# Dist$^2$: Distribution-Guided Distillation for Object Detection

Tianchu Guo[1]
tianchu.gtc@alibaba-inc.com
Pengyu Li[1]
lipengyu007@gmail.com
Wei Liu[2]
ustclwwx@gmail.com
Bin Luo[1]
luwu.lb@alibaba-inc.com
Biao Wang[2]
wangbiao225@foxmail.com

[1] Artificial Intelligence Center
DAMO Academy
Alibaba Group

[2] Work was done when
they were employed by Alibaba

**Abstract**

Knowledge distillation has been widely used as an effective technique for model compression. Previous knowledge distillation methods in object detection mainly focus on designing loss functions to minimize the feature distances between the teacher and student networks. However, after these losses converge, the imitations are still far from perfect and the distance of the feature map between teacher and student is still large. In this paper, we propose a **Dist**ribution-guided **Dist**illation method named Dist$^2$ for object detection, which concentrates on eliminating the difference of the feature distributions between the teacher and student networks. The proposed Dist$^2$ consists of a Distribution Imitation (DI) mechanism and a Flexible Imitation (FI) strategy. Specifically, the DI mechanism guarantees the feature distribution from a certain part of the student network is as close as possible to that of the teacher network. Furthermore, the FI strategy is employed, which enables the distribution imitation to perform not only on the same part but also cross different parts between the student and teacher networks. Our experimental results on COCO and PASCAL VOC datasets show that the proposed Dist$^2$ outperforms the previous state-of-the-art feature imitation methods by a large margin.

## 1 Introduction

Knowledge distillation (KD), which transfers the knowledge from a teacher model to a student one, is widely used as an effective technique for model compression [9]. Inspired by the great success achieved in classification tasks [6, 21], many methods focus on introducing KD to compress models of detection tasks. Instead of minimizing the Kullback-Leibler Divergence in classification tasks, KD methods in detection tasks perform the feature map imitation between the teacher and student models [8, 21, 28]. Li et.al transferred the knowledge from the teacher to the student by employing a feature-based loss function, which

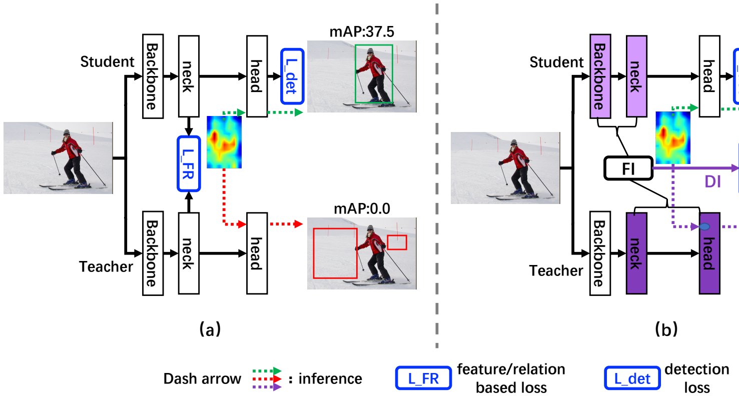

Figure 1: The comparison of existing methods and the proposed Dist$^2$. (a) Detection predictions fail if we forward the neck feature map of the student to the head of the teacher. The detection results are completely wrong even after the feature-based and relation-based losses converge to a small value, as shown in the red dash arrow. It means the distance of the feature map is far between the student and the teacher. (b) The architecture of the proposed Dist$^2$. The output feature map of the student is put forward to the teacher's part. Then the prediction of the teacher network is optimized by the detection loss. The proposed Dist$^2$ explicitly utilizes the full potential of the teacher to distillate the student making the feature distribution of the student is close to that of the teacher. Best viewed in color.

calculates the L2 distance of feature maps between the teacher and student models [10]. As the detection task is a dense prediction task [15], relation-based loss functions [16, 17, 24] are further proposed to reconstruct the structural knowledge of student's feature maps as that of teacher's feature maps.

