ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096773

TOWARDS ROBUST AUDIO-BASED VEHICLE DETECTION VIA IMPORTANCE-AWARE AUDIO-VISUAL LEARNING

*Jung Uk Kim** and Seong Tae Kim*

Department of Computer Science and Engineering, Kyung Hee University, South Korea

ABSTRACT

Although audio modality has the potential to solve various visually challenging conditions of visual modality, there are few studies on audio-based detection. This is because the audio modality itself contains less accurate spatial information. To alleviate this issue, the existing audio-based methods adopt the visual modality in the training phase to transfer more precise spatial knowledge to the audio modality. However, they do not consider the case where the visual modality is less informative. In this paper, we present a new audio-based vehicle detector that can transfer multimodal knowledge of vehicles to the audio modality during training. To this end, we combine the audio-visual modal knowledge according to the importance of each modality to generate integrated audio-visual feature. Also, we introduce an audio-visual distillation (AVD) loss that guides representation of the audio modal feature to resemble that of the integrated audio-visual feature. As a result, our audio-based detector can perform robust vehicle detection as if it were utilizing both modalities, even if it only receives audio modality as input in the inference. Comprehensive experimental results demonstrate that our method exhibits consistent improvements over the existing methods.

Index Terms— Audio-based vehicle detection, audio-visual integration, contrastive learning, deep learning

1. INTRODUCTION

Object detection has attracted a great deal of attention for decades, and it is one of the most actively studied topics in computer vision [1–5]. Since object detection is closely related to human life, it has been applied to various human-oriented research, such as autonomous driving [6] and robotics [7]. However, most object detection studies adopt visual modality (*e.g.,* RGB and thermal) as a single input [8]. It is similar to detecting object using human eye.

Thinking about it further, we can also use auditory information through the ears to detect objects. We have been frequently detecting objects using only sound without any visual information. By hearing object sounds, we can capture the location and classify object class [9]. The merit of the audio modality is that it has the ability to detect objects in visually challenging conditions [10–12], such as heavy occlusion and illumination variation. Thus, a study of effectively detecting an object using only audio modality at the inference time is an interesting field with high potential for development.

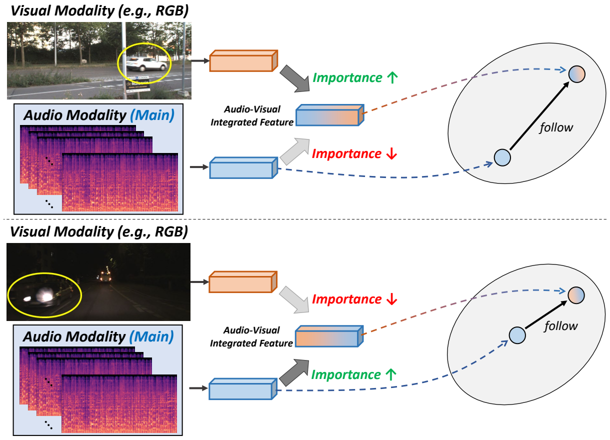
Visual Modality (e.g., RGB)
Audio Modality (Main)
Importance ↑
Audio-Visual Integrated Feature
Importance ↓
follow
Visual Modality (e.g., RGB)
Audio Modality (Main)
Importance ↓
Audio-Visual Integrated Feature
Importance ↑
follow

Fig. 1. An illustration of our main idea. Integrated audio-visual feature is encoded taking into account the importance of audio and visual modalities. Then, audio modal feature learns representation of the integrated audio-visual feature.

