Learning Lightweight Object Detectors via Multi-Teacher Progressive Distillation

Shengcao Cao [1] **Mengtian Li** [2 3] **James Hays** [4] **Deva Ramanan** [2] **Yu-Xiong Wang** [1] **Liang-Yan Gui** [1]

Abstract

Resource-constrained perception systems such as edge computing and vision-for-robotics require vision models to be both accurate and lightweight in computation and memory usage. While knowledge distillation is a proven strategy to enhance the performance of lightweight classification models, its application to structured outputs like object detection and instance segmentation remains a complicated task, due to the variability in outputs and complex internal network modules involved in the distillation process. In this paper, we propose a simple yet surprisingly effective sequential approach to knowledge distillation that *progressively transfers the knowledge of a set of teacher detectors* to a given lightweight student. To distill knowledge from a highly accurate but complex teacher model, we construct a sequence of teachers to help the student gradually adapt. Our progressive strategy can be easily combined with existing detection distillation mechanisms to consistently maximize student performance in various settings. To the best of our knowledge, we are the *first* to successfully distill knowledge from Transformer-based teacher detectors to convolution-based students, and unprecedentedly boost the performance of ResNet-50 based RetinaNet from 36.5% to **42.0%** AP and Mask R-CNN from 38.2% to **42.5%** AP on the MS COCO benchmark. Code available at https://github.com/Shengcao-Cao/MTPD.

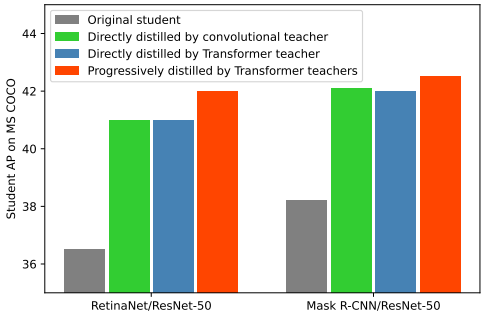


Figure 1. **Our proposed Multi-Teacher Progressive Distillation (MTPD) leads to state-of-the-art student detection performance.** When switching the teacher model from a convolution-based detector to a Transformer-based one with stronger detection performance, the student does not become more accurate, due to the architectural difference between the teacher-student pair. *Progressively distilling knowledge from multiple teacher detectors* can mitigate the capacity gap and result in the best student detection performance.

1. Introduction

Deploying deep neural models in safety-critical real-time applications is challenging, especially on devices with limited resources such as self-driving cars or virtual/augmented reality headsets. This is mainly due to the huge computational complexity and massive memory/storage demands. One effective strategy is to train lightweight architectures that have already been carefully engineered for efficient memory access, via knowledge distillation (Buciluǎ et al., 2006; Hinton et al., 2014) which is able to compress learned information from a large model into a small one.

Implementing knowledge distillation in the realm of object detection, despite existing efforts, presents its unique challenges stemming from the complex task outputs (Chen et al., 2017): Detectors operate with multi-task heads (for classification and box/mask regression) producing variable-length outputs, which differentiates detection from the single-output classification task. Therefore, distillation methods developed for classification are often not directly applicable to detection, and dedicated methods (Chen et al., 2017) need to be developed for detection in the literature (detailed discussion in Appendix E, Table 16).

Recent work (Zhang & Ma, 2021; Shu et al., 2021; Yang et al., 2022b) on detector distillation mainly considers de-

[1]University of Illinois Urbana-Champaign [2]Carnegie Mellon University [3]Now at Waymo [4]Georgia Institute of Technology. Correspondence to: Shengcao Cao <cao44@illinois.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

signing advanced distillation loss functions for transferring features from teachers to students. However, there are two unsolved challenges: 1) The *capacity gap* (Cho & Hariharan, 2019; Mirzadeh et al., 2020) between models can result in a sub-optimal distilled student even if the strongest teacher has been employed, which is undesired when optimizing the accuracy-efficiency trade-off of the student. Moreover, when distilling knowledge from Transformer-based teachers (Dosovitskiy et al., 2020; Liu et al., 2021) to classical convolution-based students, this *architectural difference* can enlarge the teacher-student gap (Figure 1). 2) Current methods assume that one target teacher has been selected. However, this meta-level optimization of *teacher selection* is neglected in the existing literature of detector distillation. In fact, finding a pool of strong teacher candidates is easy, but trial-and-error may be necessary before determining the most compatible teacher for a specific student.

To address these challenges, we propose a framework that learns a lightweight detector via **Multi-Teacher Progressive Distillation (MTPD)**: 1) We find sequential distillation from *multiple teachers arranged into a curriculum* significantly improves knowledge distillation and bridges the teacher-student capacity gap caused by different architectures. As shown in Figure 1, even with huge architectural difference, MTPD can effectively transfer knowledge from Transformer-based teachers to convolution-based students, while previous methods cannot. 2) For the teacher selection problem, we design a heuristic algorithm for a given student and a pool of teacher candidates, to *automatically determine the order of teachers* to use in the distillation procedure. This algorithm is based on the analysis of the representation similarity between models, which does not require prior knowledge of the specific distillation mechanism.

To summarize, our **main contributions** are:

- We propose a framework for learning lightweight detectors through Multi-Teacher Progressive Distillation (MTPD), which is simple yet effective and general. We develop a principled method to automatically design a sequence of teachers appropriate for a given student and progressively distill it.
- MTPD is a *meta-level* strategy that can be easily combined with previous efforts in detection distillation. We perform comprehensive empirical evaluation on the challenging MS COCO dataset and observe consistent gains, regardless of the distillation loss complexity (from a simple feature-matching loss in Table 3 to the most advanced, sophisticated losses in Figure 4).
- MTPD learns lightweight RetinaNet and Mask R-CNN with state-of-the-art accuracy, even in *heterogeneous backbone and input resolution* settings. Perhaps most impressively, for the first time, we investigate heterogeneous distillation from Transformer-based teacher detectors to a convolution-based student, and find progressive distillation is the key to bridge their gap (Figure 1, Table 5).
- We empirically show that the improvement comes from better generalization rather than better optimization. The knowledge transferred from multiple teachers leads the student to a more flat minimum, and thus help the student *generalize* better (Figure 5).