Although employing feature-based or relation-based loss functions have achieved promising results for KD, their imitations of feature maps between teacher and student models are far from perfect. As shown in Fig.1-(a), the detection predictions still fail if we forward the neck feature map of the student to the head of the teacher, even after the feature-based and relation-based losses converge to a small value. The ideal distillation is that the student's feature map is the same as that of the teacher, therefore the student feature map can be correctly predicted by the teacher's head. Thus, we argue that the student feature maps fail to imitate the teacher's. This demonstrates only using feature-based and relation-based loss is not enough to perform an efficient feature imitation between the teacher and student models. Moreover, most of these feature-based or relation-based methods mainly focus on the knowledge distillation of the backbone and neck parts for an object detector. This does not make full use of the teacher model and is incomplete for effectively transferring knowledge from the teacher to the student.

To tackle the above problems, we propose a **dist**ribution-guided **dist**illation method named Dist$^2$ for knowledge distillation on object detection, as shown in Fig.1-(b). The proposed Dist$^2$ consists of a Distribution Imitation (DI) mechanism and a Flexible Imitation (FI) strategy. In the DI mechanism, we directly forward the student's feature map to the teacher to make the final predictions. Then optimized the prediction of the teacher. This explicitly requires when given the same input image, features of the student should be the same as those of the teacher, so that the teacher can make correct predictions even with the student feature map. By employing our DI mechanism, the proposed Dist$^2$ not only makes full use of the teacher network but also guarantees the feature distribution from a certain part of the student network is as close as possible to that of the teacher network. Based on the DI mechanism, we further extend it to a FI strategy , which utilizes distributions of different part feature maps in the teacher network for knowledge distillation, e.g. put student's featuer map to different parts of the teacher. Benefiting from the FI strategy, the distribution imitation of the proposed Dist$^2$ can be performed not only on the same part but also cross different parts between the student and teacher networks.

In Summary, the main contributions of this work are:

- A distribution imitation (DI) mechanism, which not only makes full use of the teacher network but also guarantees the feature distribution from a certain part of the student network is as close as possible to that of the teacher network.
- A flexible imitation (FI) strategy that performs distribution imitation not only on the same part but also cross different parts between the student and teacher networks.
- A distribution-guided distillation method Dist$^2$ for object detection, which can achieve state-of-the-art performance on both COCO and PASCAL VOC datasets. The proposed Dist$^2$ is robust on multiple detection framework, heterogeneous backbone, and different framework between teacher-student pairs.

# 2 Related Work

## 2.1 Object Detection

Object detection plays an important role in computer vision and attracts many researchers to make tremendous strides on it. Current mainstream object detection methods contain two-stage and one-stage detectors. The two-stage detectors give accurate results but are time-consuming, such as Faster-RCNN [19] et.al. To get an efficient detector, one-stage detector [5, 14, 23] has been proposed. Among them, RetinaNet[13] introduces focal loss to balance the positive samples and negative samples, which makes the result of the one-stage detector is comparable to two-stage detectors. To avoid turning the pre-defined anchor size, Tian et.al [23] designs the anchor-free one-stage object detector, reaching a good performance with remarkable efficiency. Thus, one-stage detectors with knowledge distillation method are widely employ for fast, efficient detection task.

## 2.2 Knowledge Distillation

Knowledge distillation(KD) is a method that transfers the information from a teacher model to a student model, which is widely used for model compression. It is first proposed by Hinton [9] and achieved a good performance on the image classification task. The knowledge contains many types, such as soft targets of the output [9], the intermediate feature maps [21], and the instance relationship [18]. Dai et.al [2] classifies knowledge distillation methods into three categories. The first type is the response-based method [9], which minimizes the Kullback-Leibler Divergence of the soft output of the last layer. The second type is the feature-based method [21], which mimics the intermediate feature between the teacher and the student in terms of L2 loss. The third type is the relation-based method [16, 17, 24],

