Target Detection Model Distillation Using Feature Transition and Label Registration for Remote Sensing Imagery

Boya Zhao, Qing Wang, Yuanfeng Wu, *Senior Member, IEEE*, Qingqing Cao, and Qiong Ran

***Abstract*—Deep convolution networks have been widely used in remote sensing target detection for various applications in recent years. Target detection models with many parameters provide better results but are not suitable for resource-constrained devices due to their high computational cost and storage requirements. Furthermore, current lightweight target detection models for remote sensing imagery rarely have the advantages of existing models. Knowledge distillation can improve the learning ability of a small student network from a large teacher network due to acceleration and compression. However, current knowledge distillation methods typically use mature backbones as teacher and student networks are unsuitable for target detection in remote sensing imagery. In this article, we propose a target detection model distillation (TDMD) framework using feature transition and label registration for remote sensing imagery. A lightweight attention network is designed by ranking the importance of the convolutional feature layers in the teacher network. Multiscale feature transition based on a feature pyramid is utilized to constrain the feature maps of the student network. A label registration procedure is proposed to improve the TDMD model's learning ability of the output distribution of the teacher network. The proposed method is evaluated on the DOTA and NWPU VHR-10 remote sensing image datasets. The results show that the TDMD achieves a mean Average Precision (mAP) of 75.47% and 93.81% on the DOTA and NWPU VHR-10 datasets, respectively. Moreover, the model size is 43% smaller than that of the predecessor model (11.8 MB and 11.6 MB for the two datasets).**

***Index Terms*—Deep neural network, feature transition, label registration, model distillation, remote sensing, target detection.**

I. Introduction

REAL-TIME target detection is an essential task of intelligent remote sensing satellite systems, which are exhibiting rapid development [1]. Target detection models based on deep convolutional neural networks [2], [3] can extract features effectively and have resulted in breakthroughs in remote sensing image processing [4], [5]. However, models with better performances typically have deeper neural network structures and a large number of parameters, increasing the model's inference time and requiring extensive computational resources. Thus, it is challenging to achieve real-time data processing using intelligent remote sensing satellite data.

Many lightweight neural networks using efficient and lightweight backbones have been proposed to minimize the computational resources, such as SqueezeNet [6], MobileNet [7], and ShuffleNet [8]. Network compression is also an effective approach for reducing the number of model parameters and the computational cost. Knowledge distillation refers to network acceleration and compression by transferring knowledge from a larger teacher network to a smaller student network. The learning ability and generalization performance of the student model are usually lower than that of the teacher network, but the number of model parameters and the computational cost are lower. The goal of knowledge distillation is to transfer useful knowledge in the teacher network to the student network to improve the capabilities of the student network. The knowledge in knowledge distillation methods has three types: relationship, response, and feature knowledge.

Hinton *et al.* [9] first proposed a knowledge distillation method. The output classification probabilities of a large teacher network were transferred to the student network to improve the classification accuracy of the latter. FitNets [10] utilizes middle-layer features of the teacher network as knowledge and uses the differences in specific middle-layer features between teacher and student networks as a feature loss for training, improving the feature extraction capabilities of the student network. He *et al.* [11] exploited the differences between teacher and student networks and compressed features of the teacher network into information for the student network by using an autoencoder. An affinity distillation module was proposed to capture the long-range dependency by calculating the nonlocal interactions in the entire image. Pairwise [12] and holistic distillation [13] were proposed for dense prediction. The pairwise distillation method distills pairwise similarity [14], [15] by establishing a static graph. Subsequently, the holistic distillation, which uses adversarial training, distills the overall knowledge to the student network.

Manuscript received 24 May 2022; revised 22 June 2022; accepted 29 June 2022. Date of publication 5 July 2022; date of current version 18 July 2022. This work was supported in part by the National Key R&D Program of China under Grant 2021YFA0715203 and in part by the National Natural Science Foundation of China under Grant 62001455 and Grant 41871245. *(Corresponding author: Yuanfeng Wu.)*

Boya Zhao, Yuanfeng Wu, and Qingqing Cao are with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the School of Electronic, Electrical, and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: zhaoby@aircas.ac.cn; wuyf@radi.ac.cn; caoqingqing@hnu.edu.cn).

