsequent SOTA AVSD model. That is when the output of the AVC classifier is 0, which means the current audio is inconsistent with the current visual, suggesting removing the AV flow from fusing it with V flow because, for an inconsistent audio-visual fragment, the output of AV flow tends to significantly inferior to the output of V flow, thus fusing AV with V MUST degenerate the overall performance. When the output of the AVC classifier is 1, the whole training process is completely identical to the original SOTA AVSD model, where both AV flow and V flow are simultaneously considered. The entire data flow of our AVC-aware AVSD model can be expressed as:

$$\begin{aligned} \mathrm{OUTPUT} &\leftarrow \tilde{\mathbf{Lc}} \cdot \mathrm{Fuse}(\mathrm{AV}, \mathrm{V}) + (1 - \tilde{\mathbf{Lc}}) \cdot \mathrm{V}, \\ \tilde{\mathbf{Lc}} &= \mathrm{AVC_{cls}}(\mathrm{AV}, \mathrm{V}) \in \{0, 1\}, \end{aligned} \quad (2)$$

where $\mathrm{AVC_{cls}}$ represents the AVC classifier, and $\tilde{\mathbf{Lc}}$ is the binary prediction regarding AVC of the current input V and AV.

The training process of our AVC-aware AVSD model consists of two tasks, *i.e.*, 1) the conventional audio-visual saliency detection task, which takes the saliency labels (Ls) as GT, and

[4]To match the fps of video clips (25~30) and the audio length, we resort to Adobe Premiere CC, a professional video editing software, to align the mismatched audio and visual durations.

[5]https://github.com/MengkeSong/SCDL

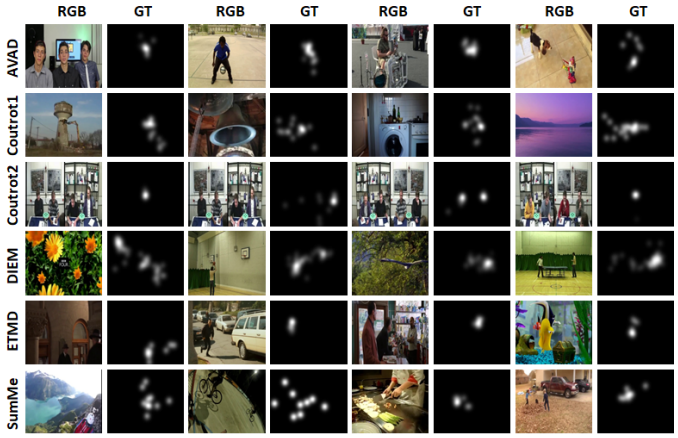


Fig. 20: Demonstration of the differences regarding the scene contents of six wide-used datasets of AVAD, Coutrot1, Coutrot1, DIEM, ETMD, and SumMe.

2) the newly added AVC classifier training, which takes the AVC labels (Lc) as GT. Thus, the overall loss function L_{all} can be detailed as:

$$L_{all} = (1-\rho)\cdot L_{cls} + \rho \cdot L_{avsd}, \tag{3}$$

where L_{cls} is a typical cross-entropy loss targeting the training of AVC classifier, L_{avsd} is the **K**ullback-**L**eibler (KL) divergence loss, the most widely-used loss function in AVSD model training, and ρ is a balancing factor which we empirically assign it to 0.5.

In the testing phase, the exact data flows are dynamically controlled by the AVC classifier, identical to the training phase.

In brief, the major highlight of our approach is its generic design, which can serve any existing bi-stream SOTA AVSD models as the plug-in and promote their performances persistently. Though a more fancy network design could bring additional performance gain, we shall leave it to future work to stay the main focus of our topic.

VI. Quantitative Verifications

A. Datasets

There exist six publicly available datasets in our AVSD research field, including DIEM [150], AVAD [141], Coutrot1 [151], Coutrot2 [152], SumMe [153], and ETMD [154]. Different from the conventional VSD sets, the eye fixations in these six sets are collected in the audio-visual environment. In contrast, in the VSD sets, the eye fixations are simply collected without audio information. We briefly introduce these six sets here, and more details can be found via the links of Table III. Some qualitative demonstrations can be found in Fig. 20.

The DIEM set consists of 84 film clips, covering 26 films, including commercials, documentaries, game trailers, movie trailers, music videos, and news clips. The video scenes in this set are generally complex with strong background interference.

The AVAD set targets at exploring the effects of the highly correlated audio and motion on eye movements. The authors of this set tested the human eyes fixation on 45 video sequences. These tested sequences are 5 to 10-second video clips containing various scenes, *e.g.*, instrumental playing, dancing, and dialogue.

The Coutrot set includes Coutrot1 and Coutrot2 subsets. The dynamic nature scenes in the Coutrot1 set are divided into 4 visual categories: single moving objects, multiple moving objects, natural landscapes, and human faces. The Coutort2 set's scenes are all conversations, and it can be found that the fixations are most likely to be located on the speaker's face.

The SumMe set contains 25 unstructured videos collected from videos taken by users, whose lengths range from 1 minute to 6 minutes. Since all videos in this set are homemade, the corresponding background sounds tend to be very noisy, and most of them are irrelevant to the salient objects, making the audio-visual fusion process very challenging.

The ETMD set contains 12 videos, which are all collected from 6 existing Hollywood movies. Each video in this set ranges from 3 to 3.5 minutes, whose contents mainly consist of action scenes and dialogues.

