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

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

In this supplementary material, we provide additional experimental results that support the novelty of the contributions made and evaluate our architectural design choices. Particularly, we first provide further details regarding the data collection methodology of our Multimodal Audio-Visual Detection (MAVD) dataset in Sec. 1. We then compare the performance of different EfficientDet variants in our proposed MM-DistillNet framework in Sec. 2.1. In Sec. 2.2, we evaluate the performance of our framework with sound from a varying number of microphones as input. Subsequently, we demonstrate the capabilities of our multi-teacher single-student framework to distill knowledge from RGB images into other modalities that are complementary to sound in Sec. 2.3. We then present ablation studies on the influence of various hyperparameters in our proposed MTA loss function in Sec. 2.4 and we compare the performance of our MTA loss with other widely employed knowledge distillation losses in Sec. 2.5. Subsequently, we present additional results on our proposed self-supervised pretext task for the audio student in Sec. 3. Then, we present results of our framework in low illumination conditions such as nighttime and dusk in Sec. 4. Finally, we extend our qualitative results with numerous examples in Sec. 5. We made the code and models of our approach publicly available at http://rl.uni-freiburg.de/research/multimodal-distill.

1. MAVD Dataset

Our approach employs four different synchronized modalities, including three visual: depth, thermal and RGB, and additional sound. We collected our MAVD dataset using a car with a rack of sensors mounted on the roof as shown in Fig. 1. Stereo images were captured using a pair of FLIR Blackfly 23S3C configured to a resolution of 1920×650 pixels and the thermal images were captured at the same resolution using a pair of FLIR ADK cameras. We employ a targetless calibration method [8] for aligning the camera images by formulating a misalignment minimization problem. The miss-alignment is computed as the difference between the gradients of the calibrated RGB image and the transformed thermal image, in the RGB coordinate frame. However, due

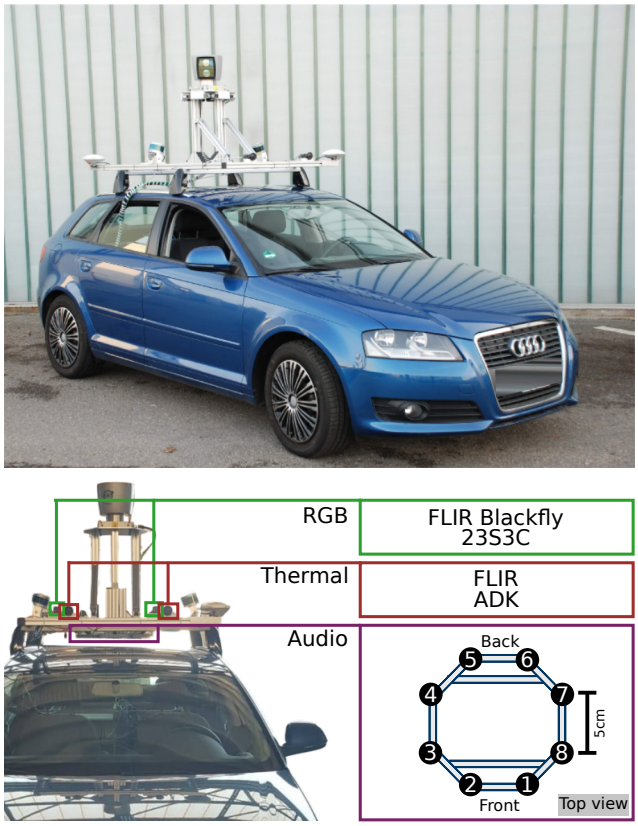

Figure 1. Top: The data collection vehicle that we use for our MAVD dataset. Bottom: Closeup view showing the 8-microphone array, stereo cameras and thermal cameras mounted on the roof of the vehicle. The vehicle also contains LiDAR, IMU and GPS, which we also collected for our dataset.

*Equal contribution

The authors would like to thank Johan Vertens for assistance in the data collection.

to the high dimensional nature of this problem, there are ambiguities that cannot be easily resolved without prior information. To this end, the method from [8] resolves these ambiguities through a pre-calibration of the camera intrinsics by a pre-processing that align predominant edges of objects that are common to the RGB and thermal cameras.