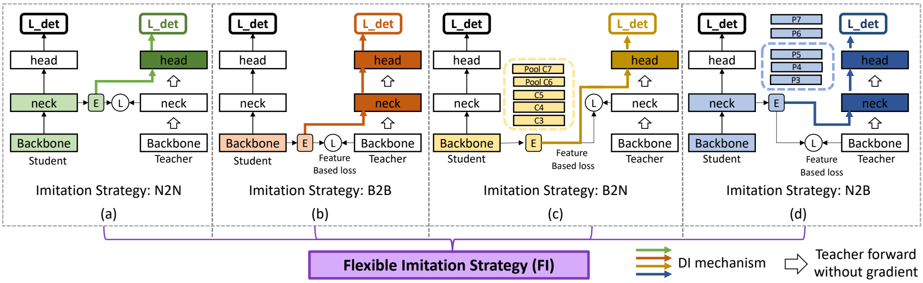

Figure 2: All the imitation strategies of the proposed Flexible Imitation (FI) strategy. The light color parts in the student net denote the feature map generator. The dark color parts in the teacher net denote the detection discriminator. Best viewed in color.

which employs structure constraints to guarantee the feature relationship between teacher and student is similar.

### 2.3 Knowledge Distillation on Object Detection

Recently, knowledge distillation is introduced into object detection tasks to obtain a high precision and lightweight model [10]. Knowledge distillation on detection tasks focuses on feature map imitation. The mimicking methods [10, 21] align the feature map in pixel level by using the L2 loss function. As the object detection task is a dense prediction problem [15], the structure information in the feature map should be considered. Thus, relation-based distillation methods are combined with feature-based methods. Liu et.al [15] proposes pairwise distillation methods to align a static affinity graph, capturing both short and long structure information among different locations in teacher and student networks. Zhang et.al [29] uses a non-local module [25] to capture the relation information. Dai et.al [2] constrains the relation between positive samples and background negative samples. All of them demonstrate that introducing the relation-based distillation helps feature map imitation and improves the performance. However, existing methods focus on minimizing the feature distances between the teacher and student networks. It causes the feature imitation far from perfect. In this paper, we study the distillation of the teacher's and student's feature distributions to reveal their importance for more efficient knowledge transfer.

## 3 Proposed Approach

In this section, we first introduce the preliminary of the knowledge distillation on object detection. Second, the proposed distribution imitation (DI) mechanism and the flexible imitation(FI) strategy are shown in Section 3.2 and Section 3.3. Finally, we describe the entire pipeline of the proposed Dist$^2$ in Section 3.4.

## 3.1 Preliminary

The general object detector contains three parts, i.e. the backbone, the neck, and the head. Given the input image $x$, the output of the detector is computed as,

$$[P,t] = F_{head}(F_{neck}(F_{backbone}(x))) \tag{1}$$

where the $F_{head}, F_{neck}, F_{backbone}$ denote the Convolutional Neural Networks (CNNs). The $P$ and $t$ denote the predicted classification scores and bounding box localizations, respectively. To train the object detectors, the parameters of the $F_{head}, F_{neck}, F_{backbone}$ are optimized by minimizing the following loss function,

$$L_{det} = L_{cls}(P, P^{gt}) + L_{reg}(t, t^{gt}) \tag{2}$$

where the $L_{cls}$ describes the classification loss function and $L_{reg}$ is the regression loss function. In our implementation, the $L_{reg}$ and $L_{reg}$ take the form of focal loss [13] and GIoU [20] respectively.

