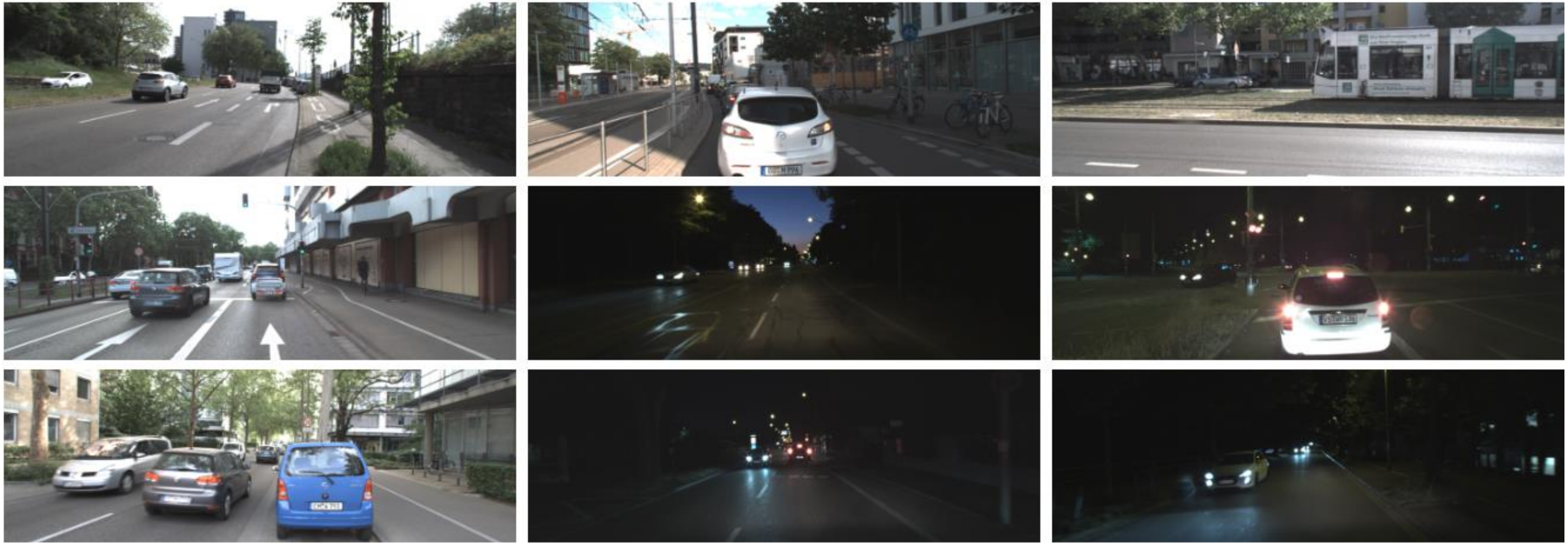
There is More than Meets the Eye: Self-Supervised Multi-Object Detection and Tracking with Sound by Distilling Multimodal Knowledge