We employ a stereo rectification and undistortion method with radtan distortion coefficients of (-0.20077378832448342, 0.06858744821624758, -8.318933053823812e-05, 0.0006149164090634826) and camera intrinsics of (1010.5596834500378, 1010.1723409131672, 975.7863331505446, 297.2804298854754). Every RGB and thermal image pair in our dataset is GPS clock synchronized with nano-second precision. The same clock timestamp is used to identify the central audio frame corresponding to the thermal and RGB images. We then identify the frame number in the audio clip using a sampling rate of 44 100 Hz, and sample 1 second around the RGB-thermal timestamp. In order to obtain the depth image, we use the network proposed by Zhang *et al.* [10] with the left and right images from the stereo rig. We set the maximum disparity to 192 and apply a jet color map to leverage the ImageNet pre-trained weights for initializing the EfficientDet backbone. We made our dataset publicly available at http://rl.uni-freiburg.de/research/multimodal-distill.

2. Extended Ablation Study

2.1. EfficientDet Compound Coefficient Selection

EfficientDet is a family of object detection models proposed by Tan *et al.* [7] which contains eight different architectural configurations that trade-off performance and runtime. We evaluated the performance of the variants to identify their suitability for our framework. To this end, we created a large dataset by combining Microsoft COCO [5], PASCAL VOC [2], and ImageNet [1]. Subsequently, we removed all the scenes that do not contain at least one vehicle and retained the bounding boxes of those that contain cars or moving vehicles. We then trained EfficientDet D0-D7 to detect the objects in this combined dataset. In Tab. 1, we present the performance in terms of the Average Precision (AP) at $IoU = 0.5$, as well as the inference time and the Floating Point Operations Per Second (FLOPS). We chose EfficientDet D2 as the backbone of our framework, given that it presents a higher improvement in the average performance with a lesser increase in the inference time. Nevertheless, our framework provides the flexibility to adopt any of the other variants as a direct drop in replacement.

2.2. Influence of Number of Microphones

Our proposed MM-DistillNet framework exploits complementary cues from different modalities such as RGB, depth, and thermal images during training to incorporate multimodal information into a single audio network. The input to the audio network is from a microphone array that captures ambient sounds. We employ multiple monophonic microphones due to the promising results that it has demonstrated for sound source localization [6]. Fig. 1 shows the microphone array that we employ mounted on our data collection vehicle. In order to estimate the number of microphones that are essential to reliably localize objects, we analyze the performance of our MM-DistillNet for varying number of microphones in a balanced subset of the dataset.

Namely, we hypothesize that there is a relationship between the number of microphones and the complexity of the scene, as measured by the number of vehicles in the environment. Our MAVD dataset contains varying number of vehicles in each scene, with a maximum of 13 vehicles in a single scene. We performed experiments to analyze the improvement that we can achieve by using more number of microphones in the array. To do so, we need to have balanced number of vehicles in the dataset. Therefore, we apply an under-sampling approach to ensure that the number of examples with varying number of vehicles are balanced.

EfficientDet Variant	AP@ 0.5	FLOPS	Inference Time (ms)
D0	0.5165	2.5B	30.04
D1	0.6870	6.1B	34.76
D2	**0.7974**	**11B**	**39.99**
D3	0.8134	25B	59.45
D4	0.8680	55B	89.93
D7	0.9200	325B	388.90

Table 1. Performance comparison of different EfficientDet variants for predicting bounding boxes of vehicles in Microsoft COCO [5], PASCAL VOC [2], and ImageNet [1], with the associated AP at 0.5 IoU.

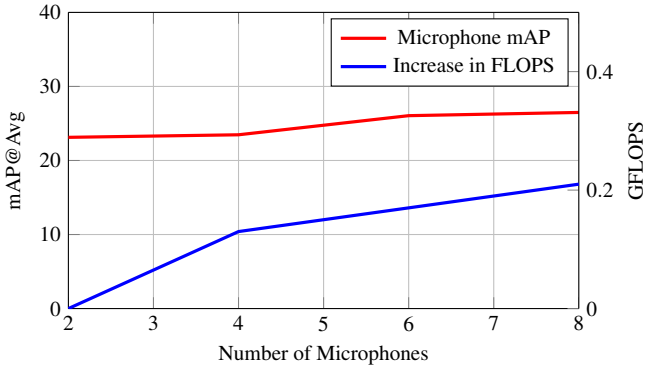

Figure 2. The left axis (blue line) shows the performance of the network vs the number of microphones. The right axis (red) shows the GLOPS increase caused by using N microphones. It can be seen, that more channels improve the performance in the given task with negligible impact in FLOPS.