Despite the high potential of audio modality, there are few studies that have adopted audio modality as the main input for object detection. One of the main reasons is that spatial information encoded through the audio modality is less accurate than the visual modality [13]. Usually, spatial information of the audio modality is simply formed based on the difference in sound arrival time of stereo or 8-channel microphone arrays [14]. Thus, recent audio-based detectors have attempted to adopt the visual modality to learn more accurate spatial knowledge during training phase [9, 13]. StereoSoundNet [13] is introduced to detect objects using stereo sound only. In StereoSoundNet, RGB image is adopted for the visual modality to transfer the spatial knowledge of RGB images to the audio modality. MM-DistillNet [9] performs a knowledge distillation to transfer the combined spatial knowledge of the RGB, thermal, and depth modalities to the audio modality. However, as mentioned earlier, in visually challenging conditions (*e.g.,* occlusion or illumination variation),

*Corresponding author (ju.kim@khu.ac.kr).

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-00155911, Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University))

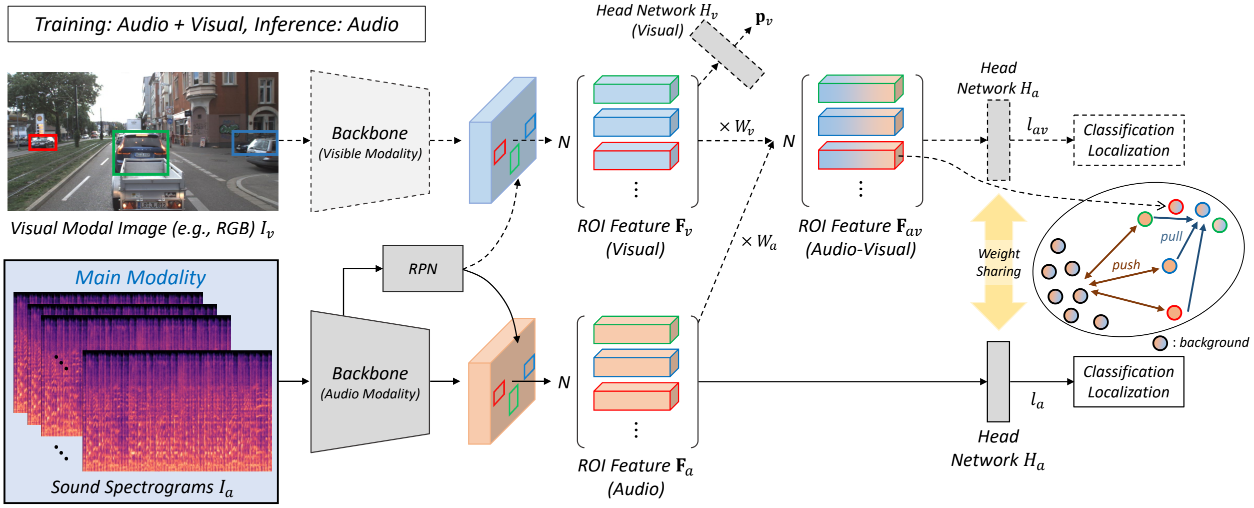

Fig. 2. The overall architecture of the proposed method. In the training phase, visual modality (*e.g.,* RGB) is additionally adopted for generating integrated audio-visual feature to guide the audio modal feature. In the inference phase, vehicle detection is conducted based on the single audio modality (sound spectrogram). Dotted lines are considered only in the training phase.

directly transferring spatial knowledge of the visual modality can limit the performance of the audio-based detectors.

Humans can use multimodal information [15–17]. Especially, humans recognize objects by looking at them and hearing their sounds at the same time, that is combining audio-visual modalities. This procedure is called audio-visual integration [18–20]. In this step, humans emphasize one modality (audio or visual) if it is helpful when recognizing an object, in the opposite case, it is suppressed [21]. The integrated audio-visual modalities contain information that can complement the single modality (audio or visual). So, providing combined knowledge of the audio-visual modalities to the audio modality can be more effective than the previous methods [9, 13] that only transfer knowledge of the visual modality.

In this paper, we introduce a novel audio-based object detector that learns spatial knowledge from audio-visual modalities. We focus on the vehicle class frequently observed in the real world. Audio-based object detection assumes that only audio modal information is provided in the inference phase. The illustration of our main idea is depicted in Fig. 1. We select RGB modality to represent the visual modality. First, we combine two modal features according to the importance of each modality. For example, as shown in Fig. 1 (lower part), in low illumination condition, visual modal feature contains less important spatial cues to detect vehicles. In this case, we encode the integrated audio-visual feature by suppressing the visual modal feature and enhancing the audio modal feature.

