Temporal Self-ensembling Teacher for Semi-supervised Object Detection

Cong Chen†, Shouyang Dong†, Ye Tian, Kunlin Cao,
Li Liu, *Senior Member, IEEE* and Yuanhao Guo⋆, *Member, IEEE*

***Abstract*—Object detection is playing a key role in multimedia content analysis and understanding. However, recent development of deep learning-based object detection is hindered by a limited amount of labeled data. Therefore, this paper focuses on the problem of semi-supervised object detection (SSOD) and aims to make good use of unlabeled data to boost performance. Recently, while knowledge distillation (KD) has been widely used for semi-supervised image classification, we remain faced with the following obstacles for an empirical adaptation of KD in SSOD. (1) The teacher model serves a dual role as a teacher and a student, such that the teacher predictions on unlabeled images may be close to those of the student, which limits the upper bound of the student. (2) The data imbalance issue caused by the large quantity of consistent predictions between the teacher and student hinders an efficient knowledge transfer from teacher to student. To mitigate these issues, we propose a novel SSOD model called Temporal Self-Ensembling Teacher (TSET). In this model, we devise a temporally evolved teacher model. First, our teacher model ensembles its temporal predictions for unlabeled images under stochastic perturbations. Second, our teacher model ensembles its model weights with those of the student model by an exponential moving average (EMA), which allows the teacher to gradually learn from the student. These ensembling strategies ensure data and model diversity, and lead to better teacher predictions for unlabeled images. In addition, we adapt the focal loss to formulate the consistency loss for handling the data imbalance issue. Together with a thresholding method, which eliminates confident background predictions, the focal loss automatically reweights the inconsistent predictions. This preserves the whole knowledge from unlabeled images, especially for objects that are difficult to detect. The mAP of our model reaches 80.73% and 40.52% on the VOC2007 test set and the COCO2014 *minival5k* set, respectively, and outperforms a strong fully supervised detector by 2.37% and 1.49%, respectively. Furthermore, the mAP of our model (80.73%) sets a new state-of-the-art performance in SSOD on the VOC2007 test set. The source code is made available at http://github.com/syangdong/tse-t.**

***Index Terms*—Semi-supervised object detection, deep convolutional neural networks, knowledge distillation, temporal self-ensembling, focal loss**

C. Chen and K. Cao are with Keya Medical Technology, ShenZhen, 518116.
S. Dong is with Software Department at Cambricon, Beijing, 100010.
Y. Tian is with Hippocrates Research Lab at Tencent, Shenzhen, 518052.
Li Liu is with the Colloge of System Engineering, National University of Defense Technology, China and is also with Center for Machine Vision and Signal analysis at the University of Oulu, Finland.
Email: li.liu@oulu.fi
Yuanhao Guo is with Institute of Automation, Chinese Academy of Sciences, Beijing, 100190.
Email:yuanhao.guo@ia.ac.cn
† C. Chen and S. Dong are equal contributors.
⋆ Y. Guo is the correspondence author

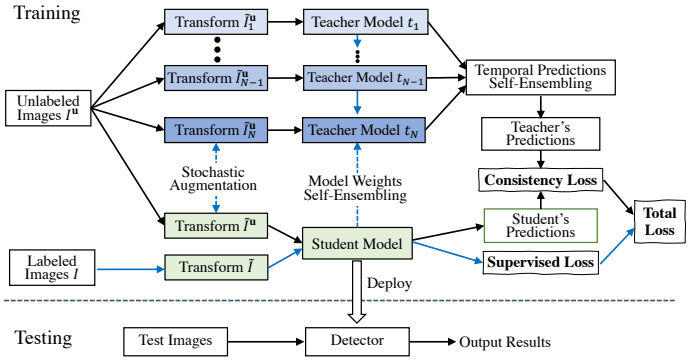

Fig. 1. **The Framework of the Proposed TSET Model for SSOD.** At training time, unlabeled images under stochastic transformations like random horizontal flip are predicted by a series of consecutive teacher models. The results are then aligned and ensembled to obtain the teacher predictions which are subsequently used as targets to regularize the training of the student model. We keep the teacher model evolved using an EMA which results in temporally diverse teacher model. At testing time, the trained student model is deployed for object detection for unseen images.

I. Introduction

Object detection is the cornerstone of computer vision, as many high-level vision tasks fundamentally rely on the ability to recognize and localize visible objects. Object detection thus touches many areas of artificial intelligence and information retrieval, such as image/video search, data mining, question answering, autonomous driving, medical diagnosis, and robotics [1]–[6]. The recent resurgence of interest in artificial neural networks, in particular deep learning, has tremendously advanced the field of generic object detection; and in the past few years, a large number of detectors [1], [7]–[13] have sprung up to improve the detection performance from aspects like accuracy, efficiency, and robustness.

Current state-of-the-art detectors [7]–[9], [11], [13] are learned in a fully supervised fashion, which requires large-scale labeled data with many high-quality object bounding box annotations or even segmentation masks. Gathering bounding box annotations or segmentation masks for every object instance is time consuming and expensive, especially when the training dataset contains a huge number of images or even videos, as it requires intensive efforts of experienced human annotators or experts (*e.g.*, medical image annotation) [1], [14]–[18]. Furthermore, manual bounding box/segmentation mask labeling may introduce a certain amount of subjective bias. In addition, the generalizability of fully supervised detectors is limited. Moreover, there are massive amounts of

unlabeled images that are acknowledged as valuable, and the key is how to make good use of them [19]–[24].

The time-consuming and expensive annotation of accurate bounding boxes of object instances is sidestepped in weakly supervised object detection (WSOD), which only utilizes image-level annotations that show the presence of instances of an object category [25]–[29]. WSOD methods may achieve relatively good performance if provided with a large number of image-level annotations; however, the performance is hardly competitive compared to their fully supervised counterparts. Considering a generic situation in object detection, we have a limited number of labeled images [14], [15], but a huge number of unlabeled images (*e.g.*, the massive amounts of unlabeled images available from the Internet). Semi-supervised learning (SSL), which falls between supervised and unsupervised learning, has shown promising results to reduce the gap between supervised and unsupervised learning. SSL has been extensively studied in image classification problems [30]–[32] and other multimedia applications [33], [34], but has received significantly less attention in object detection. In this work, our main focus is SSL for object detection, *i.e.*, semi-supervised object detection (SSOD).

Classical deep learning-based SSL methods use the maximum predictions for unlabeled images as pseudo-labels to improve the classification performance of neural networks [35]. The recently developed knowledge distillation (KD) [36], [37] aims at training a lightweight student model regularized by a cumbersome teacher model, and is originally used for deep model compression but later widely applied to analyze multimedia content [38], [39] and to solve SSL problems. Quite a few KD-based SSL methods have been proposed [19], [40], and the key to these methods is to construct a well-performing teacher to obtain stable and reliable predictions when trained with unlabeled images. The teacher predictions for unlabeled images can be used as targets (well-posed logits or soft labels) to regularize the training of the student in order to obtain similar predictions on these unlabeled images, thereby yielding a well-trained student that has a performance close to that of the teacher. This can be implemented by using the consistency loss to set a consistency regularization between the teacher and student predictions, which routinely takes the form of mean squared error (MSE) loss.

So far, however, only a few works have applied similar ideas in a challenging task like SSOD [23], [41], [42]. The main challenges are as follows. (1) The teacher model in these KD-based methods often serves a dual role as a teacher and a student. In image classification, it is sufficient to solely handle a unique prediction per image; but for object detection, which is a more complicated task since we must simultaneously classify and localize an object, such a teacher model may produce predictions that are very close to those of the student, which sets an upper bound for the student's performance. (2) The predictions in object detection are rather dense during training because an object can present at every location in an image, and an image may contain multiple objects. Given unlabeled images, an SSOD model uses the teacher predictions as targets to regularize the training of the student. A direct adoption of the widely used consistency loss from SSL is hampered in SSOD because the large quantity of consistent predictions between the teacher and student may suppress the inconsistent predictions that actually contain information useful for training the student. A recent model, the CSD model [23], tackles this problem with a background elimination (BE) method, which simply thresholds out almost all possible backgrounds. However, there are several limitations of the CSD model. Given unlabeled images, (1) the teacher and student are identical, which may limit the performance of the student; and (2) the BE method removes too much background, which may lead to a performance drop in SSOD because useful knowledge encoded in difficult objects may be neglected.

