Gradient-Guided Knowledge Distillation for Object Detectors

Qizhen Lan[1,2] and Qing Tian[1,2*]
[1]Dept. of Computer Science, Bowling Green State University, Ohio, USA
[2]Dept. of Computer Science, University of Alabama at Birmingham, Alabama, USA
{qlan, qtian}@uab.edu

Abstract

Deep learning models have demonstrated remarkable success in object detection, yet their complexity and computational intensity pose a barrier to deploying them in real-world applications (e.g., self-driving perception). Knowledge Distillation (KD) is an effective way to derive efficient models. However, only a small number of KD methods tackle object detection. Also, most of them focus on mimicking the plain features of the teacher model but rarely consider how the features contribute to the final detection. In this paper, we propose a novel approach for knowledge distillation in object detection, named Gradient-guided Knowledge Distillation (GKD). Our GKD uses gradient information to identify and assign more weights to features that significantly impact the detection loss, allowing the student to learn the most relevant features from the teacher. Furthermore, we present bounding-box-aware multi-grained feature imitation (BMFI) to further improve the KD performance. Experiments on the KITTI and COCO-Traffic datasets demonstrate our method's efficacy in knowledge distillation for object detection. On one-stage and two-stage detectors, our GKD-BMFI leads to an average of 5.1% and 3.8% mAP improvement, respectively, beating various state-of-the-art KD methods. Our codes are available at: https://github.com/lanqz7766/GKD.

1. Introduction

Over the past few years, deep learning models have achieved remarkable success in a variety of domains, including computer vision [11, 12, 31]. Object detection is one of the most critical tasks in computer vision and has seen growing demand in various applications, such as autonomous driving, surveillance, and medical imaging. However, high detection performance often comes at the cost of large and complex neural architectures, which results in slow inference speed on devices without powerful GPUs. To address this problem, various neural network compression techniques have been proposed, such as pruning [8, 33], quantization [19, 27], and knowledge distillation [14, 18]. In Knowledge Distillation (KD), a smaller, lightweight student model mimics the behavior of an unwieldy pre-trained teacher model to achieve comparable or even superior results. The information transferred across the models is usually referred to as "dark knowledge" due to its blackbox nature. Feature-based KD is one of the most popular KD types, which aims to minimize the difference between the teacher's intermediate feature representations and those of the student.

Most of the existing knowledge distillation methods in computer vision are designed for image classification [14, 18, 33, 39]. In the past few years, researchers have started to explore how KD can be effectively applied to object detection. Most state-of-the-art KD methods in object detection use feature-based approaches where the student is trained to mimic the teacher's plain or human-selected features. These methods aim to explore which parts of the teacher's features provide the most informative knowledge for the student to distill. For example, [32] and [36] respectively use the Gaussian Mask and the "fine-grained" imitation mask to select a broader distillation area. [10] distills the foreground and background separately. [37,40] leverage highly activated features and non-local modules to guide the student and distill the global relation of pixels, respectively. However, few studies have considered how these features contribute to the final detection outcome. Unlike previous approaches, we propose a novel gradient-guided knowledge distillation (GKD) method that incorporates gradient information to weigh the importance of features. The gradients of the detection loss function with respect to the model's features provide information about the features' contribution to the final detection performance. By using the task gradients to weigh the importance of features during knowledge distillation, we can effectively transfer knowledge that is more relevant to the task at hand and has a greater impact on the model's performance. To the best of our knowledge, this is the first work that utilizes gradients to weight the

*Corresponding author.

importance of features for knowledge distillation in object detection tasks. Moreover, we argue that foreground objects, including their surrounding pixels with abundant contextual information, should receive special attention during KD. Unlike [36] that distills pixels around the foreground object with fixed weights, we use a FlatGauss Mask (FGM) to assign the highest weight to the pixels within the ground truth bounding boxes and gradually decrease the weight of surrounding pixels as the distance from the center point increases. We also find that feature imitation at multiple granularities helps with the KD. In summary, the main contributions of this paper are as follows:

- We introduce a novel gradient-guided knowledge distillation (GKD) method that utilizes gradient information to weigh the importance of features so that the student model can focus on the more valuable knowledge that is relevant to the final detection. As far as we know, this is the first time that gradients are leveraged as a knowledge filter in knowledge distillation for object detectors.
- We present bounding-box-aware multi-grained feature imitation that takes bounding boxes and their contextual information into consideration during KD and performs distillation along different feature dimensions.
- Our KD method's efficacy is tested on both one-stage and two-stage detectors with different backbones our method achieves an average 4.7 and 3.7 mAP boost on the KITTI and COCO Traffic datasets, respectively, outperforming state-of-the-art KD methods.