B. Evaluation Metrics

Five quantitative metrics have been widely used in the saliency detection field. Since the objective of measuring the saliency detection performance in an audio-visual environment is almost the same as the conventional saliency detection field, all these five metrics can be directly used here, and we shall briefly introduce them. These metrics include 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).

CC is a method to measure the linear correlation between the prediction saliency (S) and the ground truth (GT), which can be formulated as:

$$\mathrm{CC}(\mathrm{S},\mathrm{GT}) = \frac{cov(\mathrm{S},\mathrm{GT})}{\sigma(\mathrm{S})\cdot\sigma(\mathrm{GT})}, \tag{4}$$

where *cov* denotes the covariance, and σ is the standard deviation.

SIM measures the similarity between two distributions. Given S and GT as input, SIM first normalizes them respectively, then measures the minimum values pixel-by-pixel (denoted by i). This process can be detailed as:

$$\mathrm{SIM} = \sum_i min\Big\{\mathcal{Z}(\mathrm{S})_i, \mathcal{Z}(\mathrm{GT})_i\Big\}, \tag{5}$$

where $\mathcal{Z}$ and *min* respectively denote the normalization operation and minimum operation.

AUC measures the area under the **r**eceiver **o**perating **c**haracteristic (ROC) curve, which has been widely used to evaluate the maps by saliency models. Given an image and

TABLE III: Details of the existing AVSD sets.

Datasets	Year	Videos	Viewers	Frames	Links
DIEM [150]	2010	84	42	78,167	[Link]
AVAD [141]	2016	45	16	9,564	[Link]
Coutrot1 [151]	2013	60	72	25,223	[Link]
Coutrot2 [152]	2014	15	40	17,134	[Link]
SumMe [153]	2019	25	10	109,788	[Link]
ETMD [154]	2019	12	10	52,744	[Link]

TABLE IV: Quantitative comparisons between our method with other fully-/weakly-/un-supervised methods on all 6 datasets. The best result is marked in **bold** font. * means that the target models (*e.g.*, STANet*, STAViS*, and AVINet*) are trained by the whole pipeline in Fig. 19 with AVC classifier; # denotes that the target models (*i.e.*,, STANet#, STAViS#, and AVINet#) are trained by removing the AV classifier model, and their OUTPUTs are manually reformulated by using **Lc** (the newly annotated binary labels, see Eq. 7) as the indicator, which represents the ideal situation. 'A.': AUC-J, 'S.': SIM, 's.': s-AUC, 'C.': CC, 'N.': NSS, 'Un-s.': Un-supervised.

Means	Datasets	AVAD [141]					DIEM [150]					SumMe [153]					ETMD [154]					Coutrot1 [151]					Coutrot2 [152]				
	Methods	A.↑	S.↑	s.↑	C.↑	N.↑	A.↑	S.↑	s.↑	C.↑	N.↑	A.↑	S.↑	s.↑	C.↑	N.↑	A.↑	S.↑	s.↑	C.↑	N.↑	A.↑	S.↑	s.↑	C.↑	N.↑	A.↑	S.↑	s.↑	C.↑	N.↑
Un-s.	ITTI [155]	.688	.170	.533	.131	.611	.663	.217	.583	.137	.555	.666	.151	.559	.097	.436	.661	.127	.582	.083	.425	.616	.178	.529	.082	.319	.694	.142	.530	.040	.331
	GBVS [156]	**.854**	.247	.572	**.337**	**1.556**	**.830**	**.318**	.605	**.356**	**1.277**	**.808**	.221	.567	**.272**	**1.134**	**.856**	.226	.613	**.299**	**1.398**	**.798**	**.253**	.526	**.272**	**1.055**	.819	**.189**	.577	**.183**	1.071
	SBF [157]	.833	**.272**	.576	.308	1.489	.759	.292	.608	.301	1.081	.783	**.228**	.590	.230	1.023	.805	**.232**	.641	.262	1.298	.726	.187	.530	.215	.789	**.827**	.152	.583	.131	**1.101**
	AWS-D [158]	.825	.221	**.589**	.304	1.378	.733	.250	**.612**	.301	1.128	.747	.192	**.603**	.186	.853	.754	.161	**.664**	.181	.907	.729	.214	**.581**	.207	.872	.783	.170	**.590**	.146	.842
Weakly-supervised	GradCAM++ [159]	.777	.273	.559	.255	1.217	.732	.216	.583	.271	.778	.774	.217	.593	.225	.924	.575	.124	.157	.576	.736	.704	.137	.537	.210	.511	.733	.114	.567	.168	.625
	WSS [160]	.858	.292	**.592**	.347	1.655	.803	.333	.620	.344	1.293	.812	.245	.589	.279	1.098	.854	.277	.661	.334	1.650	.772	.247	.547	.233	.975	.835	.208	.578	.192	1.178
	MWS [161]	.834	.272	.573	.309	1.477	.806	.336	.628	.350	1.308	.808	.237	.607	.258	1.155	.833	.237	.649	.293	1.425	.743	.231	.528	.201	.798	.839	.188	.581	.168	1.197
	STANet [75]	.873	.334	.580	.438	2.018	.861	.391	.658	.469	1.716	.854	.294	.627	.368	1.647	.908	.318	.682	.448	2.176	.829	.306	.542	.339	1.376	.850	.247	.597	.273	1.475
	STANet*	.879	**.341**	.584	.439	2.068	.891	**.392**	.662	**.498**	2.016	.870	.323	.631	.382	1.662	.922	.319	.701	.464	2.326	.837	.315	.550	.341	**1.394**	.887	.264	.602	.336	1.915
	STANet#	**.881**	**.341**	**.585**	**.442**	**2.070**	**.892**	.390	**.665**	**.498**	**2.019**	**.873**	**.325**	**.632**	**.384**	**1.663**	**.925**	**.323**	**.704**	**.467**	**2.328**	**.840**	**.318**	**.551**	**.346**	1.392	**.888**	**.266**	**.605**	**.339**	**1.921**
Fully-supervised	DeepVS [62]	.896	.391	.585	.528	3.010	.840	.392	.625	.452	1.860	.842	.262	.612	.317	1.620	.904	.349	.686	.461	2.480	.830	.317	.561	.359	1.770	.925	.259	.646	.449	3.790
	ACLNet [58]	.905	.446	.560	.580	3.170	.869	.427	.622	.522	2.020	.868	.296	.609	.379	1.790	.915	.329	.675	.477	2.360	.850	.361	.542	.425	1.920	.926	.322	.594	.448	3.160
	STAViS [147]	.919	.457	.593	.608	3.180	.883	.482	.674	.579	2.260	.888	.337	.656	.422	2.040	.931	.425	.731	.569	2.940	.868	.393	.584	.472	2.110	.958	.511	.710	.734	5.280
	STAViS*	.925	.460	.599	.623	3.252	.896	.484	.683	.582	2.499	.903	.393	.634	.460	2.102	.946	.454	.758	.620	3.406	.864	.398	.590	.487	2.203	.959	.523	.731	.738	5.396
	STAViS#	.927	.463	.597	.622	3.255	.899	.485	.684	.581	2.497	.904	.397	.635	.463	2.107	**.948**	**.457**	.764	**.623**	3.401	.869	.395	.592	.489	2.210	.963	.524	.732	.739	5.401
	AViNet [148]	.931	.499	.663	**.689**	3.740	.901	.504	.722	.637	2.540	.900	.350	.697	.470	2.420	.931	.410	.740	.576	3.070	.891	.427	.638	.561	2.710	.953	.477	.739	.738	5.730
	AViNet*	.932	.509	.691	.678	3.756	.905	.516	.786	**.645**	**2.637**	.909	.400	.699	.491	2.529	.944	.447	.761	.616	3.437	**.899**	**.431**	.644	.573	2.770	.963	.579	.742	.806	**5.993**
	AViNet#	**.936**	**.511**	**.694**	.679	**3.759**	**.906**	**.518**	**.797**	.643	2.634	**.913**	**.405**	**.702**	**.492**	**2.535**	.946	.448	**.765**	.617	**3.441**	.894	.428	**.647**	**.579**	**2.774**	**.967**	**.581**	**.746**	**.809**	5.990