In the present work, we aim at a simple but generic solution to alleviate the above issues and improve SSOD. To this end, we propose the ***T***emporal ***S***elf-***E***nsembling ***T***eacher model, coined ***TSET***. We present the framework of our model in Fig.1. The TSET model is devised on top of the KD framework, which consists of a teacher and a student model. Both the teacher and student are initiated from a pretrained detection network in a fully supervised manner. During semi-supervised training, the teacher obtains the predictions for both the category and location of all possible objects presented in the unlabeled images. The student also obtains its detections for these unlabeled images. The KD framework aims to minimize the dissimilarity between teacher and student predictions by using consistency loss. During testing, the trained student model is deployed for object detection in new images.

Based on the above framework, our TSET model targets two main goals.

(1) Our first goal is to enhance the performance of the teacher model on object detection in unlabeled images. To this end, we devise a temporally updated teacher model that is asynchronous from the training of students.

Specifically, instead of using a constant teacher as proposed in the original KD-based methods [21], [36], our TSET model devises a teacher that ensembles its temporal predictions from consecutive training epochs for the unlabeled images under stochastic perturbations (random transformations like horizontal flip). This type of data augmentation and temporal predictions ensembling strategy has been used to effectively improve the prediction accuracy in SSL problem [40]. Moreover, our teacher model ensembles its temporal model weights with the student model weights, which allows the teacher to gradually learn from the student. This is implemented by using an exponential moving average scheme (EMA) [19], [24]. In this way, the evolution of the teacher model is decoupled from the training of the student, which prevents the teacher from obtaining similar predictions as the student.

These self-ensembling strategies together increase data and model diversity, thus yielding stable and reliable teacher predictions for unlabeled images, which can then be used as better targets to train the student. The proposed TSET model substantially distils knowledge of multiple image geometric transformations from a well-trained teacher to the student. In other words, the student is guided to imitate the behavior of the teacher by its predictions on unlabeled images implemented in the form of a consistency loss, thus leading the student's performance approaching that of the teacher.

(2) Our second goal is to solve the data imbalance problem by preserving sufficient difficult objects. The BE method used in the CSD model [23] simply eliminates most low-confidence predictions on unlabeled images. These predictions may include some difficult but informative object examples, which prevents the encoded knowledge to be distilled to the student. To solve this problem, we employ a customized detection loss, *i.e.*, the focal loss [8] to formulate the consistency loss between teacher and student predictions; this preserves useful information from unlabeled images as much as possible.

Specifically, the data imbalance problem in SSOD is caused by the large number of well-matching object detections between the teacher and student. Such detections include consistent foreground predictions and consistent background predictions. The accumulation of a large quantity of consistent predictions is more prevalent in the training loss, which suppresses the contribution of informative training examples, *i.e.*, the inconsistent predictions. For example, in an unlabeled image, the teacher predicts an object with a confidence score of 0.6, and the student gives a prediction score of 0.3 or predicts it as background. This informative training example will be either ignored by the BE method [23] or suppressed because it is a minority in training examples. To solve this data imbalance problem, we first eliminate the confident background predictions according to a large threshold. This helps to remove many consistent background predictions and preserve sufficient objects that are difficult to detect. We then adapt the focal loss to formulate the consistency loss in our SSOD setup, which automatically consolidates the contribution of the inconsistent predictions, *i.e.*, the poor-matching predictions between the teacher and student.

We evaluate the performance of our TSET model on two standard large-scale benchmarks: PASCAL VOC [14] and MSCOCO [15]. Both evaluation results show that the TSET model can obtain remarkable improvements compared to the baseline model, *i.e.*, the fully supervised detection model only using labeled images. Specifically, the mAP of our model is 80.73% and 40.52% on the VOC2007 test set and COCO2014 *minival5k* set, respectively, thus outperforming the baseline by 2.37% and 1.49%. It should be noted that our model sets a new state-of-the-art performance in SSOD on the VOC2007 benchmark. We summarize our contributions as below.

(1) We formally employ the KD framework in the SSOD task and construct a well-trained teacher to regularize the training of a student using unlabeled images.

(2) We propose the TSET model, which ensembles temporal predictions of teacher model and updates the teacher model by ensembling the student model weights. This model produces better targets to train the student, but does not significantly increase computational complexity.

(3) We adapt focal loss to solve the data imbalance problem, which results in an efficient and effective usage of unlabeled images in SSOD.

The rest of the paper is organized as follows. We review related works in Section II. We elaborate on the proposed model in Section III. We present our experimental results in Section IV. Finally, in Section V we conclude our work and present several potential directions for future work.

II. Related Works

In this section, we review the topics related to our work, including object detection (Section II-A), semi-supervised learning (Section II-B), semi-supervised object detection (Section II-C) and mutual learning (Section II-D).

A. Object Detection

Object detection is one of the most active research topics in the computer vision community [43]. Hundreds of well-performing detectors have been developed. In this work, we focus on the generic object detection models using deep learning [1]. The pioneering R-CNN uses deep learning methods to extract features in the conventional object detection pipeline [44]. The Fast-RCNN [45] and Faster-RCNN [7] initiate the study on typical two-stage detectors, and successfully complete object detection with an end-to-end deep learning architecture. The FPN [46] and RetinaNet [8] improve the feature representation for object detection by using a decoder-like feature pyramid. Moreover, one-stage detectors like SSD [11] have been developed, which generate dense predictions using fully convolutional neural networks. This type of method is much faster than two-stage detectors, and one extraordinary example is YOLO and its advanced versions [10], [47], [48]. In Mask-RCNN [9] a multi-task network is proposed to integrate the object detection and semantic segmentation, which reshapes the instance segmentation with an end-to-end manner. All the methods above use popular anchor boxes to encode the object bounding box, which leads to a translation-invariant detection, and makes regression easier. Recently developed anchor-free detectors [12], [49]–[51] reformulate the object detection task as a key point detection and grouping task. This line of detectors reduces the quantity of output but still results in comparable performance.

B. Semi-Supervised Learning

SSL is one important category of machine learning techniques [30]–[32], and trains a machine learning model by using both labeled and unlabeled data. The key to SSL methods is to generate a training target for the unlabeled data. The Γ model [52] adapts the autoencoder network [53] to solve the SSL problem. This model passes an unlabeled image and its noisy version through the encoder. The training target is set as the clean image and its latent features. For the noisy image, the decoder reconstructs its clean version and the clean latent features. Instead, the Π model [54] minimizes the prediction difference of the same unlabeled data with various stochastic transformations through perturbed networks. Sufficient randomizations like dropout and randomized data augmentation are used during training with unlabeled data, which results in better generalization for testing. Since the emergence of the knowledge distillation network [36], SSL has been reshaped by the teacher–student model architecture. A well-posed prediction for the unlabeled data can be obtained by a strong pretrained teacher model, and this prediction is used as a target to guide the training of the student model [40]. The strategy of ensembling multiple networks

