electronics

MDPI

Article

Shared Knowledge Distillation Network for Object Detection

Zhen Guo [1,2,*], Pengzhou Zhang [1,*] and Peng Liang [2]

[1] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China
[2] China Unicom Smart City Research Institute, Beijing 100048, China; liangp@chinaunicom.cn
* Correspondence: cathy.guozhen@cuc.edu.cn (Z.G.); zhangpengzhou@cuc.edu.cn (P.Z.)

Abstract: Object detection based on Knowledge Distillation can enhance the capabilities and performance of 5G and 6G networks in various domains, such as autonomous vehicles, smart surveillance, and augmented reality. The integration of object detection with Knowledge Distillation techniques is expected to play a pivotal role in realizing the full potential of these networks. This study presents Shared Knowledge Distillation (Shared-KD) as a solution to overcome optimization challenges caused by disparities in cross-layer features between teacher–student networks. The significant gaps in intermediate-level features between teachers and students present a considerable obstacle to the efficacy of distillation. To tackle this issue, we draw inspiration from collaborative learning in real-world education, where teachers work together to prepare lessons and students engage in peer learning. Building upon this concept, our innovative contributions in model construction are highlighted as follows: (1) A teacher knowledge augmentation module: this module is proposed to combine lower-level teacher features, facilitating the knowledge transfer from the teacher to the student. (2) A student mutual learning module is introduced to enable students to learn from each other, mimicking the peer learning concept in collaborative learning. (3) The Teacher Share Module combines lower-level teacher features: the specific functionality of the teacher knowledge augmentation module is described, which involves combining lower-level teacher features. (4) The multi-step transfer process can be easily optimized due to the minimal gap between the features: the proposed approach breaks down the knowledge transfer process into multiple steps, which can be easily optimized due to the minimal gap between the features involved in each step. Shared-KD uses simple feature losses without additional weights in transformation, resulting in an efficient distillation process that can be easily combined with other methods for further improvement. The effectiveness of our approach is validated through experiments on popular tasks such as object detection and instance segmentation.

Keywords: shared knowledge network; knowledge distillation; object detection; cross-layer distillation

check for updates

Citation: Guo, Z.; Zhang, P.; Liang, P. Shared Knowledge Distillation Network for Object Detection. *Electronics* **2024**, *13*, 1595. https://doi.org/10.3390/electronics13081595

Academic Editor: Silvia Liberata Ullo

Received: 4 March 2024
Revised: 14 April 2024
Accepted: 18 April 2024
Published: 22 April 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In the fast-changing world of 5G and emerging 6G networks [1–3], precise and efficient object detection in deep neural networks (DNNs) is crucial. Object detection is essential in various applications, including autonomous vehicles, smart surveillance, and augmented reality, where models must be both accurate and computationally efficient.

Despite the remarkable success of deep neural networks in various tasks [4–6], their widespread adoption is hindered by the high computational costs due to the large number of parameters. To address this issue, several methods [7–10] have been proposed to reduce the computational cost of deep learning models. One effective technique is Knowledge Distillation (KD), which involves transferring knowledge from a high-capacity teacher to a low-capacity student model. This knowledge transfer enhances the accuracy-efficiency tradeoff of the student model during runtime. This Shared Knowledge Distillation approach is a powerful methodology for refining object detection models, ensuring they can handle current network infrastructures and meet future 5G/6G network demands.

The initial approach of Knowledge Distillation (KD) [11] utilizes the logit outputs of the teacher network as a source of knowledge. To enhance this knowledge transfer, feature distillation techniques [7,8] have been introduced to encourage the student network to emulate the intermediate features of the teacher network. Subsequent studies [8,12–16] have focused on extracting and aligning informative features through various loss functions and transformations. However, these methods primarily concentrate on feature pairs within the same layer of the teacher–student network, disregarding the potential advantages of cross-layer feature transfer. The dissimilarities in shape and semantics between cross-layer features present optimization challenges and may result in information loss during feature transformations. Recent research has explored meta-learning approaches to identify optimal cross-layer feature pairs, which adds complexity to the optimization process. On the other hand, a different study [17] suggests that the front layer features of the teacher network are more valuable for student training and proposes a complex residual feature and fusion module for cross-layer distillation. In contrast, AFD [18] argues that the last layer features of the teacher network contain more relevant knowledge and proposes self-attention strategies to align cross-layer features in the spatial dimension, and whereas these approaches achieve performance improvements by leveraging cross-layer feature knowledge, their reliance on intricate feature transformations and matching strategies limits their practical usability.

In order to tackle these challenges, we introduce a straightforward and efficient framework called Shared Knowledge Distillation (Shared-KD), as shown in Figure 1. In contrast to existing methods of cross-layer feature distillation, our approach proposes a novel two-step process for decomposing the original cross-layer feature supervision from teachers to students. This process includes identical-layer distillation between teacher and student networks and cross-layer distillation within the student network itself. The first step focuses on identical-layer distillation, which shares similarities in shape and semantics between the teacher and student networks. The second step involves utilizing the hierarchical features of the online student network, which exhibit close optimization and semantic properties. To enhance the efficiency of distillation, we employ simple l_2 distances for the feature mimicking loss and utilize spatial pooling and channel cropping to align feature shapes without the need for complex feature transformations. This efficient feature transfer also helps mitigate knowledge reduction. Shared-KD offers three key advantages: (1) Our framework sheds new light on decoupling cross-layer distillation using multi-step strategies. (2) Shared-KD enhances the effectiveness of KD methods in overcoming unstable optimization issues, leveraging the full knowledge of teacher features to achieve significant performance improvements. (3) Shared-KD incorporates a simple feature alignment component without introducing additional parameters. In contrast, other cross-layer distillation techniques require complex feature transformations and optimizations, increasing training time and resource requirements. Shared-KD can potentially expand the application of KD and facilitate further research in this area. Shared Knowledge Distillation techniques improve the efficiency and adaptability of deep neural networks for object detection, paving the way for advancements in real-time applications and services within the paradigm of next-generation networks.

