There is More than Meets the Eye: Self-Supervised Multi-Object Detection and Tracking with Sound by Distilling Multimodal Knowledge

Francisco Rivera Valverde* Juana Valeria Hurtado* Abhinav Valada
University of Freiburg
{riverav, hurtadoj, valada}@cs.uni-freiburg.de

Abstract

Attributes of sound inherent to objects can provide valuable cues to learn rich representations for object detection and tracking. Furthermore, the co-occurrence of audiovisual events in videos can be exploited to localize objects over the image field by solely monitoring the sound in the environment. Thus far, this has only been feasible in scenarios where the camera is static and for single object detection. Moreover, the robustness of these methods has been limited as they primarily rely on RGB images which are highly susceptible to illumination and weather changes. In this work, we present the novel self-supervised MM-DistillNet framework consisting of multiple teachers that leverage diverse modalities including RGB, depth and thermal images, to simultaneously exploit complementary cues and distill knowledge into a single audio student network. We propose the new MTA loss function that facilitates the distillation of information from multimodal teachers in a self-supervised manner. Additionally, we propose a novel self-supervised pretext task for the audio student that enables us to not rely on labor-intensive manual annotations. We introduce a large-scale multimodal dataset with over 113,000 time-synchronized frames of RGB, depth, thermal, and audio modalities. Extensive experiments demonstrate that our approach outperforms state-of-the-art methods while being able to detect multiple objects using only sound during inference and even while moving.

1. Introduction

Human perception is deceptively effortless, we can sense people behind our backs and in the dark, although we cannot remotely see them. This elucidates that perception is inherently a complex cognitive phenomenon that is facilitated by the integration of various sensory modalities [11]. “There is More than Meets the Eye” aptly summarizes this complexity of our visual perception system. Modeling this ability using learning algorithms is, however, far from being solved.

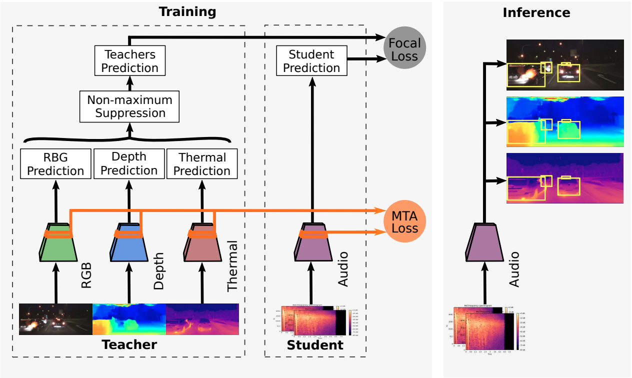

Figure 1. Our proposed cross-modal MM-DistillNet distills knowledge exploiting complementary cues from multimodal visual teachers into an audio student. During inference, the model detects and tracks multiple objects in the visual frame using only audio as input.

The natural co-occurrence of modalities such as images and audio in videos provides strong cues for supervision that can be exploited to learn more robust perception models in a self-supervised manner. Attributes of sound inherent to objects in the scene also contain rich time and frequency domain information that is valuable for grounding sounds within a visual scene. In this sense, the characteristics of sound are complementary and correlated to the visual information [40]. Cross-modal learning from images and sound exploits this natural correspondence between audio-visual streams that represent the same event. As a result, the integration of sound with vision enables us to use one modality to supervise the other as well as to use both modalities to supervise each other jointly [8, 6, 58].

Generally, training models to detect objects requires large amounts of groundtruth annotations for supervision. However, we can train models to recognize objects that produce sound without relying on labeled data by jointly leveraging audio-visual learning using the teacher-student strategy [22]. With this approach, numerous works [1, 49, 8] have used the audio-visual correlation to localize sounds sources. Moreover, recent work [46] has shown that we can exploit this

[*] Equal contribution.

audio-visual synchronicity to detect and track an object over the visual frame. Thus far, this promising capability has only been shown to be feasible in scenarios where the camera is static and for detecting a single object at a time using stereo sound and metadata containing camera pose information as input. Moreover, it distills knowledge only from models trained with RGB images, which are highly susceptible to perceptual changes such as varying types, scales, and visibility of objects, domain differences in terms of weather, illumination, and seasonality, among many others. Addressing these challenges will enable us to employ the system for detection and tracking in a wide variety of applications.

In this work, we present the novel self-supervised Multi-Modal Distillation Network (MM-DistillNet) that provides effective solutions to the aforementioned problems. Our framework illustrated in Fig. 1 consists of multiple teacher networks, each of which takes a specific modality as input, for which we use RGB, depth, and thermal to maximize the complementary cues that we can exploit (appearance, geometry, reflectance). The teachers are first individually trained on diverse pre-existing datasets to predict bounding boxes in their respective modalities. We then train the audio student network to learn the mapping of sounds from a microphone array to bounding box coordinates of the combined teachers' prediction, only on unlabeled videos. To do this, we present the novel Multi-Teacher Alignment (MTA) loss to simultaneously exploit complementary cues and distill object detection knowledge from multimodal teachers into the audio student network in a self-supervised manner. During inference, the audio student network detects and tracks objects in the visual frame using only sound as an input. Additionally, we present a self-supervised pretext task for initializing the audio student network in order to not rely on labor-intensive manual annotations and to accelerate training.

To facilitate this work, we collected a large-scale driving dataset with over 113,000 time-synchronized frames of RGB, depth, thermal, and multi-channel audio modalities. We present extensive experimental results comparing the performance of our proposed MM-DistillNet with existing methods as well as baseline approaches, which shows that it substantially outperforms the state-of-the-art. More importantly, for the first time, we demonstrate the capability to detect and track objects in the visual frame, from only using sound as an input, without any meta-data and even while moving in the environment. We also present detailed ablation studies that highlight the novelty of the contributions that we make. Finally, we make our dataset, code and models publicly available at http://rl.uni-freiburg.de/research/multimodal-distill.

