This ICCV workshop paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

Self-training and multi-task learning for limited data: evaluation study on object detection

Hoàng-Ân Lê Minh-Tan Pham
IRISA, Université Bretagne Sud, UMR 6074, 56000 Vannes, France
{hoang-an.le,minh-tan.pham}@irisa.fr

Abstract

Self-training allows a network to learn from the predictions of a more complicated model, thus often requires well-trained teacher models and mixture of teacher-student data while multi-task learning jointly optimizes different targets to learn salient interrelationship and requires multi-task annotations for each training example. These frameworks, despite being particularly data demanding have potentials for data exploitation if such assumptions can be relaxed. In this paper, we compare self-training object detection under the deficiency of teacher training data where students are trained on unseen examples by the teacher, and multi-task learning with partially annotated data, i.e. *single-task annotation per training example. Both scenarios have their own limitation but potentially helpful with limited annotated data. Experimental results show the improvement of performance when using a weak teacher with unseen data for training a multi-task student. Despite the limited setup we believe the experimental results show the potential of multi-task knowledge distillation and self-training, which could be beneficial for future study. Source code and data splits are at* https://lhoangan.github.io/multas

1. Introduction

Besides the impressive capability in solving complicated problems, deep learning is well-known for being computationally expensive and highly data-demanding. The former, due to the complex architectural model, limits the deployment on low-capacity edge devices while the latter constrains its generalizability and robustness.

The data scarcity problem is mostly due to expensive annotating efforts as raw unlabeled data are practically everywhere [4, 15]. Thus amelioration is often studied under a different training paradigm such as self-, weakly- or unsupervised learning. Self-training is a weakly-supervised method based on knowledge distillation or teacher-student models [9]. The idea is to train a network, called *student* using the combination of the same labeled data, with which a (usually) more cumbersome *teacher* network is trained, and new unlabeled data whose pseudo-labels are provided by the teacher's predictions.

Self-training typically assumes that the same training data of the teacher is part of the student training set [29, 37] and a sufficiently-trained teacher model can perform generally well on *unseen* data, making fewer errors than correct predictions [29]. However, the teachers' training data are not always available but only the teacher's pre-trained weights due to copyright issues or confidentiality concerns [23, 26]. To put an emphasis on the limit of training data, in this paper, we consider the deficiency scenario where the teacher is trained on *small* amount of data and that the student training data have not been seen by the teacher, *i.e*. the teacher-student training sets are disjoint.

On the other hand, multi-task learning assumes that simultaneously optimizing multiple related targets for the same input helps the network to extract common features, learn salient interrelationships, and improve performance. Such assumption is usually perceived as demanding and put extra burdens to data preparation as it multiply the required annotations by the number of tasks and thus multiply the efforts to maintain training set with consistent annotations for all tasks. In the context of limited data, we consider the multitask partially annotated scenario [17], where each task annotation sets are disjoint, *i.e*. each image is annotated for a single task and there are no images containing both-task annotations. This setup is data efficient and would be an alternative method to ameliorate data scarcity if a network could exploit task interrelationships to improve performance without requiring all-task annotations per data example.

The contributions of the paper are as follows. (1) We extensively study the performance of standard object detection with self-training under limited annotated data. The situation is studied together with feature-imitation knowledge distillation as a way to mitigate the lack thereof. The student's performances with respect to deficient teacher's training status are observed by gradually reducing the training size. (2) We briefly examine cross-task scenarios between

object detection and semantic segmentation for knowledge distillation where the teacher and student are trained for different but related tasks. (3) We evaluate the multi-task framework under partially annotated scenario for data exploitation and show the usefulness of combining both framework to improve the network performance. Although the paper does not provide novel architectural contribution or improvement over state-of-the-art we believe the provided experimental results are important for insights and understanding on possible approach for data scarcity.

2. Related Work

2.1. Knowledge distillation

The idea of knowledge distillation (KD) is based on the separability of the training and inference process. Consequently, a network can learn from the outputs of a larger model and get improved without complicated modification in deployment. Since the pioneering publication of Hinton *et al.* [9], it has attracted various studies and more understanding was obtained: "a good teacher is patient and consistent" [1], an intermediate-sized teaching assistance network helps better when the complexity gap is large between the teacher and student [25], and how knowledge distillation can be applied in a self-supervised context with contrastive loss [31], *etc*.