2. Related Work

Knowledge distillation for classification: The idea of training a shallow student network with supervision from a deep teacher was originally proposed by Buciluǎ et al. (2006), and later formally popularized by Hinton et al. (2014). Different knowledge can be used, such as response-based knowledge (Hinton et al., 2014), and feature-based knowledge (Romero et al., 2015; Ahn et al., 2019). Several multi-teacher knowledge distillation methods have been proposed (Vongkulbhisal et al., 2019; Sau & Balasubramanian, 2016), which usually use the average of logits and feature representations as the knowledge (You et al., 2017; Fukuda et al., 2017). Mirzadeh et al. (2020) show that an intermediate teacher assistant, decided by architectural similarities, can bridge the gap between the student and the teacher. We find: 1) Extending Mirzadeh et al. (2020) to detection where teacher architectures are diverse is challenging. 2) Classification-oriented distillation (Romero et al., 2015; Ahn et al., 2019) is not directly applicable to detection. 3) Using a sequence of teachers, instead of their ensemble (You et al., 2017; Fukuda et al., 2017), is more effective. A more detailed discussion that compares our approach with prior work on progressive distillation, multi-teacher distillation, online distillation, deep mutual learning, and other distillation mechanisms is in Appendix E.

Object detection and instance segmentation: A variety of convolutional neural network (CNN) based object detection frameworks have been proposed, and can be generally divided into single-stage and two-stage detectors. Typical single-stage methods include YOLO (Redmon et al., 2016; Redmon & Farhadi, 2018) and RetinaNet (Lin et al., 2017b), and two-stage methods include Faster R-CNN (Ren et al., 2014) and Mask R-CNN (He et al., 2017). Recently, several multi-stage models are proposed, such as HTC (Chen et al., 2019a) and DetectoRS (Qiao et al., 2021). These detection frameworks achieve better detection accuracy with better feature extraction backbones and more complicated heads, which are more computationally expensive.

Knowledge distillation for detection and segmentation: Dedicated distillation methods are proposed to train efficient object detectors for this task different from classification. Chen et al. (2017) first use knowledge distillation to enforce the student detector to mimic the teacher's predictions. More recent efforts usually focus on learning from the teacher's features, rather than final predictions. Various distillation

mechanisms have been proposed to leverage the impact of foreground and background objects (Wang et al., 2019; Guo et al., 2021a), relation between individual objects (Zhang & Ma, 2021; Dai et al., 2021), or relation between local and global information (Yang et al., 2022a;b). Different from the methods that distill from a single teacher, we study multi-teacher distillation where an ordered sequence of teachers is required, and we find that a simple feature-matching loss is adequate to significantly boost student accuracy.

3. Approach

In **Multi-Teacher Progressive Distillation (MTPD)**, we propose to progressively distill a student model S with a pool of N teachers $\mathcal{P} = \{T_i\}_{i=1}^{N}$. Typical object detectors consist of four modules: (1) backbone, which extracts visual features, such as ResNet (He et al., 2016) and ResNeXt (Xie et al., 2017); (2) neck, which extracts multi-level feature maps from various stages of the backbone, such as FPN (Lin et al., 2017a) and Bi-FPN (Tan et al., 2020); (3) optional region proposal network (RPN) used in two-stage detectors; and (4) head, which generates final predictions for object detection and segmentation. We denote the output feature maps of the *neck* as F^{Net}, where Net can be either the student model S or one of the teachers $T_i \in \mathcal{P}$. With neck modules like FPN, the feature maps can be multi-level.

MTPD is a *general meta-strategy* for detector distillation that progressively learns a student using a sequence of teachers. Here, to examine this meta-strategy without involving sophisticated distillation mechanisms, we introduce a simple feature-matching distillation for a single teacher T_i in Section 3.1. Then we discuss progressive distillation with multiple teachers from $\mathcal{P}$ in Section 3.2.

3.1. Preliminary: Single Teacher Distillation via Simple Feature Matching

In order to learn an efficient student detector S through distillation, we encourage the feature representation of a student to be similar to that of the teacher (Chen et al., 2017; Yang et al., 2021). To this end, we minimize the discrepancy between the feature representations of the teacher and the student. Without bells and whistles, we simply minimize the L2 distance between F^{T_i} and F^S:

$$L_{\text{distill}} = \left\| F^{T_i} - r(F^S) \right\|_2^2, \tag{1}$$

where $r(\cdot)$ is a function used to match the feature map dimensions of the teacher and the student.

We define $r(\cdot)$ as follows:

- (Homogeneous case) If the numbers of channels and the spatial resolutions are both the same between T_i and S, $r(\cdot)$ is an identity function.
- (Heterogeneous case) If the numbers of channels are different but the spatial resolutions are the same, we use 1×1 convolutional filters as $r(\cdot)$. If the spatial resolutions are different but the numbers of channels are the same, we use an upsampling layer as $r(\cdot)$. If both the numbers of channels and spatial resolutions are different, we compose the convolutional and upsampling layers as $r(\cdot)$.

Note that the mapping $r(\cdot)$ is only required at training time and thus *does not add any overhead* to the inference. Overall, our loss function can be written as:

$$L = \lambda L_{\text{distill}} + L_{\text{detect}}, \tag{2}$$

where λ is a balancing hyper-parameter and L_{detect} is the detection loss based on the ground truth labels. Compared with state-of-the-art detection distillation approaches (Zhang & Ma, 2021; Shu et al., 2021; Yang et al., 2022a;b), which introduce more complex designs of the distillation loss, this feature-matching distillation is simpler and does not require running the heads of the teacher model. Our distillation loss is illustrated in Figure 2-**Left**.

3.2. Progressive Distillation with Multiple Teachers

The overall aim of knowledge distillation is to make a student mimic a teacher's output, so that the student is able to obtain similar performance to teacher's. However, the capacity of the student model is limited, making it hard for the student to learn from a highly complex teacher (Cho & Hariharan, 2019). To address this issue, multiple teacher networks are used to provide more supervision to a student (Sau & Balasubramanian, 2016; You et al., 2017). Unlike previous methods which distill knowledge from the ensemble of logits or features simultaneously, we propose to distill feature-based knowledge from multiple teachers *sequentially*. Our key insight is that instead of mimicking the ensemble of all feature information together, the student can be distilled more effectively by the knowledge provided by one teacher each time. This progressive knowledge distillation approach can be considered as designing a curriculum (Bengio et al., 2009) offered by a sequence of teachers, as illustrated in Figure 2-**Right**.