Our proposed method is extensively evaluated through experiments on detection and segmentation tasks to validate its effectiveness. The experimental results demonstrate the superiority of our approach, surpassing other existing methods by a considerable margin. Shared-KD consistently improves accuracy across various neural network architectures and data augmentation techniques. For instance, when applied to the object detection task on the MS-COCO dataset, Shared-KD outperforms other methods, such as RetinaNet and Faster R-CNN, by significantly improving the average precision (AP). These results demonstrate the generalizability and robustness of our approach.

In summary, we make the following principle contributions in this paper:

- Through analysis and exploration of feature gaps and roles in distillation, it is evident that cross-layer feature gaps within the student network are significantly smaller than

those between the student and teacher. This observation motivates us to propose a new Shared Knowledge Distillation (Shared-KD) framework.

- Our Shared-KD technique minimizes the shared features between the teacher–student layer and the cross-features within the student. This achieves cross-layer distillation without complex transformations.
- Our Shared-KD outperforms other state-of-the-art feature-distillation methods on various deep models and datasets, achieving superior performance and training acceleration.

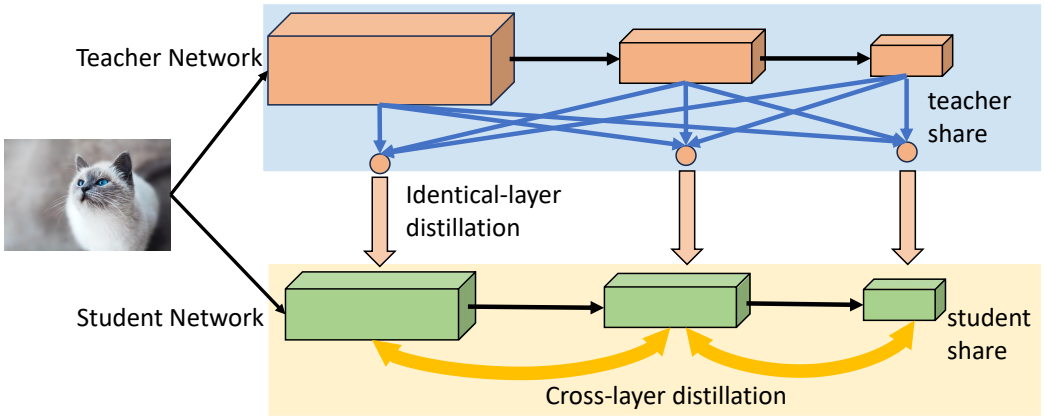

Figure 1. A schematic overview of our Shared-KD, including inter-layer and intra-layer parts. During the training phase, Shared-KD utilizes the same-layer distillation between teacher–student and the cross-layer distillation within students.

2. Related Work

2.1. Object Detection

There are three typical types of CNN-based object detection networks: two-stage detectors, one-stage detectors, and anchor-free detectors. Faster R-CNN [19] is a popular two-stage detector algorithm that proposes a framework for object detection. The system comprises a Region Proposal Network (RPN) for generating region proposals and a detection network for object classification. The YOLO family [20–22] is widely used for one-stage detectors. It treats detection as a regression problem, predicting bounding boxes and associated class probabilities. FCOS [23] is a typical anchor-free method that achieves real-time detection while maintaining competitive accuracy, making it suitable for various applications. It utilizes a set of predefined feature points across the object. It predicts the object's bounding box parameters directly at these locations, eliminating the need for predefined anchor boxes and offering flexibility and improved accuracy in object localization. In comparison to the existing literature, this paper introduces a unique contribution to the field of Knowledge Distillation for object detection, and whereas previous works have explored various methods of Knowledge Distillation, such as using soft logit outputs or intermediate feature representations, this paper proposes a novel approach called Shared-KD. Shared-KD addresses the limitations of existing feature-distillation methods by conditioning the knowledge transfer on intermediate feature representations, ensuring well-aligned structures between the teacher and student models. This is achieved through densely connected teacher generation. Additionally, Shared-KD enables direct knowledge transfer through dense cross-layer feature connections from the student to the teacher model, eliminating the need for feature matching. This simplifies the distillation process and enhances the efficiency of knowledge transfer. By leveraging intermediate feature representations and establishing direct connections between the teacher and student models, Shared-KD offers a promising alternative to traditional feature-distillation methods. It is a practical and efficient solution for Knowledge Distillation in object detection, as it eliminates the complexity of dimension reduction transformations and the reliance on stage-wise feature matching.