its ground-truth eye fixations, the fixated points are regarded as the positive set, and others are regarded as the negative set. Then, the computed saliency map is binarized into salient and non-salient regions using a hard threshold. The AUC-Judd (AUC-J) computes two items: 1) the true positives from all the saliency map values above a threshold at fixated pixels and 2) the false positive rate as the total saliency map values above a threshold at non-fixated pixels. The s-AUC samples the negatives from fixated locations of other images/frames. This sampling scheme can be greatly influenced by center bias and border cuts.

The NSS is designed to evaluate a saliency map over fixation locations. Given a saliency map S and a binary fixation map GT, NSS is defined as:

$$\mathrm{NSS} = \frac{1}{\mathrm{M}} \sum_{i}^{\mathrm{M}} \hat{\mathrm{S}}_i \cdot \mathrm{GT}_i, \quad \mathrm{M} \leftarrow \sum_i \mathrm{GT}_i, \quad \hat{\mathrm{S}} \leftarrow \frac{\mathrm{S} - \mu}{\sigma}, \tag{6}$$

where μ and σ are the mean and standard deviation of the predicted saliency map. This metric is calculated by taking the mean scores assigned by the unit normalized saliency map (with zero mean and unit standard deviation) at human eye fixations.

C. Quantitative Evidences towards the Effectiveness of the proposed AVC Classifier

As we have mentioned, our approach is generic and compatible with almost all existing bi-stream SOTA AVSD models. The proposed AVC classifier can be intergraded into the target model using a few code modifications. To verify this issue, we have tried to deploy our AVC classifier into 3-top tier SOTA AVSD models, including STANet [75], STAVIS [147], and AVINet [148]. We shall incorporate our AVC classifier into more SOTA models, yet, in the AVSD research field, most of the existing papers haven't released their codes. Also, *w.r.t.* the model training, we follow the widely-used training/testing split [147] over all 6 datasets. To demonstrate the superiority of our approach, we have compared the upgraded versions of the three target models (denoted by *) with 12 other SOTA methods, including 4 unsupervised methods, 4 weakly-supervised methods, and 4 fully-supervised methods. For a fair comparison, we use either the code implementations with default parameter settings or saliency maps provided by the authors. Specifically, we refer to the numeric results reported in the papers for others without codes.

As is shown in Table IV, all three upgraded target models (denoted by * highlighted by PINK color) can achieve persistent performance improvements. For example, our method can make an average of 1.9%, 1.5%, and 2.7% performance improvement generally of STANet, STAVIS, and AVINet, respectively, in terms of the AUC-J metric on six widely-used benchmark datasets. Also, the promoted model STANet* outperforms all weakly-supervised methods significantly, and AVINet* performs the best among all fully-supervised methods. The reason is that the AVSD benchmark datasets equipped with the newly proposed AVC classifier can filter out the unrelated audio-visual pairs so that the side effects from those mismatched audio-visual fragments can be avoided.