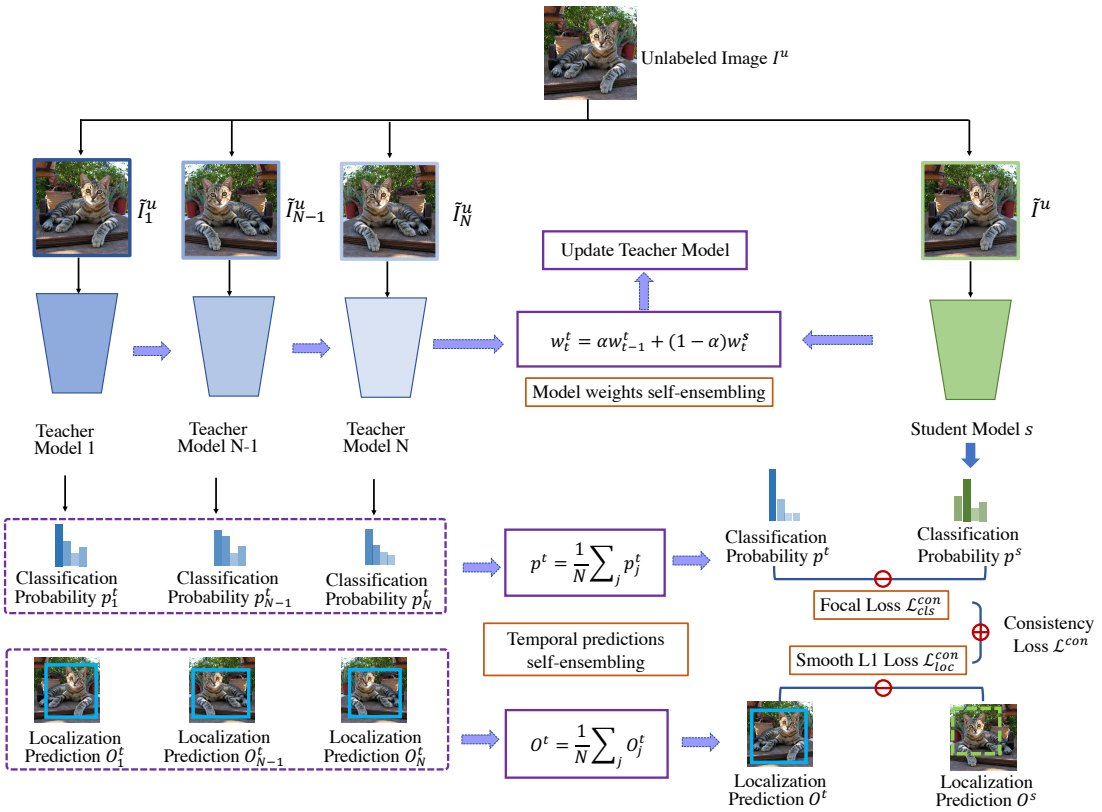

Fig. 2. **A Detailed Graphical Illustration for the Proposed TSET Model**. Our method is made on top of the KD framework which consists of a teacher and a student model. Our TSET model is devised to ensemble the temporal teacher predictions and ensemble the teacher model weights with the student model weights. These self-ensembling methods yield better targets for unlabeled images which can better retrain the student to improve its performance. We use orange bounding boxes to indicate our main contributions.

is an effective strategy to produce more accurate predictions [16]. In SSL, the temporal ensembling model [40] improves teacher predictions for the unlabeled data by accumulating the historical predictions during training, which combines the model diversity for better generalization of unlabeled data. Instead of ensembling the predictions, the mean teacher model [19] ensembles the teacher and the student model weights to yield a dynamic teacher, which results in a temporally updated teacher model. These ensembling strategies aggregate sequential predictions from the latest training epochs or model weights, which solely involves a teacher and a student during training, rather than ensembling multiple networks; naturally, this reduces the computation complexity. In comparison, our TSET model generalizes the self-ensembling strategy: it first ensembles a sequence of historical predictions of the teacher, and then ensembles the teacher model weights with the student (similar to the mean teacher model [19]).

C. Semi-Supervised Object Detection

The main obstacle of SSOD is to produce well-posed training targets, *i.e.*, the dense bounding-boxes associated with categories. A successful application of deep learning techniques on SSOD is seen with the CSD model [23], which adapts the Π model [54] to construct the consistency regularization for detecting the unlabeled image and its augmentation. To solve the problem that loss may be easily dominated by backgrounds, the BE method is proposed to remove all possible background predictions. The STAC model [42] proposes a simple but effective strategy which uses strong image augmentations to enforce consistency of the teacher and student. A very recent semi-supervised method employs a proposal-based learning scheme for two-stage object detectors [41]. For the original data and its noisy counterpart, the method uses a self-supervised proposal learning module to learn consistent perceptual semantics in feature space, and to learn consistent predictions. Meanwhile, omni-supervised object detection uses two-stage detectors, and a bounding box voting strategy generates a hard-label teacher prediction [21]. Compared to these works, our model has advantages in the following respects. (1) We propose a generalized ensembling strategy, thus improving the teacher predictions on unlabeled images. (2) Instead of using a pseudo-label as a target to train the student, we use a soft-label, which is proven to be a more informative and efficient training strategy [37]. (3) We use focal loss to solve the data imbalance problem caused by dense predictions in SSOD.

D. Mutual Learning

Joint training of the teacher and student allows for a better generalization of both the teacher and student. Rather than using a one-way knowledge distillation from a strong pretrained teacher to a weak student, the DML [55] jointly trains several students that are later ensembled to become the teacher model. However, the efficiency of DML may decrease

when multiple students are deployed. To solve this problem, the ONE learning strategy devises a multi-branch network that shares the low-level layers [56]. The ensemble of the multiple branches act as the teacher, and a single network is deployed. The mutual learning method is also used in training the low-bitwidth model [57]. This method proposes a new training strategy that jointly learns a full-precision model and a low-precision model. Additional guidance losses are added at intermediate feature layers, which forces the two models to generate similar feature maps. Our TSET model adapts a different mutual learning strategy, the EMA, to update the teacher model. We choose this strategy because we expect a mild transition from supervised to unsupervised training, and we expect stable teacher predictions for unlabeled images as robust targets to train the student model.

III. Methodology

In this section, we first present our SSOD setup in Section III-A, then elaborate the reasons for the selection of the baseline detector in Section III-B, and finally present our proposed TSET model in Section III-C.

A. SSOD Problem Setup

The overall framework of our proposed SSOD model is illustrated in Fig. 1. The objective of our proposed SSOD model is to distil knowledge from geometrically transformed unlabeled images without training a large set of models. Our pipeline (Fig. 1) involves the following steps:

(1) pretrain a fully supervised object detector on a labeled dataset, and use it to initialize the teacher and the student;

(2) apply the teacher model to a number of geometric transformations of unlabeled samples to generate detections for the unlabeled samples;

(3) ensemble multiple teacher predictions on the unlabeled data to automatically generate training targets for student;

(4) retrain the student on the union of the manually labeled data, and automatically labeled data;

(5) update the teacher by ensembling its temporal model weights and current student model weights.

Our model is based on the KD framework, which consists of a teacher and a student model. Both models are initiated from a typical one-stage detector such as the RetinaNet [8], and are pretrained using a certain number of labeled images in a fully supervised manner. During semi-supervised training, the teacher predictions on unlabeled data are used as annotations to retrain the student in an unsupervised manner. The labeled data is also used to train the student in a supervised manner to leverage the unsupervised training.

The unsupervised retraining of the student model using unlabeled data is achieved by a consistency loss, which routinely takes the form of minimizing the MSE between the teacher and student predictions. In this way, the teacher model distils useful knowledge, *i.e.*, the object category and localization in unlabeled images, to retrain the student model. In other words, the knowledge encoded by the teacher is decoded in such a way that the student back-propagates the gradients to optimize its parameters; this enables better generalization for unlabeled images. After training, the student model should achieve comparable performance to that of the teacher, and is deployed to detect objects in unseen images during inference time.

To obtain better teacher predictions on unlabeled data, we propose the TSET model, which is detailed in Fig. 2. TSET includes two types of self-ensembling strategies. (1) We run a temporal series of the teacher model on different geometric transformations (*e.g.*, horizontal flipping) of an unlabeled image, and then ensemble the predictions as the final training targets for the student. Such multiple temporal models ensembling takes advantage of the different generalizability of the models on different data transformations, which improve the teacher predictions on unlabeled data by a large margin from the student predictions. (2) We ensemble the temporal model weights of the teacher with the student model weights using the EMA so that the teacher can gradually learn from the student to not only enhance its performance but also increase its temporal diversity.

Now we formally define the SSOD problem setup. Suppose we are given a dataset of M images $\mathcal{D} = \{\mathbf{I}_i\}_{i=1}^{M}$. For a labeled image $\mathbf{I}_i^l \in \mathcal{D}$, $\boldsymbol{y}_i = [P_{i,x}, P_{i,y}, P_{i,w}, P_{i,h}, c_i]^T$ is the ground truth label vector that defines the category $c_i \in [1, ..., C]$ of an object and specifies the pixel coordinates $(P_{i,x}, P_{i,y})$ of the center of object bounding box together with its width and height $(P_{i,w}, P_{i,h})$ in pixels. For an unlabeled image $\mathbf{I}_i^u \in \mathcal{D}$, we use the teacher prediction as the target. For instance, a detected object bounding box O from an unlabeled image is represented as $\boldsymbol{z}_i^t = [O_{i,x}, O_{i,y}, O_{i,w}, O_{i,h}, p_i]^T$, where p_i is the class probability. Below, we drop the superscript i unless it is needed.

In SSOD, we aim to promote the performance of the student model regularized by the teacher model using the unlabeled images. This is achieved in an unsupervised manner, and defined as consistency loss between the teacher and student predictions $\mathcal{L}^{con}(\boldsymbol{z}^s, \boldsymbol{z}^t)$, where $\boldsymbol{z}^s$ is the student prediction, which is formulated in the same way as the teacher prediction. If only the unlabeled images are used to retrain the student model, there may be an ill-posed convergence behavior of the student. For a balanced semi-supervised training, we also employ the labeled images in the form of a supervised loss $\mathcal{L}^{sup}(\boldsymbol{z}^s, \boldsymbol{y})$. We will define $\mathcal{L}^{con}$ and $\mathcal{L}^{sup}$ in the following sections. The objective is to optimize the student model to minimize both the consistency loss and supervised loss.

$$\mathcal{L} = \mathcal{L}^{sup}(\boldsymbol{z}^s, \boldsymbol{y}) + \mu_1 \mathcal{L}^{con}(\boldsymbol{z}^s, \boldsymbol{z}^t) \tag{1}$$

We use the hyper-parameter μ_1 to leverage the contribution of the supervised loss $\mathcal{L}^{sup}$ and consistency loss $\mathcal{L}^{con}$. The selection of μ_1 will be discussed later in Section IV.