When distilling the knowledge from the teacher, the existing methods such as feature-based loss, perform on the output feature maps of the student and the teacher. The distillation progress optimizes the following formula,

$$L_{feat}(feat^S, feat^T) = \frac{1}{K} \| f_{trans}(feat^S) - feat^T \|_2^2 \tag{3}$$

where the $feat^S$ and $feat^T$ represent the student's and teacher's output feature maps. For example, if the feature map is the output of the neck part, the $feat^S$ is equal to $F^S_{neck}(F^S_{backbone}(x))$. The K denoting the feature map's size is $C \times W \times H$, where C, W, and H mean the feature channel, the width, and the height, respectively. The $f_{trans}$ represents a transfer layer to map the channel dimension of the student, i.e. $C'$ to the teacher's, i.e.$C$, which is implemented as a $1 \times 1$ convolution layer.

## 3.2 Distribution Imitation(DI) Mechanism

Inspired by the distribution generative method, the DI mechanism contains a feature map generator, which is part of the student denoted as $G_S$, and a detection discriminator, which is a part of the teacher denoted as $D_T$. The DI mechanism is to minimize the following formula,

$$\min_{G^S} DI = E_{x \sim p_{data}(x)} L_{det}(D^T(G^S(x)), (P^{gt}, t^{gt})) \tag{4}$$

As the $D^T$ is fixed, minimizing the $DI(G^S)$ is optimizing the parameter of $G^S$. Therefore, the DI will converge only on the situation that the $p_{G^S(x)} = p_{feat^T}$. The $p_{feat^T}$ is the feature distribution of the teacher's net. It means the output of $G^S$ can be perceived by $D^T$ and the distribution of the student feature map is close to that of the teacher.

In detail, as shown in Fig.2-(a), the backbone and the neck part of the student are treated as the feature generator $G^S$, denoted as the light green color. The teacher's head is treated as the detection discriminator $D^T$ denoted as the dark green color. The "E" denotes a transferring layer (eual to $f_{trans}$ in Eq. 3), which is a $1 \times 1$ convolution to align the student's channel dimension to the teacher's. The transferred feature map of the student's neck is put directly to the teacher's head. If the output of the feature map in the student's neck can be perceived by the teacher's head, it means that the feature map of the student's neck learn the knowledge of the teacher's neck.

The DI mechanism not only makes full use of the teacher network but also guarantees the feature distribution from a certain part of the student network is as close as possible to that of the teacher network.

### 3.3 Flexible Imitation(FI) Strategy

The effect of detection discriminator is to provide distribution information, Thus, extending it to multiple detection discriminators provides diverse distribution information and rich supervision. It motivated us to propose the Flexible Imitation (FI) Strategy, which makes the DI mechanism perform on not only the same part but also cross different parts between the student and teacher.

**The reason** why the FI strategy works on object detection tasks is described as follows. The backbone's feature can be detected by the head [14] directly. It means the backbone's feature contains semantic information, and it can be supervised by the head. Thus, the distribution imitation can be performed not only on the same part but also cross different parts between the teacher and student through the transferring layer "E".

There are four imitation strategies in the FI. Each of the imitation strategies is denoted as "X2Y". It means that the distribution of the X part in the student net imitates the distribution of the Y part in the teacher net. There are four types of the FI, i.e. N2N, B2B, B2N, and N2B. The "N" means the neck part and the "B" means the backbone part. They are described in Fig. 2. The generator is drawn in light color and the discriminators are denoted in dark color. More details are shown in the supplementary material.

### 3.4 The Proposed Dist$^2$

In summary, the proposed Dist$^2$ consists of the DI mechanism and all the strategies of FI.

The total loss of our Dist$^2$ is shown below,

$$L = L_{det}(P,t),(P^{gt},t^{gt}) + \lambda_{feat} \cdot \Sigma_{is} L_{feat}(feat_{is}^{S}, feat_{is}^{T}) + \lambda_{DI} \cdot \Sigma_{is} DI^{is} \tag{5}$$

where the $is$ is denoted as imitate strategy and $is \in [N2N, B2B, B2N, N2B]$. The $\lambda_{feat}$ and $\lambda_{DI}$ are the corresponding loss weights.

## 4 Experiment

In this section, first, the evaluation dataset and our implementation details will be described. Then the comparison results with the existing methods will be detailed in Section 4.2. The analysis of the DI mechanism and FI strategy will be shown in the next two subsections.