On the one hand, knowledge distillation has been studied to accommodate task-oriented information such as object localization [35], background-region knowledge [7], and teacher-student agreement [34] for object detection, object detection in remote sensing data [13], or multi-task depth-semantic segmentation knowledge distillation [16], *etc*. On the other hand, it is at the core of the self-training paradigm which seeks to expand networks' learning capacity using unlabeled data. The idea is that a (student) network can be improved by training with the predictions of a pre-trained larger (teacher) network on unlabeled data points. For object detection, Radosavovic *et al.* proposes a data-distillation model [29] which feeds various transformations of an unlabeled input image to a well-trained teacher and uses the prediction ensemble to train a student network. Similarly, Zoph *et al.* [37] studies the interaction between training methods and data augmentation and compares self-training against pre-training using unlabeled ImageNet images. Diverging from the previous work, we explore the student network performance with respect to deficient teacher training data and examine the behavior of feature-imitation knowledge distillation in combination with self-training under such condition.

2.2. Multi-task learning

The main target of multi-task learning is to infer simultaneously various aspects of a single input image. The general idea is that the features for predicting each aspect or *task* of the same image should be overlapping, and by optimizing them in the same model, using techniques such as attention mechanisms [21] or gating strategies [2], the network can pickup the interrelationships that benefits and complements one another [18, 24]. Different task combinations would require different approaches to bring out the shared information the tasks [3], inspiring different cross-task studies [24, 32].

Attempts have been made to relax the requirement for joint annotations of all tasks as they put extra burdens to data preparation. Semi-supervised learning methods such as [5, 10] lessen the data dependency and allow learning from unlabelled data, yet all-task annotations per training sample are still required.

Multi-task partially learning, where each input image is annotated with *only* one of the tasks has been studied by Li *et al*. [17] for spatially dense tasks such as semantic segmentation and depth prediction. The dense annotations of one task are projected to a joint task-space and provide supervised signals for training the other task. This approach, however, is not immediately applicable for sparse-annotation task such as object detection in this paper.

3. Implementation

We follow the setup and architectures described by Zhang *et al.* [34] for knowledge distillation and perform experiments on the ResNet family. We employ the ResNet50 (in place of the ResNet34) backbone with PAFPN [22] neck for teacher and ResNet18 backbone with FPN neck for student. The first max-pooling layer of ResNet is removed as in ScratchDet [36] and the context enhancement module is added as in ThunderNet [28]. The parameter ratio between the teacher and the student is 1.61.

For object detection, the Retina-style [20] prediction head is employed following [34] with number of convolution blocks reduced from 4 to 2. The intermediate feature channels are set to 256. The detection head includes 2 identical sub-networks (except for the last layer) for localization and classification.

The multi-scaled FPN features are aggregated for semantic segmentation using the module by Kirillov *et al*. [12]. The features at each scale are passed through a sequence of convolution and 2x-upsampling modules until being one-fourth of the input size. The subsequent feature maps are element-wise added and upsampled to the input size. The intermediate features are all 128-channel. The regular cross-entropy with softmax loss is employed.

The supervised training of the teacher and student model is performed by the regular object detection losses, including the balanced L1 loss for bounding-box localization [27] and the quality focal loss [19] for classification. The teacher is also trained with mutual guide matching [33], instead of

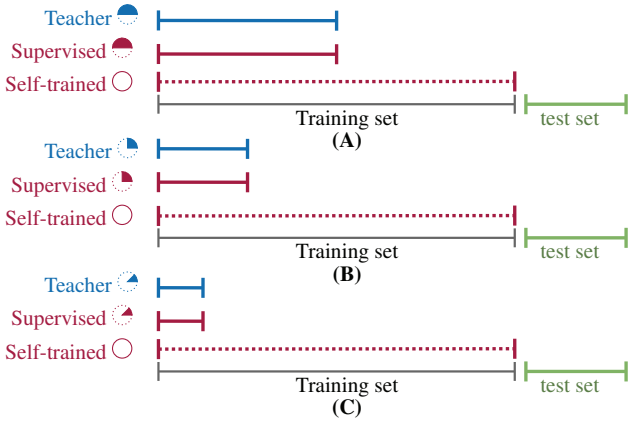

Figure 1. Teacher-student data splits with overlapping data. Continuous lines indicate availability of annotated ground truths while the dotted lines indicate the lack thereof.