2.2. Knowledge Distillation

The concept of Knowledge Distillation (KD) [24–28] involves using learned knowledge, such as logits, feature values, and sample relations, from a high-capacity teacher to guide the training of a student model. Early pioneering works [11,29] use soft logit outputs of the pre-trained teacher as the extra supervision to guide the training of the student, in addition to the ground truth labels. Then, various feature-distillation methods [7,8,30], which rely on the intermediate feature representations, are proposed. Additionally, relation distillation methods have been developed to explore the relationships and higher-order dependencies captured by the teacher model's logits or intermediate features. Several studies have been conducted on the use of Knowledge Distillation in object detection. Guo et al. [31] propose FKD, which uses attention masks to identify foreground object pixels. It also incorporates a non-local module to facilitate student learning through pixel relations. Yang et al. [32] introduce FGD, a method that effectively leverages both focal and global data, enhancing learning by emphasizing the relations among pixels. These approaches are effective in enhancing the Knowledge Distillation process. However, existing feature-distillation methods often require dimension reduction transformations and different distance metrics to match the feature maps of the student and teacher models, which may result in the loss of valuable information. Additionally, these methods rely on different feature loss objectives and weight factors to balance the loss terms. In contrast, Shared-KD presents a new approach to Knowledge Distillation by conditioning on intermediate feature representations. Shared-KD ensures well-aligned structures between the teacher and student models through densely connected teacher generation. Furthermore, it allows for direct knowledge transfer through dense cross-layer feature connections from the student to the teacher model, eliminating the need for feature matching. The Shared-KD framework simplifies the process of Knowledge Distillation and enhances the efficiency of knowledge transfer. It achieves this by leveraging intermediate feature representations and establishing direct connections between the teacher and student models. Shared-KD offers a promising alternative to traditional feature-distillation methods. Shared-KD is a practical and efficient solution for Knowledge Distillation due to its elimination of complex transformations and reliance on stage-wise feature matching. In summary, whereas existing methods have made progress in Knowledge Distillation for object detection, Shared-KD presents a novel approach that addresses the limitations of these methods by enabling direct knowledge transfer, well-aligned structures, and a simplified and efficient distillation process without the need for complex feature matching and transformations.

3. Shared Knowledge Distillation

This section begins with a review of feature-distillation methods using a general formulation. Next, we present the formulation and insights of our Shared Knowledge Distillation (Shared-KD). Finally, we combine Shared-KD with logits KD.

3.1. Knowledge Distillation

Firstly, we will recap the Knowledge Distillation method proposed by Hinton et al. [11]. This widely used method involves using category similarity as a guide for student networks. To regularize the network's learning, temperature is introduced to soften the initial categorical information, also referred to as 'dark knowledge'. The output probability of the teacher network and student network can be calculated as Equations (1) and (2).

$$p_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)} \tag{1}$$

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)} \tag{2}$$

where T represents temperature, which can adjust the softening of the output probability. z_i z_j are the logits input for softmax, p_i represents the output probability of each category

in the teacher network, and q_i represents the output probability of each category in the student network.

Cross-entropy between the distilled teacher and student models calculates the soft loss. The hard loss is calculated based on the predicted values of the student model and the true values.

$$\mathcal{L}_{\text{soft}} = -\sum_i p_i \cdot \log(q_i) \tag{3}$$

$$\mathcal{L}_{\text{hard}} = -\sum_i y_i \cdot \log(q_i) \tag{4}$$

The ground truth label (also known as the hard target label) for the i-th sample is represented by y_i. The teacher and student models' predicted probability are represented by p_i and q_i, respectively. The total loss function is calculated as below.

$$\mathcal{L} = \alpha \cdot L_{\text{soft}} + (1 - \alpha) \cdot L_{\text{hard}} \tag{5}$$

The loss function used in Knowledge Distillation involves two types of losses: soft target loss L_{soft} and hard target loss L_{hard}. The former guides the student to replicate the teacher's probability distribution, whereas the latter reflects the guidance from the actual ground truth labels. The parameter α balances the effect of these two losses. During the Knowledge Distillation process, the student receives both the challenging and soft target knowledge. The loss function can be written as follows:

$$\mathcal{L} = (1 - \alpha)\mathcal{L}_{\text{CE}}(y, \sigma(z_S)) + 2\alpha T^2 \mathcal{L}_{\text{CE}}(\sigma(\frac{z_S}{T}), \sigma(\frac{z_T}{T})) \tag{6}$$

where the loss of the cross-entropy is represented by $\mathcal{L}_{\text{CE}}$. The softmax function is represented by σ. $\mathbf{y}$ represents the ground truth label. The output logits of the student and teacher networks are denoted by z_S and z_T. The balancing hyperparameter is represented by α.

3.2. Conventional Cross-Layer Distillation

We then revisit the general formulation of cross-layer distillation methods to understand our approach better. Current feature distillation approaches [7,8] encourage the student model to mimic the intermediate features of the teacher model by explicitly optimizing the feature distillation loss. To achieve this, we minimize the loss for a target student model S with middle-level features ϕ_S and its teacher T with features ϕ_T, as defined below:

$$\mathcal{L}_{\text{FD}} = \mathcal{D}_f(T_s(\phi_S), T_t(\phi_T)) \tag{7}$$

where T_s and T_t are the student and teacher transformation to align feature dimensions (eg, channel and spatial). $\mathcal{D}_f(\cdot)$ is the distance function measuring the difference in intermediate features.

In conventional frameworks, the feature loss is typically combined with the task loss during training. A pre-trained and fixed teacher model guides the student model. Let x denote the training data and $\mathcal{Q}$ denote a set of layer location pairs for feature distillation; S_L and T_L are the layers of student and teacher networks, respectively. The general objective function can be defined as:

$$\mathcal{L}_{\text{S}} = \mathcal{L}_{\text{CE}} + \lambda \sum_{q \in \mathcal{Q}} \mathcal{D}_f\left(T_s^q(\phi_S), T_t^q(\phi_T)\right) \tag{8}$$

$$\mathcal{Q} = \{(s_l, t_l) \mid \forall\, s_l \in [1, \ldots, s_L],\, t_l \in [1, \ldots, t_L]\}, \tag{9}$$

$$\mathcal{L}_{\text{CE}} = \mathcal{L}(\theta_S, x) \tag{10}$$

where $\mathcal{L}_{\mathrm{CE}}$ is the standard cross-entropy loss function and θ_S is the parameters of the student model. λ is a tunable weighting factor to balance loss terms, which is usually initialized to a relatively large value and decays during training.

3.3. Formulation of Shared Knowledge Distillation

The aim of our Shared-KD is to divide the initial cross-layer feature supervision from teachers to students into two steps: the same-layer distillation between teachers and the cross-layer distillation within students. The first step resembles the same level of interaction between teacher and student in terms of structure and semantic gap. The second step is based on the hierarchical features of the online student, which have similar optimization and semantic properties.