The crucial question is: *What is the optimal order $\mathcal{O}$ of the teachers when distilling the student?* A brute-force approach might search over all orders and pick the best (that produces a distilled student with the highest validation accuracy). However, the space of permutation orders grows exponentially with the number of teachers, making this impractical to scale. Therefore, we propose a principled and efficient approach based on a correlation analysis of each model's learned feature representation.

First, we quantify the dissimilarity between each pair of models' representations, as a proxy for their capacity gap. Representation (dis)similarity (Raghu et al., 2017; Wang et al., 2018; Kornblith et al., 2019) has been studied to under-

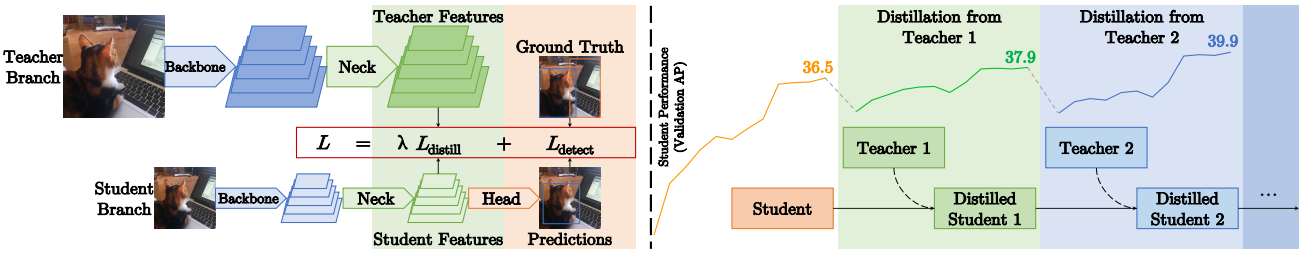


Figure 2. **Multi-Teacher Progressive Distillation (MTPD) for object detectors. Left**: For each teacher-student pair, the training target consists of two parts: L_{distill} minimizes the discrepancy between the neck feature maps of the student and the current teacher, and L_{detect} is the original detection loss based on the ground truth. **Right**: We use a *sequence* of teacher models to distill the lightweight student detector. The sequence of teachers forms a curriculum. Using a suitable sequence of teachers can significantly boost the student model's performance. The representative performance curve illustrates that MTPD improves the COCO validation AP of ResNet-50 backboned RetinaNet student first from 36.5% to 37.9% using HTC (Teacher 1), and then from 37.9% to 39.9% using DetectoRS (Teacher 2).

stand the learning capacity of neural models. In our setting, we find a linear regression model is adequate for measuring the representation dissimilarity. Given two trained detectors A and B, we freeze their parameters, and thus fixing the feature representations. Then we learn a linear mapping $r(\cdot)$, implemented by a 1×1 convolutional layer at each feature level, as specified in the heterogeneous case in Section 3.1. $r(\cdot)$ is trained to minimize L_{distill}, so it can transform A's features to approximate B's features. After training $r(\cdot)$, we evaluate it by L_{distill} on the validation set, and denote the validation loss as the *adaptation cost* $\mathcal{C}(A, B)$. This metric can be a proxy of the capacity gap between two models: When $\mathcal{C}(A, B)$ is zero, a linear mapping can transform A's features to B's, and there is no additional knowledge from B. When $\mathcal{C}(A, B)$ is large, it is more difficult to adapt A's representation to B's. Note that the adaptation cost is non-symmetric – it is relatively easier to adapt a high-capacity model's representations to a low-capacity model's representations, than the other way around.

We design a heuristic algorithm, **Backward Greedy Selection (BGS)**, to acquire a near-optimal distillation order $\mathcal{O}$ automatically (see pseudo-code in Algorithm 1 and illustration in Figure 3). Suppose the maximum number of teachers to be selected is limited by k (which can be arbitrarily decided according to desired training time), and we aim to find a teacher index sequence α no longer than k. We construct the teacher order *backwards*: The best performing teacher is set as the final target T_{α_k}; before the final teacher, we use another teacher, which has the smallest adaptation cost $\mathcal{C}(\cdot, T_{\alpha_k})$ to that final teacher, as the penultimate teacher $T_{\alpha_{k-1}}$. We repeat this procedure to find preceding teachers, until: 1) when trying to select T_{α_j}, we find the transfer costs from remaining teachers to the next teacher $\mathcal{C}(\cdot, T_{\alpha_{j+1}})$ are all larger than the transfer cost from the trained student to the next teacher $\mathcal{C}(S, T_{\alpha_{j+1}})$; or 2) we reach the given maximum step limit k. Intuitively, the resulting sequence of teachers bridges the gap between the student model and the teacher, with an increasingly difficult curriculum. Section 4.1 and Appendix A demonstrate the efficacy of BGS.

Our teacher order design approach is efficient and scalable. In fact, the main computation overhead is the optimization of a set of tiny linear mappings ($\mathbb{R}^{256} \mapsto \mathbb{R}^{256}$ for FPN-based detectors). In our setting, this process requires about 3 GPU hours per student model, *a fraction of the hundreds of GPU hours needed for distillation*. If more teacher candidates are added, we can first generate feature maps only once for each teacher. Then we optimize pair-wise linear mappings using only 10%-20% GPU hours, ensuring a near-linear time consumption increase relative to the number of teachers.

Since MTPD is a meta-level strategy, it can be integrated with previous designs of distillation mechanisms, without much efforts. Starting with a student detector and a pool of candidate teachers, we can first select a subset of teachers and design their distillation order. In place of the simple feature matching loss, we then apply a more advanced distillation mechanism with each teacher sequentially to train the student detector.

4. Experiments

We study the efficacy and generalizability of our proposed MTPD from multiple perspectives. First of all in Section 4.1, we use a controlled experiment to demonstrate that BGS consistently produces teacher orders that are near-optimal compared with all possibilities. Then in Sections 4.2 and 4.3, we apply MTPD along with the simple feature-matching loss (Section 3.1) to show that this strategy alone brings significant gains to knowledge distillation. Since our contribution of progressive distillation is orthogonal to previous efforts in designing distillation mechanisms, in Section 4.4 we then combine MTPD with state-of-the-art distillation mechanisms to maximize the student performance, and we show that MTPD is the key to the success of distillation from Transformer-based teachers to convolution-based students. Finally in Section 4.5, we understand the performance gain

Table 1. **Configuration and COCO performance of the teacher and student detectors.** We investigate a variety of models with heterogeneous input resolutions, backbones, necks, and head structures. '1×' input resolution refers to the standard 1333 × 800 resolution, and '0.25×' means 333 × 200 resolution. 'R-' backbones are ResNets with different number of layers.