Teachers	Student	mAP@ Avg
RGB	Sound	57.25
RGB	Thermal	56.70
RGB, Depth, Thermal	Sound	61.62
RGB, Depth, Thermal	RGB	81.12
RGB, Depth, Thermal	Thermal	**81.98**

Table 2. Our framework distills the knowledge from multimodal teachers trained on dataset where supervision is available, to improve the learning of a single student network. Thermal and RGB modalities show significant improvement over their single teacher counterparts.

We compute the performance of the audio student using multiples of two number of microphones, in order to always consider the microphones that are further away from each other in the hexagonal array.

Fig. 2 shows that increasing the number of microphones consistently improves the performance of the model. Therefore, we utilize sound from all the eight monophonic microphones that are available in our octagonal setup for our experiments. Nevertheless, given that each microphone adds an additional $768 \times 768 \times 1$ input to the network, we computed the additional overhead in terms of the increase in number FLOPS in the network. Fig. 2 shows that the increase in FLOPS in the network due to the addition of a microphone is negligible compared to the overall FLOPS of EfficientDet as shown in Tab. 1.

2.3. Distillation to Different Student Modalities

By taking advantage of the co-ocurrence of all the modalities present in our dataset (audio, RGB, thermal and depth), it is straightforward to interchange the input modality of the student network from audio to any of the other modalities as described in Sec. 3 of the main paper. By doing so, we can exemplify not only how our method is modality independent, but also how it can further improve the performance of existing object detection frameworks.

In the other experiments, we selected RGB, depth, and thermal images as the teacher modalities and sound as the student modality for our framework. This enables us to tackle limitations of visual modalities such as occlusions and sensor sensitivity (poor performance of RGB cameras during nighttime as well as the limited sensitivity of thermal cameras during the day). Nevertheless, our approach to transfer the knowledge from multiple pre-trained modality-specific networks to a student network, is input agnostic. Under this perspective, instead of employing sound as input to the student network, we can also use conventional RGB, depth, or thermal modalities as input to our framework. Tab. 2 presents results with different modalities as input to the student network. It can be seen that the performance against the single teacher is substantially improved. Particularly, using the thermal modality as input to the student network provides the best performance, as it provides substantial cues for vehicle detection in both day as well as night recordings. Whereas, using RGB as input to the student during night suffers due to low illumination conditions. Nevertheless, using a thermal modality requires expensive hardware and is subject to visual limitations like occlusion. For this reason, we employ audio in our MM-DistillNet as an alternative to the traditional visual inputs used in autonomous driving.

2.4. Influence of Hyperparameters in MTA Loss

We define our proposed Multi-Teacher Alignment (MTA) loss function as

$$L_{MTA} = \beta * \sum_{j} KL_{div}\left(\frac{Q_s^j}{\left\|Q_s^j\right\|_2}, \frac{Q_t^j}{\left\|Q_t^j\right\|_2}\right), \tag{1}$$

where $Q_s^j = F_{avg}^r(A_s)$ and $Q_t^j = \prod_i^N F_{avg}^r(A_{t_i})$ are the student and multi-teacher attention maps respectively. Additionally, while computing the KL_{div} as a measure of the difference between probability distributions between the teachers and the student, we can apply a temperature t to the softmax as proposed by Hinton *et al.* [4]. This temperature in the softmax computation is added to adapt the confidence on each individual probability distribution. To this end, we can expand the above notation using the student normalized activation $S_x = \frac{Q_s^j}{\left\|Q_s^j\right\|_2}$ and the integrated normalized teacher attention $T_x = \frac{Q_t^j}{\left\|Q_t^j\right\|_2}$ as

$$KL_{div}\left(S_x||T_x\right) = \sum_{x \in \mathscr{X}} \frac{e^{\frac{S_{x,i}}{t}}}{\sum_{k=1}^{n} e^{\frac{S_{x,k}}{t}}} \log\left(\frac{\frac{e^{\frac{S_{x,i}}{t}}}{\sum_{k=1}^{n} e^{\frac{S_{x,k}}{t}}}}{\frac{e^{\frac{T_{x,i}}{t}}}{\sum_{k=1}^{n} e^{\frac{T_{x,k}}{t}}}}\right) \tag{2}$$