Next, we introduce an audio-visual distillation (AVD) loss to perform knowledge distillation from the audio-visual feature to the audio modal feature. Since our audio-based vehicle detector follows the two-stage detection framework [2], we obtain various audio-visual features from the vehicle candidates. In this step, the role of AVD loss is to encode feature representation of the candidates in the audio modality to follow that of the integrated audio-visual modalities. As a result, our audio-based detector can perform robust vehicle detection even when only audio modal input (car sound) is provided. Comprehensive experimental results demonstrate the effectiveness of the proposed method.

2. PROPOSED METHOD

Fig. 2 shows the overall architecture of the proposed audio-based vehicle detector. In the training phase, we use both audio and visual modalities. Please note that visual modality is utilized only in the training phase. 8-channel sound spectrogram I_a and 3-channel visual modal image I_v pass through each modal-specific backbone network (*i.e.,* ResNet-50 [22]) to encode image feature. For 8-channel sound spectrogram input, we modify the input channel of ResNet-50 *conv1* from 3 to 8. Since the main purpose of our detector is to use single audio modality in the inference phase, Region Proposal Network (RPN) receives image feature of the audio modality to estimate N candidate regions, called Region-of-Interests (ROIs). Then, ROI features of two modalities $\mathbf{F}_a = \{\mathbf{f}_{a_i}\}_{i=1}^{N}$ ($\mathbf{f}_{a_i} \in \mathbb{R}^{w \times h \times c}$) and $\mathbf{F}_v = \{\mathbf{f}_{v_i}\}_{i=1}^{N}$ ($\mathbf{f}_{v_i} \in \mathbb{R}^{w \times h \times c}$) are extracted based on the N ROIs. The head network H_a receives $\mathbf{F}_a$ to conduct classification and localization.

In addition, integrated audio-visual ROI features $\mathbf{F}_{av} = \{\mathbf{f}_{av_i}\}_{i=1}^{N}$ are generated by combining $\mathbf{F}_a$ and $\mathbf{F}_v$. The head network H_a also receives $\mathbf{F}_{av}$ so that $\mathbf{F}_{av}$ can be a reference feature that can guide $\mathbf{F}_a$. In the inference time, our detector only receives I_a. In the following subsections, we describe how to generate $\mathbf{F}_{av}$ and how to guide $\mathbf{F}_a$ using $\mathbf{F}_{av}$.

2.1. Integrated Audio-Visual ROI Feature

We aim to generate an audio-visual feature that can provide useful knowledge to the audio modality. To this end, we combine $\mathbf{F}_a$ and $\mathbf{F}_v$ to generate $\mathbf{F}_{av}$ so that $\mathbf{F}_{av}$ can have multimodal knowledge about the vehicle. However, simply combining audio-visual features has the following problem (please see Fig.1). When a vehicle is visually occluded, visual modality lacks the spatial information of vehicle. At this time, ROI feature of the visual modality can hinder when combining it with the ROI feature of audio modality. Conversely, if the audio modality is of poor quality, it can interfere when generating an audio-visual ROI feature. Thus, we encode $\mathbf{F}_{av}$ taking into account the importance of each modality.