Teacher Share Module. The module is shown in Figure 2. The higher-level features are resized to match the shape of the lower-level features. Then, two features from different levels are concatenated to produce two H × W attention maps. These attention maps are element-wise multiplied with the corresponding features and added to yield the final output. The Teacher Share Module generates different attention maps dynamically based on input features, allowing for flexible aggregation of the two feature maps. The adaptive fusion method is considered superior to direct sum because it combines feature maps from different network stages containing diverse information. This allows for a more reasonable aggregation of low- and high-level features that may focus on different partitions. The use of attention maps facilitates this process.

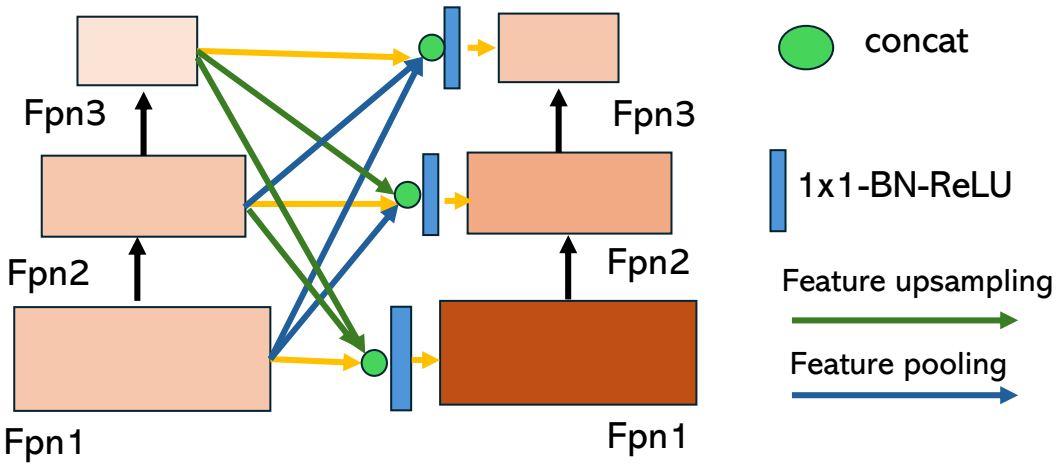

Figure 2. Detailed structure of the Teacher Share Module. It uses privileged self-features from different layers of a student network in addition to useful information from the teacher layer.

Identical-layer distillation. Features in the same layer of the teacher–student model typically share semantic features and shape size. Therefore, simple feature permutations often exist with the same-layer distillation, effectively preserving the teacher model's useful feature knowledge. The loss of inter-layer Shared-KD can be expressed as follows:

$$\mathcal{L}_{\mathrm{ident}} = \frac{1}{m}\sum_{i=1}^{L-1}\mathcal{D}_f\left(T_s^q(\phi_S), T_t^q(\phi_T)\right) \tag{11}$$

$$\mathcal{D}_f\left(T_s^q(\phi_S), T_t^q(\phi_T)\right) = \|T_s^q(\phi_S) - T_t^q(\phi_T)\|_2^2 \tag{12}$$

where m denotes the number of pair loss, L is the number of layers of selected features, we use l_2 distance as D_f, and T_s represents feature alignment. In particular, we use a pooling operation and channel cropping to align features in spatial and channel dimensions without complex transformation.

$$\mathcal{L}_{\text{cross}} = \frac{1}{m} \sum_{i=1}^{L-1} \sum_{j=1}^{L} \mathcal{D}_f\left(T_{s_i}(\phi_{S_i}), T_{s_j}(\phi_{S_j})\right) \tag{13}$$

$$\mathcal{D}_f\left(T_{s_i}(\phi_{S_i}), T_{s_j}(\phi_{S_j})\right) = \|(T_{s_i}(\phi_{S_i}) - T_{s_j}(\phi_{S_j}))\|_2^2 \tag{14}$$

where m denotes the number of pair loss, L is the number of layers of selected features, we use l_2 distance as D_f, and T_s represents feature alignment. In particular, we use a pooling operation and channel cropping to align features in spatial and channel dimensions without complex transformation.

Shared Knowledge Distillation. Overall, in our Shared-KD method, we train the student network with three losses:

$$\mathcal{L}_{\text{Shared-KD}} = \mathcal{L}_{\text{CE}} + \alpha(\mathcal{L}_{\text{ident}} + \mathcal{L}_{\text{cross}}) \tag{15}$$

where α is the weighting factor used to scale the losses. Ablation studies are introduced to demonstrate their effectiveness and robustness. The process of our method is summarized in Algorithm 1.

Algorithm 1 Shared Knowledge Distillation for Object Detection

Input: Teacher: T, Student: S, Input: x, label: y, hyper-parameter: α

1: Using S to obtain the feature ϕ_S and output $\hat{y}$ of Input x
2: Using T to obtain the feature ϕ_T of Input x
3: Calculating the original loss of the model: $\mathcal{L}_{\text{CE}}$
4: Calculating the distillation loss in Equation: $(\mathcal{L}_{\text{ident}} + \mathcal{L}_{\text{cross}})$
5: Using $\mathcal{L}_{\text{Shared-KD}} = \mathcal{L}_{\text{CE}} + \alpha(\mathcal{L}_{\text{ident}} + \mathcal{L}_{\text{cross}})$ to update S

Output: S

4. Experiments

This section first evaluates our approach for the object detection task on MS-COCO. Then, comprehensive ablation experiments are performed to analyze the key design in our Shared-KD. As a novel logit offline approach, the main competitor of Shared-KD is the FGD [32]. Thus, we conduct detailed experimental comparisons between them and also compare their performance with recent advanced KD methods. To ensure fair comparisons, we use the public codes of these approaches with the same training and data preprocessing settings throughout the experiments.

