LGD: Label-Guided Self-Distillation for Object Detection

Peizhen Zhang,†1 Zijian Kang,‡2 Tong Yang,† Xiangyu Zhang,‡1 Nanning Zheng,‡1 Jian Sun †1

1MEGVII Technology, ‡Xi’an Jiaotong University
{zhangpeizhen, yangtong, zhangxiangyu, sunjian}@megvii.com, kzj123@stu.xjtu.edu.cn, nnzheng@mail.xjtu.edu.cn

Abstract
In this paper, we propose the first self-distillation framework for general object detection, termed LGD (Label-Guided self-Distillation). Previous studies rely on a strong pretrained teacher to provide instructive knowledge that could be unavailable in real-world scenarios. Instead, we generate an instructive knowledge based only on student representations and regular labels. Our framework includes sparse label-appearance encoder, inter-object relation adapter and intra-object knowledge mapper that jointly form an implicit teacher at training phase, dynamically dependent on labels and evolving student representations. They are trained end-to-end with detector and discarded in inference. Experimentally, LGD obtains decent results on various detectors, datasets, and extensive tasks like instance segmentation. For example in MS-COCO dataset, LGD improves RetinaNet with ResNet-50 under $2 \times$ single-scale training from 36.2% to 39.0% mAP (+2.8%). It boosts much stronger detectors like FCOS with ResNeXt-101 DCN v2 under $2 \times$ multi-scale training from 46.1% to 47.9% (+1.8%). Compared with a classical teacher-based method FGFI, LGD not only performs better without requiring pretrained teacher but also reduces 51% training cost beyond inherent student learning. Codes are available at https://github.com/megvii-research/LGD.

Introduction
Knowledge distillation (KD) (Romero et al. 2015; Hinton, Vinyals, and Dean 2015) is initially proposed for image classification and obtains impressive results. Typically, it is about transferring instructive knowledge from a pretrained model (teacher) to a smaller one (student). Recently, KD applied to the fundamental object detection task, has aroused researchers' interests (Li, Jin, and Yan 2017; Wei et al. 2018; Wang et al. 2019; Zhang et al. 2020; Dai et al. 2021; Guo et al. 2021; Zhang and Ma 2021; Yao et al. 2021). Existing works achieve respectable performance but the choice of teacher is sophisticated and inconsistent among them. One common ground is that they all require a heavy pretrained teacher as it is discovered by recent works (Zhang and Ma 2021; Yao et al. 2021) that distillation efficacy could be enhanced with stronger teachers. Yet the pursuit for an ideal teacher could scarcely be satisfied in real-world applications, since it might take tons of efforts on trial and error (Peng et al. 2020). Instead, the issue that “KD for generic detection without pretrained teacher” is barely investigated.

To alleviate the pretrained teacher dependence, teacher-free schemes are proposed like (a) self-distillation, (b) collaborative learning and (c) label regularization, where instructive knowledge could be cross-layer features (Zhang et al. 2019), competitive counterparts (Zhang et al. 2018) and modulated label distribution (Yuan et al. 2020), etc. However, these methods are designed for classification and are inapplicable to detection since the latter has to handle multiple objects with different locations and categories but singe image classification. Lately, LabelEnc (Hao et al. 2020) extends traditional label regularization by introducing location-category modeling with an isolated network. It produces label representations with which the student features are supervised. Though it obtains impressive results, we find the improvement saturates (Figure 3) as detector grows stronger, e.g., with larger backbones and multi-scale training. We conjecture this is because labels themselves describe only object-wise categories and locations, without consider-
In short, our contributions are three-fold:

1. We propose a new self-distillation framework for general object detection. Unlike previous methods that use a convolution network as teacher, LGD generates instructive knowledge on-the-fly without pretrained teacher and improves the detection quality under limited training cost.

2. We introduce inter-and-intra relation to model a new instructive knowledge, rather than simply extract existent relation from student and teacher for distillation.

3. The proposed method outperforms previous teacher-free SOTA with higher upper limit and is better than classical teacher-based method FGFI in strong student settings. Beyond inherent student learning, it saves 51% training time against the classical teacher-based distillation.

