# Localization Distillation for Dense Object Detection Supplementary Materials

Zhaohui Zheng[1*], Rongguang Ye[2*], Ping Wang[2], Dongwei Ren[3], Wangmeng Zuo[3], Qibin Hou[1†], Ming-Ming Cheng[1]
[1]TMCC, CS, Nankai University [2]School of Mathematics, Tianjin University
[3]School of Computer Science and Technology, Harbin Institute of Technology

## 1. Implementation Details

The training is carried on 8 GPUs with batch size 2 images per GPU. The total training epochs are set to 12 for $1\times$ training schedule and 24 for $2\times$ training schedule following most previous work. The initial learning rate is 0.01 and a linear warm-up strategy is used for the first 500 iterations. The learning rate decreases by a factor of 10 after the 8-th and 11-th epoch for $1\times$ training schedule, while after the 16-th and 22-th epoch for $2\times$ training schedule. For ablation studies, we adopt single-scale training and test with $1333 \times 800$ resolution.

We also provide experiment results on another popular object detection benchmark, *i.e.*, PASCAL VOC [1]. We use VOC 07+12 training protocol, *i.e.*, the union of VOC 2007 trainval set and VOC 2012 trainval set (16551 images) for training, and VOC 2007 test set (4952 images) for evaluation. The initial learning rate is 0.01 and the total training epochs are set to 4. The learning rate decreases by a factor of 10 after the 3-rd epoch. We report results in terms of AP and 5 mAPs under different IoU thresholds.

## 2. More Method Studies

**Individual LD.** To investigate the performance of LD, we do not use the ground-truth bounding boxes for training the student detector by disabling the bounding box regression loss $\mathcal{L}_{\text{reg}}$ and the distribution focal loss $\mathcal{L}_{\text{DFL}}$ in Eqn. (5). From Table 1, we can see that when we only use LD within the main distillation region to train the student ResNet-18, we attain 36.4 AP and 39.3 $\text{AP}_{75}$. These results are surprisingly higher than those using $\mathcal{L}_{\text{reg}}$ and $\mathcal{L}_{\text{DFL}}$ (the first row), reflecting the capability of LD in helping the student to learn rich localization knowledge. With both $\mathcal{L}_{\text{reg}}$ and $\mathcal{L}_{\text{DFL}}$ added, only 0.1 AP is increased, indicating that the probability distribution learned by the teacher detector is a better localization supervision than the ground-truth bounding box annotations.

Table 1. Evaluation of **Individual LD** on the main distillation region. The results are reported on MS COCO val2017. **R**: ResNet [3].

| Teacher | Student | $\mathcal{L}_{\text{LD}}$ | $\mathcal{L}_{\text{reg}}$ and $\mathcal{L}_{\text{DFL}}$ | AP | $\text{AP}_{75}$ |
|---|---|---|---|---|---|
| **R**-101 | **R**-18 | | ✓ | 35.8 | 38.2 |
| | | ✓ | | 36.4 | 39.3 |
| | | ✓ | ✓ | 36.5 | 39.3 |

**Self-LD.** In KD, the student model $\boldsymbol{S}$ is generally lighter than the teacher model $\boldsymbol{T}$ so as to learn compact and efficient models. Recently, self-distillation has been consistently observed to have a positive effect in classification [2, 5]. For object detection, it is also positive to see that Self-LD with $\boldsymbol{S} = \boldsymbol{T}$ can also bring performance gains. For the reason why self-distillation can facilitate the model accuracy, [4] firstly unveils the mystery by theoretically analyzing self-distillation in Hilbert Space. It has been proven that a few rounds of self-distillation can reduce over-fitting since it induces the regularization. However, continuing self-distillation may lead to under-fitting, and we simply perform self-LD once. As listed in Table 2, Self-LD on the main distillation region consistently boosts the performance by +0.3 AP, +0.3 $\text{AP}_{75}$ for ResNet-18, +0.5 AP, +0.7 $\text{AP}_{75}$ for ResNet-50, and +0.6 AP, +0.7 $\text{AP}_{75}$ for ResNeXt-101-32x4d-DCN. Self-LD shows the universality of our LD that the localization knowledge can still be transferred when the teacher model is with the same scale as the student model.