Qing Wang and Qiong Ran are with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China (e-mail: 2019210520@mail.buct.edu.cn; ranqiong@mail.buct.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3188252

Target detection tasks need to classify and locate a certain target at the same time [16]. Chen *et al.* [17] applied knowledge distillation to target detection by designing classification, regression, and feature losses to train the student network. The output of the region proposal network (RPN) and recursive cortical network from the teacher network were utilized to calculate classification and regression losses, respectively. Fine-grained feature imitation [18] improves the feature-level distillation by a region estimation method, which estimates the region close to the target instead of the entire feature map. Zhang and Ma [19] focused on two knowledge distillation problems in target detection tasks. The first is the imbalance between the number of foreground and background pixels, and the second is the lack of the relationship between different pixels during training. Attention-guided distillation and nonlocal distillation were proposed to address these problems. The attention-guided distillation method finds the important pixels in foreground targets [20] using an attention mechanism and ensures that the student network focuses on these features. Nonlocal distillation enables the student network to learn the features of an individual pixel and the relationship between different pixels captured by non-local modules. Salehi *et al.* [21] proposed multiple intermediate hints for anomaly detection and location to leverage the teacher network knowledge. General instance distillation and the general instance selection module [22] were proposed to exploit feature-based, relation-based, and response-based knowledge, significantly improving the performance of the student network. Guo *et al.* [23] proposed a distilling target detector using decoupled features since the feature information derived from regions excluding targets is essential for training the student network. They found that the knowledge learned by the teacher network consisted of features from neck and proposals from the classification head and determined the total loss consisting of the feature loss, classification loss, RPN loss, and regression loss. The neuron selectivity transfer [24] regards knowledge distillation as a distribution matching problem. The knowledge distillation loss was calculated by minimizing the maximum mean difference between the distributions of the neuron selectivity patterns between teacher and student networks, improving the accuracy of knowledge distillation training. Activation boundaries [25] were proposed combining teacher/student transformation, distillation feature position, and a distance function. The teacher network features were transformed by margin ReLU activation functions.

Existing knowledge distillation methods for target detection tasks have used mature backbones as teacher networks (e.g., ResNet [26], VGGNet [27]) and student networks (e.g., ShuffleNet [8], MobileNet [7]). However, these methods do not fully utilize the knowledge of target detection models for remote sensing imagery. It is also crucial to balance detection accuracy and efficiency to achieve real-time data processing of intelligent remote sensing satellite data.

We propose an oriented target detection method based on knowledge distillation (TDMD) using the multiscale context and enhanced channel attention (MSCCA) method [5]. To deploy efficient target detection models on resource-constrained devices. The lightweight attention network uses a channel attention mechanism that extracts important channels from the teacher network. The parameters of these channels are the initial convergence point of the lightweight attention network. Furthermore, multiscale feature transition on the feature pyramid is performed to constrain the feature maps. Label distribution registration is proposed to constrain the output distribution of the lightweight attention network. The TDMD is evaluated on the DOTA [28] and NWPU VHR-10 [29] datasets. Experimental results show that the proposed TDMD achieves a mean average precision (mAP) of 75.47% and 93.81% mAP on the DOTA and NWPU VHR-10 datasets, respectively. The model size is 43% smaller than the predecessor model (11.8 MB and 11.6 MB for the two datasets).

The rest of this article is organized as follows. Section II describes the TDMD architecture. Section III presents the datasets and experimental results. Section IV describes the ablation study conducted for various hyperparameters. Finally, Section V concludes this article.