**Related Work**

**Detection KD with Pretrained Teachers**

Unlike classification, knowledge transfer for object detection is more challenging. In detection, models are asked to predict multiple instances with diversified categories distributed at different locations in the image. (Li, Jin, and Yan 2017) proposed Mimic to distill activations within the region proposals predicted by RPN (Ren et al. 2015). (Chen et al. 2017) introduced weighted cross-entropy and bounded regression loss for enhancing the performance. To further exploit the context information of the distilling regions around the objects, (Wang et al. 2019) extended the ground-truth box regions by anchor-assigned ones. For learning adapted sampling weight for different knowledge, (Zhang et al. 2020) proposed PAD with uncertainty modeling. Besides intermediate feature hints, (Dai et al. 2021) involved the prediction map distillation obeying the assignment rules and relation distillation (Park et al. 2019) upon their defined general instances. Instead of focusing on foreground regions only, (Guo et al. 2021) decoupled the fore/back-ground knowledge transfer. To facilitate region-agnostic distillation, (Zhang and Ma 2021) proposed dual distillation with spatial-channel-wise attention. To resolve the feature resolution mismatching in cross-layer distillation and mitigate the misaligned label assignment, (Yao et al. 2021) introduced G-DetKD. Above methods mainly conducted feature-based distillation which is followed in this work. Whereas, they are designed for settings with strong...
Teacher-free Methods

Beyond traditional KD with pretrained teacher, there are teacher-free schemes that could be divided into three categories: (1) self-distillation (2) collaborative learning and (3) label regularization. (1) self-distillation excavates instructive knowledge from model itself. For instance, (Yang et al. 2019; Kim et al. 2020) used previously saved snapshots as teachers. In (Zhang et al. 2019), network was divided into sections that deeper layers were used to teach the shallower ones. In MetaDistiller (Liu et al. 2020), the knowledge stemmed from one-step predictions. (2) Collaborative learning involves multiple students to boost each other. (Zhang et al. 2018) proposed deep mutual learning (DML) where student networks with identical architecture learned collaboratively. (Lan, Zhu, and Gong 2018) proposed ONE by considering ensemble learning in branch-granularity. In KDCL (Guo et al. 2020), predictions were fused together as instructive knowledge. Likewise in (Chen et al. 2020a), ensemble logits of multiple students were aggregated to distill another. (Furlanello et al. 2018) proposed Born-Again Network (BAN) that leveraged information from last generations to distill the next. (3) For label regularization, (Yuan et al. 2020) proposed tf-KD for regularized label distributions to distill the next. (3) For label regularization, (Yuan et al. 2020) proposed tf-KD for regularized label distribution beyond label smoothing (Szegedy et al. 2016). How-ever, above methods were designed for classification only.

Recently, there have been newly-built label regularization methods (Mostajabi, Maire, and Shakhnarovich 2018; Hao et al. 2020) using an isolated network to explicitly model labels as features for supervision, w.r.t. semantic segmentation and detection. They obtained impressive results. In (Hao et al. 2020), dense color maps with category and location information were constructed and fed into an auto-encoder-like network to fetch label representations. However, they considered each object modeling separately which was suboptimal. Instead, we propose to generate instructive knowledge by inter-object and intra-object relation modeling to form a self-distillation scheme with higher upper limit.

Method

As shown in Fig. 2, we illustrate the modules in LGD as follows: (1) An encoder that computes label and appearance embeddings. (2) An inter-object relation adapter that generates interacted embeddings given label and appearance embeddings of objects. (3) An intra-object knowledge mapper that back-projects interacted embeddings onto feature map space to obtain instructive knowledge for distillation.

