Supplementary Material for GLAMD: Global and Local Attention Mask Distillation for Object Detectors

*Younho Jang[1],*Wheemyung Shin[1],Jinbeom Kim[2],†Simon Woo[2],†Sung-Ho Bae[1]

Kyung Hee University[1], Sungkyunkwan University[2]
{2014104142, wheemi, shbae}@khu.ac.kr, {kjinb1212, swoo}@g.skku.edu

1 Experiments

1.1 Generality of Our Method in Small Models

We test our method for detectors with a small backbone network to ensure that it works well in small models. At the same time, experiments are conducted in the Pascal VOC dataset [1] to check whether the proposed method can be generally applied to datasets other than the COCO [5]. We evaluate our method on RetinaNet [4] and Faster-RCNN [6]. We designate ResNet34 [3] and ResNet18 as backbone models for the teacher and student models, respectively. As shown in Table 1, the distilled models outperform the non-distilled student models in every case. In Faster R-CNN, our model's performance is 1.95 AP higher than the baseline student's one, and in RetinaNet, our model achieves 2.57 AP improvement compared to the baseline student.

Table 1. Experimental results of the proposed method in Pascal VOC. We choose the ResNet34 and ResNet18 as backbone models for the teacher and student model.

Model	Faster R-CNN			RetinaNet		
	Teacher	Student	Ours	Teacher	Student	Ours
AP	57.28	53.81	55.76	79.63	75.57	78.14

1.2 Results with Extremely Large Patch Sizes

We explore our method's parameter sensitivity by evaluating the performance under more intense parameter settings. In this section, we further evaluate the patch size of 20 and 30 that are far larger than the default setting of 7. In Table 2,

* Equal contribution.
† Corresponding author.

the model's performance decreases as the patch size increases. This result is because the large patch sizes are not suitable to catch local details, resulting in a similar performance with the global attention method [8]. This experimental result proves that the local patches extracting fine-grained knowledge are our method's key improvement to achieve high performance.

Table 2. Experimental results of the proposed method in extremely large patch size.

Patch size	**7**	9	11	20	30
AP	**40.0**	39.8	39.7	39.4	39.3

1.3 Effects of Hyper-parameters

We conduct several ablation studies to investigate the influence of the three balancing parameters: α, β, and γ. Figure 1 reports the performance of the student networks under different hyper-parameter settings. For all experiments in Figure 1, we use RetinaNet [4] with ResNet50 [3] as student models and RetinaNet with ResNext101 [7] as teacher models. We train and evaluate the models on the COCO dataset. Even in the worst reported case, it shows only 0.2 AP drop compared to the default setting, showing GLAMD is hardly sensitive to the choice of the hyper-parameters.

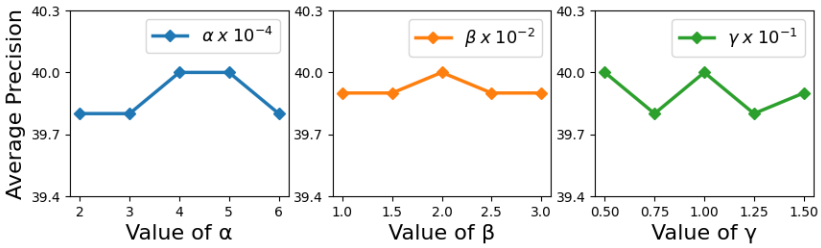

Fig. 1. Performance under different hyper-parameter settings.

1.4 Additional Comparison with SOTA

we conduct additional comparison experiments with DeFeat [2], a SOTA KD method for object detection models. We evaluate the performance of DeFeat with three different detectors on the COCO dataset. In Table 3, GLAMD generally performs better than DeFeat.

Table 3. Comparison with DeFeat KD method.

Model	Faster-50	Retina-50	Cascade-50
DeFeat	40.6	39.4	42.8
GLAMD	**40.8**	**40.0**	**43.0**

1.5 Further Ablation Study on GLAM with each KD Method

To show more experimental results on the effectiveness of GLAM, we perform further ablation studies for each KD method with and without GLAM including the experiment of applying only GLAM. For all the experiments in the table, we use the RetinaNet [4] detector and the COCO dataset. As shown in Table 4, with GLAM, applying any single KD method performs better than the result of applying all KD methods without GLAM. In addition, GLAM achieves 1.2 AP gain when all the KD methods are applied. These results indicate that GLAM boosts up distillation performance when collaborating with KD methods. In the case of using GLAM without any KD method, our method results in a similar performance to the baseline. This is because exploiting only GLAM does not affect the detector loss.

Table 4. Results for GLAM influence experiments for each distillation method.