II. Methods

As shown in Fig. 1, the TDMD is a knowledge transfer framework for target detection in remote sensing images. It is composed of 1) a lightweight attention network, 2) a multiscale feature transition module, and 3) label distribution registration. An attention mechanism is adopted in the lightweight attention network to reduce the number of channels, significantly reducing the computational cost and number of model parameters. The multiscale feature transition module uses the teacher's feature as the standard and ensures that the feature pyramid of the lightweight attention network is similar to that of the teacher network to maintain the efficiency of the convolutional features. Label distribution registration uses the outputs of the teacher network to constrain the outputs of the lightweight attention network using the cross-entropy loss.

A. Lightweight Attention Network

Remote sensing target detection models are typically customized to the application scenario. The lightweight attention network is based on the MSCCA [5], an effective remote sensing target detection model. The channel attention mechanism has been used in convolutional neural networks to determine the importance of the feature layers. In this article, the enhanced channel attention (ECA) module is used for this task.

For any given feature map $X \in \mathbb{R}^{H\times W\times C}$, the channel attention is defined as follows:

$$S(X) = \sigma W_2(\delta W_1(F_{gp}(X))) \tag{1}$$

where H and W are the height and width of the feature map; C is the number of channels of the feature map; F_{gp} represents the global average pooling; W_1 and W_2 represent two fully connected layers; σ is the Sigmoid function; δ is the ReLU function. At last, $S \in \mathbb{R}^{1\times 1\times C}$ denotes the weights of the feature channels. The ECA modules are inserted after each feature layer to obtain their weights.

The lightweight attention network removes feature channels with a low S to reduce the number of model parameters and the computational cost. The resulting network is identical