Label-appearance Encoder

(1) Label Encoding: For each object, we concatenate its normalized ground-truth box $\{\tilde{x}_1, y_1, \tilde{x}_2, y_2\}$ and one-hot category vector to obtain a descriptor. The object-wise descriptors are passed into a label encoding module for refined label embeddings $L = \{l_i \in \mathbb{R}^C\}_{i=0}^N$, where $i$ indicates object index, $C = 256$ is the intermediate feature dimension, and $N$ is the object number. $i = 0$ indexes the context object. To introduce basic relation modeling among label descriptors and maintain a permutation-invariant property, we adopt the classical PointNet (Qi et al. 2017) as the label encoding module. It processes the descriptors by a multi-layer perceptron (Friedman et al. 2001) with local-global modeling by a spatial transformer network (Jaderberg et al. 2015). Also, the label descriptors are similar to point set that is accustomed to PointNet (bounding boxes could be viewed as points in 4-dimensional Cartesian space). Empirically, using PointNet as encoder behaves slightly better than MLP or transformer encoder (Vaswani et al. 2017) (Table 4). We further replace the BatchNorm (Ioffe and Szegedy 2015) with LayerNorm (Ba, Kiros, and Hinton 2016) to adapt the small-batch detection setting. Notably, the above 1D object-wise label encoding manner is more efficient than that in LabelEnc. The LabelEnc constructs an ad-hoc color map $e = \mathbb{R}^{H \times W \times K}$ to describe labels where $(H, W)$ and $K$ are input resolution and object category number respectively $(HWK \gg C)$. The color map is processed by an extra CNN and pyramid network for 2D pixel-wise representations $L' = \{l'_i \in \mathbb{R}^{H_p \times W_p \times C}, 1 \leq p \leq P\}$. $P$ refers to the number of pyramid scales (Lin et al. 2017a) that $(H_p, W_p)$ denotes feature map resolution at scale $p$.

(2) Appearance Encoding: Beyond label encoding, we retrieve compact appearance embeddings from feature pyramid of student detector that contains appearance feature of perceived objects. We adopt a handy mask pooling to extract object-wise embeddings from the feature maps. Specifically, we pre-compute the object-wise masks: $M = \{m_i\}_{i=1}^N = \mathbb{R}^{H \times W}$ covering the entire image. For each object $i (0 \leq i \leq N)$, $m_i \in \mathbb{R}^{H \times W}$ is a binary matrix whose values are set as 1 inside the ground-truth region and 0 otherwise. The mask pooling is conducted concurrently for all pyramid levels, at each of which, object-wise masks at input level are down-scaled to corresponding resolution to become scale-specific ones. At $p$-th scale, the appearance embedding $a_p \in \mathbb{R}^C$ is obtained by calculating channel-broadcasted Hadamard product between the projected feature map $F_{proj}(X_p) \in \mathbb{R}^{H_p \times W_p \times C}$ and down-scaled object mask $m_{proj} \in \mathbb{R}^{H_p \times W_p}$, followed by global sum pooling. $F_{proj}(\cdot)$ is a single $3 \times 3$ conv layer. Thus, we collect appearance embeddings: $A_p = \{a_i \in \mathbb{R}^C\}_{i=0}^N$ for each object at level $p$ (likewise for the other levels).

Inter-object Relation Adapter

Given label and appearance embeddings, we formulate the inter-object relation adaption by a cross-attention process. In Fig. 2, this process is executed at every student appearance pyramid scale to retrieve the interacted embeddings. We omit the pyramid scale subscript below for brevity.

During the cross attention, a sequence of key and query tokens are leveraged in calculating KQ-attention relation for aggregating value to obtain attention outputs. For achieving the label-guided information adaption, we exploit the ap-
pearance embeddings $A$ at current scale as query, and the scale-invariant label embeddings $L$ as key and value. The attention scheme measures the correlation between lower-level structural appearance information and higher-level label semantics among objects then reassembles the informative label embeddings for dynamic adaption. Before conducting attention, the query, key, and value are transformed by linear layers $f_{Q}$, $f_{K}$ and $f_{V}$, respectively. We then computed the interacted embeddings $u_{i} \in \mathbb{R}^{C}$ for $i$-th object by weighting each transformed label embedding $f_{V}(l_{j})$ by label-appearance correlation factor $w_{ij}$.

$$u_{i} = \sum_{j=0}^{N} w_{ij} f_{V}(l_{j})$$  \hspace{1cm} (1)$$

$w_{ij}$ is calculated by a scaled dot-product between $i$-th appearance embeddings $a_{i}$ and $j$-th label embeddings $l_{j}$ followed by a softmax operation:

$$w_{ij} = \frac{\exp(f_{Q}(a_{i}) \cdot f_{K}(l_{j})/\tau)}{\sum_{k=0}^{N} \exp(f_{Q}(a_{i}) \cdot f_{K}(l_{k})/\tau)}$$  \hspace{1cm} (2)$$

where $\cdot$ is the notation for inner product and $\tau = \sqrt{C}$ is the denominator for variance rectification (Vaswani et al. 2017).