2. Related Works

2.1. Object Detection

Object Detection is a fundamental task in computer vision and is more challenging than classification since it involves both localization and classification of objects in an image. Over the past decade, convolutional neural networks (CNNs) have achieved remarkable success in this domain. There are three main categories of CNN-based object detection methods: two-stage detectors, anchor-based one-stage detectors, and anchor-free one-stage detectors. Two-stage detectors, such as [2, 11, 30], follow a two-step process. Initially, they employ a region proposal network (RPN) to generate region proposals, which are then refined and classified in a subsequent stage. Two-stage detectors tend to have higher accuracy compared to one-stage detectors at the expense of longer inference time. Anchor-based one-stage detectors [22, 24, 29] directly predict the category and bounding box of objects from feature maps and are thus more efficient than two-stage detectors. That being said, they use a large number of pre-defined anchor boxes as reference points, which results in additional computation. To reduce such computation, anchor-free one-stage detectors [7, 34, 38] directly predict the critical points and placements of objects without the use of anchor boxes, at the risk of sacrificing accuracy. The difficulty lies in the fact that objects can appear in various shapes, sizes, and orientations within an image, making it hard to detect the different variations of objects. Furthermore, object detection requires a more robust feature representation, as the features need to be capable of identifying objects in different locations and scales. Given the difficulty of the object detection task and the need for robust feature representation, it is important to find more valuable information from the features in order to improve the performance of object detection models. Our proposed method addresses this issue by incorporating gradient information to weight the importance of features, providing a more fine-grained measure of feature importance.

2.2. Knowledge Distillation

Knowledge distillation is a model compression technique proposed by [14]. In its original version, the output probabilities or logits of a pre-trained teacher network serve as soft labels to guide the learning of a smaller student network for classification tasks. Since then, there have been many KD works (e.g., [13, 35, 39]) that further improve the vanilla KD's performance in classification tasks. Relatively speaking, fewer works have applied knowledge distillation to object detection. Chen *et al*. [3] first apply knowledge distillation to object detection by distilling knowledge from the neck features, the classification head, and the regression head. Nevertheless, not all features in the teacher model are useful and relevant. Naively distilling all the features may mislead the student model. How to select the most valuable features for knowledge distillation in object detection is an active research area. Li *et al*. [18] choose the features sampled from the region proposal network (RPN) to improve the performance of the student model. Wang *et al*. [36] propose the fine-grained mask to distill the regions near the ground-truth bounding boxes. Sun *et al*. [32] utilize Gaussian masks to assign more importance to bounding boxes and surrounding regions for distillation. Such methods attempt to find the most informative spatial locations while ignoring the channel-wise feature selection. Guo *et al*. [10] show that both the foreground and background play important roles for distillation, and distilling them separately benefits the student. Dai *et al*. [5] distill the locations where the performances of the student and teacher differ most. All the above-mentioned methods try to infer the most informative spatial regions for knowledge distillation (e.g., the foreground or background). However, they do not consider the differences in importance across different feature channels and how the features contribute to the final detection. FKD [40] and FGD [37] incorporate non-local modules and

consider both spatial and channel attention. However, their feature importance is only based on the magnitude of activation, which is not directly related to final detection. Li [16] introduced a self-supervised distillation technique that employs more salient features to direct less salient ones within a layer. Additionally, this method utilizes deep-layer semantic features to guide shallow layers. Liu *et al*. [25] propose an N-to-one distillation framework. This approach extends the student model's last CNN layer to encompass N channels, where each channel corresponds to one teacher's feature segment, and they collectively guide the student. Another approach by Li and Zhe [17] fuses features from multiple layers to form a proxy teacher. This proxy teacher enables bidirectional distillation between the students and the fusion component. Dong *et al*. [6] propose to optimize the student's architecture and expedite the distillation process. Unlike those works, we propose gradient-guided knowledge distillation, which assigns larger weights to features that contribute more to the final detection. Unlike previously mentioned KD approaches, we explicitly make use of the gradient information that comes almost for free during the backpropagation process.

3. Methodology

Most state-of-the-art feature-based KD methods have the student model directly mimic the teacher model's plain features. Recently, some works like [37] and [40] direct more focus to channels/locations that are highly activated. Unlike previous approaches, we propose a novel gradient-guided knowledge distillation (GKD) method that gives special attention to knowledge contributing to the final detection performance. In addition, we will present how to incorporate bounding box and context information in multi-grained feature-based knowledge distillation.