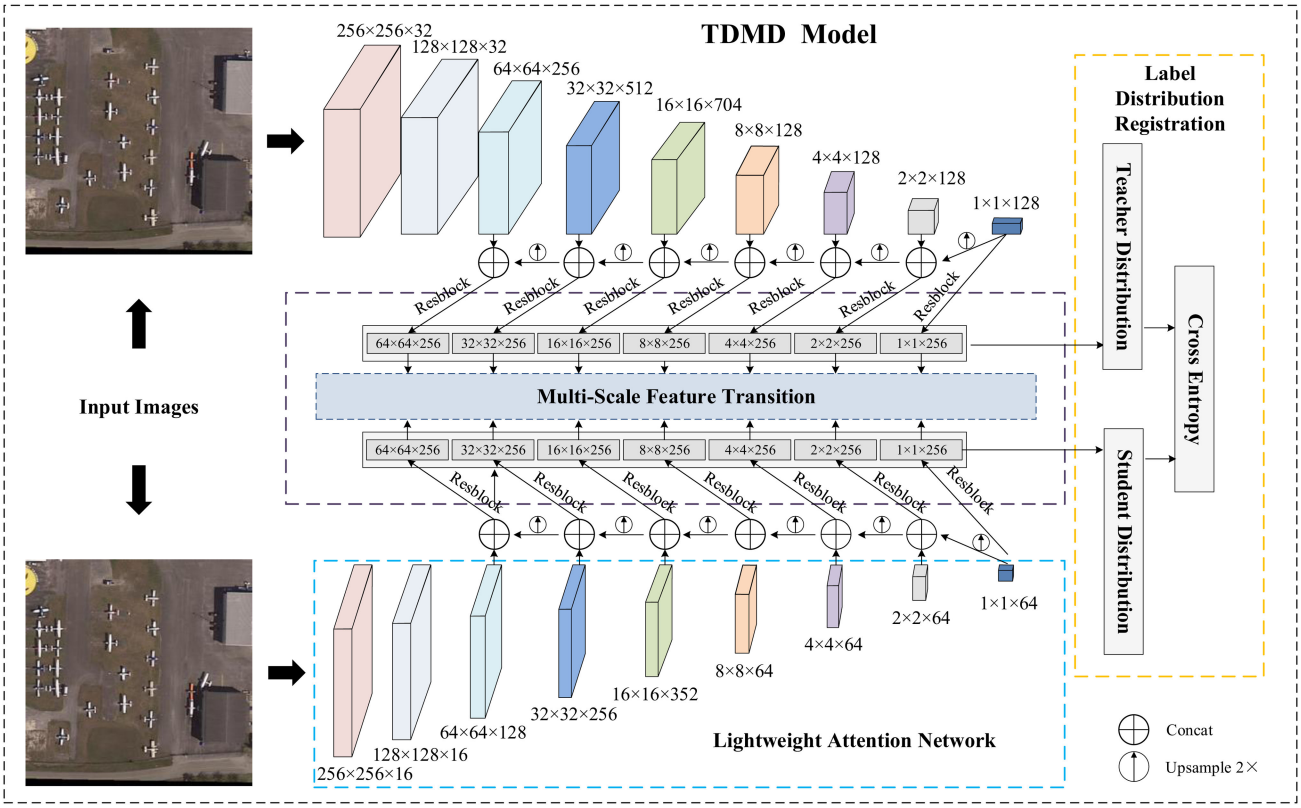

Fig. 1. TDMD architecture with lightweight attention network, multiscale feature transition, and label distribution registration.

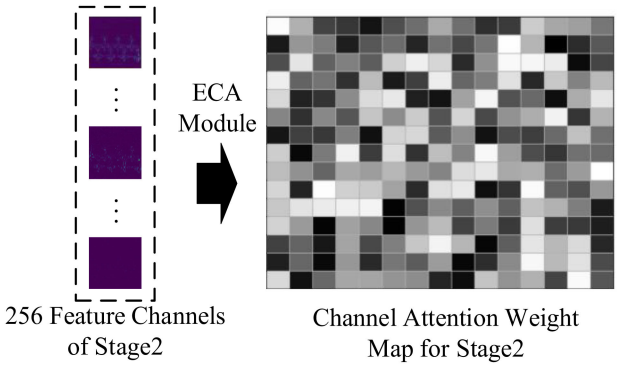

Fig. 2. Channel attention weight map of Stage 2 output features in MSCCA.

to the MSCCA. Table I lists the structure of the MSCCA and lightweight attention network of the TDMD for different lightweight rates. The lightweight rate is an index to represent the model parameter reduction. For example, lightweight rate (50%) indicate that the feature channels in the lightweight attention network is reduced to 50% of the original network. The bold values are the differences between each setting, which represent the number of feature channels. Stage 2, Stage 3, Stage 4, FP 5, FP 6, FP 7, and FP 8 are integrated into the feature pyramid.

Fig. 2 shows the channel weight map for the stage 2 layer in the MSCCA model; there are 256 channels. The ECA module calculates the attention weights of the features. The channel attention map contains 256 values, which indicate the importance

TABLE I
LIGHTWEIGHT ATTENTION NETWORK WITH VARIOUS LIGHTWEIGHT RATE

Network structure		MSCCA	TDMD (50%)	TDMD (25%)
Stage	**Layer**	**Output Feature Map**		
Stage 0	Stem Layer×1	128×128×**32**	128×128×**16**	128×128×**8**
Stage 1	Dense Layer×3	64×64×**256**	64×64×**128**	64×64×**64**
Stage 2	Dense Layer×4	64×64×**256**	64×64×**128**	64×64×**64**
Stage 3	Dense Layer×8	32×32×**512**	32×32×**256**	32×32×**128**
Stage 4	Dense Layer×6	16×16×**704**	16×16×**352**	16×16×**176**
FP5	1×1 Conv×1 3×3 Conv×1 FC×2	8×8×**128**	8×8×**64**	8×8×**32**
FP6	1×1 Conv×1 3×3 Conv×1 FC×2	4×4×**128**	4×4×**64**	4×4×**32**
FP7	1×1 Conv×1 3×3 Conv×1 FC×2	2×2×**128**	2×2×**64**	2×2×**32**
FP8	1×1 Conv×1 3×3 Conv×1 FC×2	1×1×**128**	1×1×**64**	1×1×**32**

Algorithm 1: Multi-Scale Feature Transition.