the regular intersection-over-union thresholds for a boost of performance. The student self-training is performed by *soft* supervised loss with the target given by a teacher network: bounding-box localization uses the same balanced L1 loss while classification uses focal loss [20] multiplied by $5e3$ to be in the similar range with localization[1]

For feature-imitation knowledge distillation, the feature maps after the neck layers are used (*cf.* [34]). The KD losses to be studied include (1) the trivial Mean Square Error (MSE) between the teacher's features and the projected maps of the student's, (2) the PDF-Distil [34] measuring teacher-student disagreement, and (3) DeFeat [7] distilling foreground and background regions separately.

For optimization, each task branch is trained alternately every iteration: (1) After a mini-batch of input images with single-task annotations is passed through the network, the loss function(s) of the corresponding task is computed and back-propagating gradients through the task branch and the shared encoder; (2) a mini-batch of the other-task images is passed through immediately in the next iteration; (3) only after mini-batches from both tasks have been fed in and gradients accumulated are the network parameters updated. Each experiment is trained for 70 epochs with early stopping using the validation set.

4. Experiments

4.1. Datasets

Unless stated otherwise, we follow the conventional split of the Pascal VOC [6] dataset that keeps out the test split (4,952 images) of the VOC07 set for validation, and uses the rest, training and validation split of both VOC07 and VOC12 set (16,551 images) for training [34]. We also show the results on the aerial vehicle detection VEDAI dataset [30] for the first experiment. The VOC evaluation metric for object detection is employed, including AP50, AP75 and mAP [6]. For VEDAI, we report results on AP10 and AP25 due to the very small-sized objects, following the suggestions of [14]. The IOU score [11] is reported for the semantic segmentation task. A detailed training split for a specific experiment will be described at each experiment.

[1]The weight is hard-coded and is the same for all self-training experiments based on the initial loss values of localization and classification (*e.g.* 1.25 and 0.0002-0.0003, respectively), without being exhaustively tested.

		VOC			VEDAI	
Model		AP50	AP75	mAP	AP10	AP25
Teacher	◓	80.18	63.58	58.24	81.04	80.91
Supervised	◓	77.14	54.23	51.08	74.85	74.85
+MSE	◓	78.42	57.18	53.40	73.23	73.19
+PDF	◓	**78.97**	57.67	54.18	75.66	75.63
+DeFeat	◓	78.86	**58.15**	**54.30**	**75.90**	**75.90**
Self-trained	○	79.95	60.06	55.64	76.71	76.69
+MSE	○	79.51	60.61	55.53	75.19	75.19
+PDF	○	**80.10**	**61.22**	**56.47**	**78.58**	**78.57**
+DeFeat	○	79.57	60.78	55.84	74.32	74.32

Table 1. Comparing supervised students trained on a half training set (indicated by ◓) with self-trained students using only teacher's predictions on the full training set (indicated by ○). The teacher's performance is added for reference.

4.2. Self-training for data exploitation

In the first experiment, we confirm the usefulness of a teacher with unseen data points, and consequently the idea of self-training. In particular, we withhold the annotations of a half ◡ of the training set and train a supervised *teacher* network using the other half ◓. For the student, two cases are compared, being trained (1) with the half annotated data ◓ as the teacher's using the available annotations (hard labels) and (2) with the entire training set ○ using only teacher's prediction (soft labels) (Fig. 1A).

The results in Table 1 shows that using soft label of the full dataset with possibly noisy soft labels is more beneficial than using less data with ground truth targets. PDF distillation shows superiority for self-trained methods while DeFeat distillation is better for supervised methods.

The following subsections show ablation studies with the reduced sizes of the available training set, weakening teachers performance and subsequently the self-trained students.