Model	Input Res.	Backbone	Neck	Head	AP Box	AP Mask	Runtime (ms)
Teachers							
I	1×	R50	FPN	Mask R-CNN	38.2	34.7	51
II	1×	R50	FPN	FCOS	38.7	-	36
III	1×	R50	FPN	HTC	42.3	37.4	181
IV	1×	R50+SAC	RFP	HTC (DetectoRS)	49.1	42.6	223
V	1×	R50+SAC	RFP	Mask R-CNN	45.1	40.1	142
Students							
I	1×	R50	FPN	RetinaNet	36.5	-	43
II	1×	R50	FPN	Mask R-CNN	38.2	34.7	51
III	1×	R18	FPN	Mask R-CNN	33.3	30.5	29
IV	0.25×	R50	FPN	Mask R-CNN	25.8	23.0	17

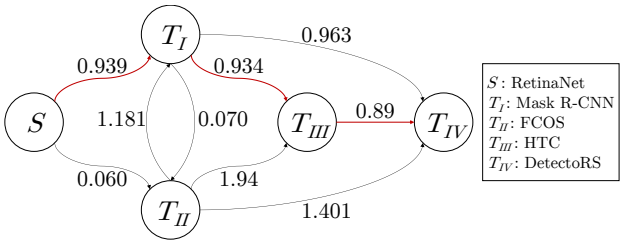


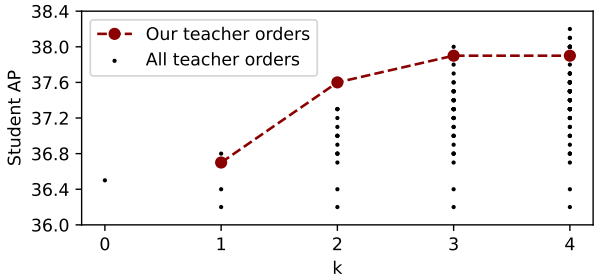
Figure 3. **Adaptation costs among models.** The number on each directed edge is the adaptation cost metric described in Section 3.2. Some edges are not shown for visual clarity. The red path is suggested by BGS when $k = 3$ teachers are selected: 1) We use the best performing Teacher IV as the final teacher in the sequence; 2) use the teacher closest to Teacher IV, which is Teacher III, as the second teacher; and 3) use the teacher closest to Teacher III, which is Teacher I, as the first teacher.

of MTPD by analyzing the training loss dynamics.

Student and teacher models: To investigate the impact of different teacher models and their combinations, as shown in Table 1, we construct a variety of teacher-student pairs from a set of widely-used object detection and instance segmentation networks, including RetinaNet (Lin et al., 2017b), Mask R-CNN (He et al., 2017), FCOS (Tian et al., 2019), HTC (Chen et al., 2019a), and DetectoRS (Qiao et al., 2021). They have a wide range of runtime and detection performance. We select ResNet-50 backboned RetinaNet and Mask R-CNN as the student models (Students I & II), due to their low latency, simple structure, and wide application, for single-stage and two-stage object detection respectively. More advanced models such as DetectoRS have better detection performance, but require much more training/inference time, so we use them as teachers. We also consider lightweight variants of Mask R-CNN as students, which have a smaller backbone (Student III) or a reduced input resolution (Student IV).

Table 2. **Comparison of the teacher order suggested by BGS with all other orders under limited training budgets (Li et al., 2020b).** k denotes the maximum number of used teachers. **Top**: We show some statistics of all possible student AP performance and the ranking of the student using our distillation order. **Bottom**: We visualize the comparative advantage of our teacher orders (red dots) over all other orders (black dots). Some black scatter points overlap due to the same student AP. BGS consistently produces highly competitive distillation orders of teachers.

k	Suggested teacher order	Student AP	All student AP range	Ranking in all orders
1	IV	36.7	[36.2, 36.8]	2 / 4
2	III→IV	37.6	[36.2, 37.6]	1 / 16
3	I→III→IV	37.9	[36.2, 38.0]	2 / 40
4	I→III→IV	37.9	[36.2, 38.2]	7 / 64



Datasets and evaluation metrics: We mainly evaluate on the challenging object detection dataset MS COCO 2017 (Lin et al., 2014), which contains bounding boxes and segmentation masks for 80 common object categories. We train our models on the split of `train2017` (118k images) and report results on `val2017` (5k images). We report the standard COCO-style Average Precision (AP) metric and end-to-end latency (from images to predictions) as the runtime. We also evaluate on another object detection dataset Argoverse-HD (Chang et al., 2019), and a more challenging evaluation protocol in streaming perception (Li et al., 2020a). These results are in Appendix D.

Baselines: Our main contribution is *orthogonal* to previous methods: We leverage a sequence of teachers to distill the student, instead of designing a sophisticated distillation loss to better transfer knowledge from one single teacher. Since we are studying a new setting where multiple teachers are available, which is missing in previous literature, we mainly focus on the *absolute improvements* – the performance of our distilled student models compared with the original student models and with the performance upper-bound of the teacher models.

4.1. Searching for Near-Optimal Teacher Orders

As we have discussed in Section 3.2, finding the optimal order of teachers for MTPD takes factorial time complexity. To acquire a near-optimal teacher order, we propose

Table 3. **Homogeneous distillation of COCO detectors,** where students with ResNet-50 backbones are distilled from teachers with ResNet-50 backbones. We report the detection ('Box') and segmentation ('Mask') APs, and we compare our student produced by MTPD with the off-the-shelf ('OTS') student and the student trained longer. MTPD significantly improves the detection AP over the 'OTS' student by $\mathbf{3.4}\%$ for RetinaNet and $\mathbf{3.2}\%$ for Mask R-CNN, and outperforms the baselines.