We employ this loss in addition to the focal loss L_{focal} to optimize our MM-DistillNet framework as

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

With this formulation, we have two hyperparameters in the MTA loss function that can be selected according to the specific task: the exponential value r, which controls the relevance of small valued activations in contrast to large valued activations, and the softmax temperature t. We performed experiments to study the influence of these two hyperparameters on the performance of the audio student network. Results from this experiment is shown in Tab. 3. We present the mean average precision of our best recipe using a single RGB teacher and the audio student network.

We observe that in this setting, a value of $r = 4$ and $t = 9$ provides the best result. This suggests that putting

r	t	mAP@ Avg
4	9	60.13
4	6	61.40
3	4	64.08
3	2	64.30
2	6	64.30
2	4	66.05
2	7	66.43
1	3	67.68
1	4	67.74
1	9	69.67
2	9	**69.72**

Table 3. Ablation study of hyperparameters r and temperature t in the proposed MTA loss function. The object detection performance is shown for an audio student with knowledge distilled from a RGB teacher network.

Input	mAP@ Avg	
	Ranking Loss [3]	MTA Loss (Ours)
RGB	56.37	**57.25**
Thermal	55.82	**56.70**
Depth	45.85	**55.41**

Table 4. Our MTA loss function was formulated to distill knowledge from multiple teachers to a single student. Nevertheless, it yields superior performance than Ranking loss employed by Gan *et al.* [3] for knowledge distillation in a single teacher-student scenario.

more importance to higher valued activations rather than low-valued activations improves the overall performance of the network.

2.5. Ranking Loss vs. MTA Loss

The proposed MM-DistillNet framework exploits the co-occurrence of different modalities representing a scene. Our MTA loss facilitates this process by aligning the intermediate representations of multiple teachers to that of a single student. The previous state-of-the-art [3] method employs the Ranking loss to distill knowledge from a single teacher to a single student network. To further demonstrate the capabilities of our proposed MTA loss, we compare its performance with Ranking loss for distillation of knowledge from a single modality-specific teacher to an audio student network. We performed this experiment for all the teacher modalities considered in this work. Although our loss function is designed to align different intermediate representations of modality-specific networks, we observe that it outperforms the Ranking loss formulation even in the single teacher setting.

3. Evaluation of Audio Student Pretext Task

Our MM-DistillNet contains multiple modality-specific teachers and a single student network. Each of the networks are composed of the EfficientNet backbone which has to be

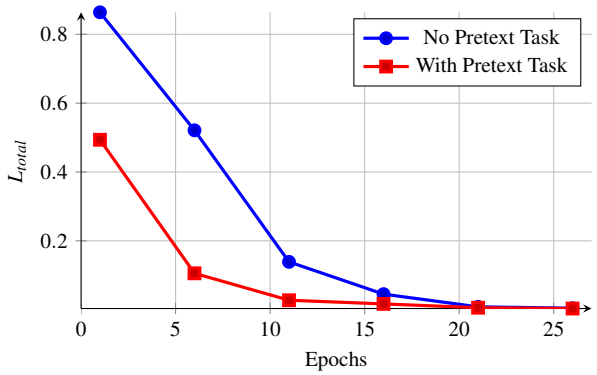

Figure 3. Comparison of training convergence of our MM-DistillNet, with and without the initialization of the student network with weights of the model trained on our proposed self-supervised pretext task.