2. Related Work

In recent years, several deep learning methods [6, 45, 23] have exploited the natural relationship between the co-occurrent vision and sound events found in video sequences. Some of these works rely on groundtruth annotations and propose supervised approaches to learn joint audio-visual embeddings by transferring knowledge between the modality-specific networks. Various tasks such as audio classification [34], lip reading [2], face recognition [31], and speaker identification [30] have been tackled using these techniques. Another set of approaches exploit self-supervision and they do not rely on any manual annotations. These methods exploit audio-vision synchronicity and learn representations of one modality while using the other counterpart modality. For example, Hershey *et al.* [21] use audio data as a supervision signal to learn visual representations, and Aytar *et al.* [8] propose SoundNet that uses visual imagery as supervision for acoustic scene classification.

More related to our work are methods that use audio-visual correspondence and vision data as a supervisory signal for localizing sound in a given visual input. This task is typically tackled using a pre-trained visual network as supervision [3, 6], by generating a common audio-visual representation [56, 32, 7, 35] or by using an attention mechanism [40, 37]. StereoSoundNet [19] performs object detection and tracking of a single-vehicle using a stereo microphone and camera pose information as the input. While on the other hand, Adanur *et al.* [1] and Ma *et al.* [28] demonstrate the advantages of using multiple microphones for spatial detection. Our proposed MM-DistillNet also performs object detection and tracking in the visual frame using only sound from a microphone array, allowing the system to detect multiple vehicles at the same time without using any camera pose information and while moving in the environment.

Given the low spatial resolution of sound, it is extremely complex and arduous to manually label audio for object localization over the visual domain. Recent techniques [56, 32, 19, 29, 45] address this problem by leveraging a vision-teacher's knowledge to supervise and generate the labels to train an audio-student network. Similarly, to reduce the groundtruth label dependency, our approach exploits the co-occurrence of modalities as a self-supervised mechanism to obtain groundtruth annotations. However, all of the aforementioned methods only use RGB images from the visual domain, which are highly susceptible to illumination and weather changes. To address this issue, several approaches have been proposed to leverage multiple modalities such as RGB, depth, and thermal images to exploit complementary cues by fusing them at the input or at the feature level [20, 44, 10]. Although these methods have substantially improved the performance of object detection and semantic segmentation in challenging perceptual conditions, they are still constrained by modality limitations such as range-of-vision or occlusions. Moreover, adding new modalities also increases the labeling effort and these fusion techniques typically require all the modalities to be present during inference time, both of which increase the overall system overhead. As opposed to these techniques, we propose a methodology to incorporate the knowledge from multiple pre-trained modality-specific teacher networks into an audio student network that learns from unlabeled videos and only uses audio during inference. Our approach exploits complementary features from the alternate modalities while training, in an effort to improve the robustness of the overall system without increasing the overhead at inference.

Besides for generating pseudo groundtruth labels, we employ the modality-specific teacher networks to guide the training of the audio student network via knowledge distillation. Previous works [43, 48, 13, 55] use the knowledge from the output logits by softening the labels. Our approach is more related to [38, 52, 51, 4], which transfers the knowledge from intermediate layers through an alignment loss function. Similar to [33, 57], our approach distills knowledge from multiple teachers. However, our framework does not merely average a dual loss among the teachers, rather it aligns the features of the intermediate teacher-student layers using a probabilistic approach. We show that the conditional knowledge given a synchronized set of modalities can improve the student network's performance.

In our approach, each modality-specific teacher distills object detection knowledge to the audio student, which can be categorized as cross-modal knowledge distillation. Salem *et al.* [39] proposes to distill the knowledge from the logits of a group of different visual modalities. Do *et al.* [15] and Zhang *et al.* [53] employ attention maps to combine the different modalities. These approaches require all the modalities, both during training and inference. Whereas our framework aims to disentangle the need for all the modalities during inference time. Alayrac *et al.* [5] recently propose to address this problem by creating an embedding with a contrastive pairwise loss that facilitates the downstream task. Nevertheless, their approach tackles cross-modal representations that are significantly different. Whereas, our approach distills knowledge from modality-specific teachers that are aligned at the object level, so the information from a common task is distilled into a complementary modality using our proposed MTA loss and the focal loss. We evaluate the performance of existing multi-teacher distillation losses with our proposed strategy in the ablation study.

3. Technical Approach

In this section, we detail our MM-DistillNet framework for distilling the knowledge from a set of pre-trained multimodal teachers into a single student that employs an unlabeled modality as input. We choose RGB, depth, and thermal images as the teacher modalities, and audio from an 8-channel monophonic-microphone array for the student. Specifically, our goal is to learn a mapping from spectrograms of ambient sounds to bounding box coordinates that indicate the vehicle location in the visual space. In our framework illustrated in Fig. 2, each pre-trained modality-specific teacher predicts bounding boxes that indicate where the vehicles are located in their respective modality space. These predictions are fused to obtain a single multi-teacher prediction, which is then used as a pseudo label for training the audio student network. To effectively exploit the complementary cues from the modality-specific teachers, we employ our proposed Multi-Teacher Alignment (MTA) loss to align the intermediate representations of the student with that of the teachers. In the rest of this section, we first describe the architecture of the teacher-student networks and the teacher pre-training procedure, followed by the novel pretext task that we propose for better initializing the audio student. We then describe the methodology that we propose for distilling knowledge from multimodal teachers to a single student and finally, the approach we employ to track vehicles over successive frames.

