# 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

**Abstract.** Knowledge distillation (KD) is a well-known model compression strategy to improve models' performance with fewer parameters. However, recent KD approaches for object detection have faced two limitations. First, they distill nearby foreground regions, ignoring potentially useful background information. Second, they only consider global contexts, thereby the student model can hardly learn local details from the teacher model. To overcome such challenging issues, we propose a novel knowledge distillation method, GLAMD, distilling both global and local knowledge from the teacher. We divide the feature maps into several patches and apply an attention mechanism for both the entire feature area and each patch to extract the global context as well as local details simultaneously. Our method outperforms the state-of-the-art methods with 40.8 AP on COCO2017 dataset, which is 3.4 AP higher than the student model (ResNet50 based Faster R-CNN) and 0.7 AP higher than the previous global attention-based distillation method.

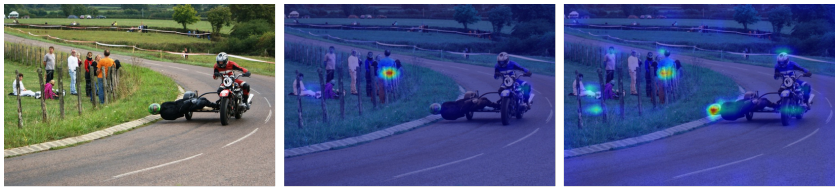
**Keywords:** Knowledge Distillation, Object Detection

## 1 Introduction

Recent advancements in deep convolutional neural networks have achieved remarkable success in various applications, especially for visual tasks such as image classification [29, 11, 34] and object detection [27, 17, 18, 31, 24–26, 2, 10]. With their high performance, current deep-learning-based methods have been integrated and deployed for a wide range of real-world applications such as CCTV surveillance, autonomous driving, and unmanned store. Although recent deep learning models have demonstrated promising results, deploying deep-learning-based applications on mobile or edge devices is still challenging. This is because of limited computing resources on devices. To address this issue, model compression techniques such as weight pruning [9, 15], model quantization [14], and Knowledge Distillation (KD) [13] have been introduced.

In particular, KD is one of the most promising methods for reducing the parameters of deep Convolutional Neural Networks (CNN) models while effectively achieving high performance. The KD method is formalized by Hinton *et al.* [13]

* Equal contribution. † Corresponding author.



(a) Original (b) Global attention (c) Ours

**Fig. 1.** Visualization of the attention masks generated by (b) Zhang *et al.* [36], and (c) GLAMD (Ours). Zhang *et al.* [36] focus only on a single small region, ignoring the other important regions. On the other hand, our attention mask successfully represents the other important local regions (people, bikes, etc.).

which uses the prediction logits of a large and cumbersome teacher model to train a lightweight student model. Hence, the soft labels from the teacher model can help the student model to mimic the teacher model's decision, producing improved performance with the small number of parameters in the student model.

However, most distillation methods [28, 30, 13, 32, 22] developed for classification are not suitable for object detection tasks because of the class imbalance problem in object detection and the absence of localization knowledge in the previous KD methods. For instance, hint learning [28] is proposed to distill the teacher model's intermediate feature maps, however it does not transfer the teacher's classification and localization information of bounding boxes. To solve this issue, Chen *et al.* [3] introduce a method of distilling feature, classification, and localization information for object detectors. Nevertheless, the method in [3] still does not effectively distill the teacher's information due to the imbalance between foreground and background. To reduce overwhelmingly large background data and further distill only from the informative foreground regions, Wang *et al.* [33] propose the mask-based feature distillation method that filters out the background regions based on ground truth. This method still has the problem of providing uniform weights to target regions regardless of the importance. Hence Zhang *et al.* [36] propose to apply an attention mechanism on a global feature map to generate a mask with soft weights, where the mask allows to deliver knowledge from the selective regions with high importance.

However, we find that considering only global feature contexts can lose important knowledge in the teacher's features because of the following two major drawbacks. First, they mainly consider the foreground regions while hardly providing attention to background regions. Ignoring the background area is not ideal because there can be valuable knowledge in the background for object detection [8]. Therefore, the key enhancement for improving distillation performance in object detection tasks is carefully selecting informative regions from both background and foreground, effectively balancing and leveraging all information from them. Second, since the global mask-based methods only focus on a few global contexts of the entire features, some important local details that are evenly distributed across the entire regions can be ignored. For example, Zhang

*et al.* [36] apply the softmax function on the global area to generate a mask that provides substantial weights to a single foreground object and barely provides attention to the other objects and background regions, as shown in Figure 1 (b).

To overcome the aforementioned limitations, we propose GLAMD, Global and Local Attention Mask Distillation for object detector, a novel patch-based attention mechanism that considers the both global contexts and local details of the teacher's features. GLAMD creates global and local attention masks by applying an attention mechanism to global features and local features divided by patches. The generated mask is then applied to the intermediate features, classification output, and regression output to distill the teacher's knowledge more efficiently.