Specifically, for more robust attention modeling, the paradigm actually involves $T$ set of concurrent operations termed heads to obtain partial interacted embeddings in parallel. By concatenating the partial interacted embeddings from all heads and applying a linear projection $f_{P}$, we obtain interacted embeddings $E = \{e_{i} \in \mathbb{R}^{C}\}_{i=0}^{N}$ for all objects:

$$e_{i} = f_{P}([u_{i}^{1}; u_{i}^{2}; \ldots; u_{i}^{T}])$$  \hspace{1cm} (3)$$

where $[;]$ denotes the concatenation operator that combines the partial embeddings along the channel dimension. The resulting embeddings are also scale-sensitive as the appearance embeddings. As aforementioned, we obtain interacted embeddings across scales by iterating over all feature scales.

Technically, above computation is accomplished by means of multi-head self attention (MHSA) (Vaswani et al. 2017). Note that our framework is decoupled to the specific choice. As will be shown in this paper, LGD shows the efficacy even with the naive transformer. It is likely to perform even better by using advanced variants like focal transformer (Yang et al. 2021) but that is beyond the scope.

**Intra-object Knowledge Mapper**

To make the 1D interacted embeddings applicable to widely-used intermediate feature distillation (Li, Jin, and Yan 2017; Wang et al. 2019) for detection, we map the interacted embeddings onto 2D feature map space to fetch instructive knowledge. Naturally, for each pyramid scale $p$, $(1 \leq p \leq P)$, the resolutions of resulting maps are confined to be identical with corresponding student feature maps.

Intuitively, since spatial topology is not maintained in label encoding for compact representations (Sec. ), it is important to recover the localization information for each object to achieve alignment in geometric perspective. Naturally, object bounding box regions serve as good heuristics. We fill each object-binding interacted embedding within its corresponding ground-truth box region on a zero-initialized feature map. In practice, for each object $i$, we acquire its feature map at $p$-th scale by calculating matrix multiplication between the vectorized object mask $m_{i} \in \mathbb{R}^{H_{p} \times W_{p} \times 1}$ and the projected, interacted embedding $e_{i}$. All these object-wise maps are added up to a unified one followed by a refinement module $F_{ref}(\cdot)$ to form the instructive knowledge:

$$X_{p}^{I} = F_{ref}\left(m_{0}F_{ctx}^{T}(e_{0}) + \mathcal{G}\left(\sum_{i=1}^{N} m_{i}F_{inst}^{T}(e_{i})\right)\right)$$  \hspace{1cm} (4)$$

where $F_{ctx}(\cdot)$ and $F_{inst}(\cdot) \in \mathbb{R}^{1 \times C}, (1 \leq i \leq N)$ are the transposes of projected context and normal object interacted embeddings, respectively. Both $F_{ctx}(\cdot)$ and $F_{inst}(\cdot)$ are single fc layers. $\mathcal{G}(\cdot)$ is a single $3 \times 3$ conv layer. $F_{ref}(\cdot)$ starts with a $relu$ followed by three $3 \times 3$ conv layers. Thus, we collect the instructive knowledge $X^{I} = \{X_{p}^{I} \in \mathbb{R}^{H_{p} \times W_{p} \times C}\}_{p=1}^{P}$ at all scales.

Beyond applicability consideration, the above mapping implies a spirit of *intra-object* regularization (Yun et al. 2020; Law and Deng 2018; Chen et al. 2020b) which enforces activation neurons inside the same foreground region on student appearance representations to be close (through subsequent distillation in Equation 5). Moreover, these instructive representations will be supervised with detection loss for ensuring the representation capability (Equation 6).