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

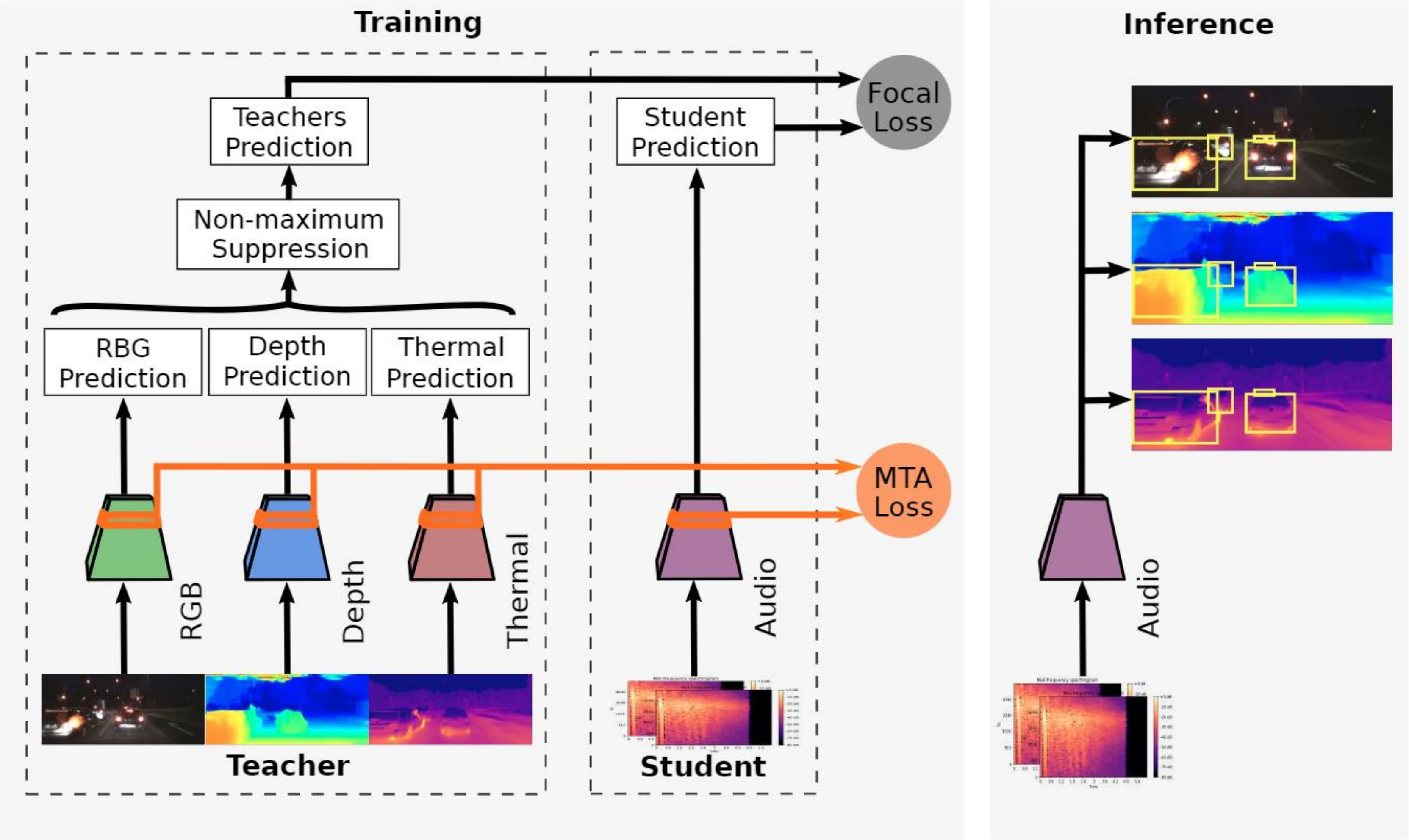
CVPR VIRTUAL JUNE 19-25

Presented by Diego Machado

Motivation



Method

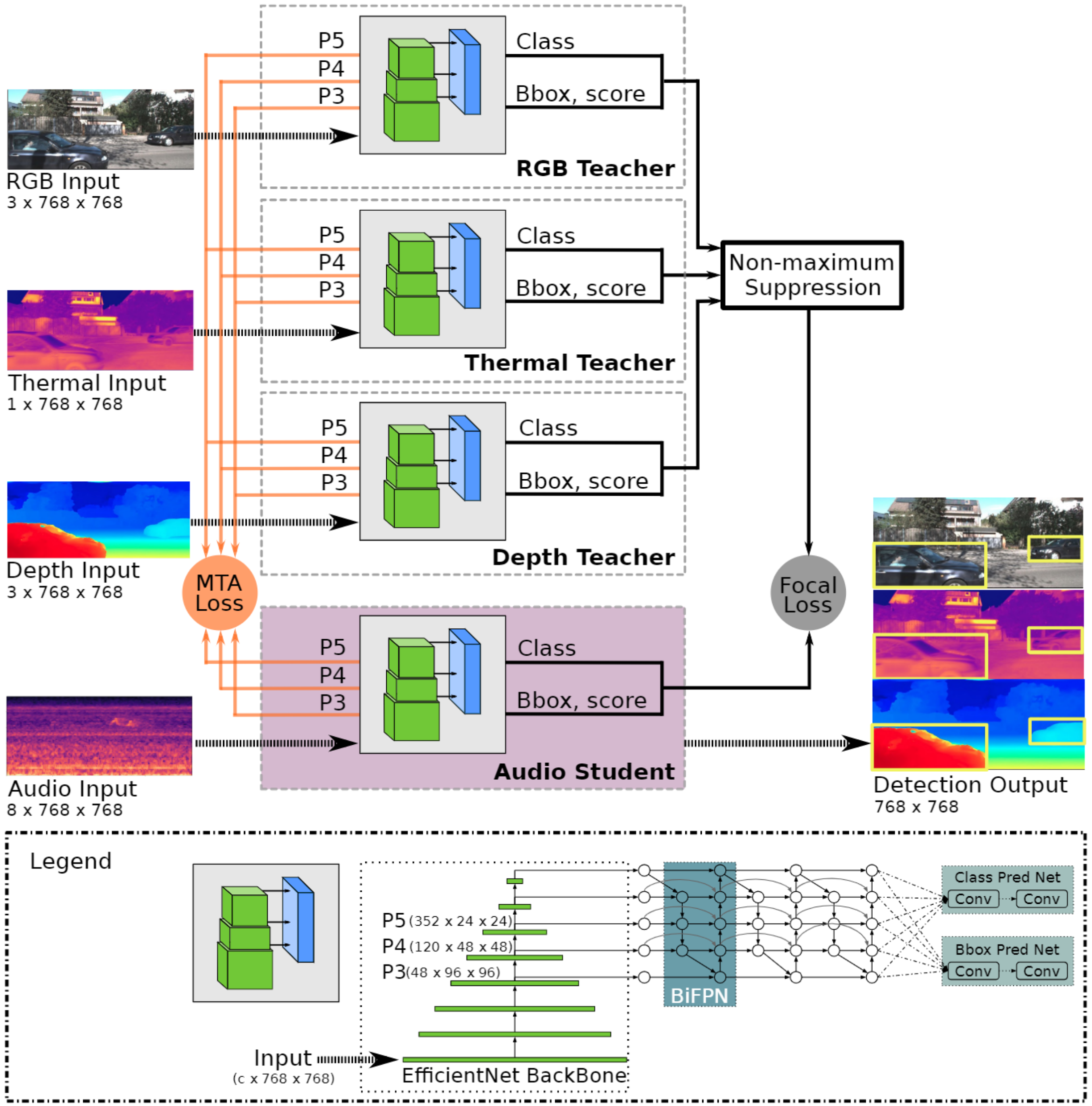