Reducing supervising data

Following up from the previous experiment, we study the student's behavior with different ratios of seen and unseen data by the teacher. In particular, we reduce the available annotations from a half ◓ to a quarter ◔ and an eighth

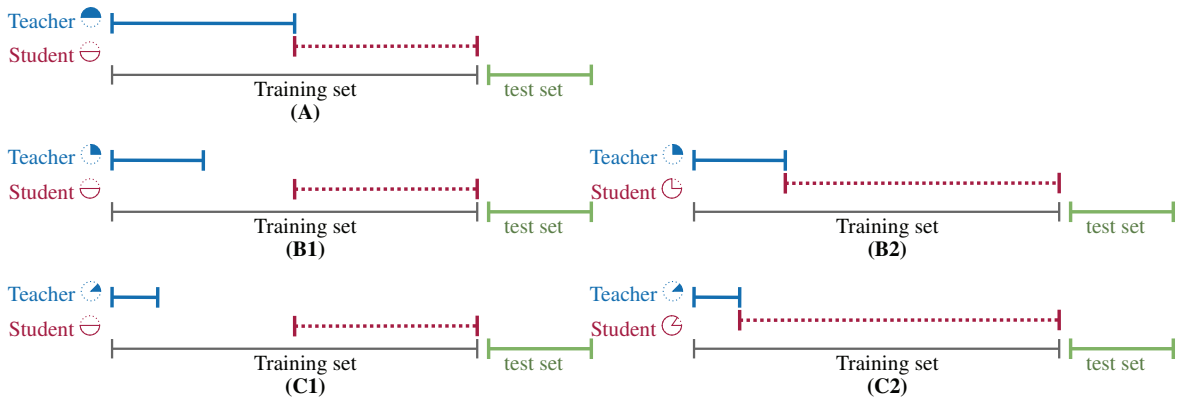

Figure 2. Various scenarios of teacher-student disjoint subsets

Teacher		58.24	53.11	46.42
Supervised	◓	51.08	47.16	39.82
+MSE	◓	53.40	48.88	41.83
+PDF	◓	54.18	**50.13**	**42.81**
+DeFeat	◓	**54.30**	49.54	42.19
Self-trained	○	55.64	51.82	46.59
+MSE	○	55.53	51.41	46.21
+PDF	○	**56.47**	**51.93**	**46.59**
+DeFeat	○	55.84	51.47	46.02

Table 2. Comparing mAP of supervised students trained on annotated data of various size indicated by the column headers (◓ a half, ◔ a quarter, and ◭ an eighth of the original VOC07-12 training set) with those self-trained using only the teachers' predictions on the full training set ○. The teachers, whose performances are shown for reference, are trained using the same partial subsets indicated in the column headers.

of the overall training set (Fig. 1B,C), thus increasing the uncertainties of the teacher's predictions used as soft labels for training the student. The results are show in Table 2.

Reducing the amount of annotations results in weaker teachers and, consequently, decreases the general performance. While the supervised students suffer from the decline of available annotations, the students self-trained with the teacher's soft labels can maintain its performance on par with or even surpass its teacher which has 1.6x more parameters and was trained in a fully supervised way.

Self-training in the absence of teacher training data

In this section, we study the student's performance in the complete absence of teacher's training data. While the teacher training split are kept the same as in the previous experiments, the student is trained on disjoint subsets as shown in Fig. 2. In Table 3, we compare the students self-trained on the same second half subset ◒ using teachers trained with various disjoint subsets (Fig.2A, B1, C1).

Teacher		58.24	53.11	46.42
Supervised	◓	51.08	47.16	39.82
Self-trained	◒	52.86	49.63	45.09
+MSE	◒	53.83	49.72	44.85
+PDF	◒	**54.26**	**50.64**	**45.63**
+DeFeat	◒	53.58	49.90	44.54

Table 3. mAP of students self-trained on unseen data by various teachers as in the previous experiments (column headers). The students are self-trained using the same second training half ◒.

It is trivial that the weaker the teachers, the less performance the students and the general performance is lower than that of the previous experiment as the teachers could not generalize well with unseen data points. The feature distillations still show the benefit, especially PDF, bringing the performance closer to the teacher's level.