Figure 1 illustrates the attention masks generated by the previous global attention method [36] and our patch-based attention method. Compared to the global attention mask that focuses only on one person in Figure 1 (b), the local-patch mask generated by our method in Figure 1 (c) covers other informative objects such as people and a bike. Since the mask is generated by applying an attention mechanism at the both global and local levels, we call the proposed mask Global and Local Attention Mask (GLAM) in this work. With the proposed GLAM, our method jointly considers the detailed information from the background and foreground. As a result, ResNet50 based Faster R-CNN with GLAMD achieves 40.8 AP on COCO2017 dataset [19], which is a 3.4 AP improvement over the baseline and 0.7 AP higher than the previous global attention mask method [36]. Our main contributions are summarized as follows: (1) we propose an attention-based distillation method that effectively incorporates a local perspective to overcome the limitation of the global attention mask that focuses on small areas of the image; and (2) we present quantitative and qualitative results and ablation studies of our distillation method on various object detection models, including two-stage, one-stage, and anchor-free detectors in the COCO dataset, achieving the state-of-the-art performance.

## 2 Related Work

### 2.1 Object Detection

Recently, object detection models have been developed as two-stage detectors [27, 2, 10] as well as one-stage detectors [25, 26, 18, 16]. First, two-stage detectors, such as Faster-RCNN [27], utilize the region proposal network (RPN) and refinement procedure of bounding boxes. While two-stage detectors retain a high detection accuracy, their computational complexity precludes them from being used for real-time detection. In comparison to two-stage detectors, one-stage detectors such as RetinaNet [18] have lower latency since they extract bounding boxes straight from the feature map.

These anchor-based models achieve successful results in object detection by using predefined anchors. However, predefined anchors bring a huge number of outputs, resulting in substantial computational costs. Anchor-free models [31,

37] have been proposed to further reduce the computational cost by directly predicting critical bounding box information. As a result, anchor-free models are lighter than anchor-based models. Nevertheless, the detection performance of these models is proportional to their model size. Due to the models' enormous computational complexity, deploying detection models to mobile devices with low computing and storage capacity has been challenging. Therefore, model compression techniques such as weight pruning, model quantization, and knowledge distillation have been proposed to address such issues.

### 2.2 Knowledge Distillation

KD is a compression method for enhancing the small student model performance by using output from the large teacher model. As a result, extracting useful information from the teacher in the distillation process has become critical. In general, there are three different distillation approaches: response-based [13], feature-based [28, 1, 35, 12], and relation-based [32, 20, 30, 22]. The response-based distillation by Hinton *et al.* [13] selects the teacher's softmax logits and teaches the student by transferring the dark knowledge of the teacher. The feature-based distillation by Romero *et al.* [28] attempts to improve the performance of the student network by matching the teacher's intermediate features to the student's features. The relation-based distillation by Park *et al.* [22] uses information from several sample images in a mini-batch to calculate distance and angle relationships among the features.

However, there is still an issue with applying the aforementioned distillation approaches to object detection models since each local region contributes differently to student models' training. To address this issue, previous research employs a selective distillation method focusing on training-relevant local regions by applying masks. Wang *et al.* [33] focus on the area of the anchor boxes with a larger IoU with ground-truth than the flexible thresholds. Dai *et al.* [5] propose a distillation mask focusing on discriminative instances by calculating differences between the outputs of the teacher and the student. On the other hand, Zhang *et al.* [36] design a soft mask by extracting intermediate feature attention that focuses on the backbone network's concentrated regions. However, as networks often provide overwhelmingly large attention weights to a small region, the existing global softmax attention masks tend to neglect other critical regions. Therefore, we propose generating an attention mask for each local patch, which can focus on other significant regions that contain local knowledge to further improve the performance.

### 2.3 Local Patch Mechanism

Recently, local patches-based methods are widely employed in various tasks, such as image classification [7, 21] and object detection [6]. Dosovitskiy *et al.* [7] propose a transformer-based image classification model, namely ViT, which divides a single input image into several fixed-sized local patches and feeds them into the transformer module. Also, Liu *et al.* [21] design a hierarchical network