ID	Model	Method	Box						Mask					
			AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
1	RetinaNet (Student I)	OTS	36.5	55.4	39.1	20.4	40.3	48.1	-	-	-	-	-	-
2		Longer 3× training schedule	39.5	58.8	42.2	**23.8**	43.2	50.3	-	-	-	-	-	-
3		Directly distilled by Teacher IV	39.5	58.6	41.9	21.0	42.8	54.0	-	-	-	-	-	-
4		**MTPD: Teachers III→IV**	**39.9**	**59.2**	**42.7**	21.7	**43.3**	**54.1**	-	-	-	-	-	-
5	Mask R-CNN (Student II)	OTS	38.2	58.8	41.4	21.9	40.9	49.5	34.7	55.7	37.2	18.3	37.4	47.2
6		Longer 3× training schedule	40.9	61.3	44.8	**24.4**	44.6	52.3	37.1	58.3	**39.9**	18.4	39.8	51.9
7		Directly distilled by Teacher IV	41.0	61.6	45.0	23.5	44.5	54.0	37.0	58.5	39.8	17.5	39.9	51.3
8		Distilled by ensemble V+IV	39.8	60.3	43.4	22.1	43.3	52.9	35.9	57.1	38.1	18.3	39.0	49.8
9		**MTPD: Teachers V→IV**	**41.4**	**61.9**	**45.1**	23.3	**45.0**	**55.4**	**37.3**	**58.8**	39.8	**19.4**	**40.4**	**52.1**

the heuristic algorithm Backward Greedy Selection (BGS, pseudo-code shown in Algorithm 1). In this section, we validate that BGS is near-optimal. To achieve this comprehensive comparison, we distill Student I with *all orders* of teachers from the pool of Teachers I-IV. We use a reduced training budget: For each teacher, we only train the student for 3 epochs on MS COCO. We use the linear learning rate schedule, which has been shown comparably effective in a limited budget setting by Li et al. (2020b).

We first measure the adaptation costs among the student and teacher models. A visualization of the cost graph is shown in Figure 3. Following BGS, we can construct a sequence of teachers. We compare the teacher orders given by BGS against *all other* orders, via the distilled students' performance. As shown in Table 2, teacher orders suggested by BGS are consistently near-optimal in this setting. In the following sections, we use the order provided by BGS, without brute-force iterating over all possible orders. One might argue that the greedy path selection of BGS, as shown in Figure 3, is inferior to a global optimization algorithm. However, we find that BGS consistently outperforms other heuristics including global optimization algorithms (see details in Appendix A). In fact, the later teachers impact the student performance more profoundly, so we need to greedily select teachers from the sequence tail.

4.2. Distillation with Homogeneous Teachers

We start by distilling RetinaNet and Mask R-CNN with a ResNet-50 backbone (Students I & II). Here we consider *homogeneous* teachers where the numbers of channels and the spatial resolutions of feature maps are *consistent* between the student and teacher. For the RetinaNet student, we still consider the pool of Teachers I-IV, the same as Section 4.1. For the Mask R-CNN student, we should no longer use Teacher I (the student itself) or Teacher II (the single-stage teacher does not outperform the student by a large margin). To compensate for that, we include Teacher V, which can be considered as a hybrid model of the DetectoRS backbone/neck and Mask R-CNN head. Thus, the teacher pool for Mask R-CNN includes Teachers III-V. To control the total training time, we limit the number of teachers to be 2. We initialize from an off-the-shelf ('OTS') student, and sequentially distill it using 2 teachers, each with a $1\times$ training schedule. In total, the student is distilled for 24 epochs, and the training time is equivalent to a $2\times$ training schedule. In addition to the OTS student, we also compare with three other baselines: 1) the student trained with a longer $3\times$ training schedule, which is commonly supported in object detection libraries and stronger than $1\times, 2\times$ training; 2) the student *directly distilled* by the final target teacher, using a $2\times$ training schedule; and 3) the student distilled by the *ensemble* of teachers' feature maps. Detector details are listed in Table 1.

Following Section 4.1, we use BGS to determine the sequence of teachers to use for each student. For the RetinaNet student, BGS suggests teacher sequence III→IV. For the Mask R-CNN student, BGS suggests teacher sequence V→IV. Table 3 shows the distillation results on COCO. Additional results, analysis, and ablation studies of Mask R-CNN distillation are in Appendix B.

Overall performance: Our distilled student models (rows 4&9) significantly improves over the 'OTS' students (rows 1&5). The box AP of RetinaNet is improved from 36.5% to 39.9% ($+3.4\%$). The box AP of Mask R-CNN is improved from 38.2% to 41.4% ($+3.2\%$) and the mask AP of Mask R-CNN is improved from 34.7% to 37.3% ($+2.6\%$). After progressive distillation, our resulting Mask R-CNN detector has *comparable performance with HTC teacher, but much less runtime* (51ms vs. 181ms).

Comparison with baselines: First, the performance gain is not merely from a longer training schedule. Our distilled student models (rows 4&9) consistently outperform original students trained with a $3\times$ schedule (rows 2&6). Second,

Table 4. **Heterogeneous distillation of COCO detectors**, where students with smaller backbones (ResNet-18 vs. ResNet-50) or input resolutions (333×200 vs. 1333×800) are distilled with heterogeneous teachers, requiring an additional feature adaptor (Section 3.1). We report the detection ('Box') and segmentation ('Mask') APs, and compare our distilled student with its teachers (see Table 1), the off-the-shelf ('OTS') student, and the student distilled from an ensemble of the teachers. MTPD significantly improves the 'OTS' students by over $\mathbf{3}\%$ AP.

ID	Model	Backbone	Resolution	AP	
				Box	Mask
1	Student III, OTS	R18	$1\times$	33.3	30.5
2	Student III, Teacher Ensemble	R18	$1\times$	36.0	32.1
3	Student III, **MTPD**	R18	$1\times$	**37.0**	**33.7**
4	Student IV, OTS	R50	$0.25\times$	25.8	23.0
5	Student IV, **MTPD**	R50	$0.25\times$	**31.5**	**28.2**

progressive distillation using a curriculum of teachers (rows 4&9) is better than direct distillation from a strong teacher (rows 3&7), even if the total training time is the same. Additionally, we find that using a sequence of teachers (row 9), instead of their ensemble (row 8), is more effective. This shows that integrating different types of knowledge from multiple teachers is non-trivial, and our progressive approach is better than simultaneously distilling from multiple teachers. Notably, our detection performance for large objects receives the most gain (about 6% AP_L improvement for both models). We emphasize AP_L because in an efficiency-centric real-world application (e.g., autonomous driving, robot navigation), detecting nearby larger objects is more crucial than others. From a realistic perspective, better AP_L shows better applicability of our approach.

4.3. Distillation with Heterogeneous Teachers

To validate that MTPD is general, we now consider a more challenging heterogeneous scenario, where students and teachers have different backbones or input resolutions. Specifically, Student III, a ResNet-18 Mask R-CNN, is distilled with ResNet-50 teachers; Student IV, a model with reduced input resolution, is distilled with teachers trained with larger input resolutions. The results are summarized in Table 4, and additional results are included in Appendix C.