Before distillation, an adaption head $F_{adapt}(\cdot)$ is used to adapt student representations, following FitNet. We conduct knowledge transfer between the instructive representations $X^{I}_{p}$ and the adapted student features $X^{S}_{p} = F_{adapt}(X^{p})$ at each feature scale. We adopt InstanceNorm (Ulyanov, Vedaldi, and Lempitsky 2016) to eliminate the appearance and label style information for both feature maps followed by a Mean-Square-Error (MSE):

$$L_{distill} = \frac{1}{N_{total}} \sum_{p=1}^{P} \left\|X^{S}_{p} - X^{I}_{p}\right\|^{2}$$  \hspace{1cm} (5)$$

where $P$ is the total number of pyramid levels, and $N_{total} = \sum_{p=1}^{P} H_{p} W_{p} C$ indicates the total size of the feature pyramid tensors. As gradient stopping technique suggested in previous studies (Hao et al. 2020; Hoffman, Gupta, and Darrell 2016), we detach instructive representations $X^{I}$ when calculating distillation loss to avoid model collapse.

Besides the distillation loss and detection loss for optimizing student detector, we further ensure the instructive representation quality and consistency with student representations by sharing the detection head for supervision. The overall detection loss is shown below:

$$L_{det} = L_{det}^{S}(\mathcal{H}(X), \mathcal{Y}) + L_{det}^{I}(\mathcal{H}(X^{I}), \mathcal{Y})$$  \hspace{1cm} (6)$$

where $\mathcal{H}(\cdot)$ refers to the detection head. $\mathcal{Y}$ stands for the label set (boxes and categories). In summary, the total training objective is:

$$L_{total} = L_{det} + \lambda L_{distill}$$  \hspace{1cm} (7)$$
where \( \lambda \) is a trade-off for distillation term and we simply adopt \( \lambda = 1 \) throughout all experiments. For stable training, the distillation starts in 30k iterations since it could be detrimental when the instructive knowledge is optimized insufficiently (Hao et al. 2020; Liu et al. 2020). The student detector backbone is frozen in early 10k iterations under 1× training schedule and 20k for 2× training schedule.

### Experiment

#### Experiments Setup

The proposed framework is built upon Detectron2 (Wu et al. 2019). Experiments are run with batch size 16 on 8 GPUs. Inputs are resized such that shorter sides are no more than 800 pixels. We use SGD optimizer with 0.9 momentum and \( 10^{-4} \) weight decay. The multi-head attention in interobject relation adapter uses \( T = 8 \) heads following common practice. For brevity, we denote by R-50, R-101 and R-101 DCN for ResNet-50, ResNet-101 and ResNet-101 with deformable convolutions v2 (Zhu et al. 2019). Main experiments are validated on MS-COCO (Lin et al. 2014) dataset that we also testify on others: Pascal VOC (Everingham et al. 2010) and CrowdHuman (Shao et al. 2018).

**MS-COCO** is a challenging object detection dataset with 80 categories. Mean average precision (AP) is used as the major metric. Following common protocol (He, Girshick, and Dollár 2019), we use the `trainval-115k` and `minival-5k` subsets w.r.t. training and evaluation. We denote by \( 1 \times \) the training for 90k iterations where learning rate is divided by 10 at 60k and 80k iterations. By analogy, \( 2 \times \) denotes 180k of iterations with milestones at 120k and 160k. We term the single and multi-scale training by \( ss \) and \( ms \) for short.

**Pascal VOC** is a dataset with 20 classes. The union of `trainval-2007` and `trainval-2012` subsets are used for training, leaving `test-2007` for validation. We report mAP and AP50/75 (AP with overlapping threshold 0.5/0.75). Models are trained for 24k iterations with milestones at 18k and 22k.

**CrowdHuman** is the largest crowd pedestrian detection dataset, containing 23 people per image. It includes 15k and 4370 images w.r.t. training and validation. The major metric is average log miss rate over false positives per image (termed mMR, lower is better). Models are trained for 30 epochs with learning rate decayed at 24th and 27th epoch.