B. Baseline Detector

Object detectors can be classified into two categories: one-stage [8], [10], [11] and two-stage [7], [9], [58]. The main distinction is that the two-stage detectors employ a region proposal network (RPN) to explicitly generate object candidates. The non-maximum suppression (NMS) then merges the spatially duplicated prediction candidates with a certain

amount of overlap. In this work, we choose to use a one-stage detector, like RetinaNet [8], for the following reasons.

First, the objective of our TSET model is to synchronize the predictions of the student with those of the teacher so that the student can approach the performance of a well-trained teacher. Therefore, we do not employ NMS before the fine-grained object detection both in the teacher and student because many confident predictions may be suppressed and the knowledge distillation from the teacher may deteriorate the generalizability of the student model for those predictions.

Second, in two-stage detectors, it is relatively difficult to solve the matching problem in RPN between an image and its transformations. This is because the employment of NMS in RPN results in the misalignment of the region proposals for different input images. A simple solution is to only feed the original image into RPN, and use the location of the obtained region proposals to estimate the location of region proposals from the transformed image [23]. It has been empirically found that the adaptation of two-stage detectors results in worse SSOD performance than the use of one-stage detectors due to the lack of consistency regularization in RPN training [23].

C. TSET Model

1) Ensemble temporal predictions

A well-performing teacher model for SSOD should provide better predictions of objects presented in unlabeled images. These predictions should contain sufficient dissimilarities from the student predictions, such that the knowledge encoded in these objects can be fully captured and then distilled to the student. By pushing the student to obtain predictions as accurately as the teacher, the student can improve their performance on detecting all possible objects in unlabeled images. In our proposed TSET, we achieve this by ensembling the temporal teacher predictions from the latest training epochs. Because we add random perturbations for each image in each epoch, this self-ensembling process produces a large number of data combinations for teacher models at different checkpoints, and this data and temporal model diversity ensures better teacher predictions on unlabeled images.

Specifically, at training time, given an unlabeled image $\mathbf{I^u} \in \mathcal{D}$, we retrieve its previous teacher predictions from the last N epochs $\boldsymbol{z}_1^t, \cdots, \boldsymbol{z}_N^t$. The TSET model obtains the current teacher prediction by averaging these predictions:

$$\boldsymbol{z}^t = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{z}_j^t, \tag{2}$$

which can be separately denoted as ensembling the localization and classification.

$$\begin{cases} O^t = \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} O_j^t, \\ p^t = \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} p_j^t. \end{cases} \tag{3}$$

Since we augment the data, we need to align the predictions before ensembling. We implement this by tracing the image orientation during augmentation, and flip the predictions back to the original reference image. We assume that ensembling more teacher predictions from different training epochs may generate better training targets for the student. We will validate this assumption and show the effect of varying values of N on SSOD in Section IV.

2) Ensemble model weights

In the original temporal ensembling model [20], the teacher model serves as the student model as well. When given an unlabeled image and its transformations, the teacher and student predictions may be very close; for example, if the objects in the image are easy to recognize. These similar predictions contribute little to the consistency loss, which constrains the upper bound of the student model. Therefore, we expect to decouple the teacher model from the student during semi-supervised training, and keep the teacher model evolved instead of fixed (as proposed in the original KD framework). To this end, our TSET model uses a temporally updated teacher model devised to ensemble the historical teacher model weights with the current student model weights using a momentum term, which is formulated as follows:

$$w_t^t = \alpha w_{t-1}^t + (1-\alpha) w_t^s, \tag{4}$$

where w_t^t and w_{t-1}^t denote teacher model weights at current and previous training step respectively; w_t^s denotes the student model weights updated at current training step. We use subscript t to index the training step. α is a momentum parameter to leverage the contribution of previous teacher model weights and current student model weights when updating the current teacher model. Such a model weights ensembling method is also referred to as the EMA. This self-ensembling can be seen as an imitation of a real circumstance: a teacher may be biased in their existing knowledge or have a gap in their knowledge, but their student may help by conveying this missing knowledge to the teacher. We expect a robust additive knowledge transfer from the student, which can result in a fast and stable convergence of the teacher. This can be manipulated by setting a relatively large value of parameter α, for example, $\alpha = 0.99$ [24].

3) Loss functions

The training objective for object detection is to minimize the prediction errors for both classification and localization. Accordingly, we specify each term in Eq. 1 using Eq. 5

$$\begin{cases} \mathcal{L}^{sup}\left(\boldsymbol{z}^s, \boldsymbol{y}\right) &= \dfrac{1}{M_1}\displaystyle\sum\left(\mathcal{L}_{cls}^{sup} + \mu_2 \mathcal{L}_{loc}^{sup}\right), \\ \mathcal{L}^{con}\left(\boldsymbol{z}^s, \boldsymbol{z}^t\right) &= \dfrac{1}{M_2}\displaystyle\sum\left(\mathcal{L}_{cls}^{con} + \mu_2 \mathcal{L}_{loc}^{con}\right), \end{cases} \tag{5}$$

where $\mathcal{L}^{sup}$ and $\mathcal{L}^{con}$ denote the supervised loss for the labeled images and the consistency loss for the unlabeled images respectively. M_1 and M_2, are, respectively, the total predictions of labeled and unlabeled images. The hyper-parameter μ_2 balances the classification and localization loss, which will be discussed in Section IV.

In our model, we use focal loss to address the data imbalance problem during training, both in supervised and consistency loss, separately formulated as:

$$\begin{cases} \mathcal{L}_{cls}^{sup} = -(1-p^s)^{\gamma}\log(p^s), \\ \mathcal{L}_{cls}^{con} = -(|p^t - p^s|)^{\gamma} p^t \log(p^s). \end{cases} \tag{6}$$

The loss functions retain the form of standard cross entropy loss, where p^t and p^s are the teacher and student prediction probabilities, respectively, for the object class.

As for the localization, we introduce the Smooth L1 loss both for the supervised and consistency localization loss:

$$\begin{cases} \mathcal{L}_{loc}^{sup} = smooth_{L1}(\tilde{O}^s - \tilde{P}), \\ \mathcal{L}_{loc}^{con} = smooth_{L1}(\tilde{O}^s - \tilde{O}^t), \end{cases} \tag{7}$$

where $\tilde{P}$, $\tilde{O}^s$ and $\tilde{O}^t$ are the offsets from the ground truth, student prediction, and teacher prediction to the anchor boxes, respectively. We provide an example for the computation of the offsets using the teacher prediction $O^t = [O_x, O_y, O_w, O_h]$.

$$\begin{cases} \tilde{O}_x = (O_x - d_x)/d_w, \\ \tilde{O}_y = (O_y - d_y)/d_h, \\ \tilde{O}_w = \log(O_w/d_w), \\ \tilde{O}_h = \log(O_h/d_h), \end{cases} \tag{8}$$

where $d = [d_x, d_y, d_w, d_h]$ is the localization of one anchor box, and $\tilde{O}^t = [\tilde{O}_x, \tilde{O}_y, \tilde{O}_w, \tilde{O}_h]$ is the normalized teacher prediction of one object's localization.

In our settings, we use the one-stage detection network and avoid employing the NMS before model ensembling. For the unlabeled images, this encourages the emergence of a large number of pairwise teacher–student predictions that are well-matched. It should be noted that the accumulation of well-matched predictions is likely to suppress the inconsistent prediction pairs that are the minority of training examples but should be main contributors in the loss. Thus, we revise the Smooth L1 function to alleviate this effect. The updated function is formulated as follows:

$$smooth_{L1}(x) = \begin{cases} \dfrac{|x|^3}{3}, & |x| < \beta \\ |x| - \beta + \dfrac{\beta^3}{3}. & |x| \geq \beta \end{cases} \tag{9}$$

In Algorithm 1, we summarize the whole training procedure of our model in the form of pseudo-code. Here we omit the fully supervised pretraining step using labeled images, and thus we directly start from a convergent detection model and train the student model using the proposed TSET model.