initialized with pre-trained weights from large datasets to ease the optimization and achieve better convergence. Since all existing pre-trained models of EfficientNet primarily employ 3-channel RGB images as input, they cannot be used for initializing the audio student network which takes 8-channel spectrograms as input (1-channel spectrogram from each of the 8 microphones in the array). To address this problem, we propose a self-supervised pretext task that provides the audio student network with semantically rich information about the relationship between the audio and visual modalities. The goal of the pretext task is to estimate the number of vehicles present in the RGB image only using sound as input to the network. In the ablation study presented in the main paper, we show that our proposed pretext task improves the performance in terms of detection metrics. In Fig. 3 of this supplementary material, we present comparisons of the training curves for the MM-DistillNet framework, with and without the initialization of the audio student network with weights of the model trained on the proposed pretext task. We can see that the model with weighted initialized from the pretext task consistently yields a lower loss since the early stages. Moreover, the final loss is 27.55% lower than the model trained from scratch. These results demonstrate that the pretext task not only improves the performance in terms of the metrics, it also accelerates training and leads to faster convergence.

4. Evaluation in Low Illumination Conditions

In this section, we compare the performance our proposed MM-DistillNet and the previous state-of-the-art StereoSoundNet [3] in different illumination and driving conditions. StereoSoundNet is trained under the supervision of the RGB teacher, whereas our MM-DistillNet uses the RGB, depth, and thermal teachers. Tab. 5 presents results in terms of the average precision metric for each of these conditions. We can see that in every scenario our proposed

Condition	Vehicle State	Network	mAP@ Avg	mAP@ 0.5	mAP@ 0.75	CDx	CDx
Day	Static	StereoSoundNet [3]	53.48	69.10	52.50	2.77	1.51
Night	Static	StereoSoundNet [3]	38.13	49.26	34.67	4.34	3.95
Day	Driving	StereoSoundNet [3]	45.59	69.20	42.84	2.50	1.60
Night	Driving	StereoSoundNet [3]	28.56	45.18	24.43	3.86	2.77
Day	Static	MM-DistillNet	63.80	83.90	63.63	1.59	0.78
Night	Static	MM-DistillNet	75.10	89.63	73.23	1.43	0.70
Day	Driving	MM-DistillNet	55.73	81.51	52.75	1.24	0.76
Night	Driving	MM-DistillNet	48.13	67.16	46.07	1.93	1.50

Table 5. Performance comparison of MM-DistillNet in different illumination conditions as well as driving and static (data collection vehicle) states. Our framework outperform the previous state-of-the-art even in static-day conditions, where the RGB teacher performs the best.

Figure 4. Qualitative comparison of detection performance with the previous state-of-the-art StereoSoundNet [9] and our MM-DistillNet. We present a scenario with an occluded car to demonstrates the novelty of using sound which overcomes limitations of visual modalities.

MM-DistillNet substantially outperforms StereoSoundNet, thereby achieving state-of-the-art performance. We observe the largest improvement during night time conditions where StereoSoundNet significantly falls behind. It can also be observed that even during the day when the data collection vehicle is not in motion, our MM-DistillNet achieves over 10% improvement in the mAP @ Average. We observe a performance drop in both the methods from static to driving conditions, which can be attributed to the distortion of sound due to the moving data collection vehicle, wind on the microphone (note that the microphones in the array were not equipped with a wind muff) and high ambient noise conditions. We believe that high fidelity microphones and hard negative mining will help overcome this problem. This experiment demonstrates that incorporating knowledge from multiple modality-specific teachers improves the performance of the audio student, especially while the data collection vehicle is in motion and in low illumination conditions.

5. Extended Qualitative Results

In this section, we extend the qualitative evaluations of our proposed MM-DistillNet. We provide further results that demonstrate that our MM-DistillNet effectively employs the knowledge of diverse multimodal pre-trained teachers to improve the performance of vehicle detection. We first highlight how the audio modality is able to overcome the visual limitation by detecting occluded cars in Fig. 4 where the white car at time t is occluded by the vehicle that appears on the left at $t+1$. However, StereoSoundNet fails to detect the foreground car, while our MM-DistillNet precisely detects both the vehicles in the scene, despite the background car not being visible.