### 4.1 Settings

**Dataset.** The existing knowledge distillation methods and the proposed method will be evaluated on the MS COCO2017 dataset[11] and PASCAL VOC dataset [4]. The MS COCO2017 dataset is a large-scale object detection dataset, which contains 120k training images split and 5k Val images split for the test. The average precision (AP) is the measurement of all the methods on the MS COCO dataset. In the PASCAL VOC dataset, the

Table 1: Comparison with Existing Methods with FCOS [23] on MS COCO 2017 dataset.

| | Method | Imit. Strat. | mAP | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|
| FCOS [23] | ResNeXt101 (T) | - | 42.7 | 26.6 | 46.2 | 54.7 |
| | MobileV2C128 (S) | - | 30.4 | 17.0 | 33.0 | 39.3 |
| | +Hint [21] | N2N | 35.1$^{+1.1}$ | 18.0$^{+1.0}$ | 33.8$^{+0.8}$ | 40.3$^{+1.0}$ |
| | +pa [15] | N2N | 31.8$^{+1.4}$ | 17.8$^{+0.8}$ | 35.0$^{+2.0}$ | 40.4$^{+1.1}$ |
| | +NonLocal [29] | N2N | 33.3$^{+1.9}$ | 18.9$^{+1.9}$ | 36.1$^{+3.1}$ | 43.7$^{+4.4}$ |
| | **ours: +DI** | N2N | **34.7$^{+4.3}$** | **20.9$^{+3.9}$** | **37.7$^{+4.7}$** | **43.7$^{+4.4}$** |
| | **ours: +Dist$^2$** | All2All | **36.4$^{+6.0}$** | **20.5$^{+3.5}$** | **39.6$^{+6.6}$** | **46.9$^{+7.6}$** |

5k trainval images split in VOC 2007 and 16k trainval images split in VOC 2012 are for training. The 5k test images split in VOC 2007 is for the test.

**Implementation details.** The proposed method is tested on the anchor-free framework, i.e. the FCOS [23], and anchor-based framework, i.e. the RetinaNet [13]. In the teacher model, the backbone is employed the ResNeXt101[26] which is pretrained on ImageNet [3], and the neck is FPN [12], whose channel number is 256. In the student model, two different backbone models are chosen, i.e. the ResNet50 [7], and the MobileNetV2 [22] for testing generalization capability. The neck is FPN with the channel number is 128, denoted as C128.

The hyper-parameters $\{\lambda_{feat} = 0.1, \lambda_{DI} = 0.3\}$ are adopted for all the experiments in the proposed method. The mmdetection [1] is employed for all the model training. The training is conducted on 4 GPUs, with 4 images per GPU. The learning rate is 0.01. Other hyper-parameters for training on the MS COCO2017 dataset are followed the default setting in mmdetection [1].

## 4.2 Comparison with Existing Methods

**Comparison with structure based detection KD method.** This part shows the results on FCOS[23], which is an anchor-free detection framework. The student is MobileV2C128 [22]. The "+Hint" means using the feature-based distillation loss function proposed in [21]. The "+pa" in [15] and "+NonLocal" in [29] are the relation-based method. The hyper-parameters of these existing methods, i.e. the loss weights, are set following their original paper. The "ours: +DI" means our proposed DI mechanism. The existing methods distillate the knowledge on the output of the neck, which is corresponding to the imitation strategy N2N, denoted in the "Imit. Strat." column. The "Dist$^2$" means the DI mechanism with FI strategy combining all the imitation strategies, denoted as "All2All".

The experimental results of our proposed Dist$^2$ on MS COCO 2017 dataset are reported in Tab. 1. Compared with previous methods, the proposed DI mechanism can achieve state-of-the-art performance under all the measurements for the COCO dataset. Besides, our Dist$^2$ method further outperforms previous methods by a large margin. This fully demonstrates our proposed Dist$^2$, which focuses on the distillation of feature distribution between teacher and student models, can lead to a more efficient knowledge transfer than previous feature-based or relation-based methods.