3.1. Gradient-Guided Knowledge Distillation

We propose to utilize the gradients of the detection loss with respect to features to represent the features' contribution to the final detection. The features corresponding to larger gradients are more influential on the decision making and thus they deserve more attention during the knowledge distillation process. Fig. 1 illustrates the general idea of our GKD and how it guides the student model to better learn the most valuable and relevant knowledge from the teacher. Mathematically, we define the importance/weight of the k-th feature map in layer l of a detector as:

$$w_k^l = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} \frac{\partial \mathcal{L}_{task}}{\partial A_{i,j,k}^l} \tag{1}$$

where $\mathcal{L}_{task}$ denotes the total detection loss (including bounding box regression loss and classification loss), $A_{i,j,k}^l$ is the single activation value at location (i, j) in the k-th feature map of the l-th layer. We first calculate the gradients of $\mathcal{L}_{task}$, with respect to feature $A_{i,j,k}^l$. These gradients flowing back are global-average-pooled over the width and height dimensions (indexed by i and j, with max value W and H, respectively) to obtain the feature channel importance w_k^l. Then, we use w_k^l to weigh the k-th activation map A_k^l:

$$\widetilde{A_k^l} = w_k^l A_k^l \tag{2}$$

where $\widetilde{A_k^l}$ is the k-th gradient-weighted activation map of the l-th layer. These maps are then linearly combined along the channel dimension (before taking absolute values and Norm) to obtain the final target map for distillation:

$$M^l = Norm(|\sum_{k=1}^{C} \widetilde{A_k^l}|) \tag{3}$$

where $Norm$ represents the min-max normalization function. By weighting the features using these gradients, we can effectively "highlight" the features that have a larger impact on the overall detection loss. The same process can be applied to both the teacher model and the student model. The resulting target maps for the teacher and the student are $M_{\mathcal{T}}^l$ and $M_{\mathcal{S}}^l$, respectively. The goal of our gradient-guided knowledge distillation is to minimize the difference between the target maps:

$$\mathcal{L}_{GKD} = \frac{1}{HW} \sum_{l=1}^{L} \sum_{i=1}^{W} \sum_{j=1}^{H} |M_{i,j,\mathcal{T}}^l - M_{i,j,\mathcal{S}}^l| \tag{4}$$

where l indicates an intermediate layer, L is the total number of intermediate layers being considered for distillation. We use L1-norm loss instead of L2-norm loss because L2 can be more susceptible to outliers when there is a large discrepancy between the teacher and student models at the beginning of training. Using L1-norm loss encourages teacher-student consistency in more locations.

To handle objects of various scales, most modern object detectors employ Feature Pyramid Networks (FPN) [21] or its variants. In our experiments, to enhance knowledge transfer across different scales, we choose the output layers of FPN as the target layers for distillation.

3.2. Bounding-box-aware Multi-grained Feature Imitation

In object detector KD, the background features are usually less informative and overwhelming, potentially misleading the distillation process. Prior approaches attempted to employ different distillation masks to extract more valuable information. Fig. 2 illustrates the differences between our proposed FlatGauss Mask and previous methods [10, 32, 36].

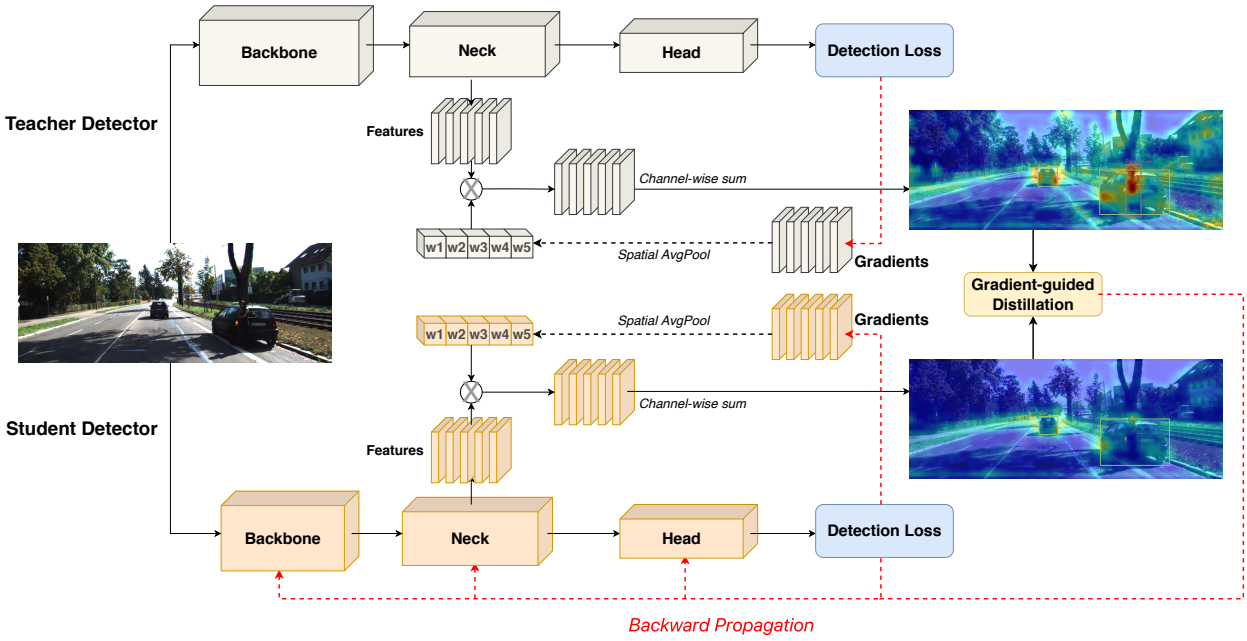

Figure 1. Illustration of the proposed Gradient-Guided Knowledge Distillation (GKD) method.