In Fig. 5, we present comparison of both methods during nighttime where poor light conditions can be observed. Our MM-DistillNet simultaneously detects multiple cars, even the distant ones, while StereoSoundNet often fails to detect beyond one vehicle. These results demonstrate the novelty of distilling multimodal knowledge in our MM-DistillNet

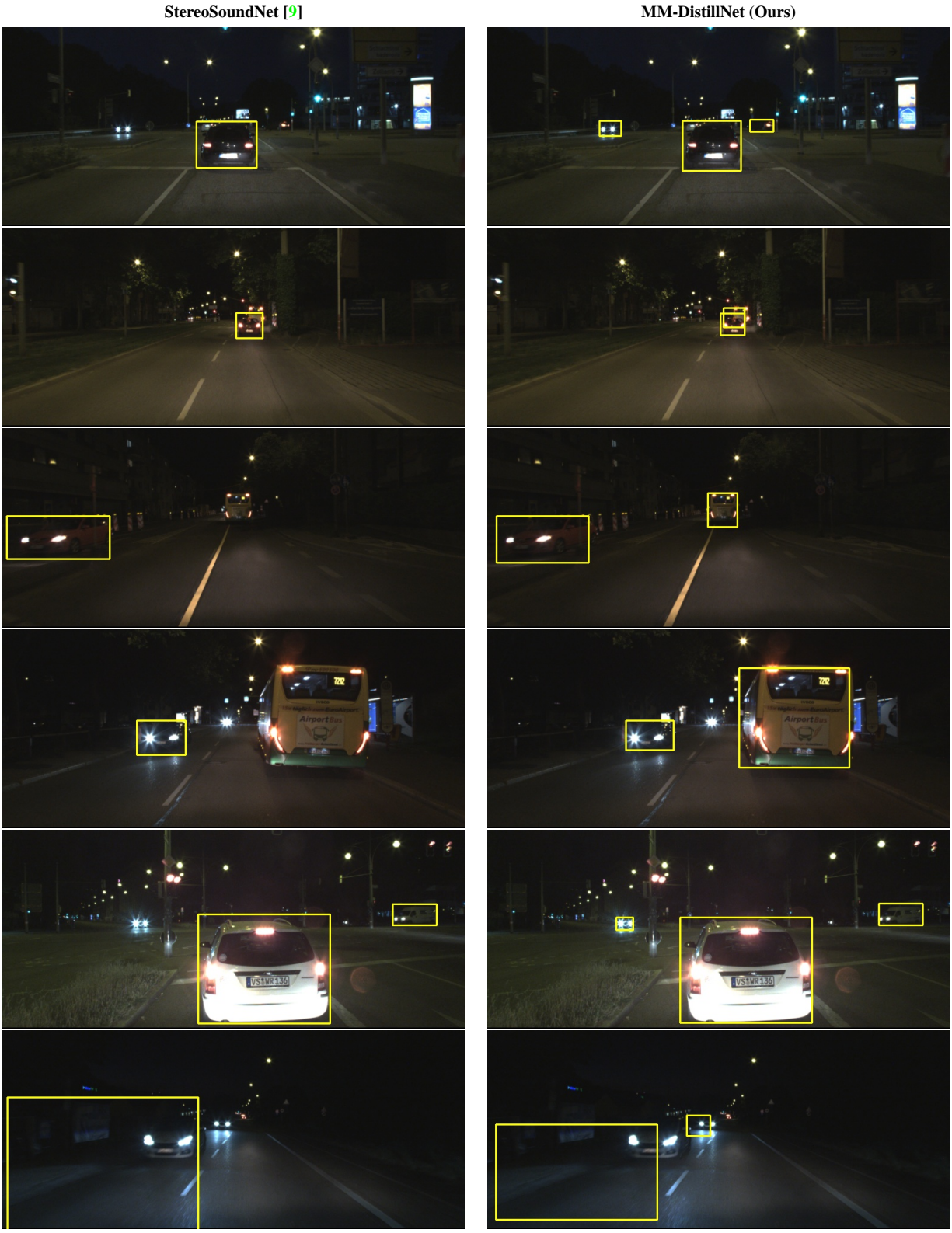

Figure 5. Qualitative comparisons of predictions in night scenes from our MM-DistillNet and the single student-teacher StereoSoundNet [9]. We present hard scenarios in poor lighting condition to demonstrate how our model does not suffer from the day to night domain gap.