In addition, it can be seen in Tab. 1 that, the improvement in $AP_L$ is more significant than that in $AP_S$, though the baseline(MobileV2C128 S) of the $AP_L$ is higher than $AP_S$. It demonstrates the DI mechanism is more conducive to improving the performance of large objects. Thus, the global structure information is important for large object detection, which is captured by the DI mechanism.

Table 2: Comparison with Existing Methods with RetinaNet [13] on MS COCO 2017 dataset.

| | Method | Imit. Strat. | mAP | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|
| Retina Net [13] | ResNeXt101 (T) | - | 41.1 | 24.0 | 44.8 | 54.1 |
| | MobileV2C128 (S) | - | 31.0 | 16.3 | 34.2 | 41.1 |
| | +Hint [21] | N2N | $31.6^{+0.6}$ | $\mathbf{17.0}^{+0.7}$ | $34.7^{0.5}$ | $42.7^{+1.6}$ |
| | +pa [15] | N2N | $31.4^{+0.4}$ | $15.5^{-0.8}$ | $33.9^{-0.3}$ | $43.8^{+2.7}$ |
| | +NonLocal [29] | N2N | $31.9^{+0.9}$ | $16.8^{+0.5}$ | $35.0^{+0.8}$ | $42.7^{+1.6}$ |
| | **ours: +DI** | N2N | $\mathbf{32.4^{+1.4}}$ | $16.7^{\mathbf{+0.4}}$ | $\mathbf{35.6^{+1.4}}$ | $44.0^{\mathbf{+2.9}}$ |
| | **ours: Dist$^2$** | All2All | $\mathbf{32.5^{+1.5}}$ | $\mathbf{17.5^{+1.2}}$ | $\mathbf{35.5^{+1.3}}$ | $\mathbf{44.0^{+2.9}}$ |

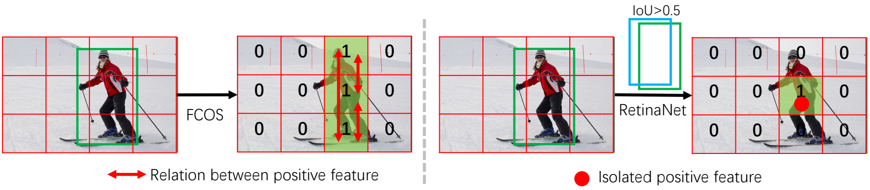

Figure 3: The positive sample definition of FCOS and RetinaNet. The FCOS promotes more points in the feature layer to be defined as positive samples. The RetinaNet contains an isolated positive sample in the feature map.

**Multiple Detection Framework.** The results evaluated on anchor-based framework, i.e. the RetinaNet[13] are shown in Tab.2. The proposed DI mechanism still outperforms the existing methods and the final mAP gain from "Dist$^2$" to the student model is about 1.5. Those results demonstrate that the proposed DI mechanism and the Dist$^2$ method can be used in multiple detection frameworks.

The gains of the anchor-based framework is lower than that of anchor-free one. It is because the definition of the positive and negative sample is different [30]. As shown in Fig.3, the positive sample points of the anchor-free method are continuous. On the contrary, in RetinaNet, the positive sample is defined according to the IoU between the anchor and the ground truth, which causes the positive sample to be **isolated**. The FI and DI provide the global relationship information of the feature map. The requirement of global relationships in RetinaNet is less than that of the FCOS. As the result, the improvement is lower than that of FCOS.

**Heterogeneous backbone.** To verify the generalization on the heterogeneous backbone, the ResNeXt101 is used as the teacher's backbone and the ResNet50C128 is employed as the student architecture. The detection framework is FCOS.