Increasing student unlabeled data

In this experiment, we increase the student unlabeled training data using the complementary subset with the teachers' as shown in Fig. 2B2, C2. For the sake of completion, we also include the performance of student trained on the 3-quarter subset ◕ while the teacher is trained on the first 1-eighth ◭.

From Table 4, there is a gradual diminution in the student performance along with the shrinkage of the teacher's training size despite the expansion of the students' training size, showing the importance of targets' accuracy and, subsequently, the teacher generalizability. Interestingly, a small reducing of the teacher training size, from ◔ to ◭ for the same student size ◕, results in a large gap in the student performance, suggesting a toleration threshold for data deficiency of the teachers. The use of knowledge distillation also helps improve the student performance, even to surpass the teacher.

Teacher	58.24	53.11	46.42	46.42
Supervised	51.08	47.16	39.82	39.82
Self-trained	52.86	51.23	46.28	46.23
+MSE	53.83	51.15	45.98	46.24
+PDF	**54.26**	**51.34**	**46.44**	**46.53**
+DeFeat	53.58	51.02	45.94	45.99

Table 4. mAP of students self-trained on unseen data by various teachers. The students' training sets are disjoint with the teachers': student on the second half and teacher on the first half, the last 3-quarter and the first quarter, the last 7-eighth and the first eighth, etc.

Teacher	68.29	71.12	72.40
(no KD)	40.58	40.58	40.58
+MSE	42.54	42.55	42.91
+DeFeat	**43.62**	**44.23**	**44.73**
(no KD)	47.41	47.41	47.41
+MSE	48.53	49.04	49.36
+DeFeat	**48.79**	**49.56**	**49.79**

Table 5. Object detection performance with knowledge distillation (KD) from various semantic segmentation teachers. The teachers are trained with disjoint semantic sets of different sizes and, and a larger set overlapping with the student training set. The teachers' IOU scores are provided for reference. Although the teachers had not been trained on the same data nor the same task with the students, the results with knowledge distillation are constantly better than without by large margins.

4.3. Cross-task knowledge distillation

So far, the teacher and student are both trained for the same task of object detection. In this experiment, the idea of cross-task KD is studied. In particular, we train a teacher for semantic segmentation and observe if a student network could be benefited for object detection.

The segmentation teacher uses the same architecture [34] with the detection teacher's as in the previous sections and has the detection head (localization and classification) replaced by an FPN segmentation head [12]. Feature-imitation KD is also done using the neck features as before. However, since the teacher's segmentation prediction is not immediately compatible with student's object detection output, the PDF-Distill method [34] cannot be applied and only the results for +MSE and +Defeat are reported. The student detection is trained using hard labels from its respective training sets.

We follow the common practice for training semantic segmentation on Pascal VOC and include the extra semantic annotations provided by SBD [8] to the training set. All the training images are randomly sampled into 2 subsets, one for detection whose semantic annotations are held back

	Detection	Segmentation
Teacher	44.73	-
Supervised	38.93	-
pretrained on	39.11	-
Self-trained	39.22	-
pretrained on	39.71	-
Self-trained	42.09	-
pretrained on	42.63	-
Multitask	**42.35**	**67.41**
pretrained on	41.75	65.11
self-training DET	41.26	**71.38**
pretrained on	**42.67**	69.83
+ PDF	**43.26**	68.36

Table 6. Compare and combine self-training with multi-task learning on partially annotated data for object detection and semantic segmentation (*i.e*. each image is annotated with only a single task). Despite not using any hard ground truths during training, self-training improves performance over the small-size supervised training. Extra data from semantic-segmentation subset provide further boost by large margins for both settings. Combining both setups and knowledge distillation see further improvements, on par with the larger network (teacher). The semantic segmentation performances are included for the sake of completeness.

or not available and the other for semantic segmentation whose bounding-box annotations are withheld, resulting in 7,558 and 7,656 respectively. Half of the detection images are further withheld to simulate the little annotated data scenario. For validation, the originally provided validation set for semantic segmentation with both task annotations are used with 1,443 images. The six images with only semantic segmentation are excluded from validation.

From Table 5, the students with knowledge distillation perform better than those without. Imitating features of a teacher, despite from different task topology, and being trained with few or many, seen or unseen data, show benefit and correlation to the student performances.