To quantify the importance of two modalities, we additionally adopt visual modal head network H_v to estimate the classification score of the vehicle class of the visual modality $\mathbf{p}_v = \{p_{v_i}\}_{i=1}^N$. Similarly, $\mathbf{p}_a = \{p_{a_i}\}_{i=1}^N$ are estimated through the audio modal head network H_a. The classification score is an indicator that determines the importance of each ROI feature [23]. The i-th weight of two modalities $W_a = \{W_{a_i}\}_{i=1}^N$ and $W_v = \{W_{v_i}\}_{i=1}^N$ can be obtained as:

$$W_{a_i} = \frac{p_{a_i}}{p_{a_i} + p_{v_i}}, \; W_{v_i} = \frac{p_{v_i}}{p_{a_i} + p_{v_i}}. \tag{1}$$

To the next, $\mathbf{f}_{av_i} \in \mathbb{R}^{w \times h \times c}$ is obtained as follows:

$$\mathbf{f}_{av_i} = G(\{W_{a_i} \cdot \mathbf{f}_{a_i}, W_{v_i} \cdot \mathbf{f}_{v_i}\}), \tag{2}$$

where $G(\cdot)$ denotes 1×1 convolution and $\{\cdot, \cdot\}$ is a concatenation operation. Through $G(\cdot)$, channel number of $\mathbf{f}_{av_i}$ become same as $\mathbf{f}_{a_i}$ and $\mathbf{f}_{v_i}$. By doing so, relatively important modality is more considered, and relatively unimportant modality is less considered when combining audio-visual modalities.

2.2. Audio-Visual Distillation (AVD) Loss

After generating $\mathbf{F}_{av}$, we try to guide $\mathbf{F}_a$ to follow feature representation of $\mathbf{F}_{av}$. To this end, we introduce an audio-visual distillation (AVD) loss. The AVD loss consists of group contrastive learning (GCL) loss and audio-visual knowledge distillation (AVKD) loss. First, for GCL loss, we borrow the idea of contrastive learning [24], which pulls positive pairs and pushes negative pairs in the feature space. In object detection, there are P positive candidates and (N-P) negative candidates. Moreover, since we handle the vehicle class only, we propose GCL loss that considers the two groups (*i.e.,* vehicle class and background). $\mathbf{F}_a$ and $\mathbf{F}_{av}$ pass through the weight-sharing head network H_a to encode vector $l_a = \{l_{a_i}\}_{i=1}^N$ ($l_{a_i} \in \mathbb{R}^d$) and $l_{av} = \{l_{av_i}\}_{i=1}^N$ ($l_{av_i} \in \mathbb{R}^d$) ($d$ is vector length). In GCL loss, i-th group contrastive losses of the audio modality and audio-visual modalities are formulated as:

$$\mathcal{L}_i^m = -log\, q_i^m = -log \frac{\sum_{j=1}^{P} exp(s(l_i^m, l_j^m)/\tau)}{\sum_{j=1}^{N} exp(s(l_i^m, l_j^m))/\tau)}, \tag{3}$$

$$s(l_i^m, l_j^m) = \frac{l_i^m \cdot l_j^m}{\left\|l_i^m\right\| \left\|l_j^m\right\|}, \tag{4}$$

where $m \in \{a, av\}$ and τ is a temperature. Therefore, GCL loss is obtained as:

$$\mathcal{L}_{GCL} = \frac{1}{P}\sum_{i=1}^{P} \mathcal{L}_i^{av} + \frac{1}{P}\sum_{i=1}^{P} \mathcal{L}_i^{a}. \tag{5}$$

GCL loss makes all P positive candidate features in audio modality similar, and so are the audio-visual modalities.

In addition, we introduce AVKD loss to guide feature representation of the audio modality to follow that of the audio-visual modalities by considering the relationship between vehicle candidates. It is represented as follows:

$$\mathcal{L}_{AVKD} = \frac{1}{P}\sum_{i=1}^{P} D_{KL}(\mathbf{q}_i^{av} || \mathbf{q}_i^{a}), \tag{6}$$

where $\mathbf{q}_i^m = (q_i^m, 1 - q_i^m)$ ($m \in \{a, av\}$) is the probability distribution from Eq. (3). Through AVKD loss, audio modal features contain the audio-visual knowledge about vehicles.

Finally, the proposed AVD loss is calculated as follows:

$$\mathcal{L}_{AVD} = \lambda_1 \mathcal{L}_{GCL} + \lambda_2 \mathcal{L}_{AVKD}, \tag{7}$$

where λ_1 and λ_2 are balancing parameter. As a result, our detector performs robust detection with more discriminative features even when only audio modal information is given.