Input: Current image A; Detection network Net; Euclidean function E.
Output: Multi-Scale feature transition loss L_{mft}.
Repeat iterations:
1 Extract each feature layer $F_{TDMD_n}$and F_{MSCCA_n} in each detection model:$F_n = Net(A)$, $n = 1 \ldots 7$ for feature pyramid.
2 Bottom-up and top-down feature fusion concatenation by a upsample function of the bilinear interpolation for each feature layer:$U_n = Upsample_2(F_{n+1}) \oplus F_n$, n= 1 …7.
3 Integrate each U_n by ResBlocks: $C_n = Res(U_n)$, n= 1 …7.
4 Compare each integrated feature layers between C_{MSCCA_n} and $C_{TDMD_n}$by E: $I_n = E(C_{MSCCA\_n}, C_{TDMD\_n})$, n= 1 …7.
5 Formulate L_{mft}:$L_{mft} = \sum_n I_n$, n= 1 …7.
6 Use L_{mft} to match the feature maps of MSCCA and TDMD.

of each feature map. The dark color represents high weights S, and the light color represents low weights S.

Pretraining is widely used in deep learning. Pretrained parameters are used in the MSCCA. Because the number of channels of the lightweight attention network is lower than that of the MSCCA, pretrained parameters are selected by channel weights. These parameters correspond to the feature channels.

B. Multiscale Feature Transition

Feature extraction is a crucial step in deep learning target detection methods. The ability of the convolutional feature determines the target detection performance. The results of many knowledge distillation methods for classification tasks have shown that student features can be constrained better by a larger teacher network. Hint layers are used in the teacher network, and guide layers of the lightweight attention network are selected to transfer the knowledge of the intermediate layer features from the teacher network to the student network. The hint layers are defined as intermediate feature layers of the MSCCA model, and the corresponding feature layers in the lightweight attention network are defined as the guide layers. As mentioned in Section II-A, the lightweight attention network is identical to the MSCCA, except for the number of feature channels. A multiscale feature pyramid is used in the MSCCA and TDMD during feature extraction. The feature pyramids of the MSCCA and TDMD have the same dimension to improve the convergence efficiency. The aim of multiscale feature transition is to ensure the consistency of the features between the TDMD and MSCCA. Algorithm 1 shows the processing flow of the multiscale feature transition algorithm. The transition procedure is based on fusion features, which adopt concatenation and bilinear interpolation functions. Then, the Euclidean distance is used for comparing the differences and formulating the loss function.

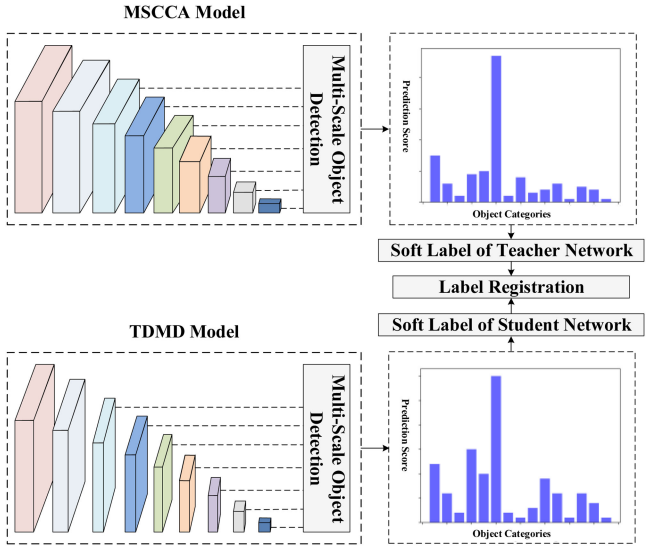

Fig. 3. Label distribution registration architecture.