Algorithm 1: Pseudocode of TSET model

f^{s*} = TSET(f^s, f^t, w^s, w^t)
Input: Training dataset $\mathcal{D}$
Pre-trained student model f^s
Pre-trained teacher model f^t
Student model weights w^s
Teacher model weights w^t
Output: Optimized student model f^{s*}
Initialization: Epochs=K, Step: $t = 0$
for $k \leftarrow 0$ **to** $K-1$ **do**
  foreach *Mini-batch* $\mathcal{D}_b$ **do**
    $\tilde{\mathcal{D}}_b = stochastic - transformation(\mathcal{D}_b)$
    # We omit the subscript b for clarity.
    $z^s = f^s(\tilde{\mathcal{D}}_b)$
    if $\tilde{\mathcal{D}}_b$ *is unlabeled* **then**
      # Align teacher predictions in previous N epochs
      $[z_1^t, \cdots, z_N^t] = f_A([z_1^t, \cdots, z_N^t])$
      # Ensemble temporal teacher predictions by Eq. 2
      $z^t = \frac{1}{N}\sum_{j=1}^{N} z_j^t$
    end
    # Compute total loss by Eq. 1
    $\mathcal{L} = \mathcal{L}^{sup}(z^s, z) + \mu_1 \mathcal{L}^{con}(z^s, z^t)$
    # Update student model by standard SGD
    $w_t^s = w_{t-1}^s - \lambda \partial \mathcal{L}/\partial w^s$
    # Update teacher model by Eq. 4
    $w_t^t = \alpha w_{t-1}^t + (1-\alpha) w_t^s$
    # For teacher prediction in next epoch
    $z_{N+1}^t = f_{w=w_t^t}^t(\tilde{\mathcal{D}}_b)$
    $t = t + 1$
  end
  # Update teacher predictions for next epoch
  $[z_1^t, \cdots, z_N^t] \leftarrow [z_2^t, \cdots, z_{N+1}^t]$
end

IV. Expreiments

In this section, we conduct experiments to evaluate our proposed TSET model. We use two standard benchmarks for object bounding box localization, the PASCAL VOC [14] and MSCOCO [15]. For comparison with existing models, we use the fully supervised RetinaNet as a strong baseline. RetinaNet has a feature pyramid network (FPN) as its backbone network, and uses the focal loss for supervised training. We also make comparisons to two state-of-the-art models, CSD [23] and STAC [42], to highlight the merits of our model under the semi-supervised setup. We implement our model based on the Mask-RCNN benchmark [59].

A. Configurations

For the PASCAL VOC benchmark, we use the VOC2007 and VOC2012 datasets, both of which consist of 20 annotated semantic object classes. Following the configuration in [23], we fix the VOC2007 test set to evaluate different models. For the MSCOCO benchmark, we choose COCO2014, which includes 80 semantic classes; and we follow the standard experimental protocol [8], [46], [60] which uses the *trainval35k* split for training, and the *minival5k* split for testing. PASCAL VOC and MSCOCO benchmarks separately contains a subset without ground truth annotations, *i.e.*, the VOC2012 test set, and the COCO unlabeled set. We use these two subsets as extra unlabeled images. In Table I, we summarize the details of the employed datasets.

We conduct all of the experiments using four NVIDIA 1080 Ti GPU cards. We use the SGD optimizer and set the batch size as 8. For the backbone network of RetinaNet, we choose to use

TABLE I
DATASETS STATISTICS

Dataset \ Fold	Train	Val	Train/Val	Test	Unlabeled
VOC2007	2,501	2,510	5011	4,952*	–
VOC2012	5,717	5,823	10,540	10,991**	–
COCO	80,000	35,000	115,000	5,000*	123,403**

* Test set in our experiments
** Extra unlabeled images in our experiments

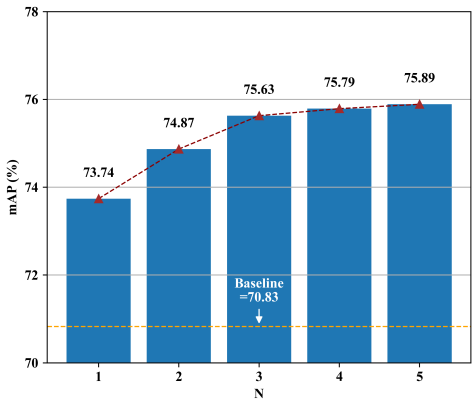

Fig. 3. **Validation of N in TSET.** X-axis indicates an N ranging from 1 to 5. Y-axis indicates the mAP of detection results. We use red plotting to show the performance tendency according to N. The horizontal dash line denotes the baseline performance of the fully-supervised detector.

ResNet-50 [61] for the experiments on the VOC dataset, and validate the performance of ResNet-50 and ResNet-101 on the COCO dataset. For all of the experiments, we use the standard metric of mean average precision (mAP) for evaluation.

When pretraining the detection model using labeled images, we use 15 epochs, and initialize the learning rate as 0.005, which is divided by 10 at both epoch 5 and epoch 8. When training the student model using unlabeled images, we use 13 epochs, and initialize the learning rate as 0.0005, which is divided by 10 at epoch 10. It has been found that for an SSOD system, once the model converges to a local minimum, it will be difficult to reach a global solution in the following training steps. Therefore, we carefully design the update strategy for μ_1. In this work, we aim for a stable transition from fully supervised training to semi-supervised training by slowly increasing the weights of the unlabeled data. We thus gradually increase μ_1 from 0.02 to 1.6 and from 0.01 to 0.08 for ResNet-50 and ResNet-101 backbone networks, respectively. As for μ_2, we choose the values of 0.07 and 0.1 for ResNet-50 and ResNet-101 backbone networks, respectively; this selection modulates the classification and localization loss at a similar scale. Finally, we set $\beta = 0.4$ as per Eq. 9.

B. Experiments on PASCAL VOC

Validation of *N*. In this experiment, we validate the effects of N in our proposed TSET on SSOD. N determines the number of historical teacher predictions used to ensemble the current teacher predictions on unlabeled images. Here, we use 2007train/val as labeled data and 2012train as unlabeled data. We leave out the 2012val to evaluate the performance of our model. We design this setup for a balance between labeled and unlabeled images. Constrained to the computation capacity of our firmware, we set N to range from 1 to 5. For the baseline model, we use the RetinaNet trained using the same labeled data, *i.e.*, 2007train/val. In our model, we keep all the configurations stable (*e.g.*, the learning rate and number of training epochs) while assigning different values to N.

From the validation results shown in Fig. 3, we can see that our model outperforms the baseline model by a large margin. An increasing N corresponds to a continuous performance improvement, which means that ensembling more historical teacher predictions benefits the retraining of the student model. We notice that a large value of N, for example, $N = 5$, may not significantly increase the mAP; however, we still choose to use $N = 5$ for the remaining experiments because it guarantees the best validation performance of our model.

Choice of background elimination threshold. In semi-supervised training, ignoring too much background may deteriorate performance. We thus design a threshold method to solve this problem. Specifically, we remove the background predictions if the estimated background probability is larger than a sufficiently large threshold. To show its efficacy in our TSET model, we compare its performance with the BE method [23]. Specifically, we compare the threshold+focal loss with the BE+focal loss in our TSET model, trained using 2007train/val as labeled images and 2012train/val as unlabeled images. Evaluated on 2007test, the mAPs of our threshold method and the BE method are 76.68% and 76.40%, respectively; thus, a validated threshold (0.9) leads to a better background elimination. As a result, for the remaining experiments, we use the threshold method in our TSET model.

Ablation study. In this experiment, we validate the effectiveness of the basic modules in our TSET model: (a) the module for self-ensembling temporal teacher predictions, which we denote as "Ensemble", (b) the module for ensembling temporal teacher model weights, which we denote as "EMA", and (c) the customized detection loss-based consistency regularization, which we denote as "Detection loss". When omitting the "EMA", we freeze the teacher model during semi-supervised training and ensemble its predictions on unlabeled images from *N* latest training epochs as targets to train the student. When omitting the "Ensemble" , we only use the teacher predictions on unlabeled images from the latest training epoch as targets to train the student. Meanwhile, when omitting the "Detection loss", we use the standard Euclidean distance to formulate consistency loss. We use "simple semi-supervised method" (S-SEMI) to refer to the semi-supervised method that does not use any of the proposed modules. In S-SEMI, we freeze the teacher model, use its predictions on unlabeled images from the latest training epoch as targets, and use MSE loss to train the student.