As shown in Tab. 3, the proposed DI mechanism with the imitation strategy N2N still outperforms the existing methods. With all the imitation strategies, the"Dist$^2$" archives 7 mAP gains compared with the original student net. Those results demonstrate that the proposed DI mechanism and the Dist$^2$ method can be used for heterogeneous backbone between teacher and student.

**Generalization on other datasets.** The results on the Pascal VOC dataset for FCOS detection framework are shown in Tab. 4.

Though the proposed DI mechanism with the imitation strategy N2N gets a comparable performance with that of the "NonLocal" method, the final Dist$^2$ using all the imitation

Table 3: Heterogeneous backbone results with FCOS on MS COCO 2017 dataset.

| Method | Imit. Strat. | mAP |
|---|---|---|
| ResNeXt101 (T) | - | 42.7 |
| ResNet50C128 (S) | - | 33.4 |
| +Hint [21] | N2N | 36.0$^{+2.6}$ |
| +pa [15] | N2N | 36.7$^{+3.3}$ |
| +NonLocal [29] | N2N | 37.5$^{+4.1}$ |
| **ours: +DI** | N2N | **38.3$^{+4.9}$** |
| **ours: +Dist$^2$** | All2All | **40.4$^{+7.0}$** |

Table 4: Generalization on PASCAL VOC dataset with FCOS.

| Method | Imit. Strat. | $AP_{50}$ |
|---|---|---|
| ResNeXt101(T) | - | 80.9 |
| ResNet50C128 (S) | - | 77.6 |
| +pa [15] | N2N | 77.8 |
| +NonLocal [29] | N2N | 78.5 |
| **ours: +DI** | N2N | 78.4 |
| **ours: Dist$^2$** | All2All | 78.7 |

Table 5: Comparison with SOTA Detection KD on FCOS.

| Method | mAP |
|---|---|
| baseline (Studnet) | 38.6 |
| +Hint [21] | 39.9$^{+1.3}$ |
| +GID [2] | 42.0$^{+3.4}$ |
| +G-DetKD [27] | 43.1$^{+4.5}$ |
| **ours: Dist$^2$** | 43.0$^{+4.4}$ |

Table 6: Evaluation results on different frameworks between Teacher-Student pairs.

| teacher | student | mAP |
|---|---|---|
| FCOS<br>(42.7) | RetinaNet(34.8) | 37.8$^{+4.4}$ |
| | FCOS(33.4) | 38.3$^{+4.9}$ |
| RetinaNet<br>(41.1) | RetinaNet(34.8) | 36.4$^{+1.6}$ |
| | FCOS(33.4) | 36.1$^{+1.3}$ |

strategies outperforms other existing methods. Those results demonstrate that the proposed DI mechanism and the Dist$^2$ method have an impressive generalization ability for different datasets.

**Comparison with other SOTA detection KD method.** Tab.5 shows the result of the comparison with existing detection KD methods. In this situation, the student backbone is ResNet50 with the FPN channel 256. Our method outperforms most of the existing method and achieves a comparable result with G-DetKD [27]. Note that the G-DetKD [27] use a more powerful teacher whose performance is 44.5. Our proposed method is a plug-in module which is compatible with existing methods.

## 4.3 Analysis of DI Mechanism

In this part, the teacher's backbone is ResNeXt101 and the student's backbone is ResNet50C128.

**Different frameworks between teacher-student pairs.**

The results in Tab.6 show that if the teacher and student come from the different detection frameworks, e.g., the teacher is FCOS and the student is RetinaNet, the performance can also be improved. It demonstrates the proposed DI mechanism can distill the high-level structure information from different detection frameworks. Also, the results show it gets better results when the student and the teacher are of the same detection framework.

**Imitation failure/success analysis.** We argue that the existing methods fail to imitate the teacher's feature map. In this part, the proposed DI mechanism will be tested to check whether it solves this problem. The feature map is the output of the student's backbone and neck part. It will be put into both teacher's head and the student's head.