2.3. Total Loss Function

The total loss function of our detector is defined as follows:

$$\begin{gathered} \mathcal{L}_{Total} = \mathcal{L}_{OD} + \lambda_3 \mathcal{L}_{AVD}, \\ \mathcal{L}_{OD} = \mathcal{L}_{RPN} + \mathcal{L}_{cls}^{a} + \mathcal{L}_{loc}^{a} + \mathcal{L}_{cls}^{av} + \mathcal{L}_{loc}^{av}, \end{gathered} \tag{8}$$

where $\mathcal{L}_{OD}$ includes RPN $\mathcal{L}_{RPN}$ [1], classification and localization losses for audio modality ($\mathcal{L}_{cls}^{a}$ and $\mathcal{L}_{loc}^{a}$) and audio-visual modalities ($\mathcal{L}_{cls}^{av}$ and $\mathcal{L}_{loc}^{av}$). λ_3 is balancing parameter.

3. EXPERIMENTS

3.1. Dataset and Implementation Details

We use Multimodal Audio-Visual Detection (MAVD) dataset [9] which is the only public dataset that contains 113,283 synchronized audio modality and RGB, thermal, and depth visual modalities for vehicle detection. The audio modal input is obtained from 8-channel monophonic-microphone array. When images are taken, sound is obtained 0.5s before and after the timestamp. It is transformed into mel-spectrogram using 80 mel-filter banks and resized to be 576×576×8. Image resolution of the visual modalty is resized to be 576×576×3.

Following [9], we adopt mean average precision (mAP). With mAP, we obtain the results when Intersection over Union (IOU) thresholds were 0.5 (mAP@0.5) and 0.75 (mAP@0.75). We also average it by varying IOU thresholds from 0.5:0.05:0.95 (mAP@Avg). Also, we adopt center distances CD_x and CD_y that measure the center error between the predicted and ground-truth bounding-box.

We adopt FPN [2] with ResNet-50 [22] as our baseline. For the 8-channel sound spectrogram, we change the input channel of ResNet-50 *conv1* from 3 to 8. We adopt stochastic gradient descent (SGD) optimizer. We use 4 GTX 1080 Ti GPUs with 32 images and learn 30 epochs. For 20 epochs,

Table 1. Detection results on MAVD Dataset. We compare our method with state-of-the-art audio-based vehicle detectors. Note that 'Visual Modality' indicates the type of the Adopted Visual Modality ($\mathcal{R}$: RGB, $\mathcal{T}$: Thermal, $\mathcal{D}$: Depth) in the training phase. $\mathcal{R}$+$\mathcal{T}$+$\mathcal{D}$ denotes three modalities with each modal-specific backbone to encode visual feature.

Method	Visual Modality	mAP@Avg↑	mAP@0.5↑	mAP@0.75↑	CD_x(%)↓	CD_y(%)↓
StereoSoundNet [13]	RGB ($\mathcal{R}$)	44.05	62.38	41.46	3.00	2.24
Pairwise loss [25]	RGB ($\mathcal{R}$)	40.45	59.72	36.73	2.98	2.20
AFD loss [26]	RGB ($\mathcal{R}$)	44.27	62.00	41.90	3.19	2.28
MM-DistillNet [9]	RGB ($\mathcal{R}$)	44.58	62.66	42.39	2.94	2.17
Ours	RGB ($\mathcal{R}$)	**58.39**	**78.91**	**56.29**	**1.31**	**1.28**
StereoSoundNet [13]	$\mathcal{R}$+$\mathcal{T}$+$\mathcal{D}$	56.16	80.03	52.96	1.46	0.80
AFD loss [26]	$\mathcal{R}$+$\mathcal{T}$+$\mathcal{D}$	58.50	82.18	55.48	1.30	0.70
MM-DistillNet [9]	$\mathcal{R}$+$\mathcal{T}$+$\mathcal{D}$	61.62	84.29	59.66	1.27	0.69
Ours	$\mathcal{R}$+$\mathcal{T}$+$\mathcal{D}$	**64.75**	**85.67**	**63.99**	**0.72**	**0.58**

Table 2. Detection results by varying the visual modal types. We fix the backbone network as FPN [2] with ResNet-50 [22].