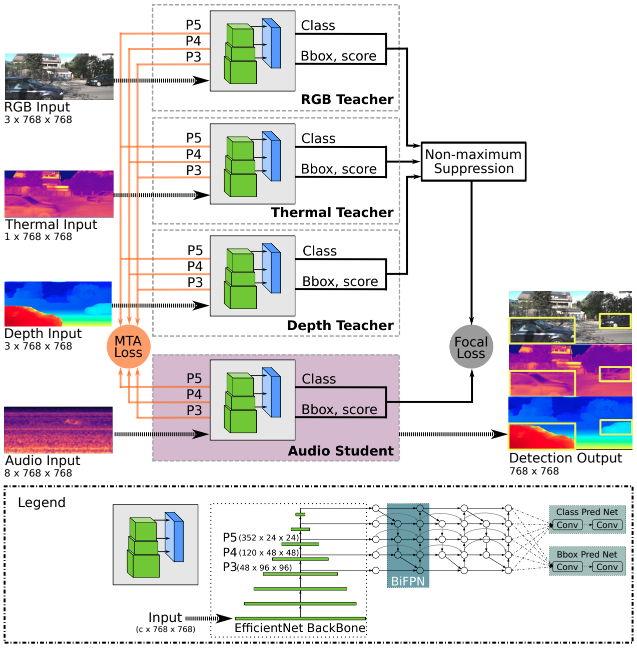

Figure 2. Our proposed MM-DistillNet framework consists of three pre-trained modality-specific teachers built upon EfficientDet-D2 that predict bounding boxes in the visual space and an audio student network that takes spectrograms of sound from a microphone array as input. By exploiting the co-occurrence of the modalities, we train the audio student to regress the bounding boxes predicted by the teachers using focal loss and our proposed MTA loss that aligns the intermediate network layers. During inference, the student detects multiple moving vehicles in the visual frame using only sound.

3.1. Network Architecture

We build upon the EfficientDet [42] architecture for the modality-specific teacher networks. EfficientDet has three main components: an EfficientNet [41] backbone, followed by a bidirectional feature pyramid network, and a final regression and classifier branch. The EfficientNet architecture uses multiple stages of mobile inverted bottleneck units [50] to extract relevant features from the input data. There are eight variants of this backbone, ranging from B0 to B7 on increasing capacity demand. This allows for trade-off performance and prediction speed, which is achieved through a compound scaling coefficient that uniformly scales the network's depth and width along with the input image resolution. To select from which stages of EfficientNet the features are extracted (and how such features are fused together), EfficientDet introduces a weighted bidirectional feature pyramid through a combination of automatic machine learning and manual tuning. The last stage of the network is a classifier and regressor branch that consist of a sequence of separable convolutions, batch normalization, and a memory-efficient swish [36].

For this work, we find that EfficientDet-D2 gives us the best speed versus performance trade-off, as demonstrated in the additional experiments in the supplementary material. It is important to note that our framework is not dependent on a specific teacher architecture, as alternate object detection networks can be readily incorporated as drop-in replacements. We use an input image resolution of 768×768 pixels, with 5 BiFPN cell repetitions, with 112 channels each. We illustrate the EfficientDet architecture in the legend shown in Fig. 2. The teacher networks in our framework are comprised of:

- **RGB teacher** that we train on COCO [26], PASCAL VOC [16], and ImageNet [14] for the *car* labels.
- **Depth teacher** that we train on the Argoverse [12] dataset using 3D *vehicle* bounding boxes mapped to 2D. Note that Argoverse does not provide direct depth/disparity data. Therefore, we generate it from stereo images using the Guided Aggregation Net [54].
- **Thermal teacher** that we train on the FLIR ADAS [18] dataset for the *car* and *other vehicle* labels.

The audio student network in our MM-DistillNet framework learns to detect vehicles as a regression problem. We adopt the same EfficientDet-D2 topology for the audio student network, which takes eight spectrograms concatenated channel-wise representing the ambient sounds from an 8-channel monophonic-microphone array, as input and predicts bounding boxes localizing vehicles in the visual reference frame. To do so, we first obtain an RGB, depth, and thermal image triplet at a given timestamp, each of which has a resolution of 1920×650 pixels. Subsequently, we select one second ambient sound clips from the microphone array, centered on the image timestamp and we generate a 80×173 pixels spectrogram for each of the eight microphones using Short-Time Fourier Transform (STFT). We further detail this procedure in Sec. 4.2. We then resize the spectrograms to a resolution of 768×768 pixels to match the input scale of the teachers. Given this 8-channel concatenated spectrograms as input, the audio student yields 4 coordinates $(x_{min}, y_{min}, x_{max}, y_{max})$ for each of the EfficientNet layers at different aspect ratios and scales (EfficientDet uses by default 3 aspect ratios $(1.0, 1.0), (1.4, 0.7), (0.7, 1.4)$ at 3 different scales $[2**0, 2**(1.0/3.0), 2**(2.0/3.0)]$.

3.2. Self-Supervised Pretext Task for Audio Student

As the input to our audio student network is an 8-channel spectrogram, we cannot leverage pre-trained weights for initializing the EfficientDet architecture, such as from models trained for image detection, which typically take a 3-channel image as the input. It has consistently been shown that models initialized with pre-trained weights from large datasets perform significantly better than models trained from scratch. More recently, self-supervised pretext tasks that learn semantically rich representations by exploiting the supervisory signal from the data itself have shown promising results, even outperforming models initialized with pre-trained weights.