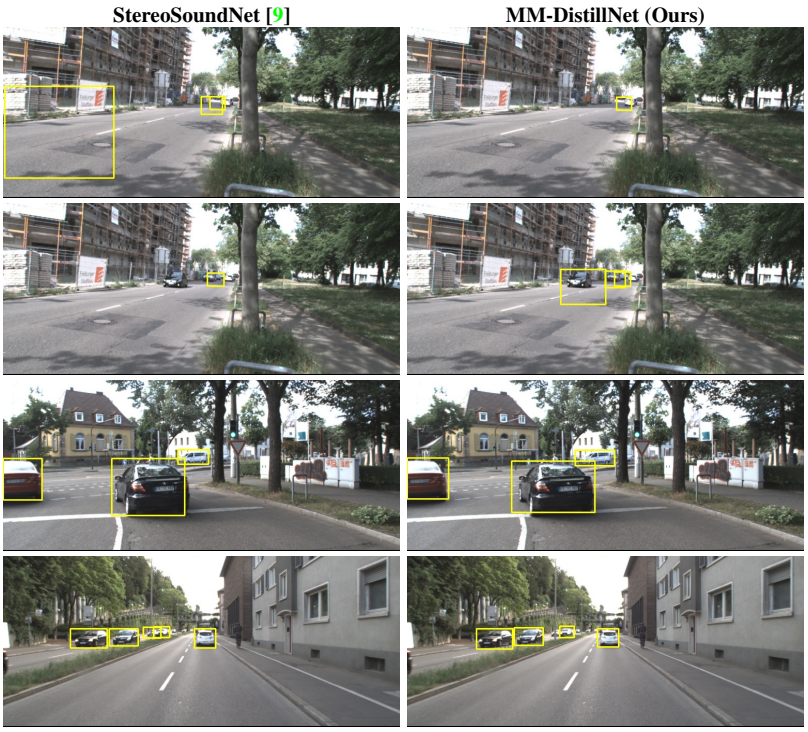

Figure 6. Qualitative comparisons of predictions of our MM-DistillNet and the previous state-of-the-art StereoSoundNet [9] in scenarios with multiple cars. We show that our network is able to simultaneously detect multiple cars, even while the data collection car is moving. Observe that our network is also able to detect very distant cars.

as it shows substantial robustness in poor light conditions, thereby successfully overcoming the limitations of distilling knowledge only from RGB images. Additionally, we qualitatively evaluate the performance of detecting multiple vehicles simultaneously in Fig. 6. Simultaneously detecting multiple vehicles with only sound is an extremely challenging task due to its low spatial resolution. By distilling knowledge from multiple modality-specific teachers, we show that it is not only feasible to detect vehicles simultaneously without relying on the arduous data labeling process, the performance of the detections also substantially improves. Further enhancing this ability will enable a broad spectrum of applications of these audio approaches in real world scenarios. Even though our MM-DistillNet is able to provide very promising results for object detection and tracking, the audio modality suffers from limitations of its own that are highlighted in examples shown in Fig. 7. We observe that occasionally, distant objects are not detected and multiple distant objects are detected as a single object. We believe that a microphone with better sensitivity as well as more examples of this phenomena in the training set will enable our approach to improve the performance in these conditions. Finally, we compare the performance of our MM-DistillNet and the modality-specific RGB, depth, and thermal teachers in Fig. 8. The results show the weaknesses and strengths of each of the selected modalities. Especially in night scenarios, we can observe how the thermal teacher contributes to the distillation of knowledge to the audio student as it reliably detects cars that are not visible in the RGB images, due to low illumination conditions.

References

[1] 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. 2

[2] 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. 2

[3] 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. 4, 5

[4] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. 3

[5] 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. 2

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

Figure 7. Qualitative comparisons of predictions from our MM-DistillNet and the previous state-of-the-art StereoSoundNet [9]. We present failure cases in these examples that include far away cars whose sounds are not sufficiently captured by the microphones.