**Inference Speed.** Since our method needs to transform the bbox representation to probability distribution, the only modification to the detector lies in the localization head output channel, which is from $H \times W \times 4$ to $H \times W \times 4n$. We also investigate this effect on model size, computations (FLOPs), and running speed (FPS) in Table 3. We can see that our method can significantly improve the performance with negligible increase on model size and FLOPs for both FCOS and ATSS. While for RetinaNet, it leads to a slight

*Equal contribution.
†Corresponding author.

Table 2. Quantitative results of **Self-LD** on the main distillation region. The results are reported on MS COCO val2017.

| Detector | Self-LD | AP | $AP_{75}$ |
|---|---|---|---|
| ResNet-18 | | 35.8 | 38.2 |
| | ✓ | **36.1 (+0.3)** | **38.5** |
| ResNet-50 | | 40.1 | 43.1 |
| | ✓ | **40.6 (+0.5)** | **43.8** |
| ResNeXt-101-32x4d-DCN | | 46.9 | 51.1 |
| | ✓ | **47.5 (+0.6)** | **51.8** |

Table 3. Effect of using our LD on model size, FLOPs and FPS. FPS is measured using a RTX 2080 Ti GPU and averaged over 3 runs.

| | LD | image size | #param. | FLOPs | FPS | AP |
|---|---|---|---|---|---|---|
| RetinaNet | | $1333 \times 800$ | 37.74M | 239.32G | 21.0 | 36.9 |
| | ✓ | $1333 \times 800$ | 39.07M | 267.64G | 19.6 | 39.0 |
| FCOS | | $1333 \times 800$ | 32.02M | 200.50G | 22.3 | 38.6 |
| | ✓ | $1333 \times 800$ | 32.17M | 203.60G | 22.4 | 40.6 |
| ATSS | | $1333 \times 800$ | 32.07M | 205.21G | 21.9 | 39.2 |
| | ✓ | $1333 \times 800$ | 32.22M | 208.36G | 21.9 | 41.6 |
| GFocal | | $1333 \times 800$ | 32.22M | 208.31G | 21.9 | 40.1 |
| | ✓ | $1333 \times 800$ | 32.22M | 208.31G | 21.9 | 42.1 |

FPS drop. This is mainly because RetinaNet uses 9 anchor boxes per location. The modification of the localization head output channel of RetinaNet is from $H \times W \times (9 \times 4)$ to $H \times W \times (9 \times 4n)$, 8 times more than those of FCOS and ATSS. As for GFocal, we obtain a free improvement.

**LD for Lightweight Detectors on PASCAL VOC.** We further conduct experiments on PASCAL VOC to check the effectiveness of our LD. Here, the main distillation region is adopted. From Table 4, we can see that our LD consistently boost the performance for the student ResNet-18. Notice that our LD performs significantly better than the baseline for the AP metrics under high IoU threshold, like $AP_{80}$ and $AP_{90}$.

## 3. More Visualization

First, we provide more detection results by GFocal and our LD in Fig. 1, from which one can see more accurate localization boxes are obtained by LD. Then, we present more detection results by GFocal and LD with different thresholds of NMS, as shown in Fig. 2. Since our LD contributes to improve localization quality of detected boxes (referring to the results of *LD '0.95' v.s. GFocal '0.95'*), redundant boxes can be suppressed by the default NMS (referring to the results of *GFocal v.s. LD*).

## References


[1] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[2] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *ICML*, pages 1607–1616, 2018.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[4] Hossein Mobahi, Mehrdad Farajtabar, and Peter L Bartlett. Self-distillation amplifies regularization in hilbert space. *arXiv preprint arXiv:2002.05715*, 2020.