4.1. Experiments on Object Detection

Implementation. We evaluate Shared-KD on MS-COCO dataset [33] and use the most popular open-source detector [34] as the strong baseline. We apply Shared-KD to the two-stage detector (e.g., Faster R-CNN [19]), one-stage detector (e.g., RetinaNet [34]), and anchor-free detector (e.g., FitNets [7]), which are widely used object detection frameworks. We choose Faster RCNN-R101 (T) as the teacher detector for the two-stage detector and Faster RCNN-R50 (S) as the student detector. For one stage detector, we choose RetinaNet-R101 (T) and RetinaNet-R50 (S) as the teacher and student detectors. For the free detector, we choose FCOS-R101 (T) and FCOS-R50 (S) as the teacher and student detectors, respectively. Following common practice [34], all models are trained with a 2× learning schedule (24 epochs). All distillation performances are e valuated in Average Precision (AP).

Comparison results. Table 1 presents a comprehensive evaluation of the proposed Shared Knowledge Distillation (Shared-KD) method against various baseline detectors and other state-of-the-art Knowledge Distillation techniques across different object detection architectures: two-stage detectors, one-stage detectors, and anchor-free detectors, as shown in Figure 3. Shared-KD achieves an impressive AP (Average Precision) of 40.6 for two-stage detectors, outperforming the baseline student model (Faster RCNN-R50) by a significant margin of 2.2. Compared to other Knowledge Distillation methods, Shared-KD demon-

strates superior performance, surpassing FitNets by 1.7 (40.6 vs. 38.9), GID by 0.4 (40.6 vs. 40.2), FRS by 1.1 (40.6 vs. 39.5), and FGD by 0.2 (40.6 vs. 40.4). These improvements are consistent across various evaluation metrics, including AP50, AP75, and AP scores for small, medium, and large objects, highlighting the robustness and effectiveness of Shared-KD for two-stage detectors. In the case of one-stage detectors, Shared-KD attains an AP of 39.4, outperforming the baseline student model (RetinaNet-R50) by a substantial 2.0. When compared to other distillation techniques, Shared-KD surpasses FitNets by 2.0 (39.4 vs. 37.4), GID by 0.3 (39.4 vs. 39.1), and FRS by 0.1 (39.4 vs. 39.3), whereas the improvements over some methods are marginal, the consistent gains across different metrics demonstrate the effectiveness of Shared-KD for one-stage detectors. Shared-KD demonstrates superior performance for anchor-free detectors with an AP of 42.2, outperforming the baseline student model (FCOS-R50) by a remarkable 3.7. Compared to other Knowledge Distillation techniques, Shared-KD achieves significant improvements, surpassing FitNets by 2.3 (42.2 vs. 39.9), GID by 0.2 (42.2 vs. 42.0), FRS by 1.3 (42.2 vs. 40.9), and FGD by 0.1 (42.2 vs. 42.1). The results highlight the effectiveness of Shared-KD in improving object detection accuracy for anchor-free detectors, consistently outperforming the baseline student models and other state-of-the-art methods. The improvements are particularly notable for anchor-free detectors, where Shared-KD achieves the highest AP gain compared to the student baseline and other methods. Overall, the experimental results demonstrate the superior performance of Shared-KD across various object detection architectures, consistently outperforming baseline student models and other Knowledge Distillation techniques. The improvements are most significant for anchor-free detectors, followed by two-stage and one-stage detectors. The robustness and effectiveness of Shared-KD are evident through consistent gains across different evaluation metrics, further solidifying its potential as a promising Knowledge Distillation approach for object detection tasks.

Table 1. Comparison with object detection KD methods on MS COCO val set.

Models	Distillation	AP	50	75	S	M	L
			Two-stage detectors				
Faster RCNN-R101 (T)	-	39.8	60.1	43.3	22.5	43.6	52.8
Faster RCNN-R50 (S)	-	38.4	59.0	42.0	21.5	42.1	50.3
Faster RCNN-R50 (S)	FitNets	38.9 (0.5↑)	59.5	42.4	21.9	42.2	51.6
Faster RCNN-R50 (S)	GID	40.2 (1.8↑)	60.7	43.8	22.7	44.0	53.2
Faster RCNN-R50 (S)	FRS	39.5 (1.1↑)	60.1	43.3	22.3	43.6	51.7
Faster RCNN-R50 (S)	FGD	40.4 (2.0↑)	-	-	22.8	44.5	53.5
Faster RCNN-R50 (S)	Shared-KD	40.6 (2.2↑)	61.6	45.0	24.5	45.6	53.7
			One-stage detectors				
RetinaNet-R101 (T)	-	38.9	58.0	41.5	21.0	42.8	52.4
RetinaNet-R50 (S)	-	37.4	56.7	39.6	20.0	40.7	49.7
RetinaNet-R50 (S)	FitNets	37.4 (0.0↑)	57.1	40.0	20.8	40.8	50.9
RetinaNet-R50 (S)	GID	39.1 (1.7↑)	59.0	42.3	22.8	43.1	52.3
RetinaNet-R50 (S)	FRS	39.3 (1.9↑)	58.8	42.0	21.5	43.3	52.6
RetinaNet-R50 (S)	Shared-KD	39.4 (2.0↑)	59.0	42.5	21.5	43.9	54.0
			Anchor-free detectors				
FCOS-R101 (T)	-	40.8	60.0	44.0	24.2	44.3	52.4
FCOS-R50 (S)	-	38.5	57.7	41.0	21.9	42.8	48.6
FCOS-R50 (S)	FitNets	39.9 (1.4↑)	58.6	43.1	23.1	43.4	52.2
FCOS-R50 (S)	GID	42.0 (3.5↑)	60.4	45.5	25.6	45.8	54.2
FCOS-R50 (S)	FRS	40.9 (2.4↑)	60.3	43.6	25.7	45.2	51.2
FCOS-R50 (S)	FGD	42.1 (3.6↑)	-	-	27.0	46.0	54.6
FCOS-R50 (S)	Shared-KD	42.2 (3.7↑)	60.9	46.1	25.7	46.7	54.1

↑ The rising arrow indicates an improvement in performance.