Heterogeneous backbones: Student III has a ResNet-18 backbone and about half runtime as its ResNet-50 counterpart (Teacher I). We find that the proper distillation scheme for Student III is to use the sequence of (rather than ensembling) Teachers I→V→IV, which significantly improves Student III over the 'OTS' model. The box AP of Student III is improved from 33.3% to 37.0% ($+3.7\%$); and especially for large objects, AP_L is improved from 43.6% to 50.0% ($+6.4\%$).

Heterogeneous input resolutions: Although inputs with varying resolutions can be fed into most object detectors without changing the architecture, the performance often degenerates when there is a resolution mismatch between training and evaluation (Tan et al., 2020; Li et al., 2020a). If ultimately we want to apply a detector to low-resolution inputs for fast inference, it is better to use low-resolution inputs during training. On the other hand, we conjecture that teachers with high-resolution inputs may provide finer details that can assist the student. With MTPD, we investigate the improvement of a low-resolution student distilled by a sequence of teachers with high-resolution inputs. We denote the standard input resolution 1333×800 as $1\times$, and a reduced resolution 333×200 as $0.25\times$. We distill Student IV (with $0.25\times$ resolution) by a sequence of Teacher I variants ($0.5\times \rightarrow 0.75\times \rightarrow 1\times$). From Table 4, we can see substantial improvement brought by MTPD: The box AP is improved from 25.8% to 31.5% ($+5.7\%$), and the mask AP is improved from 23.0% to 28.2% ($+5.2\%$).

4.4. Generalizability to State-of-the-Art Distillation Mechanisms

Our meta-level strategy of using a sequence of teachers to progressively distill a student is independent of the choice of distillation mechanism for each teacher. We have shown MTPD can boost the simple feature-matching distillation, and in this section, we will combine MTPD with **state-of-the-art distillation mechanisms for object detection** to further improve student accuracy.

Distillation protocol: We evaluate MTPD with three most recent methods on detector distillation: CWD (Shu et al., 2021), FGD (Yang et al., 2022a), and MGD (Yang et al., 2022b). In Appendix E, we show that classification-oriented distillation is inferior to methods delicately designed for detectors. For a fair comparison, we use the *same teacher-student pairs* as them: RetinaNet/ResNet-50 and RetinaNet/ResNeXt-101 (Lin et al., 2017b) are the single-stage student and final teacher. RepPoints/ResNet-50 and RepPoints/ResNeXt-101 (Yang et al., 2019b) are the two-stage, anchor-free student and final teacher. Mask R-CNN/ResNet-50 and Cascade Mask R-CNN/ResNeXt-101-DCN (He et al., 2017) are the two-stage, anchor-based student and final teacher. Between them, we insert one medium-capacity teacher to progressively distill the student: RetinaNet/ResNet-101 for the first pair, RepPoints/ResNet-101 for the second, and Cascade Mask R-CNN/ResNet50-DCN for the third. Also for fairness, we *keep the total training epochs the same*. We set "$1\times$" training schedule for each teacher, so that the total training time is equivalent to "$2\times$," the same as previous work.

Figure 4 shows that MTPD consistently improves students' final accuracy. For example, the performance of FGD-distilled RetinaNet/ResNet-50 improves from 40.7% to 41.5% AP ($+0.8\%$), and this gain is larger than mechanism

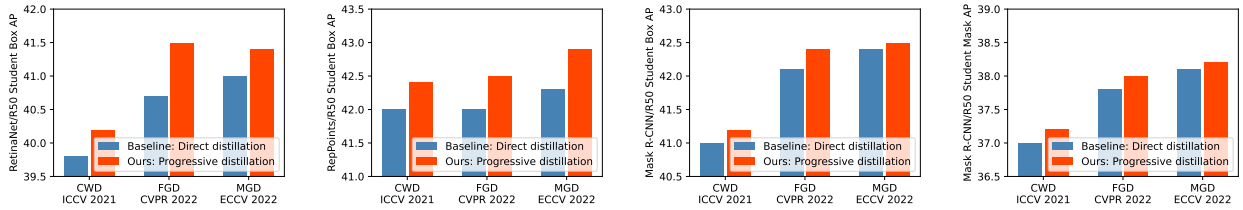


Figure 4. **MTPD consistently benefits state-of-the-art distillation mechanisms.** Using an intermediate RetinaNet/ResNet-101 teacher between RetinaNet/ResNet-50 student and RetinaNet/ResNeXt-101 teacher (**a**), RepPoints/ResNet-101 teacher between RepPoints/ResNet-50 student and RepPoints/ResNeXt-101 teacher (**b**), or Cascade Mask R-CNN/ResNet50-DCN teacher between Mask R-CNN/ResNet-50 student and Cascade Mask R-CNN/ResNeXt-101-DCN teacher (**c** for Box AP and **d** for Mask AP), we improve the direct distillation baselines by 0.2% to 0.8% AP, *without increasing training time*.

Table 5. **Distillation from Transformer-based teachers (Liu et al., 2021) to convolution-based students.** Due to the architectural difference and capacity gap, directly distilling from a stronger teacher with Swin-S backbone does not yield better students than convolution-based teachers in Figure 4. An intermediate Swin-T teacher and *progressive distillation* solve this issue without increasing training time. Compared to off-the-shelf models, our RetinaNet and Mask R-CNN students improve by $\mathbf{5.5}\%$ AP and $\mathbf{4.3}\%$ box AP, respectively.

ID	Model	Distillation	AP	
			Box	Mask
1	RetinaNet	Direct RetinaNet/Swin-S	41.0	-
2	(Student I)	**MTPD: RetinaNet/Swin-T→S**	**42.0**	-
3	Mask R-CNN	Direct MRCNN/Swin-S	42.0	37.7
4	(Student II)	**MTPD: MRCNN/Swin-T→S**	**42.5**	**38.4**

advance from FGD to MGD ($+0.3\%$). We bring performance gains to state-of-the-art detection distillation almost *for free*.