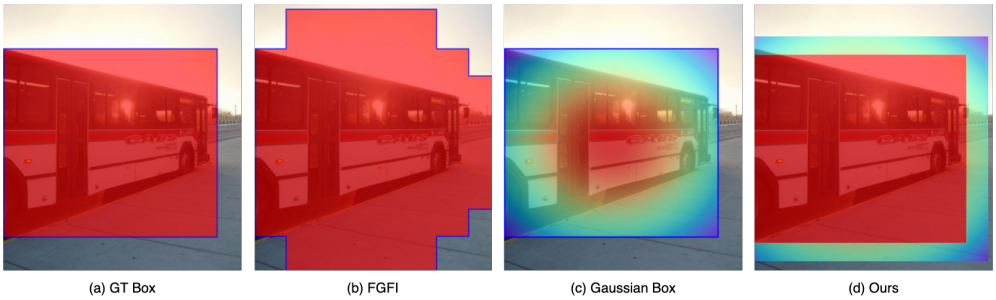

Figure 2. Popular attention regions for knowledge distillation in object detection. Different colors indicate different weights for different areas (red: high, blue: low). (a) Guo *et al*. [10] focused on the region inside the ground-truth bounding boxes and assigned different weights to the background. (b) Wang *et al*. [36] distilled the anchor-covered regions around the foreground object. (c) Sun *et al*. [32] used a Gaussian Mask to cover the ground truth bounding box for distillation. In contrast to previous methods, our approach (d) focuses on foreground objects and their neighboring pixels with gradually diminishing weights.

The previous methods either ignore the adjacent pixels or cover too many unnecessary regions. Unlike these approaches, we propose a flat gauss mask F, which is defined as:

$$F_{i,j} = \begin{cases} 1, & \text{if} \quad (i,j) \in o \\ e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{\bar{x}}+\frac{y-\bar{y}}{\bar{y}}\right)^2}, & \text{elif} \quad (i,j) \in \hat{o} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where o and $\hat{o}$ stand for the region inside the ground truth bounding box and the surrounding region, respectively. The surrounding region doubles the width and height of the original ground truth bounding box. (x, y) represent a specific point in the surrounding region. $(\bar{x}, \bar{y})$ indicates the center point of the ground truth bounding box. Eq. (5) directs enough attention to the foreground while taking the neighbouring pixels/regions into consideration as well.

We also incorporate position and channel attention (based on highly-activated features from [37]) when distilling features. The position attention mask M^P and channel attention mask M^C can be defined as follows:

$$M^P = WH \cdot softmax\left(\frac{\sum_{k=1}^{C}|A_k|}{CT}\right) \tag{6}$$

$$M^C = C \cdot softmax\left(\frac{\sum_{i=1}^{W}\sum_{j=1}^{H}|A_{i,j}|}{WHT}\right) \tag{7}$$

where A represents the plain feature maps, and W, H, C are the width, height, and channel number of A indexed by i, j, k, respectively. T is the temperature hyper-parameter introduced by [14] to modulate the distribution. Based on Eq. (5), Eq. (6), and Eq. (7), we propose our Bounding-box-aware Multi-grained Feature Imitation (BMFI) loss as follow:

$$L_{BMFI} = \sum_{k=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W} F_{i,j} M_{i,j}^{P,\mathcal{T}} M_{k}^{C,\mathcal{T}} (A_{i,j,k}^{\mathcal{T}} - A_{i,j,k}^{\mathcal{S}})^2 + \alpha(|M^{P,\mathcal{T}} - M^{P,\mathcal{S}}| + |M^{C,\mathcal{T}} - M^{C,\mathcal{S}}|) \tag{8}$$

where the subscript $\mathcal{T}, \mathcal{S}$ denotes the teacher and student detector, respectively. By adding our Gradient-guided knowledge distillation loss from Sec. 3.1, our total distillation loss is:

$$L_{KD} = L_{GKD} + \beta L_{BMFI} \tag{9}$$

The balancing hyperparameters (α in Eq. (8) and β in Eq. (9)) are empirically set in our experiments to achieve the best validation results.

4. Experiments and Results

4.1. Datasets

KITTI [9] is a 2D-object detection dataset that includes seven different types of road objects. It includes 7481 images with annotations. We split it into a training set and a validation set in the ratio of 8:2. As suggested in [1], we group similar categories into one. Specifically, we perform the following modification to the original KITTI dataset: *car, van, truck, tram* as Car, *pedestrian, person* as Pedestrian, and *cyclist* as Cyclist.