Compression in the lightweight attention only occurs in the backbone; thus, the size of ResBlock [26] and the subsequent features are not changed. There is no need to match the feature dimension between TDMD and MSCCA during multiscale feature transition. For a seven-layer feature pyramid, the dimensions of the output features after the ResBlocks are $64 \times 64 \times 256$, $32 \times 32 \times 256$, $16 \times 16 \times 256$, $8 \times 8 \times 256$, $4 \times 4 \times 256$, $2 \times 2 \times 256$, and $1 \times 1 \times 256$, respectively. The Euclidean distance is used to measure the disparities between the TDMD and MSCCA to restrain the feature maps. For feature maps $X, Y \in \mathbb{R}^{H \times W \times C}$, the Euclidean distance constraint can be formulated as follows:

$$E(X, Y) = \frac{1}{\mathrm{C}} \sum_{n=1}^{C} \|X_n - Y_n\|_2^2 \tag{2}$$

where C is the total channel number of features X and Y and n is the serial number of features X and Y. The constraint ensures the convergence of the feature pyramid of TDMD and MSCCA.

C. Label Distribution Registration

Label distribution registration is proposed to enhance the accuracy of the detection result. This step consists of hard label and soft label registrations. The hard label is the ground truth, and the soft label is the output label distribution after the Softmax function

$$y_k = \frac{e^{a_k}}{\sum_{i=1}^{N} e^{a_i}} \tag{3}$$

where y is the soft label, and$a$ is the original output of the neural network; k and N are the kth output and the total number of outputs, respectively.

As shown in Fig. 3, the prediction scores between MSCCA and TDMD are transferred to the soft labels by the Softmax function. The label registration loss is based on the soft label difference between MSCCA and TDMD. Thus, the TDMD reduces the influence of the erroneous results on the MSCCA

output. The loss function of the label distribution registration is explained in the loss function section.

In contrast to classification tasks [30]–[32], target detection produces more negative samples than positive samples, especially in remote sensing target detection. Similar to the online hard example mining (OHEM) [33], the label distribution registration only chooses backpropagation samples; the ratio of positive to negative samples is 1:3.

D. Loss Function

The loss function of TDMD is divided into two parts. The first is the detection loss L_{det}, consisting of the location loss L_{loc}, and classification loss L_{cls}. The second is the knowledge distillation loss L_{kd} consisting of the multiscale feature transition loss L_{mft} for Section II-B and the label distribution registration loss L_{ldr} for Section II-C

$$L_{\text{TDMD}} = L_{det} + L_{kd}. \tag{4}$$

Similar to the single shot multibox detector (SSD) [34], the detection loss is formulated as follows:

$$L_{det}\left(f, g, c\right) = \frac{1}{n}\left(L_{cls}\left(f, c\right) + \alpha L_{loc}\left(f, g\right)\right) \tag{5}$$

where n is the number of training anchors in each feature map; c is the classification result of the anchor; f is the feature of the anchor; g is the ground truth.

The location loss L_{loc} is based on a smooth L1 loss [35] between the ground truth and the predicted bounding box. It is defined as follows:

$$L_{loc}(f, g) = \sum_{i \in pos}^{N} \sum_{k} f_{ij}^{\text{label}} \text{smooth}L_1(\text{pre}_i^k - g_j^k) \tag{6}$$

where

$$\text{smooth}L_1 = \begin{cases} 0.5x^2(|x| \leq 1) \\ |x| - 0.5(|x| > 1) \end{cases} \tag{7}$$

where pos represents the positive samples; N is the total number of anchors; $k \in \{x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4\}$ denotes the coordinates of the four vertices of the quadrilateral ground truth for oriented target detection; $f_{ij}^{\text{label}} \in \{1, 0\}$ is the indicator for the ith anchor box to the jth ground truth of label; pre_i^k and g are the prediction box and ground truth.

The classification loss L_{cls} is determined by the multiclass Softmax function. The ratio of the positive to negative samples is 3:1, which is similar to OHEM [33]

$$L_{cls}\left(f, c\right) = -\sum_{i \in\ pos}^{N} f_{ij}^{\text{label}} log\left(\hat{c}_i^{\text{label}}\right) - \sum_{i \in\ neg} \log\left(\hat{c}_i^{\text{label}}\right) \tag{8}$$

where

$$\hat{c}_i^{\text{label}} = \frac{exp\left(c_i^{\text{label}}\right)}{\sum_{\text{label}} exp\left(c_i^{\text{label}}\right)} \tag{9}$$

where $\hat{c}_i^{\text{label}}$ is the anchor classification prediction score of the label in the ith anchor box. Moreover, the background label is 0; thus, if $i \in neg$, the $\hat{c}_i^{\text{label}}$=$\hat{c}_i^0$.