Next, we investigate how to further maximize the student performance. Due to better computation efficiency, a convolution-based (rather than Transformer-based) student is preferred. Meanwhile, Swin Transformer (Liu et al., 2021) can act as an even stronger teacher than the convolution-based teachers used in previous work. However, compared with convolution-based teachers, direct distillation from such a teacher cannot improve the student performance, even if we use the state-of-the-art method MGD. For example, RetinaNet/Swin-Small (47.1% AP) is much stronger than RetinaNet/ResNeXt-101(41.6% AP), but direct distillation from both yields the same student performance (41.0% AP). To bridge the architectural difference and capacity gap between the ResNet-50 student and Swin-Small teacher, we can utilize an intermediate Swin-Tiny teacher. As shown in Table 5, MTPD brings the best students: the performance of ResNet-50 based RetinaNet increases to 42.0% AP, and Mask R-CNN increases to 42.5% AP. We also successfully distill a Transformer-based student from convolution-based teachers in Appendix F.

4.5. Unpacking the Performance Gain: Generalization or Optimization?

We have shown that our distilled students significantly improve the accuracy on the *validation* data over off-the-shelf students. As further demonstrated in Figure 5a, the validation accuracy of the distilled student gradually increases during distillation, and achieves a higher value compared with the student trained without teachers. A natural question then arises – why is distillation helping? There are two possible hypotheses: (1) *improved optimization:* distillation facilitates the optimization procedure, leading to a better local minimum; and (2) *improved generalization:* distillation helps the student generalize to unseen data.

Improved optimization is typically manifested through a better model, a lower training loss, and a higher validation accuracy, which is exactly the case for Mask R-CNN, HTC, and DetectoRS. Consequently, one might think that distillation works in the same way. However, our investigation suggests the opposite – MTPD increases both the validation accuracy and the training loss, and therefore effectively reduces the generalization gap. In Figure 5, we compare the original RetinaNet model and the distilled student, which have the same architecture, the same latency, and are trained on the same data, but with different supervision (only ground-truth labels vs. additional knowledge distillation). To eliminate the influence of learning rate changes, we train the original student with a $3\times$ schedule and restart the learning rate at the same time with the distilled student. Interestingly, although distillation can improve the student's validation performance, the *training* detection loss of the distilled student is higher than the original student. This suggests that distillation does *not* help the optimization process to find a local minimum with a lower training loss, but rather strengthens the generalizability of the student model.

To further support this observation, we also visualize the local loss landscape (Li et al., 2018). The distilled student has a flatter loss landscape (Figure 5d) compared with

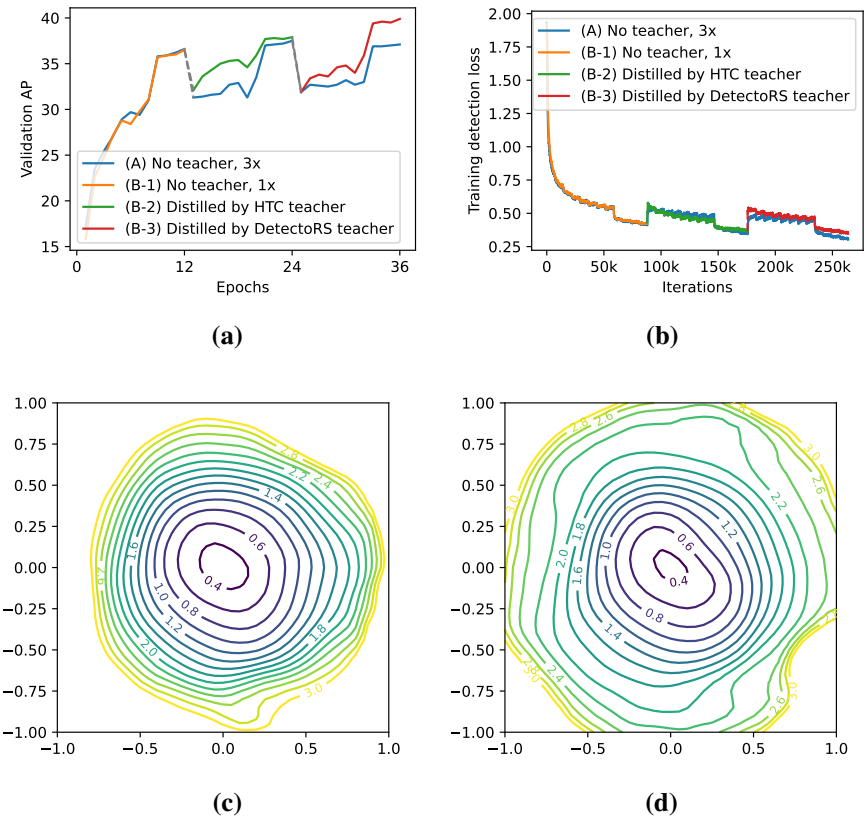


Figure 5. **Comparison of student models trained with and without teacher distillation.** We train a ResNet-50 backboned RetinaNet (Student I) with (A) a prolonged 3× training schedule (curves in blue), or (B) MTPD from HTC (Teacher III) and then DetectoRS (Teacher IV) (curves in orange-green-red). We compare the validation AP (**a**) and the training detection loss L_{detect} (**b**) of the two students during the training process. Despite its worse training loss, the distilled student can generalize better on the validation set. We also compare the loss landscapes (Li et al., 2018) of the original student (**c**) and the distilled student (**d**). Distillation can guide the student to converge to a flatter local minimum. These observations suggest that distillation *helps generalization rather than optimization*.

the original one (Figure 5c). As widely believed in the machine learning literature, flat minima lead to better generalization (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017). The observation shown in Figure 5 is illustrated for RetinaNet, but we also have a similar observation in other students. As a conclusion, knowledge distillation, which enforces the student to mimic the teachers' features, can be considered as an implicit regularization, and helps the student combat overfitting and achieve better generalization.

5. Conclusion

We present a simple yet effective approach to knowledge distillation, which progressively transfers the knowledge of a sequence of teachers to learn a lightweight object detector. Our approach automatically arranges multiple teachers into a curriculum, thus effectively mitigating the capacity gap between the teacher and student. We successfully distill knowledge from Transformer-based teachers to convolution-based students, and achieve state-of-the-art performance on the challenging COCO dataset. We also find that distillation improves generalization rather than optimization.

Limitation and future work: This work has mainly focused on empirical results and analysis. Due to the complexity of the detection task and models, we have not included theoretical understanding of the representation-based adaptation cost and better generalization resulted by distillation, but they will be our future direction. As a general approach to object detection, this work shares similar concerns with other detection techniques, such as potential misuse in enhancing surveillance systems, infringing upon privacy rights, or contributing to biased outcomes.