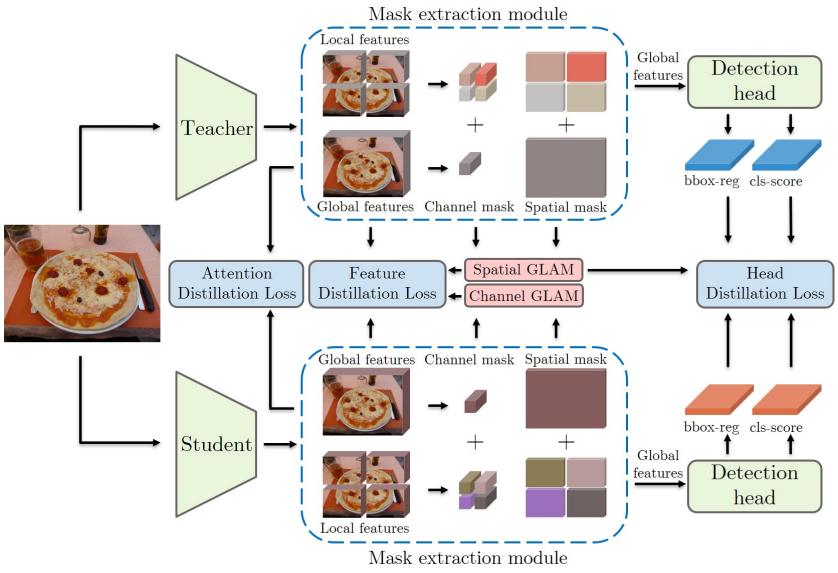



**Fig. 2.** The overall architecture of the proposed GLAMD. Local features are created by dividing the global features, which are the network's output with fixed-size patches. We then generate attention masks by taking average and the softmax function in channel and spatial directions. The proposed GLAM is generated by combining masks of the teacher and student. Finally, GLAM is applied in distillation losses.

with multiple stages that divides a local window into numerous sub-patches to calculate its attention. Their approach captures the interactions between the local windows by shifting the sub-patches over the network. Ding *et al.* [6] propose an approach to divide low-level features into local patches and then apply patch-wise channel attention to easily detect small objects that have been difficult to find in a global image. These approaches above effectively employ local patches from the input images, improving the overall network performance by recognizing local regions where the global context can hardly represent. Therefore, to overcome the limitation of the previous global mask-based distillation methods, we propose a novel mask-based distillation with local patches, which effectively distills both global and local knowledge.

## 3 Methods

Previous research [36] selectively distills features by applying global attention masks. However, [36] tends to distill only a small feature region because the global attention mask highlights a single spot, ignoring other multiple local details. To address this issue, we propose a novel mask that reflects the global and local characteristics of the features. It also can be used for feature and head distillation, as shown in Figure 2.

### 3.1 Global and Local Attention Mask (GLAM)

In this section, we describe the global and local attention masks (GLAM), a core component of the proposed GLAMD. The attention methods used in GLAM are channel and spatial attention methods denoted as $M_c$ and $M_s$, respectively. To obtain the channel attention masks, the spatial-wise average of the absolute feature elements $|x_{i,j}|$ is used in a softmax operation in the channel dimension as follows:

$$M_c(x) = HW \cdot \sigma\left(\frac{\frac{1}{HW} \cdot \sum_{i=1}^{H} \sum_{j=1}^{W} \left(|x_{i,j}|\right)}{\tau}\right), \tag{1}$$

where $\tau$ and $\sigma(\cdot)$ indicate a temperature parameter and a softmax operation, respectively, and $H$ and $W$ are the height and width of the input feature, respectively. Similarly, to obtain the spatial attention masks, the channel-wise average of the absolute feature elements $|x_k|$ is used in a softmax operation in the width and height dimensions as follows:

$$M_s(x) = C \cdot \sigma\left(\frac{\frac{1}{C} \cdot \sum_{k=1}^{C} \left(|x_k|\right)}{\tau}\right), \tag{2}$$

where $C$ is the channel of the input feature.

To generate our proposed GLAM which considers local and global perspectives, we split each output feature of Feature Pyramid Network (FPN) into $N$ local features $f_n \in \mathbb{R}^{p \times p \times C}$, where $p$ is the predetermined patch size and $n \in \{1, 2, ..., N\}$. Then, the local channel mask $L_c$ and local spatial mask $L_s$ are formulated as follows:

$$L_{c,n} = M_c(f_n^{\mathcal{T}}) + M_c(f_n^{\mathcal{S}}), \quad L_c = \psi\left(L_{c,1}, L_{c,2}, ..., L_{c,N}\right), \tag{3}$$

$$L_{s,n} = M_s(f_n^{\mathcal{T}}) + M_s(f_n^{\mathcal{S}}), \quad L_s = \psi\left(L_{s,1}, L_{s,2}, ..., L_{s,N}\right), \tag{4}$$

where $\mathcal{T}$ and $\mathcal{S}$ indicate the teacher and student, respectively, and $\psi$ denotes the concatenation operation. Similarly, given a global feature $F \in \mathbb{R}^{H \times W \times C}$, the global channel mask $G_c$ and global spatial mask $G_s$ are computed as follows:

$$G_c = M_c(F^{\mathcal{T}}) + M_c(F^{\mathcal{S}}), \quad G_s = M_s(F^{\mathcal{T}}) + M_s(F^{\mathcal{S}}). \tag{5}$$

By merging the local and global masks, our final channel and spatial attention masks, denoted as $T_c$ and $T_s$, respectively, are constructed as follows:

$$T_c = \frac{1}{2} \cdot \left(L_c + G_c\right), \quad T_s = \frac{1}{2} \cdot \left(L_s + G_s\right). \tag{6}$$

### 3.2 Feature Distillation

Typically, the teacher's features are more informative than the student's features. Therefore, we distill the intermediate features extracted from the FPN to

increase the student's performance. The feature in each stage is multiplied with the corresponding channel and spatial attention masks to selectively distill the area of interest. That is, our feature distillation loss is defined as follows:

$$\mathcal{L}_{feat} = \sum_{l=1}^{L} \left( \sum_{k=1}^{C} \sum_{i=1}^{H} \sum_{j=1}^{W} \left( F_{lkij}^{\mathcal{T}} - \phi_{adapt}(F_{lkij}^{\mathcal{S}}) \right)^2 \cdot T_{s,l} \cdot T_{c,l} \right)^{\frac{1}{2}}, \tag{7}$$

where $L$ is the number of FPN stages and the function $\phi_{adapt}$ is the $1 \times 1$ convolutional adaptation layer that matches the student's feature size to that of the teacher's feature. And, $T_{s,l}$ and $T_{c,l}$ mean spatial and channel masks of the $l$-th stage, respectively.