Mask	Feat	Cls Head	Loc Head	AP
w/o GLAM	✓	✓	✓	38.7
w GLAM	✗	✗	✗	36.8
	✓	✗	✗	39.5
	✗	✓	✗	39.0
	✗	✗	✓	38.9
	✓	✓	✗	39.7
	✗	✓	✓	39.0
	✓	✗	✓	39.6
	✓	✓	✓	**40.0**

2 Visualizations

2.1 Error Analysis

In this section, we analyze the different types of error by a baseline model and a distilled model with our GLAMD. We use RetinaNet [4] with ResNet50 [3] backbone as a baseline model and the target dataset is COCO. As shown in Figure 2, our method decreases all types of error compared to the baseline model. This

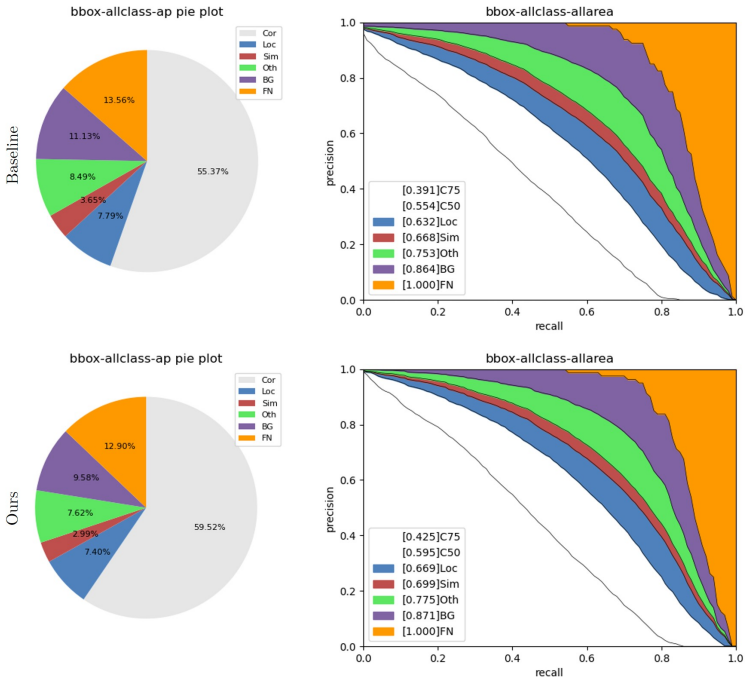

Fig. 2. Error Analysis on a baseline model and a distilled model with our method. **LOC** means localization error. **Sim & Oth** mean classification error on similar and not similar classes, respectively. **BG** means false-positive error and **FN** means false-negative error.

implies that our mask is effective to reduce both localization and classification error.

2.2 Effect on Feature Distribution at Each Feature Level

We explore how the distribution of the feature map changes after the distillation using our method. A variety of patch-wise L1 distance maps between the feature maps of teacher and student models are visualized in Figure 3. We test a global attention mask [8] and our mask on 14 image samples which are randomly selected from COCO. We observe that GLAM generally decreases the feature gap between the teacher and student in the whole region.

Furthermore, we analyze the mean of the distance maps in every FPN stage. We calculate the mean of the L1 distance map of each sample and get the average of all the sample's L1 distance mean values to represent them as a scalar value. The process is repeated for every layer in the model. As shown in Figure 4, our method generally produces decreased L1 distance maps by reducing the mean values in most cases, especially for deep layers. It means that our mask can lead