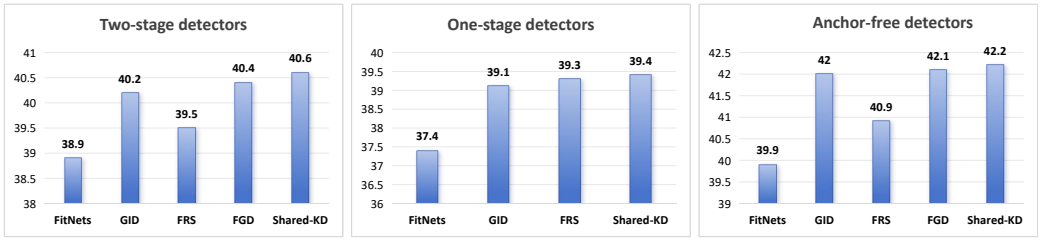

Figure 3. Comparison of our Shared-KD with other methods for different detector architectures, including two-stage detectors, one-stage detectors, and anchor-free detectors.

4.2. Instance Segmentation

We also apply our method to the more challenging instance segmentation task. We take Mask R-CNN [35] as our baseline models and distill between different backbone architectures. The models are trained on the COCO2017 training set and are evaluated on the validation set. Table 2 provides a detailed analysis of the experimental results, for instance segmentation on the MS COCO 2017 dataset. The table presents the performance of various Knowledge Distillation methods, including the proposed Shared Knowledge Distillation (Shared-KD) approach, in comparison with the baseline student model (Mask RCNN-R50) and other state-of-the-art techniques such as FKD, FGD, and MGD. In terms of overall Average Precision (AP), Shared-KD achieves an impressive score of 41.3, significantly outperforming the baseline student model by a substantial margin of 5.9 (41.3 vs. 35.4). This remarkable improvement highlights the effectiveness of Shared-KD in boosting the performance of the student model for instance segmentation tasks.

Table 2. Experiments of instance segmentation on MS COCO2017. The teacher detector is Cascade Mask R-CNN with ResNeXt-101 backbones. AP means average precision.

Models	Distillation	AP	S	M	L
Mask RCNN-R50 (S)	-	35.4	19.1	38.6	48.4
Mask RCNN-R50 (S)	FKD	37.4	19.7	40.5	52.1
Mask RCNN-R50 (S)	FGD	37.8	17.1	40.7	56.0
Mask RCNN-R50 (S)	MGD	38.1	17.1	41.1	56.3
Mask RCNN-R50 (S)	Shared-KD	41.3	23.1	45.0	55.2

Comparison results. Compared to other Knowledge Distillation methods, Shared-KD demonstrates superior performance, surpassing FKD by 3.9 (41.3 vs. 37.4), FGD by 3.5 (41.3 vs. 37.8), and MGD by 3.2 (41.3 vs. 38.1). These significant gains underscore the robustness and efficacy of Shared-KD in transferring valuable knowledge from the teacher model to the student model, enabling the student to achieve state-of-the-art performance in instance segmentation. Further analysis of the AP scores for different object scales reveals the strengths of Shared-KD in handling objects of varying sizes. For small objects, Shared-KD achieves an AP of 23.1, outperforming the baseline student model by 4.0 and other methods like FKD (19.7), FGD (17.1), and MGD (17.1). This demonstrates Shared-KD's ability to effectively capture and transfer knowledge related to small object instances, which can be challenging for traditional object detection and segmentation models. For medium objects, Shared-KD obtains an AP of 45.0, significantly surpassing the baseline student model by 6.4 and other distillation methods such as FKD (40.5), FGD (40.7), and MGD (41.1). This highlights Shared-KD's capability in accurately detecting and segmenting medium-sized objects, which are often the most common and critical instances in real-world scenarios, and whereas Shared-KD performs slightly better than the baseline student model for large objects, with an AP of 55.2 compared to 48.4, it is slightly outperformed by some other methods like FGD (56.0) and MGD (56.3). However, the overall superior performance of Shared-KD across different object scales, particularly for small and medium objects, demonstrates its robustness and generalization capabilities. In summary, the experimental results for instance segmentation on the MS COCO 2017 dataset clearly

demonstrate the effectiveness of Shared-KD in boosting the performance of the student model. Shared-KD consistently outperforms the baseline student model and other state-of-the-art Knowledge Distillation techniques, achieving significant improvements in overall Average Precision and across different object scales, particularly for small and medium objects. These impressive results highlight the potential of Shared-KD as a powerful Knowledge Distillation approach for instance segmentation tasks, paving the way for further advancements in this domain.

4.3. Ablation Study

Analysis for different components in our method. In our method, we conducted experiments to evaluate the impact of different design components, as shown in Table 3. When the identical-layer distillation component is removed, the performance drops slightly, with a decrease in overall AP to 40.1%, AP50 to 57.0%, and AP scores for small and large objects to 21.0% and 52.5%, respectively. This indicates that the identical-layer distillation plays a role in aligning the teacher and student models, contributing to better performance, particularly for small and large objects. On the other hand, removing the cross-layer distillation component also leads to a performance degradation, with a decrease in overall AP to 40.3%, AP75 to 43.9%, and AP scores for small and medium objects to 23.0% and 44.5%, respectively. This suggests that the cross-layer distillation effectively facilitates knowledge transfer between different layers of the teacher and student models, improving the detection accuracy, especially for small and medium-sized objects. The analysis demonstrates that both the identical-layer distillation and cross-layer distillation components contribute to the overall effectiveness of the proposed method, with each component playing a distinct role in enhancing object detection performance.