$$L_{focal} = -\alpha(1-pt)^{\gamma} * log(pt)$$

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

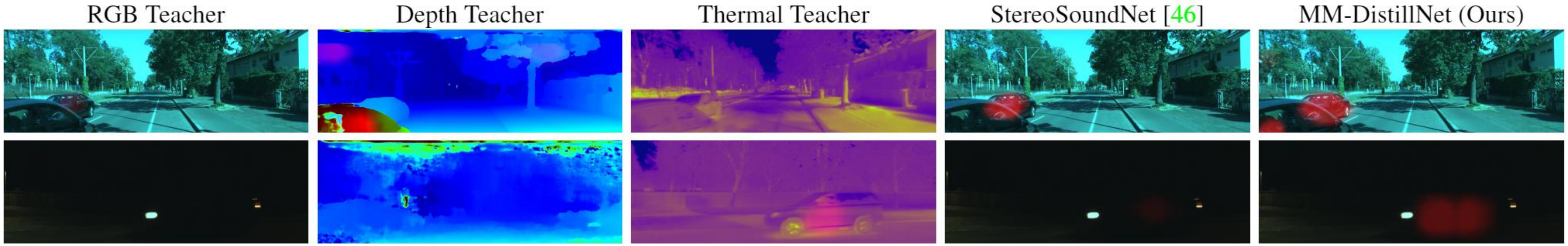
$$L_{total} = \delta * L_{focal} + \omega * L_{MTA}$$