Visual Modality	Method	mAP@ Avg↑	mAP@ 0.5↑	mAP@ 0.75↑
RGB ($\mathcal{R}$)	MTA loss [9]	55.31	80.60	55.75
	Ours	**58.39**	**78.91**	**56.29**
Thermal ($\mathcal{T}$)	MTA loss [9]	54.28	81.27	54.01
	Ours	**57.87**	**81.97**	**56.18**
Depth ($\mathcal{D}$)	MTA loss [9]	53.78	80.89	54.08
	Ours	**57.80**	**81.88**	**56.10**
$\mathcal{R}$+$\mathcal{T}$+$\mathcal{D}$	MTA loss [9]	59.65	85.03	57.20
	Ours	**64.75**	**85.67**	**63.99**

Table 3. Effect of our AVD loss and modal weight (W_a and W_v). We adopt RGB images ($\mathcal{R}$) for the visual modality.

$\mathcal{L}_{AVD}$	W_a,W_v	mAP@Avg↑	mAP@0.5↑	mAP@0.75↑
✗	✗	39.03	56.52	36.44
✓	✗	56.52	74.34	53.96
✓	✓	**58.39**	**78.91**	**56.29**

we train our method with 0.008 learning rate and decay it by 0.1. We set $\tau = 0.1$ and $\lambda_1, \lambda_2, \lambda_3 = 1$.

3.2. Performance Comparison

We compare our method with the existing audio-based vehicle detectors [9, 13, 25, 26]. Table 1 shows the results when the visual modality is one (RGB) or three (RGB, thermal, and depth). In the case of the RGB images, our method shows significant improvements. Especially, our method shows 13.81 improvement for mAP@Avg metric. Compared to the MM-DistillNet [9], our method shows the significant improvement. We also conduct experiments when the visual modalities are three (RGB, thermal, and depth). In this case, three modal-specific backbone networks encode each image feature and combine (concatenate and 1×1 conv) to generate $\mathbf{F}_v$. As shown in Table 1, even if the three modalities are provided, our method still outperforms the existing methods.

Next, since MTA loss of MM-DistillNet [9] show the best performance among the existing methods, we compare our method with MTA loss by varying type of the visual modality.

Table 4. Result comparisons when detector adopt the single visual modality or audio-visual modalities.

Input Modality	mAP@Avg↑
Visual Modality (RGB ($\mathcal{R}$))	84.81
Audio Modality + $\mathcal{R}$ (w/ W_a, W_v)	**86.29**

We fix the backbone network as FPN [2] with ResNet-50 [22]. As shown in Table 2, regardless of the type of visual modality our method still outperforms MTA loss.

Table 3 presents the ablation studies that show the effect of AVD loss and modal weight (W_a and W_v). When AVD loss is adopted, our method shows large gains. Also, considering the modal weight contributes to additional performance gains.

3.3. Discussion

To investigate whether the spatial knowledge of visual modality alone is effective or that of audio-visual modalities is effective, we compared the detection performance of the two methods. As shown in Table 4, as the detector utilizes both audio-visual modalities with the importance weight of two modalities (W_a and W_v), performance is higher than the visual modality. Through the results, we insist that multimodal knowledge has more powerful spatial knowledge than a visual modality. Therefore, audio-visual feature can convey more effective spatial knowledge to the audio modality.

