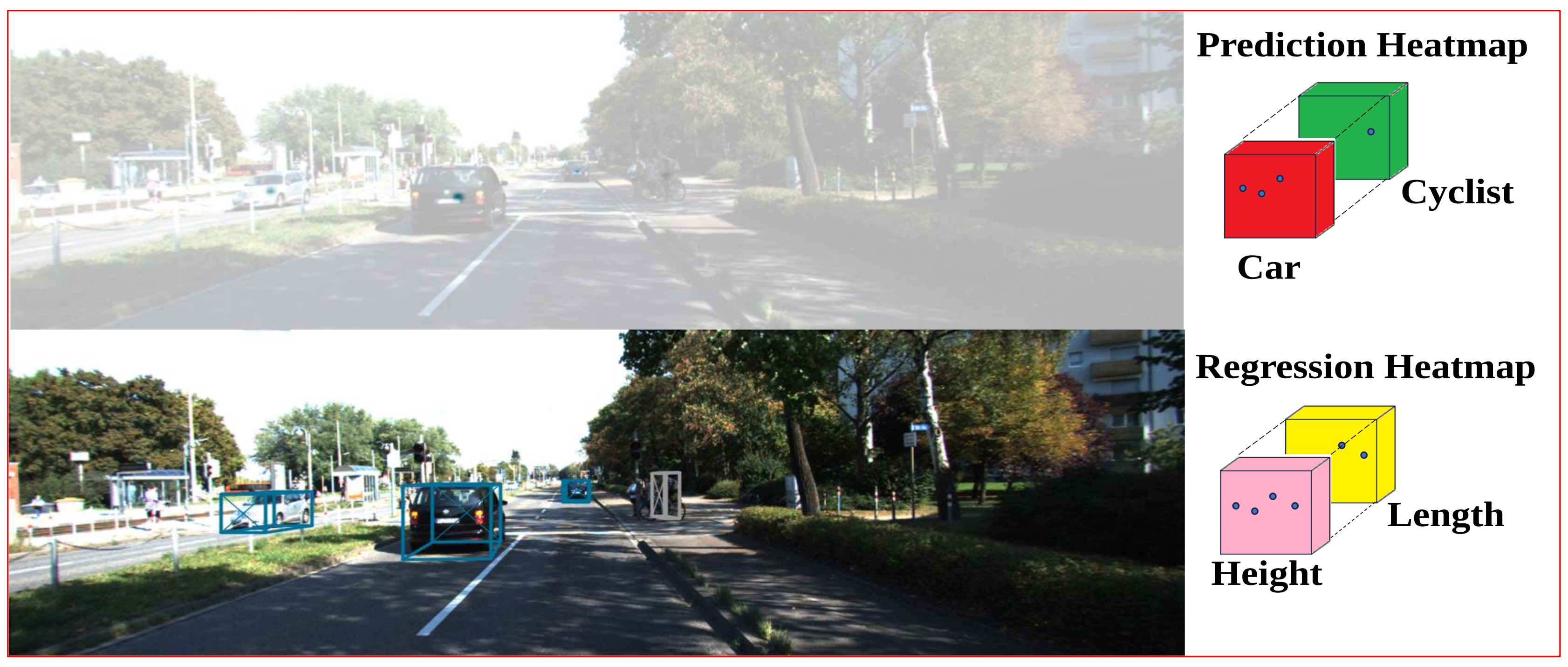
Accurate and Fast Monocular 3D Object Detection with Adaptive Feature Aggregation Centric Enhance Network

