Home Action Genome Challenge: Task 2

Team: AIST&DENSO

Kensho Hara,
AIST

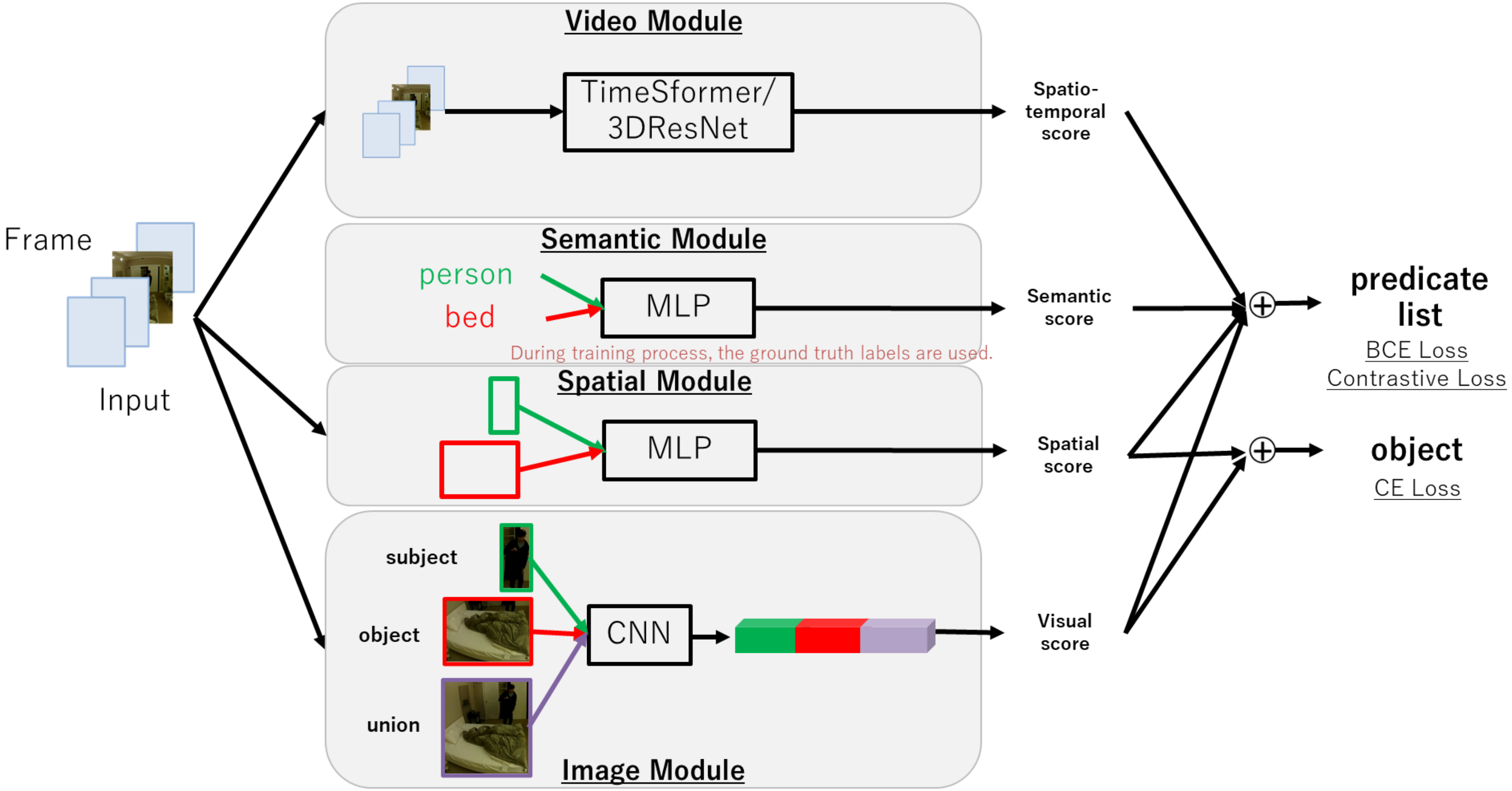
Yasufumi Kawano,
AIST, Keio University

Yoshiki Nagasaki,
AIST, Keio University

Yue Qiu,
AIST

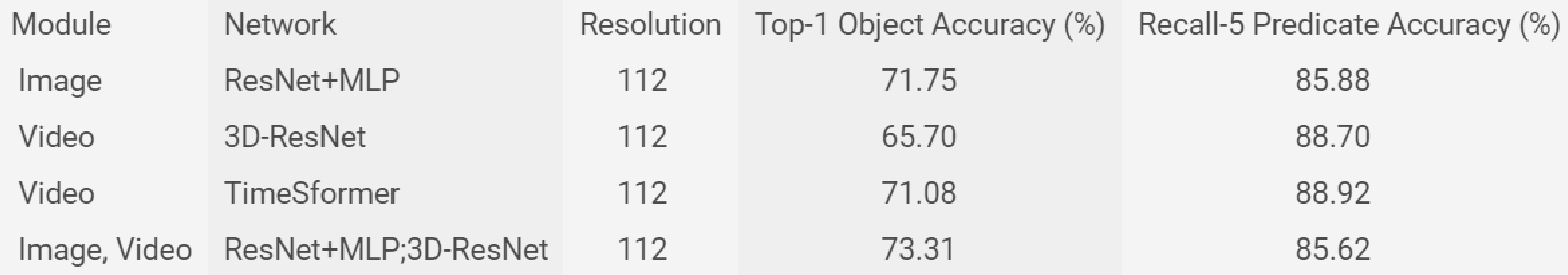
Hirokatsu Kataoka,
AIST

Overall Framework:



- ◆ The overall framework consists of image module, bounding box module, semantic module, and video module.
- ◆ The logits of each module are added together for object and predicate prediction.
- ◆ The network outputs scores for each object and predicate.
- ◆ The overall framework is inspired by RelDN [Zhang et,al., CVPR2019]

Ablation Experiments on Different Modules:

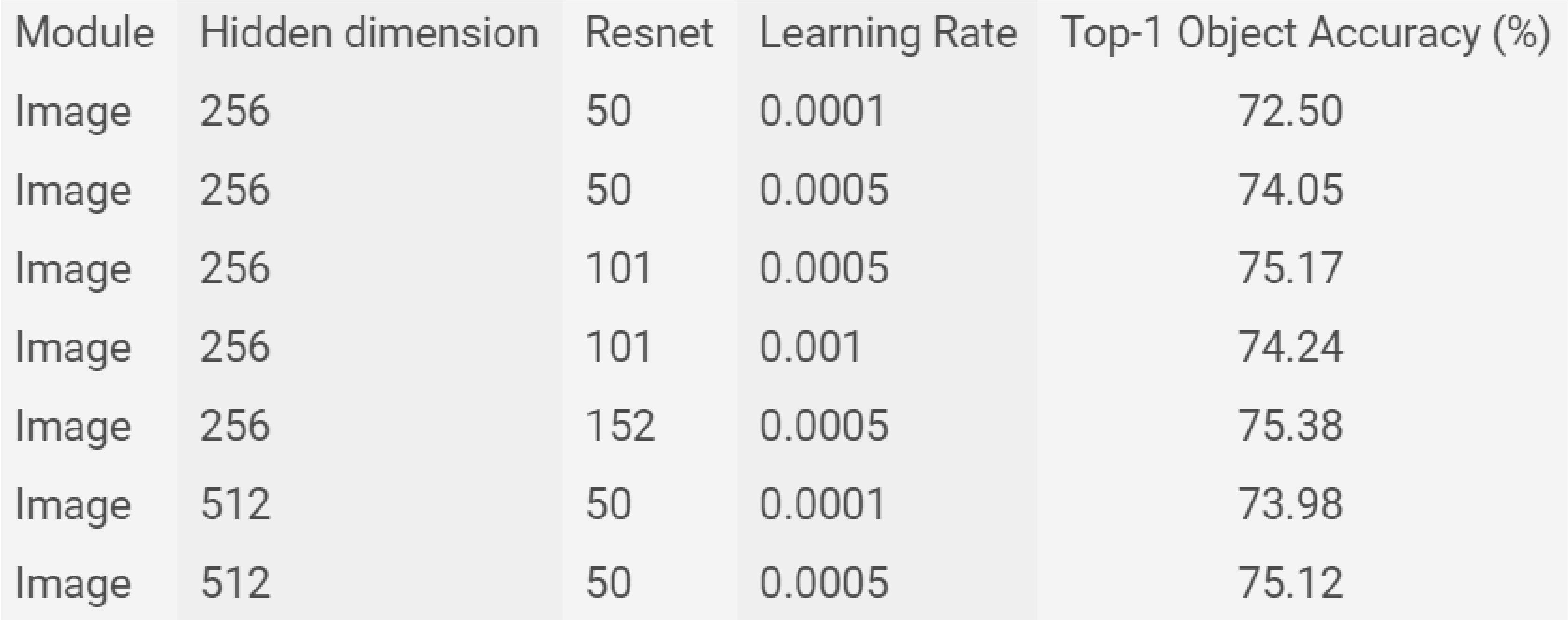


Module	Network	Resolution	Top-1 Object Accuracy (%)	Recall-5 Predicate Accuracy (%)
Image	ResNet+MLP	112	71.75	85.88
Video	3D-ResNet	112	65.70	88.70
Video	TimeSformer	112	71.08	88.92
Image, Video	ResNet+MLP;3D-ResNet	112	73.31	85.62

◆ Image module is important for obtaining better object accuracy for the current network.

◆ Video module tends to perform better for predicate prediction.