COCO-Traffic is obtained by selecting categories related to self-driving from MS COCO 2017 [23]. We keep only images containing at least one road-related object to filter out images containing only indoor objects. The selection is applied to both the training and validation sets. The COCO-Traffic dataset contains 13 traffic-related categories: *person, bicycle, car, motorcycle, bus, train, truck, traffic light, fire hydrant, stop sign, parking meter, cat, dog*.

4.2. Implementation Details

All the detection experiments are conducted in the MMDetection framework [4] using Pytorch [28]. We employed Faster-RCNN [30] as a representative of two-stage detectors and chose Generalized Focal Loss (GFL) [20] as an example of one-stage detectors. The teacher and student models (without any knowledge distillation) were trained directly using the default configuration of MMDetection [4]. The teacher models were based on a ResNet-101 backbone, and we tested two different student backbone architectures (i.e., ResNet-50 and ResNet-18). The temperature hyper-parameter T is set to 0.5. We adopt the inheriting strategy proposed in [15], where the student model is initialized with the teacher's neck and head parameters. All the models are sufficiently trained to convergence with an SGD optimizer, an initial learning rate of 0.02, momentum of 0.9, and weight decay of 0.0001. All models are evaluated in terms of mean averaged precision (mAP) with 0.5 as the Intersection over Union (IoU) threshold. For comparison, we re-implemented the following state-of-the-art KD methods in object detection: FGFI [36], FKD [40], GID [5], DeFeat [10], and FGD [37], which were published in recent years' top CV/ML conferences. All the competing knowledge distillation methods and our method are applied to FPN output layers.

4.3. Experiment Results

In our experiments, we evaluated the performance of our proposed gradient-guided knowledge distillation (GKD) method against several state-of-the-art KD methods on the KITTI and COCO-Traffic datasets using both single-stage (e.g., GFL) and two-stage (e.g., Faster RCNN) object detectors. The results on the single-stage and two-stage detectors are shown in Tab. 1 and Tab. 2, respectively.

As we can see from Tab. 1, our GKD method provides a significant boost in mAP for single-stage student detectors. Specifically, when using a ResNet-50 backbone, our GKD method achieves 4.9 and 1.8 mAP improvement on the KITTI and COCO-Traffic datasets, respectively. On the ResNet-18 backbone, the two numbers become 6.2 and 4.3. Our GKD-BMFI, which incorporates Bounding-box-aware Multi-grained Feature Imitation, outperforms all student baseline models and other state-of-the-art distillation methods. For example, on the KITTI dataset, our GKD-BMFI outperforms FGD [37] by 1.1 mAP with a ResNet-50 backbone and 2 mAP with a ResNet-18 backbone. On the COCO-Traffic dataset, it surpasses other five different KD methods by an average of 1.66 mAP with a ResNet-50 backbone and 3.14 mAP with a ResNet-18 backbone.

As shown in Tab. 2, our proposed GKD method is also effective for two-stage detectors. Specifically, when utilizing a ResNet-50 backbone on the COCO-Traffic dataset, our GKD method demonstrates a remarkable improvement of 2.6 mAP over the student-baseline and outperforms other state-of-the-art distillation methods, including FKD [40] and FGD [37], by an average of 2.35 mAP. In addition, our GKD-BMFI can further improve the distillation performance. For example, when comparing to the student-baseline with a ResNet-18 backbone on the KITTI dataset, our GKD-BMFI method demonstrates an impressive improvement of 5 mAP.

KD methods \ Student backbones	ResNet-50		ResNet-18	
	KITTI	COCO Traffic	KITTI	COCO Traffic
Teacher (w ResNet-101)	89.4	71.8	89.4	71.8
Student-baseline	85.1	67.7	81.9	61.9
FKD [40]	86.4	69.5	84.4	62.6
GID [5]	86.1	69.3	84.6	63.7
DeFeat [10]	85.4	69.3	83.3	62.7
FGD [37]	89.2	71.0	86.7	65.9
FGFI [36]	84.4	68.6	82.6	62.4
Our GKD	90.0	69.5	88.1	66.2
Our GKD-BMFI	**90.3**	**71.2**	**88.7**	**66.6**

Table 1. Performance (mAP) of different distillation methods with GFL detector [20] on the KITTI and COCO traffic datasets. (The teacher model and the student-baseline are non-distillation GFL models with ResNet-101 and ResNet-50/18 as backbones, respectively.) The highest mAP in each column is highlighted.

KD methods \ Student backbones	ResNet-50		ResNet-18	
	KITTI	COCO Traffic	KITTI	COCO Traffic
Teacher (w ResNet-101)	89.3	67.9	89.3	67.9
Student-baseline	88.9	67.5	84.1	63.1
FKD [40]	89.0	67.8	87.2	65.3
FGD [37]	88.9	67.7	87.0	64.1
Our GKD	90.6	70.1	89.0	66.5
Our GKD-BMFI	**90.8**	**70.3**	**89.1**	**66.9**

Table 2. Performance (mAP) of different distillation methods with Faster R-CNN detector [30] on the KITTI and COCO traffic datasets. (The teacher model and the student-baseline are non-distillation Faster-RCNN models with ResNet-101 and ResNet-50/18 backbones, respectively.) The highest mAP in each column is highlighted.