In addition, we distill the attention features to encourage the student in producing more effective GLAM. Hence, the extraction process of the channel and spatial attention feature can be formulated as $A_c(x) = \frac{1}{C} \cdot \sum_{k=1}^{C} x_k$ and $A_s(x) = \frac{1}{HW} \cdot \sum_{i=1}^{H} \sum_{j=1}^{W} x_{ij}$. Then, the channel attention loss is calculated by distilling both global and local channel attention features in our work. In particular, global and local spatial attention features are considered to be equivalent, as local features are formed by dividing global features into the spatial domain. As a result, in contrast to channel attention loss, spatial attention loss utilizes only global spatial attention features. Therefore, our proposed channel attention loss $\mathcal{L}_{cat}$ and spatial attention loss $\mathcal{L}_{sat}$ can be expressed as follows:

$$\mathcal{L}_{cat} = \frac{1}{2} \cdot \left( \left\| A_c(F^{\mathcal{S}}) - A_c(F^{\mathcal{T}}) \right\|_2 + \frac{1}{N} \cdot \sum_{n=1}^{N} \left\| A_c(f_n^{\mathcal{S}}) - A_c(f_n^{\mathcal{T}}) \right\|_2 \right), \tag{8}$$

$$\mathcal{L}_{sat} = \left\| A_s(F^{\mathcal{S}}) - A_s(F^{\mathcal{T}}) \right\|_2. \tag{9}$$

Finally, the overall feature attention loss is formulated by the sum of the channel and spatial attention losses, as follows:

$$\mathcal{L}_{at} = \mathcal{L}_{cat} + \mathcal{L}_{sat}. \tag{10}$$

### 3.3 Head Distillation

The response-based distillation encourages the student's outputs to mimic the teacher's. However, due to the imbalance between the foreground and background in object detection tasks, directly distilling the teacher's head outputs can cause a detrimental effect on the student's performance. Therefore, in this work, we apply spatial attention masks while performing the response-based distillation. Especially, we use the spatial attention masks from the same FPN stage in Eq. 6 to conduct the masked head distillation. The classification head loss $\mathcal{L}_{cls-head}$ can be defined as follows:

$$\mathcal{L}_{cls-head} = \sum_{l=1}^{L} \sum_{k=1}^{C} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathcal{L}_{BCE} \left( z_{lkij}^{\mathcal{S}}, z_{lkij}^{\mathcal{T}} \right) \cdot T_{s,l}, \tag{11}$$

where $z^{\mathcal{S}}$ and $z^{\mathcal{T}}$ represent the outputs of the student and the teacher classification head, respectively, and $\mathcal{L}_{BCE}$ represents the binary-cross-entropy loss.

According to Chen *et al.* [3], certain unbounded outputs of the teacher can provide incorrect guidance to the student model. To avoid the aforementioned issue, we distill the localization head using IoU loss, one of the bounded loss functions. For the localization head distillation, we use IoU loss to formulate the localization head loss as follows:

$$\mathcal{L}_{loc-head} = \sum_{l=1}^{L}\sum_{k=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathcal{L}_{IoU}\left(r^{\mathcal{S}}_{lkij}, r^{\mathcal{T}}_{lkij}\right)\cdot T_{s,l}, \tag{12}$$

where $r$ is the localization head output.

### 3.4 Overall Loss Function

We form appropriate distillation losses using the outputs of the modules in the detector and construct an overall loss by taking weighted sum with the standard classification and localization losses for the object detection task, denoted as $\mathcal{L}_{task}$. Our overall loss is defined as follows:

$$\mathcal{L} = \mathcal{L}_{task} + \alpha\mathcal{L}_{feat} + \beta\mathcal{L}_{at} + \gamma(\mathcal{L}_{cls-head} + \mathcal{L}_{loc-head}), \tag{13}$$

where $\alpha$, $\beta$, and $\gamma$ are the balancing hyper-parameters with the distillation loss and the task loss.

## 4 Experiment

### 4.1 Experiments Settings

To demonstrate the effectiveness of our method, we evaluate our method on various object detection models and compare the results with other KD methods [28, 38, 36, 33]. All experiments are implemented using the mmdetection library [4] with PyTorch framework [23] on the COCO dataset [19]. We train our model with 120k training images and test with 5k validation images from the COCO dataset. All the performances are evaluated in average precision (AP). We use 4 RTX3090 GPUs during training and use the batch size of 16.