4.4. Multi-task learning for data exploitation

Following the discovery in Sec 4.3, in this experiment we explore the possibility of using multi-task learning as data exploitation. In particular, the problem is formulated under the partially annotated data setting, where each image is annotated for only a single task, *i.e*. the annotated image sets of the two tasks are completely disjoint.

Self-training *without* provided annotated ground truths using the whole detection subset and all images (detection and segmentation subsets) are shown, with optionally pre-trained on the available detection ground truths.

The results comparing self-training on available data and multi-task learning are shown in Table 6. Agreeing with

previous studies, despite not using supervised ground truths, self-training improves results over small data-size supervised training, with large margins when the training data are expanded to cover also the semantic segmentation subset. Multitask learning seems to perform on par with self-training with the extra advantage of having also semantic segmentation prediction. Surprisingly the combination of both does not immediately yield better performance. We speculate that the mismatching due to possible errors in pseudo detection labels from self-training and ground truth semantic segmentation leads to difficulty in learning task interrelationships. Adding PDF-Distil knowledge distillation which disentangles features using prediction agreement shows further improvement.

5. Discussions and Conclusions

The paper performs extensive experiments on knowledge distillation under the self-training paradigm for object detection and compare with the partially annotated multi-task setup where extra data with only semantic segmentation annotations are used. In the scarcity of annotations, self-training from an ill-trained teacher on unseen data points shows constant favors over full supervision. Knowledge distillation, especially PDF-Distil for object detection, can further improve the performance even when the teacher is trained on a different task while multi-task training shows improvement with large margins, further boosted when combined with feature-imitation knowledge distillation.

Each setting, however, has its own limitations. While multi-task learning depends on the relationship between the two tasks and the available annotations for each member task, self-training is bound by the teacher's performance. One common problem of both setups is the domain gap between the target and extra data. The paper assumes the same domain of the extra data used for both self-training and multi-task learning, *i.e*. Pascal VOC data, so that the teacher in self-training would still perform reasonably with new data and multi-task networks could learn similar concepts in both tasks. Depending on the gap between student-teacher domains, self-training would break down and multi-task learning would have difficulty in learning the salient interrelationship. This is out of scope of the paper and could be a potentially study for further data exploitation.

6. Acknoledgements

This work was supported by the SAD 2021 ROMMEO project (ID 21007759) and the ANR AI chair OTTOPIA project (ANR-20-CHIA-0030).

References

[1] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge Distillation: A Good Teacher Is Patient and Consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10925–10934, jun 2022. 2

[2] David Bruggemann, Menelaos Kanakis, Stamatios Georgoulis, and Luc Van Gool. Automated Search for Resource-Efficient Branched Multi-Task Networks. In *British Machine Vision Conference (BMVC)*, 2020. 2

[3] David Brüggemann, Menelaos Kanakis, Anton Obukhov, Stamatios Georgoulis, and Luc Van Gool. Exploring Relational Context for Multi-Task Dense Prediction. In *IEEE/CVF Proceedings of International Conference of Computer Vision (ICCV)*, pages 15869–15878, oct 2021. 2

[4] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. NEIL: Extracting Visual Knowledge from Web Data. In *2013 IEEE International Conference on Computer Vision*, 2013. 1

[5] Zhihao Chen, Lei Zhu, Liang Wan, Song Wang, Wei Feng, and Pheng-Ann Heng. A Multi-Task Mean Teacher for Semi-Supervised Shadow Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2020. 2

[6] M Everingham, L Van∼Gool, C K I Williams, J Winn, and A Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *ijcv*, 88(2):303–338, jun 2010. 3

[7] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling Object Detectors via Decoupled Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2154–2164, jun 2021. 2, 3

[8] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998, 2011. 5

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2

[10] Abdullah-Al-Zubaer Imran, Chao Huang, Hui Tang, Wei Fan, Yuan Xiao, Dingjun Hao, Zhen Qian, and Demetri Terzopoulos. Partly Supervised Multi-Task Learning. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 769–774, 2020. 2

[11] Paul Jaccard. The distribution of the Flora in the Alpine Zone. 1. *New Phytologist*, 1912. 3