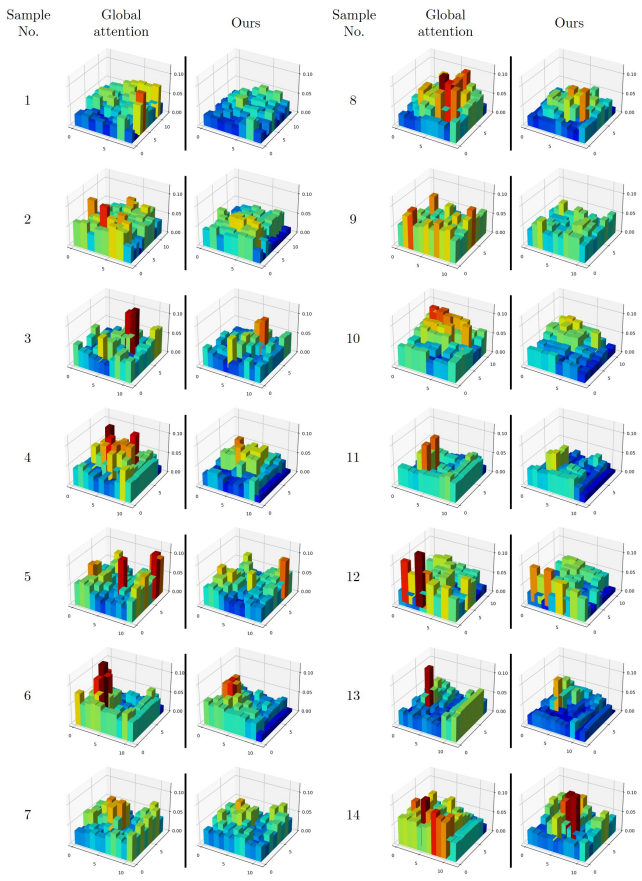

Fig. 3. Extra visualizations of the L1 distances between the local features of teacher and student for 14 samples randomly selected from COCO. The distance maps in the left are produced by a model trained with a global attention mask method and the distance maps in the right are produced by a model trained with GLAMD.

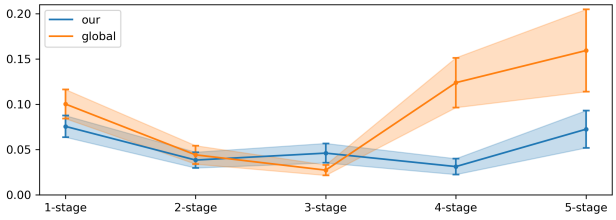

Fig. 4. Errorbar graph of the L1 distance between the feature maps of teacher and student models in every stage.

the student model to have more similar feature maps with the teacher's feature maps.

2.3 Visualization of GLAM in Each Stage

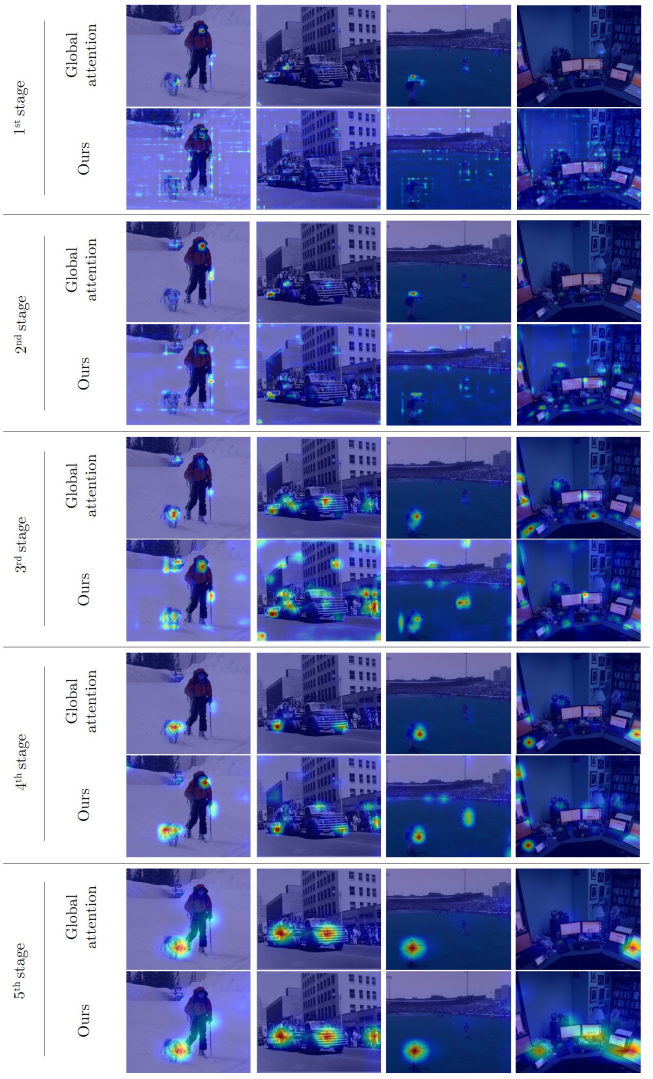

Fig. 5. Visualization of global attention masks and our GLAMs in every stage. The target samples are from COCO.

We visualize some examples of global attention masks [8] and our GLAMs on each stage of the model in Figure 5. We observe that our mask is more fine-grained in the early stage of the model, as the resolution of the feature map is higher in the early stage while the patch size is fixed to all stage. This characteristic of our mask is suitable to extract small details from the teacher's features. It is also verified that our mask generally extracts knowledge from more various objects than the global attention mask in every stage of the model.

References

1. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
2. Guo, J., Han, K., Wang, Y., Wu, H., Chen, X., Xu, C., Xu, C.: Distilling object detectors via decoupled features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2154–2164 (2021)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
4. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
5. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
6. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
7. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
8. Zhang, L., Ma, K.: Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In: International Conference on Learning Representations (2020)