Peng-Wei Lin, and Chih-Ming Hsu
National Taipei University of Technology

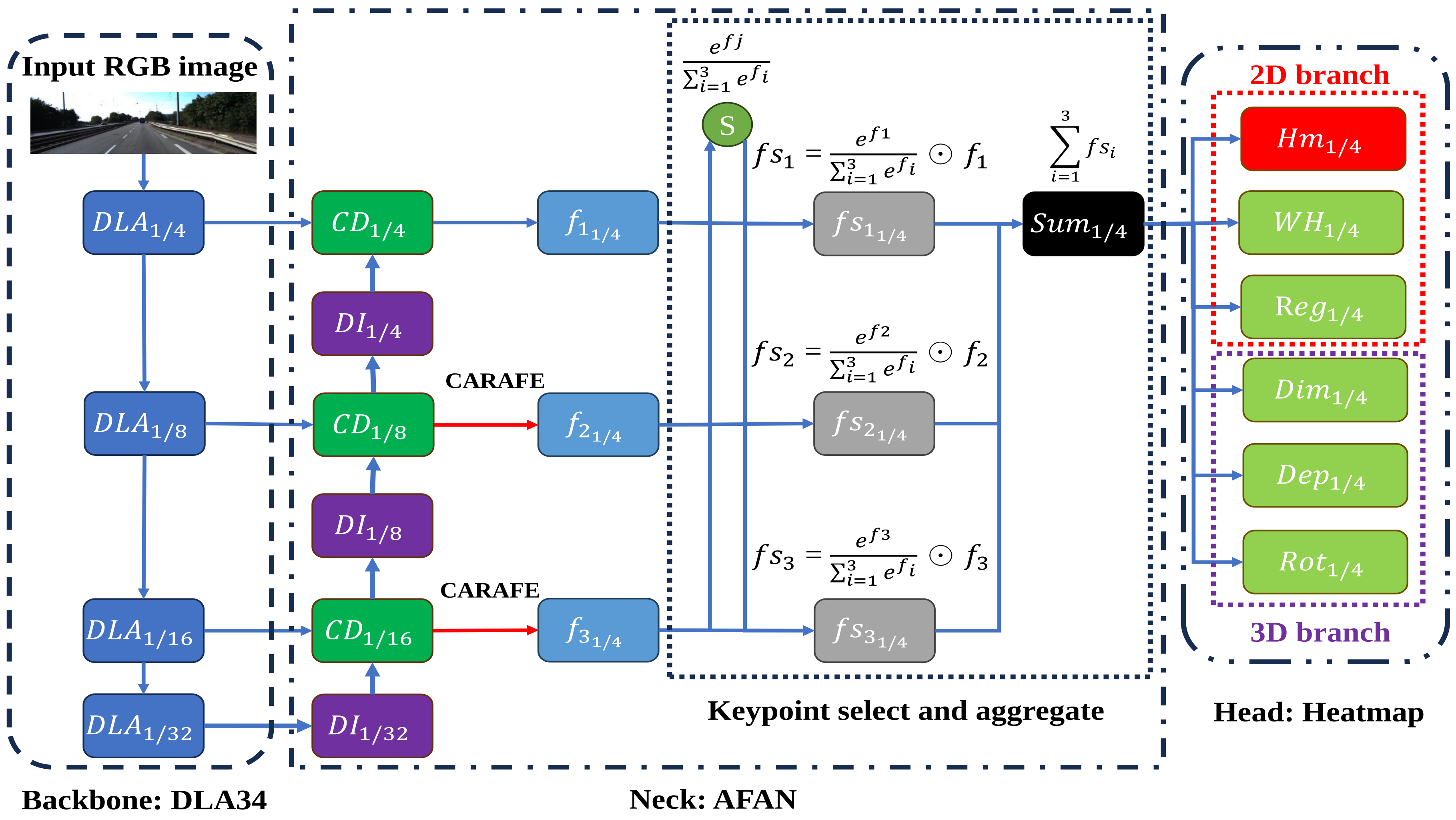
Introduction

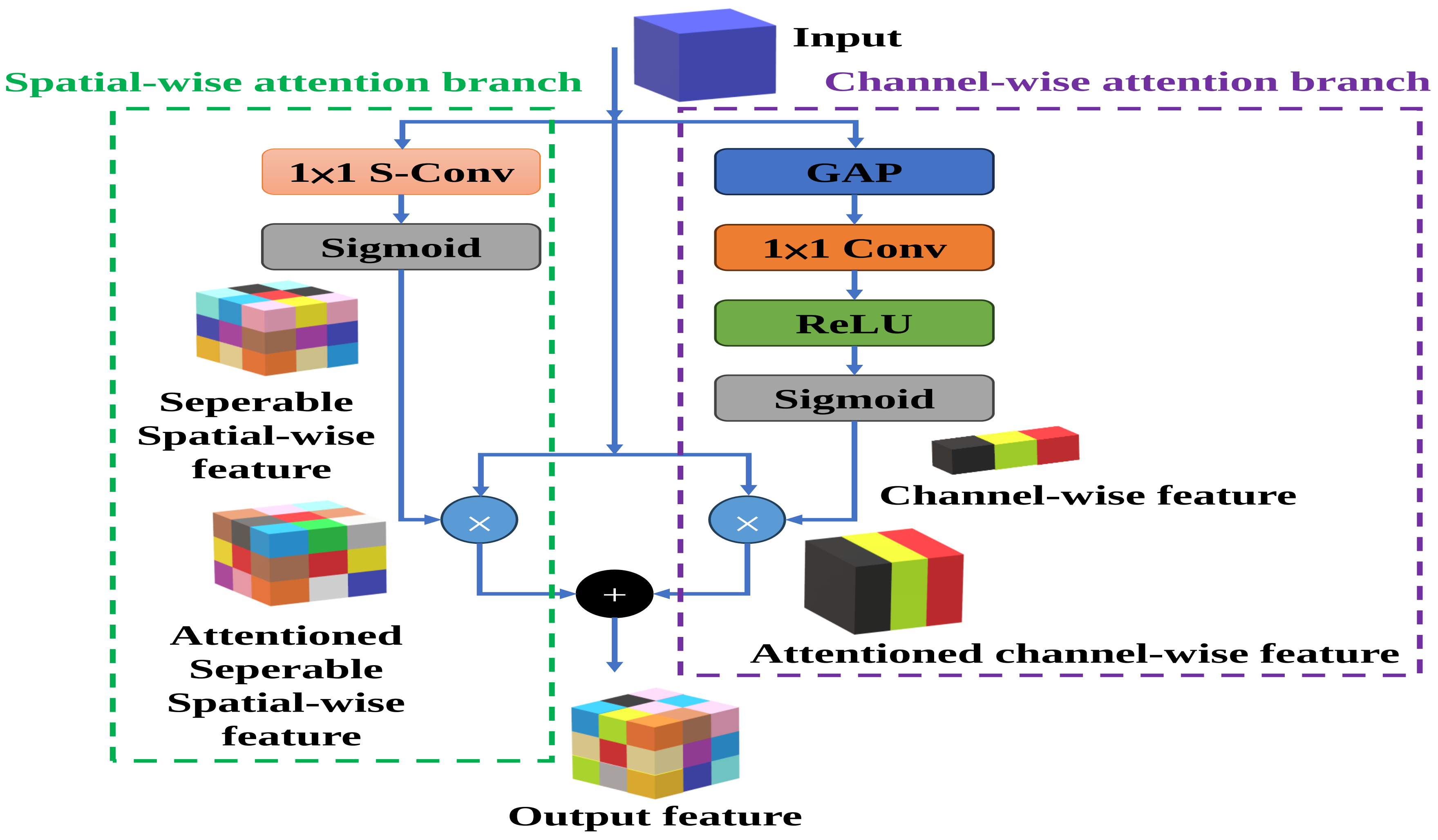
- To strike a balance between accuracy and inference speed, we propose a center-based anchor-free method inspired by CenterNet
- Our **AFAN** can aggregate multiscale features by weighting while reducing the number of calculations.
- We propose a **CEM** to accurately classify and localize objects.
- Our **AFACENet** shows the performance of the KITTI dataset, particularly for small objects, such as pedestrians and cyclists.
- Our AFACENet achieves 36FPS real-time detection speed at a resolution of 640 × 480 accelerating using an Nvidia RTX3070.