Most KD methods employ homogeneous backbone pairs. In contrast, our approach can deal with heterogeneous backbone pairs as well. As shown in Tab. 3, for faster R-CNN [30] detection, we successfully distill the knowledge from the Swin teacher backbone (a vision transformer [26]) to the student ResNet-18 backbone and achieve a promising student mAP of 68.2.

4.4. Qualitative Analysis

In Fig. 3, we visualize the gradient-guided masks from the teacher detector and different training stages of the student detector. This example comes from our experiments on the KITTI dataset using the GFL detector. By comparing the gradient-guided masks between the teacher and the students at different training stages, we can observe the student's gradual learning process and see how it tries to follow the teacher's guidance. According to the figure, the teacher detector (Fig. 3 (b)) focuses on the objects in the image (e.g cars and pedestrians) more accurately than the student detector that has only been trained for one epoch (Fig. 3 (c)). However, as our gradient-guided knowledge distillation process goes on, we can see that the student's attention becomes more and more similar to the teacher's, as seen in Fig. 3 (d). In Fig. 3 (e), we can see that the student even develops some new high-attention areas (e.g., the smaller-scale car in front of the vehicle). This potentially explains why our much smaller distilled model can sometimes surpass the teacher model (Tab. 1).

Fig. 4 illustrates a random detection example on the KITTI dataset. The qualitative results of four GFL [20] models are demonstrated. They are (from top to bottom): (a) Student baseline model, (b) FKD [40] distilled model, (c) FGD [37] distilled model, and (d) our GKD distilled model. All employ a ResNet-50 backbone. In Fig. 4, our GKD distilled model outperforms both the student baseline model and other distilled models. (a), (b), and (c) have a hard time identifying objects obscured by the wall, leading to the generation of multiple/inaccurate bounding boxes. In contrast, our method accurately detects the pedestrian and the car behind the wall with higher confidence scores.

4.5. Ablation Study

To analyze the contribution of the components of our method, we perform ablation experiments. To ensure the reliability of this study, we use two combinations of detectors and datasets. Specifically, we use GFL [20] on the KITTI dataset, and use Faster R-CNN [30] on the COCO Traffic dataset. All detectors use ResNet-50 as the student backbone. We consider the following three components in this study: our Gradient-guided Knowledge Distillation (GKD, without bells and whistles), bounding-box-aware FlatGauss Mask (FGM), and Multi-grained Feature Imitation (MFI)

Backbone	Swin-S	ResNet-18		
Faster R-CNN [30]	Teacher	Student-baseline	FGD [37]	our GKD-BMFI
mAP	75.3	63.1	67.5	**68.2**

Table 3. Performance (mAP) of different distillation methods with Faster R-CNN detector [30] on the COCO traffic datasets (The teacher model and the student-baseline are non-distillation GFL models with Swin-small and ResNet-18 as backbones, respectively.)

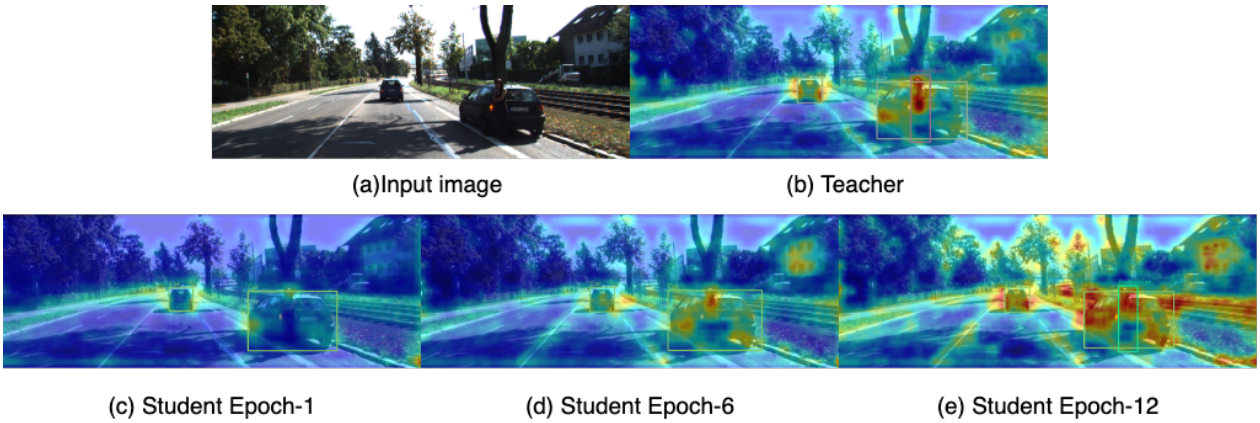

Figure 3. Visualization of the gradient-guided masks from the teacher detector and different training stages of the student detector using GKD. Different colors indicate different attention levels, with the red color representing the highest attention and the blue color representing the lowest.