For all experiments, student models are trained under $1\times$ scheduler, the default setting of the mmdetection, to train 12 epochs on the COCO dataset. We start training with a warm-up strategy during the first 2,000 iterations and perform a learning rate decay, which divides the learning rate by 10 at the 8-th and 11-th epochs. We use the SGD optimizer to train the detection model, and set the learning rate to 0.02 in Faster R-CNN and 0.01 in the rest of the models. Also, the weight decay and momentum are set to 1e-4 and 0.9, respectively. We set the hyper-parameters in Eq. 13 to $\{\alpha = 4\times10^{-4}$, $\beta = 2\times10^{-2}$, $\gamma = 1\times10^{-1}$, $\tau = 1\times10^{-1}\}$ for single-stage detectors and $\{\alpha = 7\times10^{-5}$, $\beta = 4\times10^{-3}$, $\gamma = 1\times10^{-1}$, $\tau = 5\times10^{-1}\}$ for two-stage detectors.

**Table 1.** The generalization ability of our GLAMD in various object detectors.

| Method | Scheduler | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|---|
| Faster-ResNet50 (Student) | 1× | 37.4 | 58.1 | 40.4 | 21.2 | 41.0 | 48.1 |
| Faster-ResNext101 (Teacher) | 3× | 43.1 | 63.6 | 47.2 | 26.5 | 46.9 | 56.0 |
| **GLAMD (Ours)** | 1× | **40.8** | **61.4** | **44.3** | **23.2** | **45.0** | **53.2** |
| Cascade-ResNet50 (Student) | 1× | 40.3 | 58.6 | 44.0 | 22.5 | 43.8 | 52.9 |
| Cascade-ResNext101 (Teacher) | 3× | 44.5 | 63.2 | 48.5 | 25.5 | 48.1 | 58.4 |
| **GLAMD (Ours)** | 1× | **43.0** | **61.5** | **46.8** | **24.1** | **47.3** | **56.8** |
| Mask-ResNet50 (Student) | 1× | 38.2 | 58.8 | 41.4 | 21.9 | 40.9 | 49.5 |
| Mask-ResNext101 (Teacher) | 2× | 42.7 | 62.9 | 47.1 | 23.8 | 46.5 | 56.7 |
| **GLAMD (Ours)** | 1× | **40.2** | **61.1** | **43.7** | **23.0** | **44.3** | **52.6** |
| RetinaNet-ResNet50 (Student) | 1× | 36.5 | 55.4 | 39.1 | 20.4 | 40.3 | 48.1 |
| RetinaNet-ResNext101 (Teacher) | 3× | 41.6 | 61.4 | 44.3 | 23.9 | 45.5 | 54.5 |
| **GLAMD (Ours)** | 1× | **40.0** | **59.5** | **42.5** | **22.8** | **44.0** | **53.4** |
| GFL-ResNet50 (Student) | 1× | 40.2 | 58.4 | 43.3 | 23.3 | 44.0 | 52.2 |
| GFL-ResNet101 (Teacher) | 2× | 44.9 | 63.1 | 49.0 | 28.0 | 49.1 | 57.2 |
| **GLAMD (Ours)** | 1× | **43.0** | **61.0** | **46.5** | **26.4** | **47.4** | **55.2** |
| ATSS-ResNet50 (Student) | 1× | 39.4 | 57.6 | 42.8 | 23.6 | 42.9 | 50.3 |
| ATSS-ResNet101 (Teacher) | 1× | 41.5 | 59.9 | 45.2 | 24.2 | 45.9 | 53.3 |
| **GLAMD (Ours)** | 1× | **41.0** | **59.1** | **44.3** | **23.8** | **45.1** | **52.9** |
| FCOS-ResNet50 (Student) | 1× | 36.6 | 56.0 | 38.8 | 21.0 | 40.6 | 47.0 |
| FCOS-ResNet101 (Teacher) | 1× | 39.1 | 58.3 | 42.1 | 22.7 | 43.3 | 50.3 |
| **GLAMD (Ours)** | 1× | **38.6** | **58.1** | **41.2** | **22.8** | **42.5** | **49.3** |

### 4.2 Results on Different Detection Frameworks

We evaluate the generalization ability of our proposed GLAMD on multiple detection architectures, including two-stage [27, 2, 10], one-stage [18, 16], and anchor-free [31, 37] detectors. For all the detectors, we use ResNext101 [34] or ResNet101 [11] as backbones of teachers and ResNet50 as backbones of students, respectively. As shown in Table 1, our proposed method achieves significant gains in terms of AP on all types of detection architectures. On average, our method obtains 3.2 and 2.7 AP boosts, outperforming the baseline one-stage and two-stage detectors, respectively. For anchor-free detectors (ATSS and FCOS), our method obtains APs of 41.0 and 38.6 that are comparable to the teacher models. Such results demonstrate that our method is effective in various detectors.