In addition to the target detection loss, the distillation loss L_{kd} of TDMD contains a multiscale feature transition loss and a label distribution registration loss. It is defined as follows:

$$L_{kd}(x, s, t) = L_{mft}(x) + L_{ldr}(s, t) \tag{10}$$

where x is the feature; s and t are the soft labels of student net and teacher net, respectively.

The multiscale feature transition loss uses the Euclidean distance. For each feature pyramid layer, the loss is formulated as follows:

$$L_{mft}(x) = \sum_{p}\left(\frac{1}{C}\sum_{n=1}^{C}\left\|x_{pn}^s - x_{pn}^t\right\|_2^2\right) \tag{11}$$

where C represents the number of channels of the current layer; x_{Pn}^s represents the hint layer features n in pyramid p of the MSCCA model; x_{Pn}^t represents the guide layer features n in pyramid p of the TDMD model.

The label distribution registration loss employs cross-entropy for soft labels

$$L_{ldr}(s) = -\sum_{i \in\ pos}^{N} t_i^{\text{label}} \log\left(s_i^{\text{label}}\right) - \sum_{i \in\ neg} \log\left(s_i^{\text{label}}\right) \tag{12}$$

where

$$s^{\text{label}}, t^{\text{label}} = \frac{\exp(z^{\text{label}}/T)}{\sum_{\text{label}} \exp(z^{\text{label}}/T)} \tag{13}$$

where t_i^{label} represents the soft label of the MSCCA model for the ith anchor of label; s_i^{label} represents the soft label of the TDMD model for the ith anchor of label. The output class probabilities are generated by the Softmax function, which applies to the logit output. Then, T is a soft-parameter that controls the smoothness of the probability distribution of the classes. z^{label} is the logit output of the MSCCA or TDMD model.

III. Experiments

The proposed TDMD is evaluated on the DOTA and NWPU VHR-10 datasets. The learning policies and results are described in detail.

A. Datasets

1) DOTA Dataset: DOTA is a real dataset for target detection in remote sensing imagery. It contains 2806 remote sensing images acquired from various platforms. The image resolution ranges from 800 × 800 to 4000 × 4000, and there are various targets with different orientations and shapes. A large number of small targets are labeled and classified into 15 categories: plane (PL), bridge (BD), ground-track-field (GTF), harbor (HA), large vehicle (LV), small-vehicle (SV), ship (SH), storage tank (ST), baseball field (SBF), tennis court (TC), basketball court (BC), helicopter (HC), roundabout (RA), soccer ball field (SBF), and swimming pool (SP).

The training images were cropped to 512 × 512 to maintain the input image size.