4. CONCLUSION

In this paper, we consider a situation where only audio modal information is provided in the inference phase. We present an interesting method that can perform robust vehicle detection even in such a situation. Unlike the previous methods that only transfer the spatial knowledge of the visual modality, we generate the integrated audio-visual feature to transfer the combined multimodal knowledge to the audio modality. When generating integrated audio-visual features, we consider the importance of each modality. Moreover, we propose audio-visual distillation (AVD) loss to guide the feature of the audio modality to resemble that of the audio-visual modalities. As a result, our detector can detect vehicles as if it had used both audio and visual modalities simultaneously.

5. REFERENCES

[1] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2015.

[2] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2117–2125.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 213–229.

[4] Mingxing Tan, Ruoming Pang, and Quoc V Le, "Efficientdet: Scalable and efficient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10781–10790.

[5] Jung Uk Kim, Sungjune Park, and Yong Man Ro, "Robust small-scale pedestrian detection with cued recall via memory learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 3050–3059.

[6] Jung Uk Kim, Jungsu Kwon, Hak Gu Kim, Haesung Lee, and Yong Man Ro, "Object bounding box-critic networks for occlusion-robust object detection in road scene," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2018, pp. 1313–1317.

[7] Haihua Lu, Xuesong Chen, Guiying Zhang, Qiuhao Zhou, Yanbo Ma, and Yong Zhao, "Scanet: Spatial-channel attention network for 3d object detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2019, pp. 1992–1996.

[8] Jung Uk Kim, Sungjune Park, and Yong Man Ro, "Towards versatile pedestrian detector with multisensory-matching and multispectral recalling memory," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2022, pp. 1157–1165.

[9] Francisco Rivera Valverde, Juana Valeria Hurtado, and Abhinav Valada, "There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 11612–11621.

[10] Mizuho Wakabayashi, Hiroshi G Okuno, and Makoto Kumon, "Multiple sound source position estimation by drone audition based on data association between sound source localization and identification," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 782–789, 2020.

[11] Jintao Jiang, Gerasimos Potamianos, Harriet Nock, Giridharan Iyengar, and Chalapathy Neti, "Improved face and feature finding for audio-visual speech recognition in visually challenging environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2004, pp. 873–876.

[12] Karren Yang, Bryan Russell, and Justin Salamon, "Telling left from right: Learning spatial correspondence of sight and sound," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 9932–9941.

[13] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba, "Self-supervised moving vehicle tracking with stereo sound," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 7053–7062.

[14] Wangli Hao, He Guan, and Zhaoxiang Zhang, "Vag: A uniform model for cross-modal visual-audio mutual generation," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–13, 2022.

[15] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2551–2566, 2023.

[16] Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu, "Deep partial multi-view learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2402–2415, 2020.

[17] Jung Uk Kim, Sungjune Park, and Yong Man Ro, "Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1510–1523, 2021.

[18] Tommi Raij, Kimmo Uutela, and Riitta Hari, "Audiovisual integration of letters in the human brain," *Neuron*, vol. 28, no. 2, pp. 617–625, 2000.

[19] Gemma A Calvert, Peter C Hansen, Susan D Iversen, and Michael J Brammer, "Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the bold effect," *Neuroimage*, vol. 14, no. 2, pp. 427–438, 2001.

[20] Sabri Gurbuz, Zekeriya Tufekci, Eric Patterson, and John N Gowdy, "Multi-stream product modal audio-visual integration strategy for robust adaptive speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2002, pp. 2021–2024.

[21] Pan Zhou, Wenwen Yang, Wei Chen, Yanfeng Wang, and Jia Jia, "Modality attention for end-to-end audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2019, pp. 6565–6569.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[23] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 5127–5137.

[24] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," *Proc. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 18661–18673, 2020.

[25] Yifan Liu, Changyong Shu, Jingdong Wang, and Chunhua Shen, "Structured knowledge distillation for dense prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[26] Kafeng Wang, Xitong Gao, Yiren Zhao, Xingjian Li, Dejing Dou, and Cheng-Zhong Xu, "Pay attention to features, transfer learn faster cnns," in *Intl. Conf. Learning. Representations (ICLR)*, 2020.