To further investigate the importance of our key idea, *i.e.*, the audio-visual consistency matters when performing AVSD, we have removed the proposed AVC classifier from the upgraded target AVSD models. Instead, we directly use the original versions, yet their outputs are "manually reformulated" according to our newly provided AVC labels (*i.e.*, **Lc** in Fig. 19). That is, the target model's output will be derived directly by using either AV or V, and this process can be formulated as:

$$\mathrm{OUTPUT} \leftarrow \mathbf{Lc} \cdot \mathrm{Fuse}(\mathrm{AV}, \mathrm{V}) + (1 - \mathbf{Lc}) \cdot \mathrm{V}. \tag{7}$$

where all symbols are identical to Eq. 3, and the major difference is that the $\tilde{\mathbf{Lc}}$ has been replaced by **Lc**. Actually,

TABLE V: Ablation study regarding different AVC classifiers, *e.g.*, L3Net, AVENet, and AVID. The target AVSD model used here is AVINet [148]. AVID+*ws* denotes that the AVID-based AVC classifier is trained in a weakly-supervised manner, the same as [86]. The bests are highlighted in **bold** font.

Datasets / Methods	Accuracy	AVAD [141]					DIEM [150]					SumMe [153]				
		AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
L3Net (*ours*)	82.15%	0.928	0.505	0.682	0.674	3.750	0.902	0.514	0.768	0.639	2.605	0.902	0.377	0.698	0.483	2.489
AVENet (*ours*)	85.64%	0.929	0.507	0.685	0.677	3.753	0.903	0.511	0.779	0.642	2.617	0.907	0.392	**0.699**	0.488	2.510
AVID + *ws* [89]	85.82%	0.915	0.487	0.667	0.658	3.652	0.891	0.493	0.755	0.628	2.589	0.894	0.369	0.680	0.473	2.474
AVID (*ours*)	**87.59%**	**0.932**	**0.509**	**0.691**	**0.678**	**3.756**	**0.905**	**0.516**	**0.786**	**0.645**	**2.637**	**0.909**	**0.400**	**0.699**	**0.491**	**2.529**

OUTPUT from Eq. 7 is in ideal situation, which tends to persistently outperform that from the upgraded version powered by the AVC classifier (*i.e.*, Eq. 3). The main reason is clear: our AVC classifier can not completely avoid erroneous binary predictions. The quantitative results of these ideal versions have been marked by # with BLUE background color, and the detailed results can be found in Table IV.

Further, as mentioned above, the classification accuracy of the AVC classifier will affect AVSD performance slightly. Thus, we have tested three AVC classifiers to verify this issue, *i.e.*, L3Net [86], AVENet [18], and AVID [89]. The AVID is our default setting, and the other two classifiers can be used to replace the AVID in our method, as shown in Fig. 19 (b). That is, in each experiment, we only replace the target AVSD model's AVC classifier with either L3Net, AVENet, or AVID. The experimental results have been shown in Table V.

The influence of the classification result is based on the amount of corresponding audio-visual pairs, *e.g.*, the more the corresponding audio-visual pairs are, the better the performance of the target models obtain; otherwise, the target models will degenerate into the original versions. According to the results, the AVID-based AVC classifier has achieved the best accuracy (*i.e.*, 87.59%), and thus, as expected, the corresponding AVSD performance outperforms others. In short, the higher the performance of the AVC classifier is, the better the performance of the target models obtains.

VII. Conclusions and Future Work

This paper presents the first comprehensive review covering both topics ranging from saliency detection to audio-visual fusion. Based on this extensive review, we also provided a deep insight into the audio-visual saliency detection task and reached our new claim about the importance of an AVSD model to be audio-visual consistency aware (AVC-aware). We have also devised a generic method to convert the existing AVC-unaware SOTA AVSD models to be AVC-aware. The key is the newly proposed AVC classifier, which controls the data as a plug-in flow of the bi-steam target AVSD mode to avoid side effects caused by mismatched audio-visual training fragments. Specifically, to train the proposed AVC classifier, we have newly labeled all existing publicly available AVSD datasets, equipping them with AVC labels. Lastly, we have conducted extensive experiments to verify the effectiveness of our claim. Hoping this review could draw more research attention to the AVSD research field, and the newly claimed AVC-aware issue could inspire future works in performance improvement.

Specifically, although audio-visual-based saliency detection has made notable progress over the past several decades, there is still significant room for improvement, *i.e.*, the AVSD model can only obtain limited performance. Thus, in the near future, we are particularly interested in further designing a more reasonable AVC classifier to improve the performance of audio-visual correspondence.

Acknowledgments. This research was supported in part by National Natural Science Foundation of China (62172246, 61976123), Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems (VRLAB2021A05), Youth Innovation and Technology Support Plan of Colleges and Universities in Shandong Province (2021KJ062), Taishan Young Scholars Program of Shandong Province, and Key Development Program for Basic Research of Shandong Province (ZR2020ZD44).