Accurate heatmap prediction will influence the results of object classification and localization. The regression heatmap must regress the correct values in the corresponding position.

METHODOLOGY




Input
Spatial-wise attention branch
Channel-wise attention branch
1x1 S-Conv
Sigmoid
Seperable Spatial-wise feature
Attentioned Seperable Spatial-wise feature
GAP
1x1 Conv
ReLU
Sigmoid
Channel-wise feature
Attentioned channel-wise feature
×
×
+
Output feature

- Adaptive Feature Aggregation Network
- Attention Head

EXPERIMENT

- Dataset and Metrics

Detection results of the proposed AFACENet. All detection results have their corresponding BEVs.

THE RESULT IS EVALUATED UNDER AP40 ON THE KITTI VALIDATION SPLIT. FOR CAR CLASS, THE IOU THRESHOLD ≥ 0.7. FOR PEDESTRIANS AND CYCLISTS, THE IOU THRESHOLD ≥ 0.5.

Class	Method	Easy	Moderate	Hard
Car	MonoRCNN [38]	16.94	12.00	9.46
	MonoDIS [35]	16.50	12.20	10.30
	MonoPair [39]	16.28	12.30	10.42
	M3D-RPN [40]	14.53	11.07	8.65
	PGD [41]	19.27	13.23	10.65
	DFR-Net [42]	19.55	14.79	11.04
	Zhou et al. [43]	20.15	16.09	15.59
	MonoFENet [18]	21.29	13.87	11.71
	GUPNet [44]	20.11	14.20	11.77
	MonoDTR [45]	21.99	15.39	12.73
	MDS-Net [46]	24.30	14.46	11.12
	AFACENet	21.63	19.25	16.49
Pedestrian	MonoDIS	9.50	7.10	5.70
	Zhou et al.	15.80	13.80	12.30
	M3D-RPN	4.92	3.48	2.94
	D4LCN [47]	4.55	3.42	2.83
	PGD	2.28	1.49	1.38
	MDS-Net	10.68	7.09	6.06
	AFACENet	21.30	18.17	17.49
Cyclist	MonoDIS	2.70	1.50	1.30
	Zhou et al.	2.50	2.00	2.00
	M3D-RPN	0.94	0.65	0.47
	D4LCN	2.45	1.67	1.36
	PGD	2.81	1.38	1.20
	MDS-Net	5.37	2.68	2.22
	AFACENet	23.07	14.02	13.74

CONCLUSION

- We used an **off-the-shelf ImageNet** pre-trained model. ImageNet has **a diversity feature** that can benefit the model training. However, the training model on ImageNet is time-consuming. Therefore, it is important to quickly devise a method to train and finetune an ImageNet pre-trained model.
- Our model has six output heads and loss functions. When modeling during backpropagation, the **gradient flow** is crucial for the goodness of learning. Therefore, some skills, such as the policy gradient, may further improve the training model.

HPC Asia 2024, January 25-27 2024, Nagoya, Japan