Inspired by this recent progress, we propose a simple pretext task for the audio student that counts the number of cars present in the scene. This task aims at enabling the student to learn audio representations depicting the number of vehicles in the visual field, only using an 8-channel spectrogram as input. To do so, we first use the predictions of multiple pre-trained teachers to identify the number of cars present in the image. Subsequently, we use the corresponding 8-channel spectrogram as the input to EfficientNet with an MLP classifier at its output and we train the network with the cross-entropy loss function to predict the number of cars in the scene. We then use the weights from the model trained on this pretext task to initialize the audio student network in our MM-DistillNet framework while training to detect cars in the visual frame from spectrograms of sound as input.

3.3. Knowledge Distillation from Multiple Teachers

To train the audio student network to detect vehicles in the visual frame, given the sound input, we use two different loss functions. First, we employ an object detection loss function at the final prediction of the networks, as shown in Fig. 2. Second, we use our Multi-Teacher Alignment (MTA) loss function to align and exploit complementary cues from the intermediate layers of the modality-specific teachers with the audio student. Given that we use multiple teachers, we also obtain multiple sets of bounding box predictions. Each teacher network receives only its input modality and predicts a tuple of bounding boxes, which correspond to their best individual estimation of where the vehicles are located in the visual space. There are often scenes in which each modality-specific teacher predicts a different number of bounding boxes. Therefore, we need to consolidate such predictions. To do so, we obtain three sets of tuples coming from the RGB, depth and thermal teachers, which are consolidated using non-maximum suppression with intersection over union $IoU = 0.5$. This generates a unified prediction from the modality-specific teachers, which is enforced on the student using the Focal loss (FL) [25]. Focal loss is a form of cross-entropy loss with a penalizing parameter that reduces the relative loss for well-classified examples, allowing the network to focus on the training examples that are hard to classify. The Focal loss is given by

$$L_{focal} = -\alpha(1-pt)^{\gamma} * log(pt), \tag{1}$$

where α is the weight assigned to hard examples (set to $\alpha = 0.25$) and, γ is a focusing hyperparameter to balance how much effort to put on hard to classify examples against easy background cases (set to $\gamma = 2.0$).

With our proposed MTA loss, we aim to exploit complementary cues contained in the intermediate layers of each modality-specific teacher. In order to achieve this, we train the student network in a manner such that the distribution of activations in specific layers of both the student and the multiple teachers are aligned. Particularly, we enforce the alignment of the $(p3, p4, p5)$ layers of the EfficientNet backbone, as shown in Fig. 2. To do so, we compute the distribution of activations using the attention map of each layer normalized to a $[0,1]$ range. We compute the student attention map as $Q_s^j = F_{avg}^r(A_s)$, where F_{avg} is a function that collapses the activation tensor A in its channel dimension through the average of the neuron's output at the given layer $j \in \{P3, P4, P5\}$, and r is the exponential over each of the $i-th$ elements of the vector, a hyperparameter that trades-off how much importance to give to high valued activations versus low-valued activations at a given layer.

In the case of the teacher networks, the activation distributions of each modality $P(A_{t_i}|m_i)$ indicates the confidence of each teacher that given an input modality m_i, the intermediate representations have a high likelihood of detecting a relevant key indicator of a vehicle. With this in mind, we propose to leverage the attention maps of the multiple teachers by means of the product of the modality-specific activation distributions at the selected layers. We assume that the modalities are independent and we use the chain rule of probability so that $P(A_{t_i}, ..., A_{t_N}|m_i, ..., m_N) = P(A_{t_i}|m_i) * ... * P(A_{t_N}|m_N)$. We believe this assumption is reliable as the teachers have been trained on disjoint datasets, with modalities extracted using different sensor hardware. The intuition behind this idea is to incorporate the knowledge of each modality-specific teacher in an incremental approach. We can consider each pre-trained teacher as a vehicle-sensor engine, and our loss, a mechanism of integrating new measurements in the Bayesian's context. If multiple modalities agree on a bounding box, the probability of this proposal is encouraged. Nevertheless, a modality can also propose a disjoint bounding box with a small probability, allowing the student to learn bounding boxes exclusive to a particular modality. We effectively estimate the probability of detecting a car in a scene, given the privileged knowledge of each modality. This allows for the flexibility to also incorporate other knowledge as confidence scores for each bounding box, so that we reduce the occurrence of false predictions. Therefore, we compute the multi-teacher attention map as $Q_t^j = \prod_i^N F_{avg}^r(A_{t_i})$, where i denotes each of the N considered modalities. Formally, we define our Multi-Teacher Alignment (MTA) loss as

$$L_{MTA} = \beta * \sum_j KL_{div}\left(\frac{Q_s^j}{\left\|Q_s^j\right\|_2}, \frac{Q_t^j}{\left\|Q_t^j\right\|_2}\right), \tag{2}$$

where the summation iterates over each of the selected EfficientNet layers from the inverted pyramid (e.g., p3, p4, and p5 layers), s and t stand for student and teacher, and $\beta = 0.5$ is used for loss balancing. We denote this loss function as Multi-Teacher Alignment loss, as it integrates different modalities privileging the agreement of different inputs while still considering vehicle predictions proposed by one modality. Finally, we optimize our MM-DistillNet framework with the weighted summation of the focal loss and our proposed MTA loss as

$$L_{total} = \delta * L_{focal} + \omega * L_{MTA}, \tag{3}$$

where L_{total} enforces knowledge transfer from the teachers at the output, as well as at the intermediate network layers.