References

[1] C. Chen and et al, "Depth-quality-aware salient object detection," *TIP*, vol. 30, pp. 2350–2363, 2021.
[2] C. Chen and et al, "Improved saliency detection in rgb-d images using two-phase depth estimation and selective deep fusion," *TIP*, vol. 29, pp. 4296–4307, 2020.
[3] G. Ma and et al, "Rethinking image salient object detection: Object-level semantic saliency reranking first, pixelwise saliency refinement later," *TIP*, vol. 30, pp. 4238–4252, 2021.
[4] M. Jian and et al, "Visual-patch-attention- aware saliency detection," *TCYB*, vol. 45, no. 8, pp. 1575–1586, 2015.
[5] C. Chen and et al, "Bilevel feature learning for video saliency detection," *TMM*, vol. 20, no. 12, pp. 3324–3336, 2018.
[6] C. Chen and et al, "A novel bottom-up saliency detection method for video with dynamic background," *SPL*, vol. 25, no. 2, pp. 154–158, 2018.
[7] C. Chen and et al, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion," *TIP*, vol. 26, no. 7, pp. 3156–3170, 2017.
[8] C. Chen and et al, "Robust salient motion detection in non-stationary videos via novel integrated strategies of spatio-temporal coherency clues and low-rank analysis," *PR*, vol. 52, pp. 410–432, 2016.
[9] C. Chen and et al, "A novel video salient object detection method via semi-supervised motion quality perception," *TCSVT*, 2021.
[10] C. Chen and et al, "Improved robust video saliency detection based on long-term spatial-temporal information," *TIP*, vol. 29, pp. 1090–1100, 2020.
[11] Y. Li and et al, "Accurate and robust video saliency detection via self-paced diffusion," *TMM*, vol. 22, no. 5, pp. 1153–1167, 2019.
[12] M. Jian and et al, "Integrating object proposal with attention networks for video saliency detection," *Inf. Sci.*, vol. 576, no. 1, pp. 819–830, 2021.
[13] A. K. Katsaggelos and et al, "Audiovisual fusion: Challenges and new approaches," *JPROC*, vol. 103, no. 9, pp. 1635–1653, 2015.
[14] T. Baltrusaitis and et al, "Multimodal machine learning: A survey and taxonomy," *TPAMI*, vol. 41, no. 2, pp. 423–443, 2019.
[15] R. Cong and et al, "Review of visual saliency detection with comprehensive information," *TCSVT*, vol. 29, no. 10, pp. 2941–2959, 2019.
[16] W. Wang and et al, "Revisiting video saliency prediction in the deep learning era," *TPAMI*, vol. 43, no. 1, pp. 220–237, 2021.
[17] H. Zhu and et al, "Deep audio-visual learning: A survey," *IJAC*, vol. 18, pp. 351–376, 2021.
[18] R. Arandjelovi and et al, "Objects that sound," in *ECCV*, 2018, pp. 1–17.

[19] A. Senocak and et al, "Learning to localize sound sources in visual scenes: Analysis and applications," *TPAMI*, vol. 43, no. 5, pp. 1605–1619, 2021.

[20] A. Owens and et al, "Audio-visual scene analysis with self-supervised multisensory features," in *ECCV*, 2018.

[21] S. Ma and et al, "Contrastive learning of global and local audio-visual representations," *arXiv:2104.05418v1*, 2021.

[22] Y. Zhu and et al, "Learning audio-visual correlations from variational cross-modal generation," in *ICASSP*, 2021, pp. 1–5.

[23] B. Chen and et al, "Multimodal emotion recognition with temporal and semantic consistency," *TASLP*, vol. 29, pp. 3592–3603, 2021.

[24] C.-M. Chang and et al, "Enforcing semantic consistency for cross corpus emotion prediction using adversarial discrepancy learning," *TAC*, 2021.

[25] T. Han and et al, "Focus on semantic consistency for cross-domain crowd understanding," in *ICASSP*, 2020, pp. 1848–1852.

[26] S. Wu and et al, "Self-adapted frame selection module: Refine the input strategy for video saliency detection," in *ICAPP*, vol. 13156, 2022, pp. 509–516.

[27] Z. Wang and et al, "Video saliency prediction via joint discrimination and local consistency," *TCYB*, vol. 52, no. 3, pp. 1490–1501, 2022.

[28] D. Li and et al, "You only infer once: Cross-modal meta-transfer for referring video object segmentation," in *AAAI*, 2022.

[29] M. Lan and et al, "Siamese network with interactive transformer for video object segmentation," in *AAAI*, 2022.

[30] K. Zhang and et al, "Deep transport network for unsupervised video object segmentation," in *ICCV*, 2021, pp. 8761–8770.

[31] Q. Lai and et al, "Video saliency prediction using spatiotemporal residual attentive networks," *TIP*, vol. 29, pp. 1113–1126, 2020.

[32] Y. Lee and et al, "Iteratively selecting an easy reference frame makes unsupervised video object segmentation easier," in *AAAI*, 2022.

[33] X. Xu and et al, "Reliable propagation-correction modulation for video object segmentation," in *AAAI*, 2022.

[34] Y.-W. Chen and et al, "Video salient object detection via contrastive features and attention modules," in *WACV*, 2022.

[35] Y. Lu and et al, "Depth-cooperated trimodal network for video salient object detection," *arXiv:2202.06060*, 2022.

[36] W. Zhao and et al, "Weakly supervised video salient object detection," in *CVPR*, 2021, pp. 16 821–16 830.

[37] Y. Tang and et al, "Video salient object detection via adaptive local-global refinement," *arXiv:2104.14360*, 2021.

[38] Y. Jiao and et al, "Guidance and teaching network for video salient object detection," in *ICIP*, 2021, pp. 2199–2203.

[39] X. Zhao and et al, "Multi-source fusion and automatic predictor selection for zero-shot video object segmentation," in *ACM MM*, 2021.

[40] H. Li and et al, "Motion guided attention for video salient object detection," in *ICCV*, 2019, pp. 7273–7282.

[41] S. Ren and et al, "Tenet: Triple excitation network for video salient object detection," in *ECCV*, 2020, pp. 212–228.

[42] M. Zhang and et al, "Dynamic context-sensitive filtering network for video salient object detection," in *ICCV*, 2021, pp. 1533–1543.