Modules	GKD	×	×	×	✓	✓
	FGM	×	×	✓	✓	×
	MFI	×	✓	✓	✓	×
GFL	KITTI	85.1	88.5	89.7	**90.3**	90.0
Faster R-CNN	COCO Traffic	67.5	69.0	69.9	**70.3**	70.1

Table 4. Ablation study of the three different components of our GKD-BMFI. GKD: Gradient-guided Knowledge Distillation (with no bells and whistles), FGM: FlatGauss Mask, MFI: Multi-grained Feature Imitation (M^P and M^C related). This ablation study is conducted on the KITTI/COCO Traffic dataset using GFL/Faster R-CNN with a ResNet-50 backbone, respectively.

Model	Backbones	Parameters(M)	GFLOPs	mAP (KITTI)	mAP (COCO Traffic)
GFL [20]	ResNet-101	51.03	13.79	89.4	71.8
	ResNet-50	32.04	10.05	90.3	71.2
	ResNet-18	19.09	7.61	88.7	66.6
Faster R-CNN [30]	ResNet-101	60.13	27.09	89.3	67.9
	ResNet-50	41.13	23.36	90.8	70.3
	ResNet-18	28.13	20.77	89.1	66.9

Table 5. Model complexity (with 224×224 input resolution), the mAP are of the teacher-baseline model (with ResNet-101 backbone) and our GKD-BMFI distilled student (with ResNet-50/18 backbone) on KITTI/COCO Traffic dataset, respectively.

methods. According to the results in Tab. 4, all three components play a positive role in the mAP boost, but the GKD with no bells and whistles makes the most contribution. To be more specific, GKD alone can improve the GFL detector baseline mAP from 85.1 to 90.0 on the KITTI dataset. The combination of the three components results in a 3.85 mAP improvement on average. From Tab. 4, we can also observe that using only the MFI component results in an average of 2.45 mAP improvement. By incorporating the bounding-box-aware FlatGauss Mask into MFI, we get BMFI (as described in Eq. (8)), which leads to an average of 3.9 mAP improvement over the student detector.

4.6. Efficiency

Our gradient-guided KD does not add much to the training burden as gradients come almost for free during

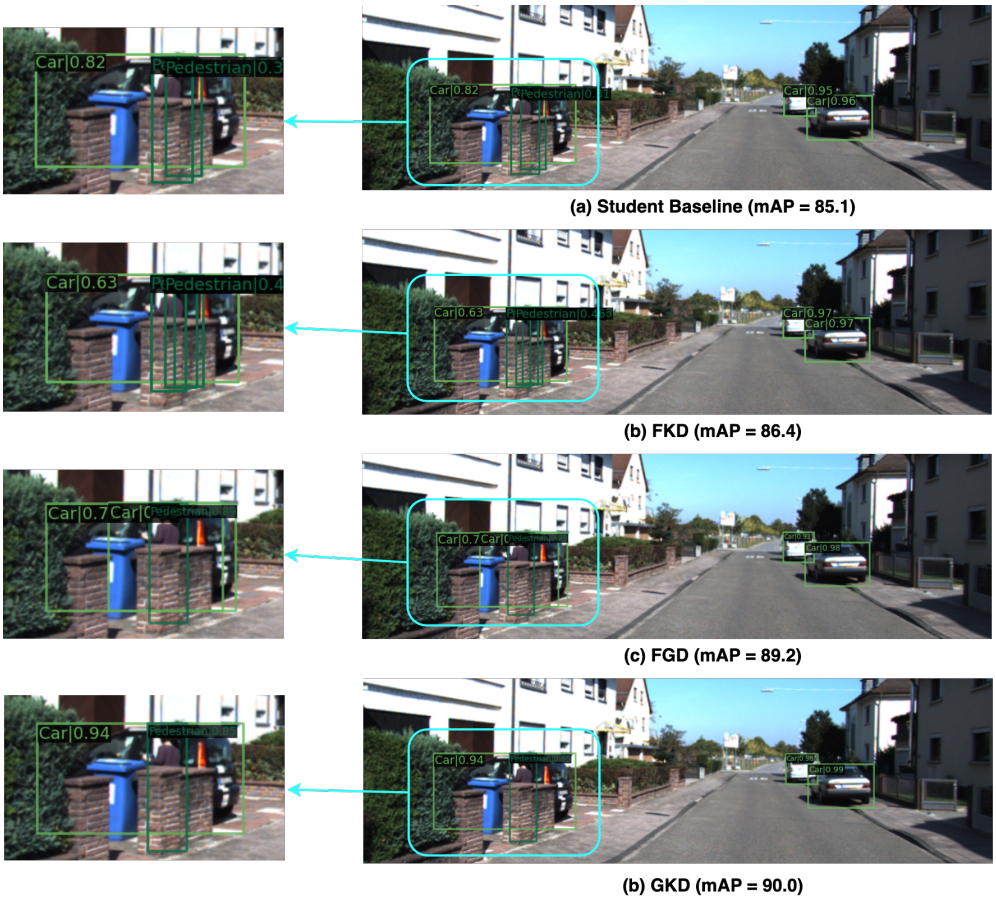

Figure 4. **Qualitative analysis** - This analysis is conducted on the KITTI dataset using the GFL [20] detector. The detection results are from (a) the Student baseline model, (b) the model distilled by FKD [40], (c) the model distilled by FGD [37], and (d) the model distilled by our GKD.