### 4.3 Comparison with Other KD Methods

To compare the results of our method with other KD methods, we evaluate our method and recent KD methods on both one-stage (RetinaNet [18]) and two-stage detectors (Faster R-CNN [27] and Cascade R-CNN [2]). The comparison

**Table 2.** Comparison with various object detection KD methods.

| Method | Scheduler | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|---|
| RetinaNet-ResNet50 (Student) | 1× | 36.5 | 55.4 | 39.1 | 20.4 | 40.3 | 48.1 |
| RetinaNet-ResNext101 (Teacher) | 3× | 41.6 | 61.4 | 44.3 | 23.9 | 45.5 | 54.5 |
| Hint learning [28] | 1× | 37.1 | 56.5 | 39.2 | 21.4 | 40.7 | 48.8 |
| Wang *et al.* [33] | 1× | 38.4 | 57.5 | 41.1 | 20.8 | 42.0 | 51.9 |
| Zhang *et al.* [36] | 1× | 39.0 | 58.1 | 41.8 | 22.3 | 42.9 | 51.7 |
| FRS [38] | 1× | 39.3 | 58.7 | 41.9 | 21.4 | 43.1 | 52.3 |
| **GLAMD (Ours)** | 1× | **40.0** | **59.5** | **42.5** | **22.8** | **44.0** | **53.4** |
| Faster-ResNet50 (Student) | 1× | 37.4 | 58.1 | 40.4 | 21.2 | 41.0 | 48.1 |
| Faster-ResNext101 (Teacher) | 3× | 43.1 | 63.6 | 47.2 | 26.5 | 46.9 | 56.0 |
| Hint learning [28] | 1× | 38.7 | 59.7 | 41.8 | 23.1 | 42.0 | 50.9 |
| Wang *et al.* [33] | 1× | 39.5 | 59.9 | 43.2 | 21.7 | 43.4 | **53.2** |
| Zhang *et al.* [36] | 1× | 40.1 | 60.8 | 43.4 | 22.9 | 44.1 | 53.1 |
| FRS [38] | 1× | 40.3 | **61.8** | 43.9 | **23.3** | 44.3 | 52.4 |
| **GLAMD (Ours)** | 1× | **40.8** | 61.4 | **44.3** | 23.2 | **45.0** | **53.2** |
| Cascade-ResNet50 (Student) | 1× | 40.3 | 58.6 | 44.0 | 22.5 | 43.8 | 52.9 |
| Cascade-ResNext101 (Teacher) | 3× | 44.5 | 63.2 | 48.5 | 25.5 | 48.1 | 58.4 |
| Hint learning [28] | 1× | 40.6 | 59.4 | 44.4 | 22.9 | 44.1 | 53.8 |
| Wang *et al.* [33] | 1× | 41.7 | 60.6 | 45.6 | 23.3 | 45.2 | 55.9 |
| Zhang *et al.* [36] | 1× | 42.4 | 60.9 | 46.2 | 23.4 | 46.2 | 56.1 |
| FRS [38] | 1× | 42.7 | 61.3 | 46.7 | **24.4** | 46.3 | 56.2 |
| **GLAMD (Ours)** | 1× | **43.0** | **61.5** | **46.8** | 24.1 | **47.3** | **56.8** |

results with other KD methods are provided in Table 2. As shown in Table 2, our approach outperforms all previous KD methods in distillation performance. In particular, the results achieved by ours are 1.0, 0.7, and 0.6 AP higher than the results from the global attention-based method [36], respectively. Also, our method outperforms the recent distillation method FRS [38]. It clearly shows that local knowledge extracted by our method effectively enhances the overall distillation performance.

### 4.4 Ablation Study

We conduct four different ablation studies to further explore the properties of our proposed method.

**Modules in GLAMD.** To verify the effect of each module in Eq. 13, we evaluate the detection performance with and without each of them in GLAMD. The result of the ablation study is presented in Table 3. Our method achieves 0.2 and 0.1 AP improvements on the classification head and regression head, respectively. When we conduct distillation from both classification and regression heads together, our method achieves 0.5 AP improvement. These results show that each part of the distillation loss in our method contributes to the perfor-

**Table 3.** Ablation study for the contribution of each module in GLAMD. "Feat", "Cls Head", and "Loc Head" indicate each distillation loss used in our model. "GLAM" indicates applying our mask (GLAM) to distillation losses.

| Feat | GLAM | Cls Head | Loc Head | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|---|---|---|
| ✗ | ✗ | ✗ | ✗ | 36.5 | 55.4 | 39.1 | 20.4 | 40.3 | 48.1 |
| ✓ | ✗ | ✗ | ✗ | 37.1 | 56.5 | 39.2 | 21.4 | 40.7 | 48.8 |
| ✓ | ✓ | ✗ | ✗ | 39.5 | 58.9 | 42.0 | 23.2 | 43.5 | 52.0 |
| ✓ | ✓ | ✓ | ✗ | 39.7 | 59.2 | 42.4 | **23.5** | **44.1** | 52.5 |
| ✓ | ✓ | ✗ | ✓ | 39.6 | 58.9 | 42.3 | 22.1 | 43.6 | 51.9 |
| ✓ | ✓ | ✓ | ✓ | **40.0** | **59.5** | **42.5** | 22.8 | 44.0 | **53.4** |

**Table 4.** Experimental results for comparing performance change according to different types of attention masks.

| Local | Global | AP | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|
| ✗ | ✗ | 38.7 | 22.7 | 42.5 | 51.8 |
| ✓ | ✗ | 39.8 | **23.3** | 44.0 | 53.0 |
| ✗ | ✓ | 38.9 | 22.4 | 42.8 | 52.0 |
| ✓ | ✓ | **40.0** | 22.8 | **44.0** | **53.4** |

mance gain, and fully utilizing our losses in the form of Eq. 12 can improve the final AP performance in a complementary manner.