[43] G.-P. Ji and et al, "Full-duplex strategy for video object segmentation," in *ICCV*, 2021, pp. 4902–4913.

[44] H. Song and et al, "Pyramid dilated deeper convlstm for video salient object detection," in *ICCV*, 2018, pp. 715–731.

[45] D. Fan and et al, "Shifting more attention to video salient object detection," in *CVPR*, 2019, pp. 8554–8564.

[46] Y. Gu and et al, "Pyramid constrained self-attention network for fast video salient object detection," in *AAAI*, 2020, pp. 10 869–10 876.

[47] C. Chen and et al, "Exploring rich and efficient spatial temporal interactions for real-time video salient object detection," *TIP*, vol. 30, pp. 3995–4007, 2021.

[48] C. Liu, "Beyond pixels: Exploring new representation and applications for motion analysis," in *MIT Ph.D. dissertation*, 2009.

[49] S. Hochreiter and et al, "Long short-term memory," *NECO*, vol. 9, no. 8, pp. 1735–1780, 1997.

[50] X. Shi and et al, "Convolutional lstm network: a machine learning approach for precipitation nowcasting," in *NIPS*, 2015.

[51] D. Tran and et al, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015, pp. 4489–4497.

[52] Y. Su and et al, "A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection," *arXiv:2203.04708*, pp. 1–12, 2022.

[53] K. Huang and et al, "Transformer-based cross reference network for video salient object detection," *PRL*, vol. 160, pp. 122–127, 2022.

[54] C. Bak and et al, "Spatio-temporal saliency networks for dynamic saliency prediction," *TMM*, vol. 20, no. 7, pp. 1688–1698, 2018.

[55] A. Dosovitskiy and et al, "Flownet: Learning optical flow with convolutional networks," in *ICCV*, 2015, pp. 2758–2766.

[56] E. Ilg and et al, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *CVPR*, 2017, pp. 1647–1655.

[57] K. Zhang and et al, "A spatial-temporal recurrent neural network for video saliency prediction," *TIP*, vol. 30, pp. 572–587, 2021.

[58] W. Wang and et al, "Revisiting video saliency: A large-scale benchmark and a new model," in *CVPR*, 2018, pp. 4894–4903.

[59] P. Linardos and et al, "Simple vs complex temporal recurrences for video saliency prediction," *arXiv:1907.01869*, pp. 1–12, 2019.

[60] J. Chen and et al, "Video saliency prediction using enhanced spatiotemporal alignment network," *PR*, vol. 109, p. 107615, 2021.

[61] R. Droste and et al, "Unified image and video saliency modeling," in *ECCV*, 2019, pp. 419–435.

[62] L. Jiang and et al, "Deepvs: A deep learning based video saliency prediction approach," in *ECCV*, 2018, pp. 625–642.

[63] X. Wu and et al, "Salsac: A video saliency prediction model with shuffled attentions and correlation-based convlstm," in *AAAI*, 2020, pp. 12 410–12 417.

[64] M. Cornia and et al, "Predicting human eye fixations via an lstm-based saliency attentive model," *TIP*, vol. 27, no. 10, pp. 5142–5154, 2018.

[65] S. Gorji and et al, "Going from image to video saliency: Augmenting image salience with dynamic attentional push," in *CVPR*, 2018, pp. 7501–7511.

[66] W. Zou and et al, "Sta3d: Spatiotemporally attentive 3d network for video saliency prediction," *PRL*, vol. 147, pp. 78–84, 2022.

[67] K. Min and et al, "Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection," in *ICCV*, 2019, pp. 2394–2403.

[68] G. Bellitto and et al, "Hierarchical domain-adapted feature learning for video saliency prediction," *arXiv:2010.01220v4*, pp. 1–12, 2021.

[69] Y. Wang and et al, "Spatiotemporal module for video saliency prediction based on self-attention," *Image Vision Comput.*, vol. 112, 2021.

[70] C. Ma and et al, "Video saliency forecasting transformer," *TCSVT*, 2022.

[71] K. Simonyan and et al, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[72] K. He and et al, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[73] Y. Cheng and et al, "Look, listen, and atend: Co-atention network for self-supervised audio-visual representation learning," in *ACM MM*, 2020, pp. 3884–3892.

[74] C. Gan and et al, "Self-supervised moving vehicle tracking with stereo sound," in *ICCV*, 2019, pp. 7052–7061.

[75] G. Wang and et al, "From semantic categories to fixations: A novel weakly-supervised visual-auditory saliency detection approach," in *CVPR*, 2021, pp. 15 119–15 128.

[76] G. D. Sad and et al, "Complementary models for audio-visual speech classification," *IJSP*, vol. 25, no. 1, pp. 231–249, 2022.

[77] F. Ma and et al, "Data augmentation for audio-visual emotion recognition with an efficient multimodal conditional gan," *Appl. Sci.-Basel*, vol. 12, no. 1, 2022.

[78] X. Qian and et al, "Audio-visual tracking of concurrent speakers," *TMM*, vol. 24, pp. 942–954, 2022.

[79] Y. Aytar and et al, "Soundnet: Learning sound representations from unlabeled video," in *NIPS*, 2016, pp. 892–900.

[80] H. Chen and et al, "Vggsound: A large-scale audio-visual dataset," in *ICASSP*, 2020, pp. 721–725.

[81] M. Muller, *Information Retrieval for Music and Motion*. Springer Berlin, Heidelberg, 2007, vol. 2.

[82] H. Alwassel and et al, "Self-supervised learning by cross-modal audio-video clustering," in *NIPS*, 2020.