3.4. Tracking

We adopt an approach similar to that of Gan *et al.* [19] for object tracking. Specifically, we leverage the detected bounding boxes, and we use the IoU values between boxes of consecutive frames to relate the objects to the same tracklet. We set the IoU threshold to 0.5 to assign two bounding boxes from different timesteps to the same object. We initialize a tracklet each time an object is detected with a confidence score higher than 0.8. The next bounding box related to that tracklet is selected by comparing it to the current frame's detection. The association process between the tracklet and a bounding box is made so that it maximizes the IoU. The tracklet is set as inactive if there are no bounding boxes with $IoU > 0.5$ in the subsequent frames. Given that our contribution with multiple modality-specific teachers improves the quality of the bounding boxes and the number of detected objects, we also expect to enhance the tracking accuracy with this method that primarily relies on IoU object matching.

4. Experimental Evaluation

In this section, we first describe the data collection methodology that we employ, followed by the protocol that we use for training the MM-DistillNet framework. We then present quantitative results comparing our approach to several strong baselines as well as the state-of-the-art. Subsequently, we present detailed ablation studies and qualitative evaluations to demonstrate the novelty of our contributions.

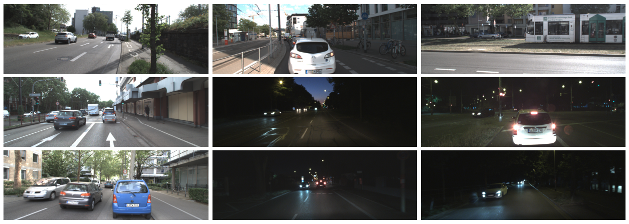

Figure 3. Example images from our MAVD dataset showing diverse scenes with multiple moving vehicles and low-illumination conditions captured with a camera mounted on a moving car.

4.1. Multimodal Audio-Visual Detection Dataset

As there are no publicly available datasets that consist of synchronized audio, RGB, depth, and thermal images, we collected a large-scale Multimodal Audio-Visual Detection (MAVD) dataset in autonomous driving scenarios. The dataset was gathered from 24 car drives during 3 months and at 20 different locations. Each drive has an average of half hour duration. We recorded data on diverse scenarios ranging from highways to densely populated urban areas and small towns. The recordings consist of high traffic density, freeway driving, and multiple traffic lights (involving transition from static to driving conditions). To capture diverse noise conditions, we recorded sounds not only during conventional city driving but also near trams and while going through tunnels.

We provide two types of scenarios, static condition in which the car is motionless and nearly 300 km of driving data. Our dataset contains three cars on average for every image (ranging from 1 to a maximum of 13 cars per scene). We only retained the images with at least one car in the scene. The subset that we use for training the detection stage contains 24589 static day images, 26901 static night images, 26357 day driving images, and 35436 night driving images, amounting to a total of 113283 synchronized multi-channel audio, RGB, depth, and thermal modalities. Additionally, the dataset also contains GPS/IMU data and LiDAR point clouds. An image showing the data collection vehicle and the sensor setup is shown in the supplementary material. The sensors that we used include an RGB stereo camera rig (FLIR Blackfly 23S3C), a thermal stereo camera rig (FLIR ADK), and eight monophonic microphones in an octagon array. The audio was recorded and stored in the 1-channel Microsoft WAVE format with a sampling rate of 44100 Hz. All the sensor data, including the microphone recordings were synchronized to each other via the GPS clock. Example scenes from the dataset are shown in Fig. 3.

4.2. Training Protocol

Data Split: We use a 60/20/20% split for training, validation, and testing. The validation split was used to perform hyperparameter optimization with Hyperband [17].

Evaluation Metric: We use the standard mean average precision metric for evaluating object detection performance and the center distance proposed by Gan *et al.* [19]. Mean average precision is the mean over classes of the interpolated area under each class's precision and recall curve. The Center distance *CDx* and *CDy* metrics indicate the prediction accuracy because the spatial information is not directly available for the audio (possible error between the predicted bounding box center and the groundtruth).

Training Setup: We train for 50 epochs with ReduceLRonPlateau learning rate scheduler and an initial learning rate of $1e-5$, weight decay of $5e-4$, $betas = (0.9, 0.999)$, and Adam optimizer. For the MTA loss, we use $r = 2.0$ and $temperature = 9.0$ as this selection of hyperparameters provided the best results for individual modalities. We provide additional details in the Supplementary Material. For our loss calculation, we set $\delta = 1.0$ and $\omega = 0.05$ as these settings provided the best performances (more details can be found in the Supplementary Material). The original resolution of all RGB/depth/thermal images is 1920×650. We resize them to be 768×768 as per [42] D2 variant. For the audio, we extract 0.5 seconds before and 0.5 seconds after the registered timestamp, an RGB image was taken. We normalize this 1-second raw waveform and further resample it on a Mel-frequency scale with 80 bins resulting in 8 $(80, 173)$ arrays. This is further normalized to [0-1] and re-scaled to 768×768×8 dimensionality.

4.3. Quantitative Results

In order to evaluate the performance of multi-teacher distillation from different modalities, we compare the performance of our MM-DistillNet with *StereoSoundNet* [19] which uses a single RGB teacher with the Ranking loss to distill the information into an audio student network. We also compare with several strong baselines: *2M-DistillNet Audio* employs a single RGB teacher with our proposed MTA loss to train an audio student network. The comparison with this baseline enables us to evaluate the performance of our proposed MTA loss over the Ranking loss. In order to evaluate the performance of using other modalities representing an object in the student network, we compare with *2M-DistillNet Depth* and *2M-DistillNet Thermal* that use an RGB teacher to train a depth student or a thermal student, respectively using our MTA loss. The comparison with these two models shows the significance of using the audio modality in the student network. Finally, we compare with *MM-DistillNet Avg* that uses a straightforward approach to combine the predictions from RGB, depth, and thermal teachers by averaging the individual modality-specific network activations. Here, we assume that all bounding boxes predicted by any of the modalities are valid (after applying non-maximum suppression with IoU=0.5). The comparison with this baseline demonstrates the utility of our MTA loss function to effectively distill multimodal knowledge from the teaches. All the aforementioned baselines use the 8-channel