We design the following different configurations of the dataset for semi-supervised training: (a) 2007train/val as labeled images and 2012train/val as unlabeled images; (b) 2007tain/val as labeled images and 2012train/val/test as unlabeled images; and (c) 2007train/val+2012train/val as labeled images and 2012test as unlabeled images. Compared to setup (a), setup (b) includes more unlabeled images, and setup (c)

TABLE II
ABLATIVE EXPERIMENTS FOR TSET ON THE VOC2007 TEST SET

Model	07train/val	12train/val	12test	Ensemble	EMA	Detection Loss	mAP
RetinaNet	✓	–	–	–	–	–*	71.56
S-SEMI	✓	SS	–	–	–	–	$71.98_{\uparrow 0.42}$
TSET	✓	SS	–	–	✓	–	$72.45_{\uparrow 0.89}$
TSET	✓	SS	–	–	✓	✓	$74.58_{\uparrow 3.02}$
TSET	✓	SS	–	✓	–	–	$75.11_{\uparrow 3.55}$
TSET	✓	SS	–	✓	–	✓	$75.46_{\uparrow 3.90}$
TSET	✓	SS	–	✓	✓	–	$76.24_{\uparrow 4.68}$
TSET	✓	SS	–	✓	✓	✓	$76.68_{\uparrow 5.12}$
S-SEMI	✓	SS	SS	–	–	–	$72.51_{\uparrow 0.95}$
TSET	✓	SS	SS	–	✓	–	$73.14_{\uparrow 1.85}$
TSET	✓	SS	SS	–	✓	✓	$75.35_{\uparrow 3.77}$
TSET	✓	SS	SS	✓	–	–	$76.05_{\uparrow 4.49}$
TSET	✓	SS	SS	✓	–	✓	$76.35_{\uparrow 4.77}$
TSET	✓	SS	SS	✓	✓	–	$76.98_{\uparrow 5.42}$
TSET	✓	SS	SS	✓	✓	✓	$77.24_{\uparrow 5.68}$
RetinaNet	✓	✓	–	–	–	–*	78.36
S-SEMI	✓	✓	SS	–	–	–	$78.52_{\uparrow 0.16}$
TSET	✓	✓	SS	–	✓	–	$78.87_{\uparrow 0.51}$
TSET	✓	✓	SS	–	✓	✓	$78.86_{\uparrow 0.50}$
TSET	✓	✓	SS	✓	–	–	$79.37_{\uparrow 1.01}$
TSET	✓	✓	SS	✓	–	✓	$79.76_{\uparrow 1.40}$
TSET	✓	✓	SS	✓	✓	–	$80.35_{\uparrow 1.99}$
TSET	✓	✓	SS	✓	✓	✓	$\mathbf{80.73}_{\uparrow 2.37}$

* RetinaNet is trained with focal loss in supervised manner. "Detection loss" here is an adapted version of focal loss for unsupervised training.

includes more labeled images.

We show the results on the VOC2007 test set in Table II. We use the abbreviation "SS" to show that the data are used as unlabeled images in semi-supervised training. For each training setup, we show the baseline performance of the supervised object detector, *i.e.*, RetinaNet, in the first row. From the results, we obtain the following observations.

(1) We first find that the performance of the S-SEMI is rather limited compared to the baseline model (RetinaNet). This indicates that a better teacher model is necessary to make full use of the unlabeled data in semi-supervised training.

(2) Compared to the RetinaNet, our TSET model obtains a large-margin performance improvement across all semi-supervised training setups. Particularly, our model shows its advantage when a limited quantity of labeled data is available. For example, when only using 2007train/val as labeled data and using 2012train/val as unlabeled data in setup (a), the mAP of our model is 76.68%, which is 5.12% better than the baseline performance .

(3) We can see that each of the basic modules in our TSET independently improves the SSOD. Furthermore, the concurrency of these basic modules results in the best performance. This means that the performance of our TSET model is not limited by the upper bound performance of each basic module; instead, the intrinsic integration of the proposed strategies leads to the dramatic improvement of our model.

(4) As for the self-ensembling strategies, the temporal teacher predictions ensembling seems to improve performance more than the temporal teacher model weights ensembling. For example, in setup (a), the mAP of solely employing the former self-ensembling strategy is 75.11% which exceeds that of solely employing the latter self-ensembling strategy by 3.01%. This result shows that ensembling the temporal teacher predictions significantly improves the teacher predictions on unlabeled images, which benefits the training for the student. The observation that solely employing the temporal teacher model weights ensembling results in a limited performance improvement implies that the teacher may learn limited knowledge from the student.

(5) Comparing the results obtained with setups (a) and (b), our model is improved when using more unlabeled images, with the mAP increasing from 76.68% to 77.24%. Comparing the results in setup (b) and (c), our model gains a large-margin improvement from 77.24% to 80.73% when using more labeled images. These results suggest that solely increasing the quantity of unlabeled images for an SSOD system may lead to the performance improvement reaching a local maximum. The employment of a certain number of labeled images may guide the detector to escape from this dilemma. The key factor is that the supervised training using the extra labeled images lifts the lower bound in our TSET model. One may notice a limitation of our model: when an object detector is better optimized using more labeled data, its performance improvement using our semi-supervised model may reach an upper bound. However, our proposed TSET model still holds

TABLE III
STATE-OF-THE-ART COMPARISON ON THE VOC2007 TEST SET

Model	Backbone Network	mAP
CSD-RFCN	ResNet-101-RFCN	74.70
STAC*	ResNet-50-FPN	$77.45_{\uparrow 2.45}$
TSET	ResNet-50-RetinaNet	$76.68_{\uparrow 1.98}$
TSET	ResNet-101-RetinaNet	$78.18_{\uparrow 3.48}$

* The model is trained with strong data augmentation.

TABLE IV
PERFORMANCE COMPARISON ON THE VOC2007 TEST SET

Model	07train/val	12train/val	12test	mAP
SSD-300	✓	–	–	70.20
CSD-SSD-300	✓	SS	–	$72.30_{\uparrow 2.10}$
RFCN	✓	–	–	73.90
CSD-RFCN	✓	SS	–	$74.70_{\uparrow 0.80}$
SSD-512	✓	–	–	73.30
CSD-SSD-512	✓	SS	–	$75.80_{\uparrow 2.50}$
Faster-RCNN*	✓	–	–	78.21
STAC*	✓	SS	–	$77.45_{\downarrow 0.76 (1.65)}$
RetinaNet	✓	–	–	71.56
TSET	✓	SS	–	$76.68_{\uparrow 5.12 (0.88)}$
RetinaNet†	✓	–	–	73.51
TSET†	✓	SS	–	$78.18_{\uparrow 4.67 (2.38)}$
RetinaNet	✓	✓	–	78.36
TSET	✓	✓	SS	**80.73**$_{\uparrow 2.37 (4.93)}$

* The model is trained with strong data augmentation. The AP of Faster-RCNN and STAC is 43.40 and $44.64_{\uparrow 1.24}$ evaluated by standard COCO metric (averaged average precision over varying thresholds of IoU).
† The model is trained using ResNet-101 as backbone network.

its potential to dramatically improve object detection when there is only a very limited number of labeled images.

Comparison on the VOC2007 test set. In Table III, we compare the performance of our TSET model with the competing SSOD models on the VOC2007 test set. All models are trained under the same setup, *i.e.*, using 2007train/val as labeled images and 2012train/val as unlabeled images. We list the backbone network that is used in the detector for a clear comparison. From the results, we see that the mAP of our model with a relatively shallow backbone network (ResNet-50) outperforms the CSD model by 1.98%. When using the same backbone network (ResNet-50), the mAP of our model drops slightly by 0.77% compared to STAC. We note that STAC uses very strong data augmentation to enhance the training in both supervised and unsupervised stages. Instead, we only use random horizontal flip as augmentation. When using a deeper backbone network (ResNet-101), our model shows the best performance among all methods. This illustrates that our model retains its potential to further improve performance when introducing proper training strategies. We think that the same augmentation as used in STAC may further improve the performance of our model.

Considering that the initialization state of the detection network may affect the performance of an SSOD model, in this experiment, we compare different SSOD models with their baseline detectors. For the detection networks, our TSET model uses RetinaNet [8]; The CSD model uses three different types of detectors, SSD-300, SSD-512 [11] and RFCN [58], and STAC uses Faster-RCNN [7]. In Table IV, we show the performance of these detectors trained in a supervised manner, as well as the performance of the SSOD models trained on top of these detectors. We observe the following key findings.