**Effectiveness of Local Attention Mask Distillation.** Due to the biased distribution of the global attention mask, KD with the global feature attention tends to distill knowledge from a single large object. On the other hand, we hypothesize that the local feature attention can extract knowledge from small objects as well. To analyze the effectiveness of each attention mask, we perform distillation with local or global attention masks individually. As shown in Table 4, local attention achieves higher AP than global attention, especially with significantly improved AP for small objects. Although the global attention result is worse than the local attention, it is noteworthy to observe that detection performance further increases by using global attention and local attention together. These results indicate that the local attention mask proposed by our GLAMD is complementary to the global attention mask to improve the performance further.

**Effect of Local Patch Size.** The primary parameter affecting local attention is the patch size. To evaluate the influence of the patch size, we alter the patch size in [3, 5, 7, 9, 11]. As shown in Table 5, the performance in AP tends to increase when the patch size decreases until the patch size of 7. This is because attention masks generated from small patches are suitable for representing fine-grained features, yet extremely tiny patches are incapable of capturing the underlying local structure. More efficient theoretical ways to determine the optimal $p$ can be further investigated as future work.

**Table 5.** Performance under different settings of the patch size $p$.

| $p$ | 3 | 5 | **7** | 9 | 11 |
|---|---|---|---|---|---|
| AP | 39.7 | 39.7 | **40.0** | 39.8 | 39.7 |

**Table 6.** Results of three different loss functions used in $\mathcal{L}_{loc-head}$.

| Loss | $L_1$ | MSE | Smooth-$L_1$ | **IoU** |
|---|---|---|---|---|
| AP | 39.9 | 39.8 | 39.8 | **40.0** |

**Comparison of Different Localization Head Distillation Loss.** We also study the impact of different types of loss functions used in Eq. 12. To verify the effectiveness of the bounded regression loss in Eq. 12, We compare the distillation performance of various loss functions including IoU loss, L1 loss, MSE loss, and smooth-L1 loss. As shown in Table 6, IoU loss produces the best result among the other losses.

### 4.5 Qualitative Analysis and Visualizations

The visualizations of the global attention mask generated by [36] and our mask generated by GLAMD are shown in Figure 3. We observe that our mask encompasses various critical regions that the global attention mask overlooks. This property of our mask results in two significant enhancements in terms of mask distribution: (1) It captures the fine-grained details from various objects. For instance, our mask pays attention to a kid in Figure 3 (a) and another polar bear in Figure 3 (b) which are considered as the background in global attention; (2) It extracts structural information such as edges and lines from objects. In Figure 3 (e), our mask provides weight to the edges of the tent and car, demonstrating that it extracts crucial clues from the local objects for solving the challenging object detection tasks.

Next, Figure 4 qualitatively compares the results produced by a model with our method and a baseline student model without KD. The results show that our method improves the detection performance by taking advantage of GLAM. As shown in Figure 4 (a) and (c), small objects neglected by the baseline model are detected after applying GLAMD, demonstrating that it is effective at enhancing the student model's capability to extract local information. Additionally, our method strengthens the student's ability to distinguish occluded objects by distilling knowledge about the object's edge, shown in Figure 4 (b) and (d).

### 4.6 Feature Similarity

We visualize patch-wise distance maps in Figure 5. For the visualization, we calculate the L1 distance between the features of the teacher and the student. Next, we average the distance values in each patch to generate the distance maps. As shown in Figure 5, the distance is much lower in every patch with GLAMD than with a global attention mask method. This means that our method encourages students to mimic the teacher's feature map more closely across all local regions owing to the local attention masks.


**Fig. 3.** Visualization of masks on COCO2017 samples. The original images are shown in the first column. Global attention masks [36] and our masks are shown in the second and the third columns, respectively.

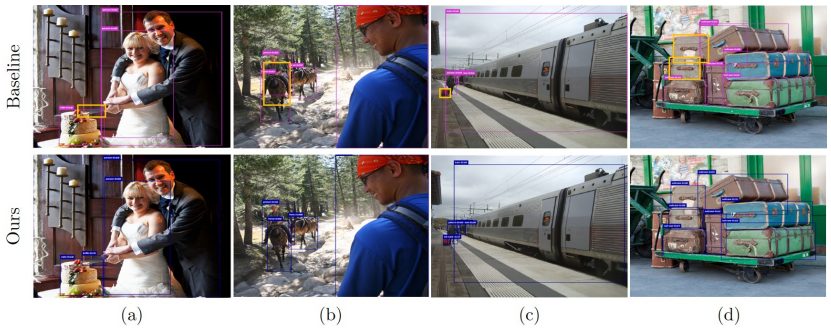



**Fig. 4.** Qualitative analysis on COCO2017. The results are produced by a model without KD (baseline) and a model distilled with GLAMD. The orange boxes in (a), (c), and (d) indicate undetected objects of the baseline detector, and the box in (b) shows a wrongly detected object.