Network	mAP@	mAP@	mAP@	CDx	CDx
	Avg	0.5	0.75		
StereoSoundNet [19]	44.05	62.38	41.46	3.00	2.24
2M-DistillNet RGB	57.25	68.01	59.15	2.67	2.13
2M-DistillNet Depth	55.41	66.83	57.30	2.60	2.10
2M-DistillNet Thermal	56.70	69.15	58.63	2.43	1.98
MM-DistillNet Avg	51.63	66.14	52.24	2.14	1.80
MM-DistillNet (Ours)	**61.62**	**84.29**	**59.66**	**1.27**	**0.69**

Table 1. Comparison of cross-modal multi-object detection performance with several baselines. '2M-DistillNet Teacher' refers to 2-modal distillation approach to train the audio student using our MTA loss. 'MM-DistillNet Avg' refers to averaging individual modality-specific teacher activations.

Loss Function	KD	mAP@	mAP@	mAP@	CDx	CDx
		Avg	0.5	0.75		
Ranking loss [19]	RGB	44.05	62.38	41.46	3.00	2.24
Pairwise loss [27]	RGB	40.45	59.72	36.73	2.98	2.20
AFD loss [47]	RGB	44.27	62.00	41.90	3.19	2.28
Avg. Ranking loss	R,D,T	56.16	80.03	52.96	1.46	0.80
Avg. AFD loss	R,D,T	58.50	82.18	55.48	1.30	0.70
Avg. MTA loss	R,D,T	59.46	82.29	56.94	1.35	0.73
MTA loss (Ours)	RGB	44.58	62.66	42.39	2.94	2.17
MTA loss (Ours)	R,D,T	**61.62**	**84.29**	**59.66**	**1.27**	**0.69**

Table 2. Comparison of various loss functions for Knowledge Distilation (KD). All the models were trained with the same MM-DistillNet architecture but with different loss functions. 'R,D,T' refers to RGB, Depth, and Thermal teachers. Avg. Loss averages the individual modality-specific teacher activations.

Approach	MOTA↑	ID Sw.↓	Frag.↓	FP↓	FN↓
StereoSoundNet [19]	16.94%	1327	1077	3696	3349
MM-DistillNet (Ours)	**26.96%**	**1078**	**1076**	**2758**	**3524**

Table 3. Comparison of tracking performance.

Model	Teacher	Student	mAP@	AP@	AP@
	Modalities	Pretext	Avg	0.5	0.75
M1	RGB	-	44.58	62.66	42.38
M2	RGB, Depth	-	42.89	62.07	39.67
M3	RGB, Thermal	-	55.81	79.84	54.67
M4	Depth, Thermal	-	44.79	65.14	41.82
M5	RGB, Depth, Thermal	-	61.10	83.81	59.07
M6	RGB, Depth, Thermal	✓	**61.62**	**84.29**	**59.66**

Table 4. Ablation study on influence of various modality-specific teachers and self-supervised pretext task of audio student.

spectrogram from the microphone array as input and are trained to perform multi-object detection.

We present quantitative comparisons using the same EfficientDet-D2 topology with pre-trained weights as detailed in Sec. 3 for all the baselines, as well as our MM-DistillNet model. Results from this experiment is shown in Tab. 1. We can observe how the knowledge of different teachers improves the performance over the previous state-of-the-art *StereoSoundNet* [19], using the same input (audio only). Furthermore, our baseline *2M-DistillNet Audio*, which also uses an RGB-teacher to train an audio student, yields superior performance than StereoSoundNet. This demonstrates that our MTA loss function outperforms the Ranking loss. Tab. 1 also elucidates that audio is a valuable modality to detect moving vehicles. We also observe that combining the prediction of individual RGB, depth, and thermal teachers using averaging does not improve the performance. Nevertheless, we can see that our proposed MM-DistillNet with our MTA loss function exploits complementary cues from the multimodal teachers and facilitates effective distillation.

We also evaluate the performance of our MTA loss against other knowledge distillation techniques. Tab. 2 compares the baseline loss used by [19], as well as the pairwise loss conventionally used for similar embedding task [5] and an attention based loss with learnable parameters [47]. We use the audio student and a single RGB teacher. Our loss is intended to distill knowledge from multiple teachers into a single student, yet, in the individual teacher case, it provides appealing results in the object detection task. In the supplementary material, we provide further comparisons against different modalities. Additionally, for multiple teachers, previous methods such as [57] proposed to average the predictions of multiple teachers. Tab. 2 shows the effect of integrating the prediction of the teachers rather than computing an average. Furthermore, Fig. 4 provides an intuition on how the predictions of multiple teachers are integrated into the student by visualizing the activations. We obtain the activation using Score-CAM [46] from the P5 layer of EfficientNet. In particular, we show that the baseline's activations are linked only to the RGB teacher, whereas our method's activation is the product of the activation of the RGB, depth and thermal teachers. Furthermore, the thermal teacher in this night setting is the privileged modality which is able to predict cars under poor light conditions. In Tab. 3, we report the multiple object tracking accuracy (MOTA), identity switches (ID Sw.), fragment (Frag.), false positive (FP) and false negative (FN) as evaluation metrics [9, 24] on a subset of our test dataset. We excluded scenes that involved tracking of multiple cars for a fair comparison with StereoSoundNet.