[83] Y. Wu and et al, "Dual attention matching for audio-visual event localization," in *ICCV*, 2019, pp. 6291–6299.

[84] Y. Tian and et al, "Unified multisensory perception: Weakly-supervised audio-visual video parsing," in *ECCV*, 2020, pp. 436–454.

[85] R. Gao and et al, "Listen to look: Action recognition by previewing audio," in *CVPR*, 2020, pp. 10 457–10 467.

[86] R. Arandjelovic and et al, "Look, listen and learn," in *ICCV*, 2017.

[87] X. Lu and et al, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in *CVPR*, 2019, pp. 3623–3632.

[88] R. J. Chen and et al, "Multimodal co-attention transformer for survival prediction in gigapixel whole slide images," in *ICCV*, 2021, pp. 3995–4005.

[89] P. Morgado and et al, "Audio-visual instance discrimination with cross-modal agreement," *arXiv:2004.12943v3*, 2021.

[90] Y. Lin and et al, "Unsupervised sound localization via iterative contrastive learning," *arXiv:2104.00315v1*, 2021.

[91] G. Verma and et al, "Learning affective correspondence between music and image," in *ICASSP*, 2019, pp. 3975–3979.

[92] R. Su and et al, "Themes informed audio-visual correspondence learning," *arXiv:2009.06573v2*, 2020.

[93] X. Wang and et al, "Learning person re-identification models from videos with weak supervision," *TIP*, vol. 30, pp. 3017–3028, 2021.

[94] J. Meng and et al, "Deep graph metric learning for weakly supervised person re-identification," *TPAMI*, 2021.

[95] X. Shu and et al, "Large-scale spatio-temporal person re-identification: Algorithms and benchmark," *TCSVT*, 2021.

[96] M. Cao and et al, "Progressive bilateral-context driven model for post-processing person re-identification," *TMM*, vol. 23, pp. 1239–1251, 2021.

[97] A. Nagrani and et al, "Seeing voices and hearing faces: Cross-modal biometric matching," in *CVPR*, 2018, pp. 8427–8436.

[98] Y. Weny and et al, "Disjoint mapping network for cross-modal matching of voices and faces," in *ICLR*, 2019.

[99] R. Wang and et al, "A novel distance learning for elastic cross-modal audio-visual matching," in *ICMEW*, 2019, pp. 300–305.

[100] A. Nagrani and et al, "Learnable pins: Cross-modal embeddings for person identity," in *ECCV*, 2018, pp. 73–89.

[101] K. Hoover and et al, "Putting a face to the voice: Fusing audio and visual signals across a video to determine speakers," *arXiv:1706.00079v1*, 2017.

[102] D. Suris and et al, "Cross-modal embeddings for video and audio retrieval," in *ECCVW*, 2018.

[103] J. Hershey and et al, "Audio-vision: Using audio-visual synchrony to locate sounds," in *NIPS*, 1999, pp. 813–819.

[104] H. Zhao and et al, "The sound of motions," in *ICCV*, 2019, pp. 1735–1744.

[105] H. Izadinia and et al, "Multimodal analysis for identification and segmentation of moving-sounding objects," *TMM*, vol. 15, no. 2, pp. 378–390, 2013.

[106] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.

[107] E. Kidron and et al, "Pixels that sound," in *CVPR*, 2005, pp. 88–95.

[108] Y. Tian and et al, "Audio-visual event localization in unconstrained videos," in *ECCV*, 2018, pp. 252–268.

[109] Y. Liu and et al, "Visual localization of non-stationary sound sources," in *ACM MM*, 2009, pp. 513–516.

[110] J. Nakajima and et al, "Incorporating audio signals into constructing a visual saliency map," in *PSIVT*, 2013, pp. 468–480.

[111] J. Ruesch and et al, "Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub," in *ICRA*, 2008, pp. 1050–4729.

[112] B. Schauerte and et al, "Multimodal saliency-based attention for object-based scene analysis," in *IROS*, 2011, pp. 1173–1179.

[113] T. Zhang and et al, "Acousticfusion: Fusing sound source localization to visual slam in dynamic environments," in *IROS*, 2021, pp. 6868–6875.

[114] D. Hu and et al, "Class-aware sounding objects localization via audiovisual correspondence," *TPAMI*, 2021.

[115] J. Chen and et al, "Multimodal fusion for indoor sound source localization," *PR*, vol. 115, 2021.

[116] H. Chen and et al, "Localizing visual sounds the hard way," in *CVPR*, 2021, pp. 16 862–16 871.

[117] A. Senocak and et al, "Learning to localize sound source in visual scenes," in *CVPR*, 2018, pp. 4358–4366.

[118] Y. Nakatani and et al, "Auditory spatial saliency and its effects on perceptual noisiness," *IEEE Access*, vol. 10, pp. 10 160–10 175, 2022.

[119] L. Wang and et al, "Functional brain networks underlying auditory saliency during naturalistic listening experience," *TCDS*, vol. 14, no. 1, pp. 156–163, 2022.

[120] L. Itti and et al, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Res.*, vol. 40, no. 10, pp. 1489–1506, 2000.

[121] C. Kayser and et al, "Mechanisms for allocating auditory attention: An auditory saliency map," *Curr. Biol.*, vol. 15, no. 21, pp. 1943–1947, 2005.

[122] C. Gan and et al, ""wow!" bayesian surprise for salient acoustic event detection," in *ICASSP*, 2013, pp. 6402–6406.

[123] T. Tsuchida and et al, "Auditory saliency using natural statistics," in *PAMCSS*, 2012, pp. 1–4.