Table 3. Results of different components in our method.

Method	AP	50	75	S	M	L
Ours	40.6	61.6	45.0	24.5	45.6	53.7
without Teacher Share Module	40.5	61.2	44.6	24.2	45.2	53.3
without identical-layer distillation	40.1	57.0	43.5	21.0	44.0	52.5
without cross-layer distillation	40.3	60.9	43.9	23.0	44.5	53.0

Compare to other cross-layer distillation techniques. In comparison with other cross-layer distillation methods, such as SemCKD [17], our proposed method, referred to as Shared-KD, demonstrates superior performance on the MS COCO validation set, as shown in Table 4. The results in the table clearly illustrate the effectiveness of our approach in improving the object detection accuracy of the RetinaNet-R50 (Student) model when distilled from the RetinaNet-R101 (Teacher) model. Our Shared-KD method outperforms SemCKD in terms of AP50 (59.0 vs. 58.5) and AP for large objects (AP_L: 54.0 vs. 52.0), indicating its superior performance in detecting objects with high confidence and handling large-scale objects. Additionally, our Shared-KD method demonstrates computational advantages over SemCKD. It requires fewer training hours (10.5 h compared to 13.8 h for SemCKD) and consumes less memory (3.8 GB compared to 4.5 GB for SemCKD), making it more efficient and resource-friendly. The performance gains achieved by our Shared-KD method can be attributed to the effective knowledge transfer from the teacher model to the student model, leveraging shared representations and feature alignment techniques.

Table 4. Comparison with other cross-layer distillation (SemCKD [17]) on MS COCO val set.

Models	Distillation	AP	50	75	S	M	L	Time	Memory
RetinaNet-R101 (T)	-	38.9	58.0	41.5	21.0	42.8	52.4	-	-
RetinaNet-R50 (S)	-	37.4	56.7	39.6	20.0	40.7	49.7	-	-
RetinaNet-R50 (S)	SemCKD [17]	38.8 (1.4↑)	58.5	49.5	22.0	43.0	52.0	13.8 h	4.5 GB
RetinaNet-R50 (S)	Shared-KD	39.4 (2.0↑)	59.0	42.5	21.5	43.9	54.0	10.5 h	3.8 GB

↑ The rising arrow indicates an improvement in performance.

Sensitivity study for hyper-parameters. As shown in Table 5, experiments are conducted to study the hyper-parameter sensitivity. We vary the weight α from 2, 5, 10, and 20 to compare their performance. Furthermore, the AP is 39.4, 38.6, 37.8, and 37.2, respectively. The results demonstrate that the weight α of 2 is the best solution for the hyper-parameter setting. These results demonstrate that our approach can achieve robust performance improvements under different hyper-parameters.

Table 5. Ablation study of individual distillation loss with RetinaNet-R50 as student, RetinaNet-X101 as teacher.

Method	AP	50	75	S	M	L
Baseline	37.4	56.7	39.6	20.0	40.7	49.7
2	39.4	58.3	42.3	22.6	43.5	51.2
5	38.6	57.9	41.0	21.6	42.0	51.8
10	37.8	57.2	40.6	21.2	41.8	51.4
20	37.2	56.8	40.1	20.8	41.2	51.0

Qualitative analysis. Figure 4 shows significant improvements of our proposed approach compared to the baseline methods and the FGD technique, showing results that are closer to the ground truth (GT) annotations. For small objects, our approach is able to more accurately localize and classify them, overcoming the challenges that often arise when dealing with small-scale instances. The specialized handling of foreground and background classes, as well as the robust feature distillation strategies, contribute to the enhanced small object detection capabilities. Furthermore, our method demonstrates a notable improvement in preventing missed detections. By decoupling the training process for different class types and employing tailored masking and loss functions, our framework is better equipped to capture the subtleties and contextual cues that are crucial for comprehensive object detection. This leads to a reduction in the number of missed detections, resulting in detection outputs that more closely align with the ground truth annotations. The qualitative analysis showcases the strengths of our approach in handling small objects and mitigating missed detections, underscoring its effectiveness in advancing the state-of-the-art in object detection tasks. The ability to produce detection results that are visually more accurate and aligned with the ground truth highlights the practical benefits of our proposed method.

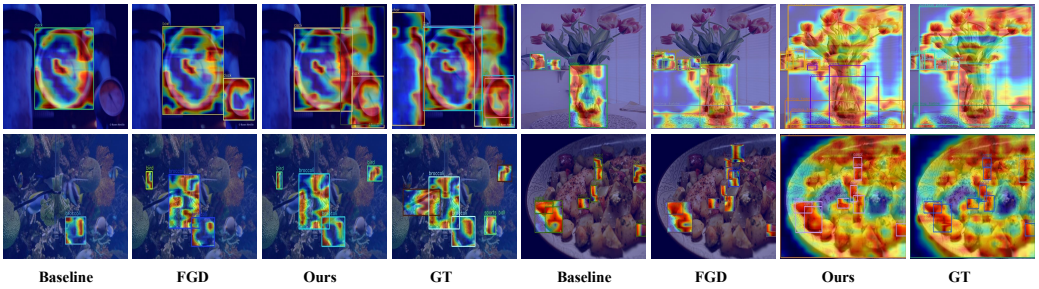

Figure 4. Qualitative analysis of baseline, FGD, our method, and GT for RetinaNet on COCO benchmarks.