4.4. Ablation Studies

In Tab. 4, we ablate the effect of incrementally adding modalities and the impact of our pretext task. It can be seen that RGB and thermal are the main contributors to performance improvement. This can be attributed to the performance of modalities in the day and night, respectively. Nevertheless, integrating depth improves performance, there-

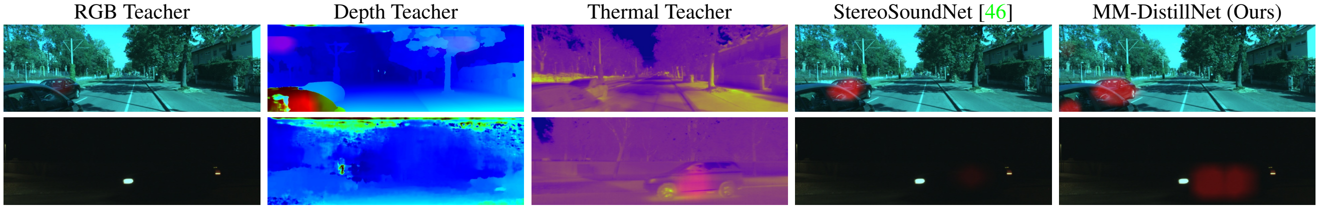

Figure 4. Comparison of activations visualization from modality-specific teachers with the previous state-of-the-art StereoSoundNet [46], and our proposed MM-DistillNet. High activations (in red) indicate regions where a vehicle is likely to be detected.

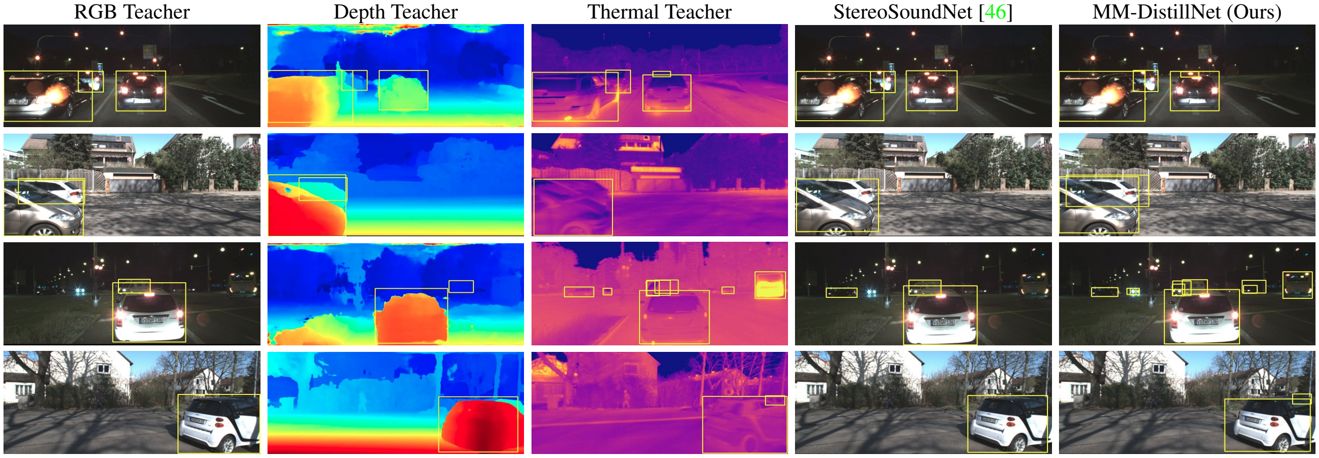

Figure 5. Qualitative comparisons of the predictions from individual modality-specific teachers with the previous state-of-the-art StereoSoundNet [46], and our MM-DistillNet. Our network consistently detects moving vehicles even in scenes where the baselines fail.

fore we still employ a depth teacher. Moreover, our pretext task shows an improvement of 0.52. Additionally, there is an average reduction of 27.55% in the loss value, indicating that the proposed weight initialization accelerates training. We refer to the supplementary material for more details.

4.5. Qualitative Evaluations

In this section, we qualitatively evaluate the performance of our proposed MM-DistillNet framework. The audio modality is able to overcome certain limitations of visual sensors, as demonstrated in Fig. 5. The first row highlights how our approach enables us to use the knowledge of the pre-trained teachers to improve the audio student's predictions. The baseline fails to predict a car that the RGB only teacher cannot see. As our model distills knowledge from all the teachers, our MM-DistillNet proactively detects the cars that are not visible to the RGB camera, in this case, coming from the thermal teacher. Our framework also facilitates better student learning, which is highlighted in the second row of Fig. 5. Even though the RGB teacher detects two cars in the image, the baseline does not learn enough cues to predict two cars. Our model uses the RGB and depth teacher to re-enforce the fact that there are two cars in the scene. Our work is not limited to two cars in the scene as in [19]. We attribute this capability to the incorporation of audio from the microphone array. Finally, the last row of Fig. 5 shows how our model can predict cars that are not visible in any of the modalities, such as occluded cars entering the scene.

5. Conclusions

This paper proposed a self-supervised framework to distill the knowledge from different expensive sensor modalities into a more accessible one. We do so by leveraging the co-occurrence of modalities and the fact that there exist pre-trained networks for object detection in the visual domain. We use a self-supervised scheme to label audio spectrograms for object detection. During training, we use RGB, depth, and thermal teachers to improve the training of a student network; this enables us to require only audio during inference time. Our results demonstrates how audio is a robust alternative to traditional sensor modalities used in autonomous driving, particularly in overcoming visual limitations. We also publicly released our large-scale MAVD dataset. We compared our approach to different baselines, including different numbers and combinations of modalities, losses, and configurations. We presented qualitative results that highlight the ability of our models to overcome visual limitations such as occlusions and thereby facilitate new applications.