[124] A. Rodriguez-Hidalgo and et al, "The robustness of echoic log-surprise auditory saliency detection," *IEEE Access*, vol. 6, pp. 72 083–72 093, 2018.

[125] A. Zlatintsi and et al, "A saliency-based approach to audio event detection and summarization," in *EUSIPCO*, 2012, pp. 1294–1298.

[126] G. Evangelopoulos and et al, "Video event detection and summarization using audio," in *ICASSP*, 2009.

[127] E. M. Kaya and et al, "A temporal saliency map for modeling auditory attention," in *CISS*, 2012.

[128] K. Lin and et al, "Improving faster-than-real-time human acoustic event detection by saliency-maximized audio visualization," in *ICASSP*, 2012.

[129] G. Evangelopoulos and et al, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *TMM*, vol. 15, no. 7, pp. 1553–1568, 2013.

[130] A. Zlatintsi and et al, "Audio salient event detection and summarization using audio and text modalities," in *EUSIPCO*, 2015.

[131] Z. Podwinska and et al, "Acoustic event detection from weakly labeled data using auditory salience," in *ICASSP*, 2019.

[132] S. Ramenahalli and et al, "Audio-visual saliency map: Overview, basic models and hardware implementation," in *CISS*, 2013, pp. 1–6.

[133] N. Sidaty and et al, "An audiovisual saliency model for conferencing and conversation videos," in *ISEI*, 2016.

[134] D. Zhu and et al, "Lavs: A lightweight audio-visual saliency prediction model," in *ICME*, 2021, pp. 1–6.

[135] S. Yao and et al, "Deep audio-visual fusion neural network for saliency estimation," in *ICIP*, 2021, pp. 1604–1608.

[136] F.-Y. Chao and et al, "Towards audio-visual saliency prediction for omnidirectional video with spatial audio," in *VCIP*, 2020, pp. 355–358.

[137] H. Akolkar and et al, "Visual-auditory saliency detection using event-driven visual sensors," in *EBCCSP*, 2015.

[138] A. Coutrot and et al, "An audiovisual attention model for natural conversation scenes," in *ICIP*, 2014, pp. 1100–1104.

[139] N. Sidaty and et al, "Toward an audiovisual attention model for multimodal video content," *NEUCOM*, vol. 259, pp. 94–111, 2017.

[140] X. Min and et al, "Sound influences visual attention discriminately in videos," in *QoMEX*, 2014, pp. 153–158.

[141] X. Min and et al, "Fixation prediction through multimodal analysis," *TOMM*, vol. 13, no. 1, pp. 1–23, 2016.

[142] X. Min and et al, "A multimodal saliency model for videos with high audio-visual correspondence," *TIP*, vol. 29, pp. 3805–3819, 2020.

[143] A. Tsiami and et al, "A behaviorally inspired fusion approach for computational audiovisual saliency modeling," *SPIC*, vol. 76, pp. 186–200, 2019.

[144] C. Parise and et al, "Cross-correlation between auditory and visual signals promotes multisensory integration," *Multisens. Res.*, vol. 26, no. 3, pp. 307–316, 2013.

[145] M. Rolf and et al, "Attention via synchrony: Making use of multimodal cues in social learning," *TAMD*, vol. 1, no. 1, pp. 55–67, 2009.

[146] H. Tavakoli and et al, "Dave: A deep audio-visual embedding for dynamic saliency prediction," *arXiv:1905.10693v2*, 2020.

[147] A. Tsiami and et al, "Stavis: Spatio-temporal audiovisual saliency network," in *CVPR*, 2020, pp. 4766–4776.

[148] S. Jain and et al, "Vinet: Pushing the limits of visual modality for audio-visual saliency prediction," *arXiv:2012.06170v2*, 2021.

[149] J. Chen and et al, "Audiovisual saliency prediction via deep learning," *NEUCOM*, vol. 428, pp. 248–258, 2021.

[150] P. K. Mital and et al, "Clustering of gaze during dynamic scene viewing is predicted by motion," *Cogn. Comput.*, vol. 3, no. 1, pp. 5–24, 2011.

[151] A. Coutrot and et al, "Toward the introduction of auditory information in dynamic visual attention models," in *WIAMIS*, 2013.

[152] A. Coutrot and et al, "How saliency, faces, and sound influence gaze in dynamic social scenes?" *J.Vision*, vol. 14, no. 8, 2014.

[153] M. Gygli and et al, "Creating summaries from user videos," in *ECCV*, 2014, pp. 505–520.

[154] P. Koutras and et al, "A perceptually based spatio-temporal computational framework for visual saliency estimation," *SPIC*, vol. 38, pp. 15–31, 2015.

[155] L. Itti and et al, "A model of saliency-based visual attention for rapid scene analysis," *TPAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[156] H. Jonathan and et al, "Graph-based visual saliency," in *NIPS*, vol. 8, 2007, pp. 545–552.

[157] D. Zhang and et al, "Supervision by fusion: Towards unsupervised learning of deep salient object detector," in *ICCV*, 2017, pp. 4068–4076.

[158] V. Leborn and et al, "Dynamic whitening saliency," *TPAMI*, vol. 39, no. 5, pp. 893–907, 2017.
[159] A. Chattopadhay and et al, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *WACV*, 2018, pp. 839–847.
[160] L. Wang and et al, "Learning to detect salient objects with image-level supervision," in *CVPR*, 2017, pp. 3796–3805.
[161] Y. Zeng and et al, "Multi-source weak supervision for saliency detection," in *CVPR*, 2019, pp. 6067–6076.