5. Conclusions

Object detection based on Knowledge Distillation has the potential to enhance the functionality and efficiency of 5G and 6G networks across diverse domains. This paper presents the Shared-KD network, a simple, effective, and new framework for addressing the challenges associated with cross-layer feature discrepancies in teacher–student networks. It has been demonstrated that the significant gaps between teacher and student models at the intermediate level pose obstacles to the success of feature distillation. By drawing inspiration from collaborative learning in education, we have proposed a knowledge augmentation module for teachers and a mutual learning module for students. This approach has allowed us to decompose the original feature supervision into two steps: identical-layer distillation between teacher and student, and cross-layer distillation within students. Through experiments on various tasks, including object detection and instance segmentation, we have shown that Shared-KD consistently outperforms other methods, achieving significant performance gains across different neural network architectures. Moreover, the simplicity and efficiency of Shared-KD, which eliminates the need for complex transformations and extra training parameters, make it a practical and versatile solution for Knowledge Distillation. We believe that the insights provided by Shared-KD and its success in improving model performance will inspire further advancements in Knowledge Distillation research and contribute to a deeper understanding of feature distillation. Within the 5G/6G next-generation network paradigm, this will pave the way for advances in real-time applications and services.

Author Contributions: Methodology: Z.G. and P.Z.; writing—original draft: Z.G.; experiment(s): Z.G. and P.L.; writing—review and editing: Z.G., P.L. and P.Z.; supervision: P.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (No. 2022YFC3302100).

Data Availability Statement: The data presented in this study are available in this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gao, J.; Wu, D.; Yin, F.; Kong, Q.; Xu, L.; Cui, S. MetaLoc: Learning to learn wireless localization. *IEEE J. Sel. Areas Commun.* **2023**, *41*, 3831–3847. [CrossRef]
2. Cao, X.; Lyu, Z.; Zhu, G.; Xu, J.; Xu, L.; Cui, S. An overview on over-the-air federated edge learning. *arXiv* **2024**, arXiv:2208.05643v1.
3. Sun, W.; Zhao, Y.; Ma, W.; Guo, B.; Xu, L.; Duong, T.Q. Accelerating Convergence of Federated Learning in MEC with Dynamic Community. *IEEE Trans. Mob. Comput.* **2023**, *23*, 1769–1784. [CrossRef]
4. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. Acm* **2012**, *60*, 84–90. [CrossRef]
5. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
6. Li, L.; Li, A. A2-Aug: Adaptive Automated Data Augmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2266–2273.
7. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. Fitnets: Hints for thin deep nets. *arXiv* **2014**, arXiv:1412.6550.
8. Zagoruyko, S.; Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv* **2016**, arXiv:1612.03928.
9. Dong, P.; Niu, X.; Li, L.; Xie, L.; Zou, W.; Ye, T.; Wei, Z.; Pan, H. Prior-Guided One-shot Neural Architecture Search. *arXiv* **2022**, arXiv:2206.13329.
10. Zhu, C.; Li, L.; Wu, Y.; Sun, Z. Saswot: Real-time semantic segmentation architecture search without training. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 26 February 2024; Volume 38, pp. 7722–7730.
11. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:1503.02531.
12. Kim, J.; Park, S.; Kwak, N. Paraphrasing complex network: Network compression via factor transfer. *arXiv* **2018**, arXiv:1802.04977.

13. Ahn, S.; Hu, S.X.; Damianou, A.; Lawrence, N.D.; Dai, Z. Variational information distillation for knowledge transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9163–9171.
14. Tung, F.; Mori, G. Similarity-preserving knowledge distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1365–1374.
15. Heo, B.; Lee, M.; Yun, S.; Choi, J.Y. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January 2019–1 February 2019; Volume 33, pp. 3779–3787.
16. Huang, Z.; Wang, N. Like What You Like: Knowledge Distill via Neuron Selectivity Transfer. *arXiv* **2017**, arXiv:1707.01219.
17. Chen, D.; Mei, J.P.; Zhang, Y.; Wang, C.; Wang, Z.; Feng, Y.; Chen, C. Cross-Layer Distillation with Semantic Calibration. *arXiv* **2020**, arXiv:2012.03236.
18. Chung, I.; Park, S.; Kim, J.; Kwak, N. Feature-map-level online adversarial knowledge distillation. In Proceedings of the International Conference on Machine Learning. PMLR, Virtual Event, 13–18 July 2020; pp. 2006–2015.
19. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv* **2015**, arXiv:1506.01497.
20. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
21. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
22. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
23. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
24. Li, L. Self-regulated feature learning via teacher-free feature distillation. In Proceedings of the European Conference on Computer Vision. Springer, Tel Aviv, Israel, 3–27 October 2022; pp. 347–363.
25. Dong, P.; Li, L.; Wei, Z. Diswot: Student architecture search for distillation without training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 11898–11908.
26. Liu, X.; Li, L.; Li, C.; Yao, A. Norm: Knowledge distillation via n-to-one representation matching. *arXiv* **2023**, arXiv:2305.13803.
27. Li, L.; Liang, S.N.; Yang, Y.; Jin, Z. Teacher-free distillation via regularizing intermediate representation. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18–23 July 2022; pp. 1–6.
28. Li, L.; Dong, P.; Wei, Z.; Yang, Y. Automated knowledge distillation via monte carlo tree search. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 17413–17424.
29. Buciluǎ, C.; Caruana, R.; Niculescu-Mizil, A. Model compression. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006; pp. 535–541.
30. Yim, J.; Joo, D.; Bae, J.; Kim, J. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4133–4141.
31. Zhang, L.; Ma, K. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 30 April 2020.
32. Yang, Z.; Li, Z.; Jiang, X.; Gong, Y.; Yuan, Z.; Zhao, D.; Yuan, C. Focal and global knowledge distillation for detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4643–4652.
33. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
34. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
35. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.