### Comparison to Teacher-free Methods

#### Detailed Comparison with State-of-the-Art.

As shown in Figure 3 and Table 1, we compare our LGD framework with the baseline and previous teacher-free SOTA, i.e., the LabelEnc (Hao et al. 2020) regularization method. We verify the efficacy on MS-COCO on three popular detectors: Faster R-CNN (Ren et al. 2015), RetinaNet (Lin et al. 2017b) and FCOS (Tian et al. 2019). Figure 3 shows the result trending as student detector grows stronger (longer periods: \( 1 \times \rightarrow 2 \times \), scale augmentations: \( ss \rightarrow ms \) and larger backbones: R-50 \( \rightarrow \) R-101 \( \rightarrow \) R-101 DCN). Our model compares favorably to or is slightly better than LabelEnc in earlier settings. For RetinaNet or FCOS R-50 at \( 2 \times ss \) setting, the baseline runs into overfitting while our method tackles that and achieves 2.8% mAP gain. Notably, as the detector setting becomes stronger, the gain of LabelEnc shrinks rapidly while ours still consistently boosts the performance. For Faster R-CNN with R-101 and R-101 DCN, LabelEnc underperforms the baseline (41.4 vs. 41.7 and 44.0 vs. 44.1). Instead, our method manage to improve and surpasses LabelEnc at around 1% mAP; verifying higher upper limit. Likewise, for RetinaNet and FCOS with R-101 and R-101 DCN, our method could steadily achieve gains of 1.2~1.5%. Note that in traditional distillation schemes, it remains unknown to find suitable teacher for such strong students.

#### Comparison with Typical Methods.

As aforementioned, teacher-free schemes other than LabelEnc are NOT designed for detection. For surplus concern, we transfer and reimplement typical methods like DML, tf-KD and BAN to detection by substituting their logits distillation with intermediate feature distillation in mainstream detection KD literature (except tf-KD). As shown in Table 2, these methods obtain slight improvement or are even harmful (tf-KD). BAN performs the best among them. It obtains 0.6% improvement on RetinaNet \( 1 \times ms \) R-50 at a cost of actual 3× training periods. However, it fails to generalize to other settings.

<table>
<thead>
<tr>
<th>Method</th>
<th>RetinaNet</th>
<th>FRCN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>36.6</td>
<td>37.4</td>
</tr>
<tr>
<td>DML†</td>
<td>37.0</td>
<td>37.4</td>
</tr>
<tr>
<td>tf-KD†</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>BAN†, ♦</td>
<td>36.8</td>
<td>38.0</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>38.3</strong></td>
<td><strong>38.5</strong></td>
</tr>
</tbody>
</table>

Table 1: Detailed comparison with previous SOTA.

Table 2: Comparison with typical teacher-free methods. † denotes our transfer to detection. ♦ denotes reporting the 3rd generation result in BAN literature which costs 3× longer training schedules far more than regular 1×. Also, it is undefined for tf-KD to experiment on RetinaNet with focal loss.
Ablation Studies

<table>
<thead>
<tr>
<th>Method</th>
<th>AP</th>
<th>APs</th>
<th>APm</th>
<th>APL</th>
<th>ΔAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>36.6</td>
<td>21.2</td>
<td>40.4</td>
<td>48.1</td>
<td>--</td>
</tr>
<tr>
<td>MLP</td>
<td>37.9</td>
<td>21.5</td>
<td>41.9</td>
<td>49.7</td>
<td>+1.3</td>
</tr>
<tr>
<td>TransEnc</td>
<td>37.9</td>
<td>21.7</td>
<td>41.6</td>
<td>50.2</td>
<td>+1.3</td>
</tr>
<tr>
<td>PointNet</td>
<td>38.3</td>
<td>23.2</td>
<td>42.0</td>
<td>50.0</td>
<td>+1.7</td>
</tr>
</tbody>
</table>

Table 4: Label Encoder Ablation

Label Encoding. In this work, we adopt PointNet (Qi et al. 2017) as the label encoding module. In fact, other modules are also applicable. We conduct comparisons on three alternations under $2 \times ms$ schedule on MS-COCO with RetinaNet based on ResNet-50 backbone. Specifically, we compare PointNet with a MLP only network, and an encoder network composed of 6 scaled dot-product attention heads (Vaswani et al. 2017), abbreviated as “TransEnc”. Similar to the handling we have done upon PointNet, we feed label descriptors into these networks to obtain label embeddings. We respectively input these label embeddings to remaining LGD modules and examine. All variants achieve good results as shown in Table 4, which demonstrates the robustness of our framework. The PointNet we finally adopt is the best among three of them, perhaps owing to its local-global relationship modeling along label descriptors.