(1) The employment of unlabeled images for the SSOD models indeed improves the performance of the object detector. In Table IV, we use vertical arrows to indicate the absolute performance increase of the SSOD model from its detection network trained using a fully supervised manner. When trained using 2012train/val as unlabeled data, the mAP of our TSET model is 76.68%; this is 0.88% better than the best CSD model, CSD-SSD-512. We report these comparisons in italics within parentheses in Table IV.

(2) It is clear that the baseline detectors have different performances. Specifically, when trained using the same labeled images, the SSD-512, Faster-RCNN, and RetinaNet obtain an mAP of 73.30%, 76.30% and 71.56% respectively. Compared to the CSD and STAC models, our TSET model improves much more from its baseline detector. For example, under the same training setup of using the 2012train/val as unlabeled images, our TSET model outperforms it baseline by 5.12%; the CSD-SSD512 model outperforms its baseline by 2.50%; and the STAC model outperforms its baseline by 1.24%. This shows that our model boosts the performance of an object detector with an ill-posed initialization.

(3) By using a deeper backbone network (ResNet-101) to obtain a well-initialized detector, the mAP of our TSET reaches 78.18%. By using more labeled images to obtain a well-initialized detector, the mAP of our TSET reaches 80.73%. This is a remarkable performance improvement, and is a new state-of-the-art performance on the VOC2007 test set under the semi-supervised setup. In this experiment, we note that the performance of RetinaNet in our model is very similar to that of the SSD-512 (73.51% vs. 73.30%) and Faster-RCNN (78.36% vs. 78.21%). This means that compared to other SSOD models, our TSET model may perform better when using the same detection network initialized with the same initialization state.

C. Experiments on MSCOCO

Considering that the MSCOCO dataset is a more challenging benchmark, we conduct experiments to determine an efficient backbone network for the detection model. Here, we choose to use ResNet-50 and ResNet-101 for comparison. We use the standard evaluation metrics for the COCO dataset to illustrate the results: AP (averaged average precision over varying thresholds of IoU), AP_{50} (AP of IoU=0.5), AP_{75} (AP of IoU=0.75), AP_S (AP for "small size" objects), AP_M (AP for "medium size" objects), and AP_L (AP for "large size" objects). As for the training data, we configure two different setups: (a) train set as labeled images and val set as unlabeled images; and (b) train/val set as labeled images and the extra unlabeled images as unlabeled data. In Table V, we show the experimental results of the fully supervised object detector and our TSET model under various training setups. From the results, we can draw the following conclusions.

TABLE V
PERFORMANCE EVALUATION FOR TSET WITH VARYING BACKBONE NETWORKS ON THE COCO *minival5k* SET

Model	Backbone	train	val	unlabeled	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet	ResNet-50	✓	–	–	34.51	53.26	36.54	17.96	37.29	46.56
TSET	ResNet-50	✓	SS	–	$35.42_{\uparrow 0.91}$	$53.88_{\uparrow 0.62}$	$37.40_{\uparrow 0.86}$	$18.87_{\uparrow 0.91}$	$40.16_{\uparrow 2.87}$	$48.70_{\uparrow 2.14}$
RetinaNet	ResNet-50	✓	✓	–	36.34	55.22	38.90	19.66	39.94	48.95
TSET	ResNet-50	✓	✓	SS	$36.96_{\uparrow 0.62}$	$55.70_{\uparrow 0.48}$	$39.42_{\uparrow 0.52}$	$19.59_{\downarrow 0.07}$	$40.76_{\uparrow 0.82}$	$50.12_{\uparrow 1.17}$
RetinaNet	ResNet-101	✓	✓	–	39.03	58.31	41.66	22.01	42.83	51.87
TSET	ResNet-101	✓	✓	SS	$40.14_{\uparrow 1.11}$	$59.58_{\uparrow 1.27}$	$42.78_{\uparrow 1.12}$	$23.93_{\uparrow 1.92}$	$44.70_{\uparrow 1.92}$	$50.99_{\downarrow 0.88}$
TSET*	ResNet-101	✓	✓	SS	**40.52**$_{\uparrow 1.49}$	**59.93**$_{\uparrow 1.62}$	**43.48**$_{\uparrow 1.82}$	**24.13**$_{\uparrow 2.12}$	**45.47**$_{\uparrow 2.64}$	**52.97**$_{\uparrow 1.10}$

* Use extra training image augmentation, *i.e.*, random image resizing.

(1) By comparing TSET with RetinaNet, we can see that our TSET model outperforms its fully supervised counterpart on the COCO dataset when using the same backbone network in the detection model and using the same training data. For example, when using ResNet-50 as the backbone network, the AP of our TSET model is 35.42% and 36.96% under training setups (a) and (b), respectively. These results are superior to those obtained by the RetinaNet trained under the same conditions, by 0.91% and 0.62%, respectively. When using ResNet-101 as the backbone network, our model outperforms the baseline method by 1.11%, which suggests that our TSET model is generic enough to improve SSOD regardless of the specific type of backbone network.

(2) When using more labeled images to train the RetinaNet on the COCO dataset, RetinaNet's performance is remarkably improved, and its AP reaches 39.03%. This thus improves the lower bound performance of our TSET model, whose AP finally reaches 40.14%. To further improve the performance of our model, we use random resizing to augment the training images. These results are shown in the last row in Table V, and are indicated by an asterisk. In this case, the AP of our TSET model hits 40.52% , thus exceeding the fully supervised baseline by 1.49%.

(3) We observe that our TSET model with a deeper backbone network like ResNet-101 trained using more labeled data can obtain a greater performance improvement on detecting small and medium size objects. Such behavior may imply that the difficult examples of small and medium objects can be properly decoded and distilled to the student under such a training setup. We will further reason and generalize this behavior on other training setups of the COCO dataset in future work.

D. Qualitative Results

In Fig. 4 (A) and (B), we visualize the detection results from the VOC2007 test set and the COCO2014 *minival5k* set. We present the results from an image obtained by the RetinaNet and our TSET model side by side for an easy comparison. We show several different cases for a fair and comprehensive comparison. Case I: our model can successfully detect the small, difficult objects. Case II: our model can alleviate the false positive detections that the RetinaNet misclassified. Case III: some examples that our model fails to detect. We show these different detection results in the top, middle, and bottom rows of Fig. 4 (A) and (B), respectively. These visual effects demonstrate the effectiveness of our TSET model to improve object detection under the semi-supervised setting. The self-ensembling strategies and employment of focal loss in our model formulate a better teacher model that yields better predictions on unlabeled images for difficult examples (such as small objects with severe shape deformations, and the objects with occlusions). Among these detection results, we find that our model fails to separate skis from each other, and fails to recognize some animals with severe occlusions, like the dog and sheep. We assume this phenomenon is caused by the class imbalance in the unlabeled images, which may be solved in future work by taking the quantity of training examples from each class into account.

V. CONCLUSIONS AND DISCUSSIONS

We propose the TSET model to tackle the challenge of SSOD. We have two fundamental goals for TSET. First, based on the KD framework, the student is regularized by the teacher to better generalize the objects in unlabeled images. Second, the student needs to intimate the whole behavioral patterns of the teacher on predicting the unlabeled images rather than only learning high-confidence predictions from the teacher. To these ends, the proposed TSET model first ensembles temporal teacher predictions and temporal teacher model weights, which increases data and model diversity. This produces much better teacher predictions than those of the student, and accordingly increases the upper bound to optimize the student. Moreover, TSET adapts the focal loss to formulate the consistency loss between teacher and student predictions. Such a method retains all useful information, such as the information encoded in low-confidence hard examples from unlabeled images, which aligns the behaviors of teacher and student and mitigates the data imbalance issue in SSOD. Experimental results show that our model sets a new state-of-the-art SSOD performance on the VOC2007 test set (mAP of 80.73%), and obtains a dramatic improvement on the COCO2014 *minival5k* set (mAP increase of 40.52%). A possible direction to further improve our work would balance between ensembling multiple heterogeneous models and training efficiency. Second, we could take the categorical balance in unlabeled images into account and apply other strong augmentations to leverage the detection of objects with large scales.

(a) Detection results from VOC2007 test set

(b) Detection results from COCO2014 *minival5k* set

Fig. 4. **Detection results comparison of TSET model and its fully supervised counterpart, the RetinaNet, on PASCAL VOC and MSCOCO datasets**. The green bounding boxes indicate the detections from the RetinaNet, and the blue bounding boxes denote the detections of our TSET model. We arrange the detection results of the same image side by side for a convenient read. For each dataset, we show the examples from following cases: The TSET model recalls difficult objects (Top row); The TSET model alleviates false positives (Middle row); The TSET may fail to detect the objects with severe occlusions (Bottom row).