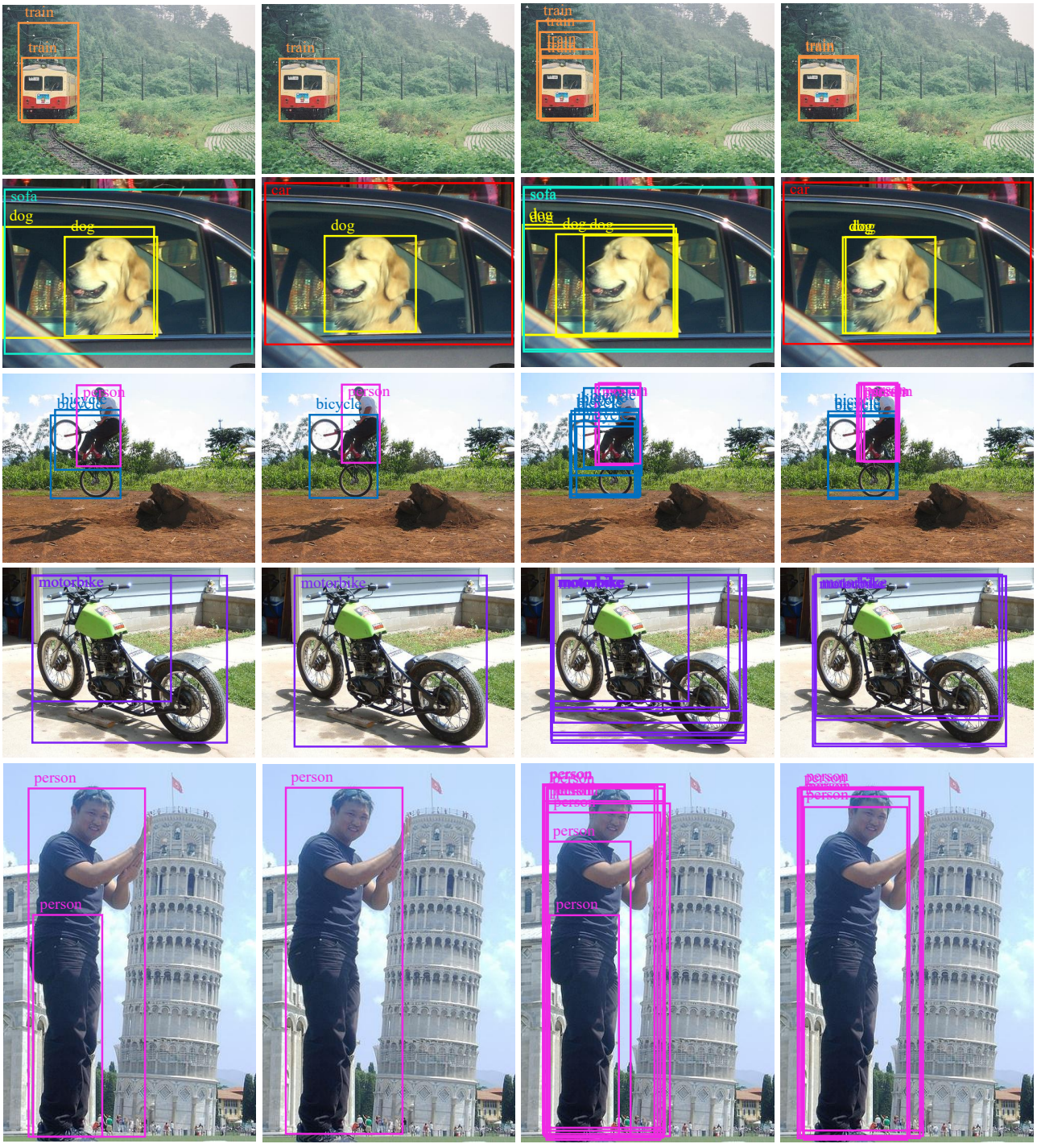
[5] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *ICCV*, pages 3713–3722, 2019.

Table 4. Quantitative results of LD on the main distillation region. The results are reported on the test set of PASCAL VOC 2007.

| Teacher | Student | AP | $AP_{50}$ | $AP_{60}$ | $AP_{70}$ | $AP_{80}$ | $AP_{90}$ |
|---|---|---|---|---|---|---|---|
| – | ResNet-18 | 51.8 | 75.8 | 72.0 | 62.9 | 46.5 | 20.6 |
| ResNet-34 | | 52.9 | 75.7 | 72.2 | **64.5** | **49.1** | 22.0 |
| ResNet-50 | | 52.8 | 75.4 | 72.1 | 64.0 | 49.0 | **23.0** |
| ResNet-101 | | **53.0** | **75.9** | **72.4** | 64.0 | 48.9 | 22.7 |
| ResNet-101-DCN | | 52.8 | 75.3 | 72.0 | 64.2 | 49.0 | 22.1 |


GFocal LD GFocal LD

Figure 1. Detection results by original GFocal and LD. Our LD makes detected boxes become more accurate.



GFocal LD GFocal ‘0.95’ LD ‘0.95’

Figure 2. Detection results by GFocal and LD. ‘0.95’ means that default NMS threshold 0.6 in GFocal is increased to 0.95. Detected boxes by original GFocal are with varying localization qualities, and redundant boxes may still survive after default NMS. In contrast, our LD contributes to improve localization quality of detected boxes, among which redundant ones are easy to be suppressed by default NMS.