Inter-object Relation Adapter. As aforementioned in Sec , the proposed method adopts the student appearance embeddings as queries and label embeddings as keys and values to involve in the guided inter-object relation modeling (here abbreviated as “Student”). We also experiment with the reverse option that using label embeddings as queries (abbreviated as “Label”). As shown in Table 5, for RetinaNet and FRCN $1 \times ss$ with R-50 as backbone, the adopted “student” mode are 0.7% and 0.5% better than “Label” mode.

Intra-object Knowledge Mapper. As specified in Equation 4, the instructive knowledge is dependent on interacted embeddings of both actual objects and virtual context. We ablate their usage in Table 6a. As expected, the context alone is not helpful since mere context provides nothing useful towards object detection. It manages to enhance the performance when combined with object embeddings (+0.3%).

Comparison with Classical Teacher-based KD

We also compare the proposed teacher-free LGD with the classical teacher-based method, FGFI (Wang et al. 2019). Experiments are conducted on RetinaNet $2 \times ms$ with backbones R-50, 101 and 101 DCN respectively. As shown in Figure 1 and Table 3, our framework performs better when student gets stronger. Towards strong detector with R-101 DCN as backbone, LGD is 0.9% and 1.4% superior to LabelEnc and FGFI. The reason why the benefits of FGFI diminish might attribute to lack of much stronger teacher (Zhang and Ma 2021; Yao et al. 2021). We believe it is possible that FGFI with larger teacher or other stronger teacher-based detection KD can outperform ours, but such teacher-presumed setting is not the design purpose of our framework.

Table 3: Results corresponding to Figure 1. Our method is effective for stronger students compared with others.

Inter-object Relation Adaption ablations with RetinaNet, Faster R-CNN and FCOS with R-50 $1 \times ss$.
Table 6: Intra-object knowledge adapter ablations.

(a) Embedding Participation

(b) Head sharing choice

Table 7: Comparison of Training Cost (hours).

Table 8: Pascal VOC.

Table 9: CrowdHuman. mMR: the lower, the better.

Table 10: Comparison on instance segmentation.

Instance Segmentation. To further validate the versatility, we conduct experiments on instance segmentation on MS-COCO. In this task, a detector is required to simultaneously localize and segment each object. We experiment on Mask R-CNN (He et al. 2017). To fully utilize the labels, we replace the object-wise box masks (Section (2)) with the segmentation masks as better spatial prior. As shown in Table 10, our method boosts 1% and 0.8% mask-box AP with respect to Mask R-CNN R-50 and 101. Please refer to the supplementary materials for more details.

Conclusion

In this paper, we propose a brand new self-distillation framework, termed LGD for knowledge distillation in general object detection. It absorbs the spirits of inter-and-intra object relationship into forming the instructive knowledge given regular labels and student representations. The proposed LGD runs in an online manner with decent performance and relatively lower training cost. It is superior to previous teacher-free methods and a classical teacher-based KD method especially for strong student detectors, showing higher potential. We hope LGD could serve as a baseline for future detection KD methods without pretrained teacher.
References


He, K.; Girshick, R. B.; and Dollár, P. 2019. Rethinking ImageNet Pre-Training. In ICCV.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. B. 2017. Mask R-CNN. In ICCV.


Hoffman, J.; Gupta, S.; and Darrell, T. 2016. Learning with Side Information through Modality Hallucination. In CVPR.


Huang, Z.; Zou, Y.; Kumar, B. V. K. V.; and Huang, D. 2020. Comprehensive Attention Self-Distillation for Weakly-Supervised Object Detection. In NeurIPS.


Li, Q.; Jin, S.; and Yan, J. 2017. Mimicking Very Efficient Network for Object Detection. In CVPR.


Lin, T.; Goyal, P.; Girshick, R. B.; He, K.; and Dollár, P. 2017b. Focal Loss for Dense Object Detection. In ICCV.


Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. FCOS: Fully Convolutional One-Stage Object Detection. In ICCV.


Wei, Y.; Pan, X.; Qin, H.; Ouyang, W.; and Yan, J. 2018. Quantization mimic: Towards very tiny cnn for object detection. In *ECCV*.


3317