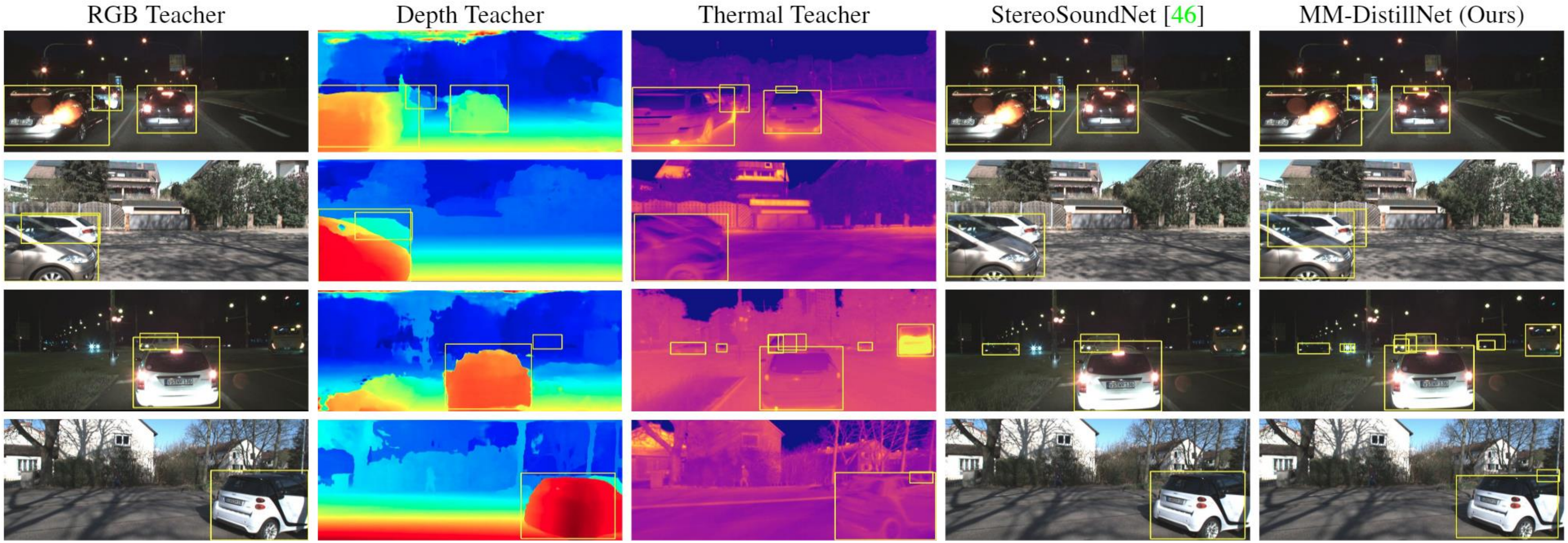
Method



P5
P4
P3
Class
Bbox, score
RGB Teacher
RGB Input
3 x 768 x 768
Thermal Teacher
Thermal Input
1 x 768 x 768
Depth Teacher
Depth Input
3 x 768 x 768
MTA Loss
Non-maximum Suppression
Focal Loss
Audio Student
Audio Input
8 x 768 x 768
Detection Output
768 x 768
Legend
P5 (352 x 24 x 24)
P4 (120 x 48 x 48)
P3 (48 x 96 x 96)
BiFPN
Class Pred Net
Conv
Conv
Bbox Pred Net
Conv
Conv
Input
(c x 768 x 768)
EfficientNet BackBone

Results


RGB Teacher
Depth Teacher
Thermal Teacher
StereoSoundNet [46]
MM-DistillNet (Ours)


RGB Teacher
Depth Teacher
Thermal Teacher
StereoSoundNet [46]
MM-DistillNet (Ours)

Results – Baseline comparison

Network	mAP@ Avg	mAP@ 0.5	mAP@ 0.75	CDx	CDx
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**

Results – Loss comparison

Loss Function	KD	mAP@ Avg	mAP@ 0.5	mAP@ 0.75	CDx	CDx
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**

Results – Ablation studies

Model	Teacher Modalities	Student Pretext	mAP@ Avg	AP@ 0.5	AP@ 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**

Results – Tracking performance

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

Extras

Method – Q (activation maps)

bone, 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.

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

Recording details

- 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.

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.

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.

3 modality input (teacher)

RGB, Thermal, Depth → EfficientDet for extracting img. features

non compression maximum to select matching bounding boxes across img. modalities

Focal loss to learn right bounding boxes

→ Joint loss

student → Audio

same time stamp
(audio → ± 0.5 sec.)

efficientDet → MTA loss → for distill knowledge / match distribution between img. & audio

same input dim as teachers (768 x 768)
↳ with 8-channels

KL Divergence of attention maps.