Ablation Experiments on Parameters: (without Video Module)



Module	Hidden dimension	Resnet	Learning Rate	Top-1 Object Accuracy (%)
Image	256	50	0.0001	72.50
Image	256	50	0.0005	74.05
Image	256	101	0.0005	75.17
Image	256	101	0.001	74.24
Image	256	152	0.0005	75.38
Image	512	50	0.0001	73.98
Image	512	50	0.0005	75.12

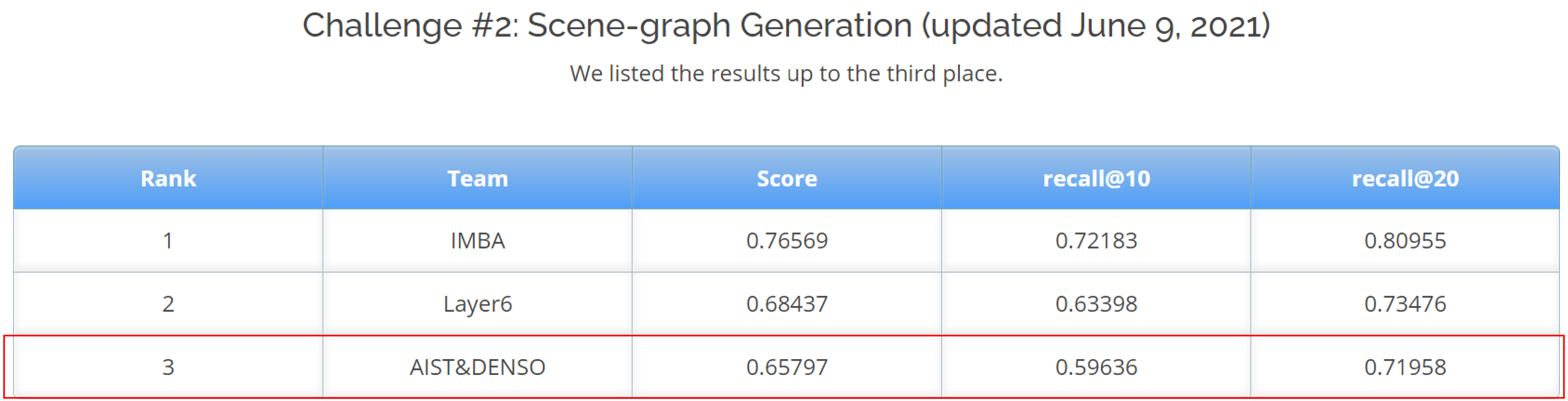
◆ Details of Image module:
Resnet module + 2-layered MLP (input dimension -> hidden dimenson -> out dimension)

◆ ResNet 152 and ResNet 101 are slightly better than ResNet 50.

Predicate prediction:

◆We compute the distribution of predicate list for each object;
◆We determine the score of each predicate list for all objects (prior scores);
◆We record the scores of each predicate for all objects (predicted scores), which is computed through the network;
◆The final score of each predicate list for each object is computed by multiplying the prior scores with the predicted scores.

Final results:



Challenge #2: Scene-graph Generation (updated June 9, 2021)

We listed the results up to the third place.

Rank	Team	Score	recall@10	recall@20
1	IMBA	0.76569	0.72183	0.80955
2	Layer6	0.68437	0.63398	0.73476
3	AIST&DENSO	0.65797	0.59636	0.71958

Home Action Genome

Our submission:

- Image-only module;
- Image resolution: 224;
- Image feature extraction: ResNet 101;
- Object and Predicate prediction: 2-layered MLP with hidden dimension of 256.

Summary:

- ◆ Image module is important for obtaining better accuracy for the current module.
- ◆ Video module tends to perform better for predicate prediction.
- ◆ The network needs to be improved for combining image and video information.

References:

- ◆ Zhang, Ji, et al. "Graphical contrastive losses for scene graph parsing." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
- ◆ Hara, Kensho, Hirokatsu Kataoka, and Yutaka Satoh. "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018.
- ◆ He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
- ◆ Rai, Nishant, et al. "Home Action Genome: Cooperative Compositional Action Understanding." arXiv preprint arXiv:2105.05226 (2021).
- ◆ Ji, Jingwei, et al. "Action genome: Actions as compositions of spatio-temporal scene graphs." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
- ◆ Bertasius, Gedas, Heng Wang, and Lorenzo Torresani. "Is Space-Time Attention All You Need for Video Understanding?." arXiv preprint arXiv:2102.05095 (2021).

References:

◆ Zhang, Ji, et al. "Graphical contrastive losses for scene graph parsing." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

◆ Hara, Kensho, Hirokatsu Kataoka, and Yutaka Satoh. "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018.

◆ He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

◆ Rai, Nishant, et al. "Home Action Genome: Cooperative Compositional Action Understanding." arXiv preprint arXiv:2105.05226 (2021).

◆ Ji, Jingwei, et al. "Action genome: Actions as compositions of spatio-temporal scene graphs." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

◆ Bertasius, Gedas, Heng Wang, and Lorenzo Torresani. "Is Space-Time Attention All You Need for Video Understanding?." arXiv preprint arXiv:2102.05095 (2021).