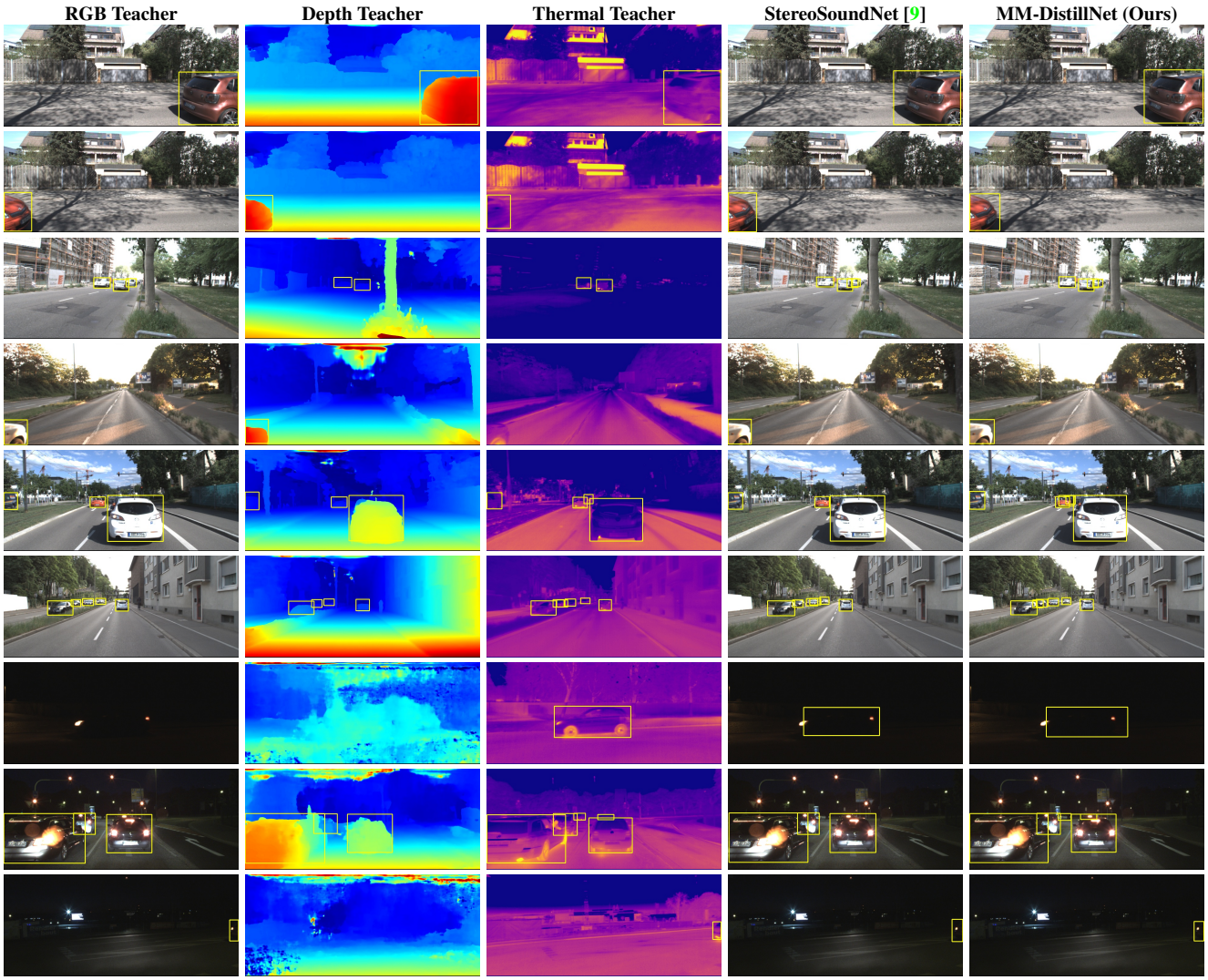

Figure 8. Qualitative comparisons of predictions from individual modality-specific teachers with the previous state-of-the-art StereoSoundNet [9], and our MM-DistillNet. Our network consistently detects moving vehicles even in the scenes where the baselines fail.

[7] 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. 2

[8] Johan Vertens, Jannik Zürn, and Wolfram Burgard. Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. *arXiv preprint arXiv:2003.04645*, 2020. 1, 2

[9] 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. 5, 6, 7, 8, 9

[10] 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. 2