backpropagation. Also, the negligible increase in training cost is incurred only once during training (offline). Here, we mainly focus on inference efficiency, which is of paramount importance to many real-world applications (e.g., autonomous driving perception). In Tab. 5, we compared different architectures in terms of FLOPs (multiply and add) and parameters. According to the results, our ResNet-50/18 distilled student model enjoys an average of 34.41%/67.90% reduction in model size and an average of 20.22%/33.70% savings in FLOPs. Also, our GKD-BMFI distilled student model with ResNet-50 backbone can even outperform the teacher model (e.g. 90.3 vs 89.3 mAP for GFL on KITTI and 70.3 vs 67.9 mAP for Faster R-CNN on COCO Traffic) while enjoying the complexity reduction.

5. Conclusion

In this paper, we have proposed a novel gradient-guided knowledge distillation (GKD) method. It leverages the gradients of the detection loss w.r.t. feature maps to identify valuable and relevant knowledge for knowledge distillation. Our GKD gives special attention to feature maps contributing more to the final detection. In addition, we have presented bounding-box-aware multi-grained feature imitation (BMFI) to further improve the distilled model's performance. Experiments on the KITTI and COCO-Traffic datasets, using various detectors and backbones, demonstrate our method's efficacy. The qualitative analysis shows that our gradient-guided knowledge distillation allows the student to get similar or even more informative attention maps than the teacher.

Acknowledgment

This research was supported by the National Science Foundation (NSF) under Award No. 2153404. This work would not have been possible without the computing resources provided by the Ohio Supercomputer Center.

References

[1] Ssd: Single shot multibox detector – train the kitti dataset, Mar 2017. 5

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 2

[3] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017. 2

[4] 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. 5

[5] 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. 2, 5, 6

[6] Peijie Dong, Lujun Li, and Zimian Wei. Diswot: Student architecture search for distillation without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11898–11908, 2023. 3

[7] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019. 2

[8] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018. 1

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 5

[10] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2154–2164, 2021. 1, 2, 3, 4, 5, 6

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 2

[12] 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. 1

[13] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1921–1930, 2019. 2

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 1, 2, 5

[15] Zijian Kang, Peizhen Zhang, Xiangyu Zhang, Jian Sun, and Nanning Zheng. Instance-conditional knowledge distillation for object detection. *Advances in Neural Information Processing Systems*, 34:16468–16480, 2021. 5

[16] Lujun Li. Self-regulated feature learning via teacher-free feature distillation. In *European Conference on Computer Vision*, pages 347–363. Springer, 2022. 3

[17] Lujun Li and Zhe Jin. Shadow knowledge distillation: Bridging offline and online knowledge transfer. *Advances in Neural Information Processing Systems*, 35:635–649, 2022. 3

[18] 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. 1, 2

[19] Rundong Li, Yan Wang, Feng Liang, Hongwei Qin, Junjie Yan, and Rui Fan. Fully quantized network for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2810–2819, 2019. 1

[20] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33:21002–21012, 2020. 5, 6, 7, 8

[21] 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. 3

[22] 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. 2

[23] 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. 5

[24] 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. 2

[25] Xiaolong Liu, Lujun Li, Chao Li, and Anbang Yao. Norm: Knowledge distillation via n-to-one representation matching. *arXiv preprint arXiv:2305.13803*, 2023. 3

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[27] Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision, pages 1325–1334, 2019. 1

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Curran Associates Inc., Red Hook, NY, USA, 2019. 5

[29] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[30] 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, 2015. 2, 5, 6, 7

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[32] Ruoyu Sun, Fuhui Tang, Xiaopeng Zhang, Hongkai Xiong, and Qi Tian. Distilling object detectors with task adaptive regularization. *arXiv preprint arXiv:2006.13108*, 2020. 1, 2, 3, 4

[33] Qing Tian, Tal Arbel, and James J Clark. Task dependent deep lda pruning of neural networks. *Computer Vision and Image Understanding*, 203:103154, 2021. 1

[34] 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. 2

[35] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374, 2019. 2

[36] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4933–4942, 2019. 1, 2, 3, 4, 5, 6

[37] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4643–4652, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[38] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9657–9666, 2019. 2

[39] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 1, 2

[40] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*, 2021. 1, 2, 3, 5, 6, 8