Acknowledgement: This work was supported in part by NSF Grant 2106825, NIFA Award 2020-67021-32799, the Jump ARCHES endowment, the NCSA Fellows program, the IBM-Illinois Discovery Accelerator Institute, the Illinois-Insper Partnership, and the Amazon Research Award. This work used NVIDIA GPUs at NCSA Delta through allocation CIS220014 from the ACCESS program.

References

Ahn, S., Hu, S. X., Damianou, A., Lawrence, N. D., and Dai, Z. Variational information distillation for knowledge transfer. In *CVPR*, 2019.

Ashraf, K., Wu, B., Iandola, F. N., Moskewicz, M. W., and Keutzer, K. Shallow networks for high-accuracy road object-detection. *arXiv preprint arXiv:1606.01561*, 2016.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *ICML*, 2009.

Buciluǎ, C., Caruana, R., and Niculescu-Mizil, A. Model compression. In *KDD*, 2006.

Cai, Z. and Vasconcelos, N. Cascade R-CNN: Delving into high quality object detection. In *CVPR*, 2018.

Chang, M.-F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., et al. Argoverse: 3D tracking and forecasting with rich maps. In *CVPR*, 2019.

Chen, G., Choi, W., Yu, X., Han, T., and Chandraker, M. Learning efficient object detection models with knowledge distillation. In *NeurIPS*, 2017.

Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., Loy, C. C., and Lin, D. Hybrid task cascade for instance segmentation. In *CVPR*, 2019a.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C. C., and Lin, D. MMDetection: Open MMLab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019b.

Cho, J. H. and Hariharan, B. On the efficacy of knowledge distillation. In *ICCV*, 2019.

Dai, X., Jiang, Z., Wu, Z., Bao, Y., Wang, Z., Liu, S., and Zhou, E. General instance distillation for object detection. In *CVPR*, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

Fukuda, T., Suzuki, M., Kurata, G., Thomas, S., Cui, J., and Ramabhadran, B. Efficient knowledge distillation from an ensemble of teachers. In *Interspeech*, 2017.

Guo, J., Han, K., Wang, Y., Wu, H., Chen, X., Xu, C., and Xu, C. Distilling object detectors via decoupled features. In *CVPR*, 2021a.

Guo, Q., Wang, X., Wu, Y., Yu, Z., Liang, D., Hu, X., and Luo, P. Online knowledge distillation via collaborative learning. In *CVPR*, 2020.

Guo, S., Alvarez, J. M., and Salzmann, M. Distilling image classifiers in object detectors. In *NeurIPS*, 2021b.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016.

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

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. In *NeurIPS Workshop*, 2014.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural computation*, 1997.

Kang, Z., Zhang, P., Zhang, X., Sun, J., and Zheng, N. Instance-conditional knowledge distillation for object detection. In *NeurIPS*, 2021.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. In *ICLR*, 2017.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *ICML*, 2019.

Lan, X., Zhu, X., and Gong, S. Knowledge distillation by on-the-fly native ensemble. In *NeurIPS*, 2018.

Li, C., Wang, Z., and Qi, H. Online knowledge distillation by temporal-spatial boosting. In *WACV*, 2022.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. In *NeurIPS*, 2018.

Li, M., Wang, Y.-X., and Ramanan, D. Towards streaming perception. In *ECCV*, 2020a.

Li, M., Yumer, E., and Ramanan, D. Budgeted training: Rethinking deep neural network training under resource constraints. In *ICLR*, 2020b.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection. In *CVPR*, 2017a.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *ICCV*, 2017b.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A ConvNet for the 2020s. In *CVPR*, 2022.

Mirzadeh, S. I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., and Ghasemzadeh, H. Improved knowledge distillation via teacher assistant. In *AAAI*, 2020.

Qiao, S., Chen, L.-C., and Yuille, A. DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *CVPR*, 2021.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *NeurIPS*, 2017.

Redmon, J. and Farhadi, A. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

Ren, S., He, K., Girshick, R., and Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2014.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. FitNets: Hints for thin deep nets. In *ICLR*, 2015.

Sau, B. B. and Balasubramanian, V. N. Deep model compression: Distilling knowledge from noisy teachers. *arXiv preprint arXiv:1610.09650*, 2016.

Shu, C., Liu, Y., Gao, J., Yan, Z., and Shen, C. Channel-wise knowledge distillation for dense prediction. In *ICCV*, 2021.

Tan, M., Pang, R., and Le, Q. V. EfficientDet: Scalable and efficient object detection. In *CVPR*, 2020.

Tian, Z., Shen, C., Chen, H., and He, T. FCOS: Fully convolutional one-stage object detection. In *ICCV*, 2019.

Vongkulbhisal, J., Vinayavekhin, P., and Visentini-Scarzanella, M. Unifying heterogeneous classifiers with distillation. In *CVPR*, 2019.

Wang, L., Hu, L., Gu, J., Wu, Y., Hu, Z., He, K., and Hopcroft, J. Towards understanding learning representations: To what extent do different neural networks learn the same representation. In *NeurIPS*, 2018.

Wang, T., Yuan, L., Zhang, X., and Feng, J. Distilling object detectors with fine-grained feature imitation. In *CVPR*, 2019.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.

Yang, C., Xie, L., Su, C., and Yuille, A. L. Snapshot distillation: Teacher-student optimization in one generation. In *CVPR*, 2019a.

Yang, J., Martinez, B., Bulat, A., and Tzimiropoulos, G. Knowledge distillation via softmax regression representation learning. In *ICLR*, 2021.

Yang, Z., Liu, S., Hu, H., Wang, L., and Lin, S. RepPoints: Point set representation for object detection. In *ICCV*, 2019b.

Yang, Z., Li, Z., Jiang, X., Gong, Y., Yuan, Z., Zhao, D., and Yuan, C. Focal and global knowledge distillation for detectors. In *CVPR*, 2022a.

Yang, Z., Li, Z., Shao, M., Shi, D., Yuan, Z., and Yuan, C. Masked generative distillation. In *ECCV*, 2022b.

Yao, A. and Sun, D. Knowledge transfer via dense cross-layer mutual-distillation. In *ECCV*, 2020.

You, S., Xu, C., Xu, C., and Tao, D. Learning from multiple teacher networks. In *KDD*, 2017.

Zagoruyko, S. and Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.

Zhang, L. and Ma, K. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *ICLR*, 2021.