The results are shown in the Tab. 7. After being trained using "+pa" and "+Nonlocal", the mAP evaluated from the teacher's head is still 0, denoted in the column "T. Head". Both of these methods fail to imitate the teacher's feature map. The mAP of "+DI" in the "T.

Table 7: Imitation failure/success analysis.

| FCOS ResNeXt101(42.7)-ResNet50C128(33.4) | | | |
|---|---|---|---|
| N2N | S.Backbone + S.Neck | | Imit. |
| | S. Head | T. Head | Failure/Success |
| +pa[15] | $36.7^{+3.3}$ | 0.0 | Failure |
| +NonLocal [29] | $37.5^{+4.1}$ | 0.0 | Failure |
| **+DI** | $\mathbf{38.3^{+4.9}}$ | **37.9** | **Success** |

Table 8: Evaluation results in each imitation Strategy of FI on COCO dataset.

| Imit. Strat. | N2N | B2B | B2N | N2B |
|---|---|---|---|---|
| +DI | **38.3**<br>**+4.9** | **38.3**<br>**+4.9** | **38.0**<br>**+4.6** | **37.9**<br>**+4.5** |

Table 9: Effect of the increasing number of imitation strategy.

| Imit. Strat.<br>Num. | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| mAP | 38.3 | 38.6 | 40.0 | **40.4** |

Head" column is 37.9. It means the DI mechanism makes the student successfully imitate the teacher's feature map.

**Sensitiveness of the loss weight.** The best weight of $\lambda_{DI}$ is set to 0.3. Other choices of the $\lambda_{DI}$ will be shown in the supplementary material.

## 4.4 Analysis of FI Strategy

In this part, the teacher's backbone is ResNeXt101 and the student's backbone is ResNet50C128. The detection framework is FCOS and the evaluated dataset is MSCOCO.

**Effect of each imitation strategy in FI.** The effect of each imitation strategy will be tested. The results in Tab.8 show that all the methods can improve the performance in any of the imitation strategies. It means that the feature imitation can be performed not only on the same part but also cross different parts between student and teacher, according to the corresponding transferring layer. It demonstrates learning distribution obtains more knowledge from the teacher.

**Effect of using more imitation strategies in FI.** With the increasing number of imitation strategies, the model needs more training epochs. All the results are trained to converge.

The results in Tab.9 show that as the number of the imitate strategies increases, the performance is constantly improved. It demonstrates the more imitation strategies are employed, the more knowledge from the teacher is learned by the student.

# 5 Conclusion

In this paper, we propose a Dist$^2$, consisting of a distillation imitation(DI) mechanism and the flexible imitation(FI) strategy. Different from the previous works, the DI mechanism distills the feature distribution of the teacher rather than minimizing the feature distance between the teacher and the student networks. In addition, the FI strategy makes the distribution imitation performs not only on the same part but also cross different parts between the student and teacher networks. The proposed method is evaluated on both COCO and PASCAL VOC datasets and outperforms the state of the art. It is also robust on multiple detection framework, heterogeneous backbone, and different framework between teacher-student pairs.

# References

[1] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[2] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2021.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[5] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017.

[6] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR, 2018.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3779–3787, 2019.

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[10] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 6356–6364, 2017.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[12] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[13] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[15] Yifan Liu, Changyong Shu, Jingdong Wang, and Chunhua Shen. Structured knowledge distillation for dense prediction. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[16] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. Knowledge distillation via instance relationship graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7096–7104, 2019.

[17] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019.

[18] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5007–5016, 2019.

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[20] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.

[21] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[22] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[23] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.

[24] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374, 2019.

[25] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[26] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[27] Lewei Yao, Renjie Pi, Hang Xu, Wei Zhang, Zhenguo Li, and Tong Zhang. G-detkd: Towards general distillation framework for object detectors via contrastive and semantic-guided feature imitation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3591–3600, 2021.

[28] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017.

[29] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*, 2020.

[30] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. 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*, pages 9759–9768, 2020.