[12] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollar. Panoptic Feature Pyramid Networks. In *IEEE/CVF Proceedings of Computer Vision and Pattern Recognition*, 2019. 2, 5

[13] Hoàng-Ân Lê and Minh-Tan Pham. Knowledge distillation for object detection: from generic to remote sensing datasets. *arXiv preprint arXiv:2307.09264*, 2023. 2

[14] Hoàng-Ân Lê, Heng Zhang, Minh-Tan Pham, and Sébastien Lefèvre. Mutual Guidance Meets Supervised Contrastive Learning: Vehicle Detection in Remote Sensing Images. *Remote Sensing*, 14(15), 2022. 3

[15] Li-Jia Li, Gang Wang, and Li Fei-Fei. OPTIMOL: automatic Online Picture collecTion via Incremental MOdel Learning.

In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 1

[16] Wei-Hong Li and Hakan Bilen. Knowledge Distillation for Multi-task Learning. In *European Conference on Computer Vision workshop (ECCVw)*, 2020. 2

[17] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Learning Multiple Dense Prediction Tasks from Partially Annotated Data. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2022. 1, 2

[18] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Universal Representations: A Unified Look at Multiple Task and Domain Learning. *arXiv preprint arXiv:2204.02744*, 2022. 2

[19] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. In H Larochelle, M Ranzato, R Hadsell, M F Balcan, and H Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21002–21012. Curran Associates, Inc., 2020. 2

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, oct 2017. 2, 3

[21] Shikun Liu, Edward Johns, and Andrew J Davison. End-To-End Multi-Task Learning With Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2019. 2

[22] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path Aggregation Network for Instance Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2018. 2

[23] Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. *arXiv*, abs/1710.07535, 2017. 1

[24] Yao Lu, Soren Pirk, Jan Dlabal, Anthony Brohan, Ankita Pasad, Zhao Chen, Vincent Casser, Anelia Angelova, and Ariel Gordon. Taskology: Utilizing Task Relations at Scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8700–8709, jun 2021. 2

[25] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved Knowledge Distillation via Teacher Assistant. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5191–5198, apr 2020. 2

[26] Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, R. Venkatesh Babu, and Anirban Chakraborty. Zero-shot knowledge distillation in deep networks. *36th International Conference on Machine Learning, ICML 2019*, 2019-June, 2019. 1

[27] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra R-CNN: Towards Balanced Learning for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[28] Zheng Qin, Zeming Li, Zhaoning Zhang, Yiping Bao, Gang Yu, Yuxing Peng, and Jian Sun. ThunderNet: Towards Real-time Generic Object Detection on Mobile Devices. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2

[29] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data Distillation: Towards Omni-Supervised Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2018. 1, 2

[30] Sebastien Razakarivony and Frederic Jurie. Vehicle detection in aerial imagery : A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34, 2016. 3

[31] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge Distillation Meets Self-Supervision. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[32] Amir R Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J Guibas. Robust Learning Through Cross-Task Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2020. 2

[33] Heng Zhang, Elisa Fromont, Sebastien Lefevre, and Bruno Avignon. Localize to Classify and Classify to Localize: Mutual Guidance in Object Detection. In *Asian Conference on Computer Vision (ACCV)*, 2020. 2

[34] Heng Zhang, Elisa Fromont, Sebastien Lefevre, and Bruno Avignon. PDF-Distil: including Prediction Disagreements in Feature-based Distillation for object detection. In *Proceedings of the British Machine Vision Conference (BMVC)*, nov 2021. 2, 3, 5

[35] Zhaohui Zheng, Rongguang Ye, Ping Wang, Dongwei Ren, Wangmeng Zuo, Qibin Hou, and Ming-Ming Cheng. Localization Distillation for Dense Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9407–9416, jun 2022. 2

[36] Rui Zhu, Shifeng Zhang, Xiaobo Wang, Longyin Wen, Hailin Shi, Liefeng Bo, and Tao Mei. ScratchDet: Training Single-Shot Object Detectors From Scratch. In *CVPR*, 2019. 2

[37] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. Rethinking Pre-Training and Self-Training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 1, 2