ACKNOWLEDGMENTS

We thank Prof. Yu Liu from International School of Information Science and Engineering, Dalian University of Technology, and Dr. Wei Chen from Leiden Institute of Advanced Computer Science, Leiden University for the insightful suggestions and proof reading on this work. We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

REFERENCES

[1] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, 2020. 1, 3

[2] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep face recognition: A survey," in *Proc. - SIBGRAPI Conf. Graph., Patterns Images, SIBGRAPI*. IEEE, 2018, pp. 471–478. 1

[3] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1907–1915. 1

[4] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, and X. Wang, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020. 1

[5] J. Song, Y. Yang, Z. Huang, H. T. Shen, and J. Luo, "Effective multiple feature hashing for large-scale near-duplicate video retrieval," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1997–2008, 2013. 1

[6] J. Yu, W. Zhang, Y. Lu, Z. Qin, Y. Hu, J. Tan, and Q. Wu, "Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval," *IEEE Trans. Multimedia*, 2020. 1

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2015, pp. 91–99. 1, 3, 5, 10

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988. 1, 3, 5, 6, 7, 10

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969. 1, 3, 5

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788. 1, 3, 5

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 21–37. 1, 3, 5, 10

[12] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 734–750. 1, 3

[13] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162. 1

[14] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010. 1, 2, 3, 7

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 740–755. 1, 2, 3, 7

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015. 1, 4

[17] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983. 1

[18] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig *et al.*, "The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale," *Int. J. Comput. Vis.*, vol. 128, pp. 1956–1981, 2020. 1

[19] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Advances Neural Inf. Process. Syst.*, 2017, pp. 1195–1204. 2, 4

[20] G. French, M. Mackiewicz, and M. Fisher, "Self-ensembling for visual domain adaptation," no. 6, 2018. 2, 6

[21] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4119–4128. 2, 4

[22] P. Goyal, D. Mahajan, A. Gupta, and I. Misra, "Scaling and benchmarking self-supervised visual representation learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6391–6400. 2

[23] J. Jeong, S. Lee, J. Kim, and N. Kwak, "Consistency-based semi-supervised learning for object detection," in *Proc. Advances Neural Inf. Process. Syst.*, 2019, pp. 10 758–10 767. 2, 3, 4, 6, 7, 8

[24] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," pp. 9729–9738, 2020. 2, 6

[25] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?-Weakly-supervised learning with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 685–694. 2

[26] Y. Tang, X. Wang, E. Dellandrea, and L. Chen, "Weakly supervised learning of deformable part-based models for object detection via region proposals," *IEEE Trans. Multimedia*, vol. 19, no. 2, pp. 393–407, 2016. 2

[27] F. Wan, P. Wei, J. Jiao, Z. Han, and Q. Ye, "Min-entropy latent model for weakly supervised object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1297–1306. 2

[28] D. Zhang, J. Han, L. Zhao, and D. Meng, "Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework," *Int. J. Comput. Vis.*, vol. 127, no. 4, pp. 363–380, 2019. 2

[29] Y. Zhang, K. Jia, and Z. Wang, "Part-aware fine-grained object categorization using weakly supervised part detection network," *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1345–1357, 2019. 2

[30] X. J. Zhu, "Semi-supervised learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2005. 2, 3

[31] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009. 2, 3

[32] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]," *IEEE Trans. Neural Netw.*, vol. 20, no. 3, pp. 542–542, 2009. 2, 3

[33] D. Mandal, P. Rao, and S. Biswas, "Semi-supervised cross-modal retrieval with label prediction," *IEEE Trans. Multimedia*, vol. 22, no. 9, 2019. 2

[34] S. Min, X. Chen, H. Xie, Z.-J. Zha, and Y. Zhang, "A mutually attentive co-training framework for semi-supervised recognition," *IEEE Trans. Multimedia*, vol. 14, no. 8, 2020. 2

[35] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, Proc. Internat. Conf. Mach. Learn.*, vol. 3, 2013, p. 2. 2

[36] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Deep Learning and Representation Learning Workshop, Proc. Advances Neural Inf. Process. Syst.*, 2015. 2, 3

[37] M. Phuong and C. Lampert, "Towards understanding knowledge distillation," in *Proc. Internat. Conf. Mach. Learn.*, 2019, pp. 5142–5151. 2, 4

[38] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2604–2613. 2

[39] M. Yuan and Y. Peng, "CKD: Cross-task knowledge distillation for text-to-image synthesis," *IEEE Trans. Multimedia*, no. 8, 2019. 2

[40] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Int. Conf. Learn. Represent.*, 2017. 2, 4

[41] P. Tang, C. Ramaiah, R. Xu, and C. Xiong, "Proposal learning for semi-supervised object detection," *arXiv preprint arXiv:2001.05086*, 2020. 2, 4

[42] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee, and T. Pfister, "A simple semi-supervised learning framework for object detection," *arXiv preprint arXiv:2005.04757*, 2020. 2, 4, 7

[43] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, 2001, pp. I–I. 3

[44] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587. 3

[45] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448. 3

[46] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125. 3, 7

[47] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7263–7271. 3

[48] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018. 3

[49] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 850–859. 3

[50] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 840–849. 3

[51] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6569–6578. 3

[52] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2015, pp. 3546–3554. 3

[53] H. Valpola, "From neural pca to deep unsupervised learning," in *Proc. Advances Neural Inf. Process. Syst.* Elsevier, 2015, pp. 143–171. 3

[54] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *Proc. Advances Neural Inf. Process. Syst.*, 2016, pp. 1163–1171. 3, 4

[55] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4320–4328. 5

[56] X. Lan, X. Zhu, and S. Gong, "Knowledge distillation by on-the-fly native ensemble," in *Proc. Advances Neural Inf. Process. Syst.*, 2018, pp. 7528–7538. 5

[57] B. Zhuang, C. Shen, M. Tan, L. Liu, and I. Reid, "Towards effective low-bitwidth convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7920–7928. 5

[58] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2016, pp. 379–387. 5, 10

[59] FAIR. (2018) Faster R-CNN and Retina network in PyTorch 1.0: Model zoo and baselines. [Online]. Available: https://github.com/facebookresearch/maskrcnn-benchmark 7

[60] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2874–2883. 7

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778. 8

Cong Chen received the M.S. degree in Computer Science from the South Central University For Nationalities, Wuhan, China, in 2010. He is a senior researcher at Keya Medical Technology. His expertise is in developing advanced algorithms of computer vision, image processing and machine learning, and further translating these techniques into engineering and products. His current research projects range from blood vessel segmentation, lung nodule detection and recognition, to weakly supervised pneumonia classification.

Shouyang Dong received his M.S. degree in Electrical Engineering from the Paris Saclay University, Paris, France, in 2018. He is currently working at Cambricon Technology as a software engineer. His research interests include computer vision, machine learning and machine learning system.

Ye Tian received his B.S. degree from the Beijing Jiaotong University, Beijing, China, in 2016 and the M.S. degree from the Johns Hopkins University, MD, US, in 2018. He is currently a senior researcher at Hippocrates Research Lab of Tencent. His research interests include video understanding, action recognition and assessment, and reinforcement learning.

Kunlin Cao received the Ph.D. and B.S. degree from University of Iowa,Iowa City, USA, in 2008 and 2012 respectively. She is currently working at Keya Medical Technology as the Principal Investigator of Artificial Intelligence. Her research interests include computer vision and medical image processing.

Li Liu (Senior Member, IEEE) received the Ph.D. degree in Information and Communication Engineering from the National University of Defense Technology, China, in 2012. She is currently a Professor with the College of System Engineering. During her Ph.D. study, she spent more than two years as a Visiting Student with the University of Waterloo, Canada, from 2008 to 2010. From 2015 to 2016, she spent ten months visiting the Multimedia Laboratory with the Chinese University of Hong Kong. From December 2016 to November 2018, she worked as a Senior Researcher with the Machine Vision Group, the University of Oulu, Finland. Her current research interests include computer vision, pattern recognition, and machine learning. Her papers have currently over 3800+ citations in Google Scholar.

Yuanhao Guo (Member, IEEE) received his Ph.D. degree in Leiden Institute of Advanced Computer at the Leiden University, Leiden, the Netherlands, in 2017 and received his M.S. degree in Information Science and Engineering at the Shandong University, Jinan, China, in 2012. He is currently an assistant professor affiliated at National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include computer vision, deep learning, image processing as well as their applications in biological and biomedical imaging.