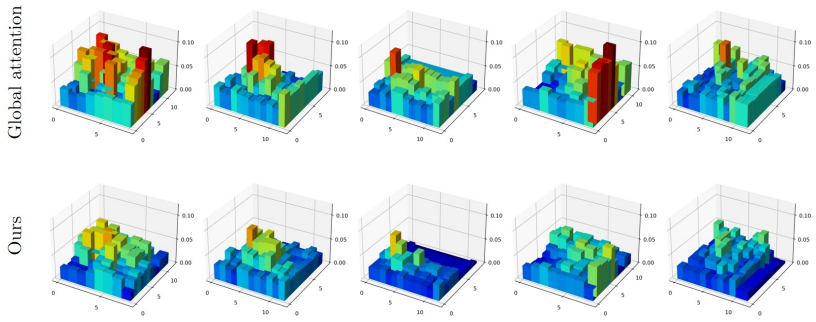



**Fig. 5.** Visualizations of the L1 distances between the local features of teacher and student. The distance maps in the top row are produced by a model trained with a global attention mask method and the distance maps in the bottom row are produced by a model trained with GLAMD.

## 5 Conclusions

In this paper, we propose GLAMD, a novel mask-based KD method for object detection that effectively applies global and local attention mechanisms to extract local details and background knowledge. To obtain local details, we divide the input features into several patches and apply attention mechanisms to each patch. Our method enables the extraction of more useful background information as well as fine-grained details from a variety of objects, resulting in much improved distillation performance. We demonstrate GLAMD's effectiveness with the various detection frameworks, outperforming other KD methods. Additionally, we conduct an extensive ablation study and analysis, showing that distilling local knowledge from various regions is crucial in object detection tasks. We expect that our work provides a turning point of conventional KD methods for object detection that focus exclusively on global knowledge to develop into more effective approaches that consider local knowledge as well.

**Acknowledgments.** This work was partially supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea Government (MSIT) (Artificial Intelligence Innovation Hub) under Grant 2021-0-02068, and partially supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00421, AI Graduate School Support Program (Sungkyunkwan University)

## References


1. Ahn, S., Hu, S.X., Damianou, A., Lawrence, N.D., Dai, Z.: Variational information distillation for knowledge transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9163–9171 (2019)
2. Cai, Z., Vasconcelos, N.: Cascade r-cnn: high quality object detection and instance segmentation. IEEE transactions on pattern analysis and machine intelligence **43**(5), 1483–1498 (2019)
3. Chen, G., Choi, W., Yu, X., Han, T., Chandraker, M.: Learning efficient object detection models with knowledge distillation. Advances in neural information processing systems **30** (2017)
4. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
5. Dai, X., Jiang, Z., Wu, Z., Bao, Y., Wang, Z., Liu, S., Zhou, E.: General instance distillation for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7842–7851 (June 2021)
6. Ding, L., Tang, H., Bruzzone, L.: Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **59**(1), 426–435 (2020)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. 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)
9. Guo, Y., Yao, A., Chen, Y.: Dynamic network surgery for efficient dnns. Advances in neural information processing systems **29** (2016)
10. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
11. 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)
12. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., Choi, J.Y.: A comprehensive overhaul of feature distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1921–1930 (2019)
13. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 **2**(7) (2015)
14. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., Kalenichenko, D.: Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2704–2713 (2018)
15. Li, H., Kadav, A., Durdanovic, I., Samet, H., Graf, H.P.: Pruning filters for efficient convnets. arXiv preprint arXiv:1608.08710 (2016)
16. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. Advances in Neural Information Processing Systems **33**, 21002–21012 (2020)

17. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
18. 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)
19. 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)
20. Liu, Y., Cao, J., Li, B., Yuan, C., Hu, W., Li, Y., Duan, Y.: Knowledge distillation via instance relationship graph. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7096–7104 (2019)
21. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
22. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3967–3976 (2019)
23. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
24. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
25. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7263–7271 (2017)
26. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
27. 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)
28. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
30. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. arXiv preprint arXiv:1910.10699 (2019)
31. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019)
32. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1365–1374 (2019)
33. Wang, T., Yuan, L., Zhang, X., Feng, J.: Distilling object detectors with fine-grained feature imitation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
34. 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)

35. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4133–4141 (2017)
36. Zhang, L., Ma, K.: Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In: International Conference on Learning Representations (2020)
37. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9759–9768 (2020)
38. Zhixing, D., Zhang, R., Chang, M., Liu, S., Chen, T., Chen, Y., et al.: Distilling object detectors with feature richness. Advances in Neural Information Processing Systems **34** (2021)