6. Acknowledgments

This work was partly funded by the Eva Mayr-Stihl Stiftung and the Graduate School of Robotics in Freiburg.

References

[1] Resul Adanur, Yıldıray Yeşilyurt, Cem Şişman, Selim Sağır, and Ismail Kaya. Deep learning for audio signal source positioning using microphone array. In *Seventh International Conference on Digital Information Processing and Communications (ICDIPC)*, 2019. 1, 2

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Asr is all you need: Cross-modal distillation for lip reading. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020. 2

[3] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. *arXiv preprint arXiv:2008.04237*, 2020. 2

[4] Gustavo Aguilar, Yuan Ling, Yu Zhang, Benjamin Yao, Xing Fan, and Chenlei Guo. Knowledge distillation from internal representations. In *AAAI*, 2020. 3

[5] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *arXiv preprint arXiv:2006.16228*, 2020. 3, 7

[6] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1, 2

[7] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018. 2

[8] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, 2016. 1, 2

[9] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 7

[10] Wolfram Burgard, Abhinav Valada, Noha Radwan, Tayyab Naseer, Jingwei Zhang, Johan Vertens, Oier Mees, Andreas Eitel, and Gabriel Oliveira. Perspectives on deep multimodel robot learning. In *Robotics Research*, pages 17–24. Springer, 2020. 2

[11] Gemma A Calvert, Michael J Brammer, and Susan D Iversen. Crossmodal identification. *Trends in cognitive sciences*, 2(7):247–253, 1998. 1

[12] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 4

[13] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 3

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*, 2009. 4

[15] Tuong Do, Thanh-Toan Do, Huy Tran, Erman Tjiputra, and Quang D Tran. Compact trilinear interaction for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 3

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 4

[17] Stefan Falkner, Aaron Klein, and Frank Hutter. Bohb: Robust and efficient hyperparameter optimization at scale. *arXiv preprint arXiv:1807.01774*, 2018. 6

[18] FLIR. Free flir thermal dataset for algorithm training. 4

[19] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2, 5, 6, 7, 8

[20] Frank Hafner, Amran Bhuiyan, Julian FP Kooij, and Eric Granger. A cross-modal distillation network for person reidentification in rgb-depth. *arXiv preprint arXiv:1810.11641*, 2018. 2

[21] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 2

[22] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. 1

[23] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[24] Yuan Li, Chang Huang, and Ram Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2953–2960. IEEE, 2009. 7

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2017. 5

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, 2014. 4

[27] Yifan Liu, Changyong Shu, Jingdong Wang, and Chunhua Shen. Structured knowledge distillation for dense prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 7

[28] Wei Ma and Xun Liu. Phased microphone array for sound source localization with deep learning. *Aerospace Systems*, 2(2):71–81, 2019. 2

[29] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. In *Advances in Neural Information Processing Systems*, 2018. 2

[30] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Learnable pins: Cross-modal embeddings for person iden-

tity. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 71–88, 2018. 2

[31] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 2

[32] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[33] SeongUk Park and Nojun Kwak. Feed: Feature-level ensemble for knowledge distillation. *arXiv preprint arXiv:1909.10754*, 2019. 3

[34] Andres Perez, Valentina Sanguineti, Pietro Morerio, and Vittorio Murino. Audio-visual model distillation using acoustic images. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2854–2863, 2020. 2

[35] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. *arXiv preprint arXiv:2007.06355*, 2020. 2

[36] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017. 4

[37] Janani Ramaswamy and Sukhendu Das. See the sound, hear the pixels. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2970–2979, 2020. 2

[38] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 3

[39] Tawfiq Salem, Connor Greenwell, Hunter Blanton, and Nathan Jacobs. Learning to map nearly anything. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019. 3

[40] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018. 1, 2

[41] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019. 3

[42] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3, 6

[43] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019. 3

[44] Abhinav Valada, Ankit Dhall, and Wolfram Burgard. Convoluted mixture of deep experts for robust semantic segmentation. In *IEEE/RSJ International conference on intelligent robots and systems (IROS) workshop, state estimation and terrain perception for all terrain mobile robots*, 2016. 2

[45] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Semantic object prediction and spatial sound super-resolution with binaural sounds. *arXiv preprint arXiv:2003.04210*, 2020. 2

[46] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 24–25, 2020. 1, 7, 8

[47] Kafeng Wang, Xitong Gao, Yiren Zhao, Xingjian Li, Dejing Dou, and Cheng-Zhong Xu. Pay attention to features, transfer learn faster cnns. In *International Conference on Learning Representations*, 2019. 7

[48] Tiancheng Wen, Shenqi Lai, and Xueming Qian. Preparing lessons: Improve knowledge distillation with better supervision. *arXiv preprint arXiv:1911.07471*, 2019. 3

[49] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 1

[50] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 4

[51] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[52] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 3

[53] Chenrui Zhang and Yuxin Peng. Better and faster: knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification. *arXiv preprint arXiv:1804.10069*, 2018. 3

[54] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 4

[55] Zizhao Zhang, Han Zhang, Sercan O Arik, Honglak Lee, and Tomas Pfister. Distilling effective supervision from severe label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3

[56] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 2

[57] Peng Zhou, Long Mai, Jianming Zhang, Ning Xu, Zuxuan Wu, and Larry S Davis. M2kd: Multi-model and multi-level knowledge distillation for incremental learning. *arXiv preprint arXiv:1904.01769*, 2019. 3, 7

[58] Jannik Zürn, Wolfram Burgard, and Abhinav Valada. Self-supervised visual terrain¨ classification from unsupervised acoustic feature learning. *IEEE Transactions on Robotics*, 2020. 1