TABLE II
DETECTION RESULTS OF DOTA DATASET

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP	Fps
YOLOv2[36]	76.9	33.8	22.7	34.8	38.7	32.0	52.3	61.6	48.5	33.9	29.2	36.8	36.4	38.2	11.6	39.2	30
RetinaNET[37]	78.22	53.41	26.38	42.27	63.64	52.63	73.19	87.17	44.64	57.99	18.03	51.00	43.39	56.56	7.44	50.39	14
R-FCN[38]	81.01	58.96	31.64	58.97	49.77	45.04	49.29	68.99	52.07	67.42	41.83	51.44	45.15	53.30	33.89	52.58	9
YOLOv3[39]	79.0	77.1	33.9	68.1	52.8	52.2	49.8	89.9	74.8	59.2	55.5	49.0	61.5	55.9	41.7	60.0	13
R^2CNN[40]	80.94	65.67	35.34	67.44	59.92	50.91	55.81	90.67	72.39	66.92	55.06	52.23	55.14	53.35	48.22	60.67	13.0
LO-Det 608[41]	89.22	66.14	31.32	55.96	70.05	71.04	84.27	90.74	75.09	81.28	44.65	59.34	59.98	65.13	48.82	66.17	**60**
DSSD[42]	**91.1**	71.8	54.6	66.4	79.0	77.2	87.5	87.6	52.1	69.7	38.0	72.6	75.4	59.4	28.9	67.4	9
ROI Trans[43]	88.53	77.91	37.63	74.08	66.53	62.97	66.57	90.50	79.46	76.75	59.04	56.73	62.54	61.29	55.56	67.74	7.1
DYOLO[44]	86.0	71.4	54.6	52.5	**79.2**	80.6	87.8	82.2	54.1	75.0	51.0	69.2	66.4	59.2	51.3	68.1	17
FPN[45]	88.7	75.1	52.6	59.2	69.4	78.8	84.5	90.6	81.3	82.6	52.5	62.1	76.7	66.3	60.1	72.0	6
FMSSD[46]	89.11	81.51	48.22	67.94	69.23	73.56	76.87	90.71	82.67	73.33	52.65	67.52	72.37	80.57	60.15	72.43	16
SCRDet[47]	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	**87.94**	86.86	65.02	66.68	66.25	68.24	65.21	72.61	7.4
DRN[48]	89.71	82.34	47.22	64.10	76.22	74.43	85.84	90.57	86.18	84.89	57.65	61.93	69.30	69.63	58.48	73.23	9.8
Pelee[49]	87.65	72.99	52.84	73.71	73.59	77.97	76.38	90.09	80.75	74.44	40.75	68.09	71.78	79.65	83.78	74.00	29.2
FR-EST[50]	89.63	81.17	50.44	70.19	73.52	77.98	86.44	90.82	84.13	83.56	60.64	66.59	70.59	66.72	60.55	74.20	—
BBAVectors[51]	88.63	84.06	52.13	69.56	78.26	80.40	**88.06**	**90.87**	87.23	86.39	56.11	65.62	67.10	72.08	63.96	75.36	11.7
R3Det[52]	89.80	83.77	48.11	66.77	78.76	**83.27**	87.84	90.82	85.38	85.51	65.67	62.68	67.53	78.56	72.62	76.47	10
SCRDET++[53]	90.05	**84.39**	**55.44**	**73.99**	77.54	71.11	86.05	90.67	87.32	**87.08**	**69.62**	68.90	73.34	71.29	65.08	**76.81**	13
TDMD-no KD	88.90	75.11	52.28	65.28	73.86	79.22	79.64	89.75	71.10	69.85	41.41	73.10	72.92	81.86	84.20	73.23	34.5
TDMD	89.62	77.31	54.12	68.30	76.01	80.44	80.00	90.21	74.51	73.60	46.92	**75.50**	**75.81**	**82.92**	**86.82**	75.47	34.5

2) NWPU VHR-10 Dataset: The NWPU VHR-10 is a public target detection dataset cropped from Google Earth and Vaihingen datasets. This dataset includes 650 labeled images. The number of samples for each class is less than that of the DOTA dataset. The NWPU VHR-10 dataset has almost no targets with an area of less than 1000 pixels. Those targets are classified into 10 types: PL, SH, ST, BD, TC, BC, ground track field (GT), HA, BR, and vehicle (VH).

B. Learning Policy

For the DOTA dataset, the feature pyramid contains seven layers, and the anchor settings are the same as MSCCA. The initial learning rate is 0.00005 with 120 000 iterations. The learning rate is subsequently reduced by one order of magnitude after every 40 000 iterations, and the total number of training iterations is 200 000. The optimization tricks have a momentum of 0.9 and a weight decay of 0.0005. The batch size is 16. In the label distribution registration training, the temperature T is 4.

The feature pyramid and anchor settings are the same for the NWPU VHR-10 dataset. The initial learning rate is 0.00005 with 80 000 iterations. The learning rate is subsequently reduced by one order of magnitude after every 20 000 iterations, and the total number of training iterations is 120 000. The optimization tricks have a momentum of 0.9 and a weight decay of 0.0005. The batch size is 16, and the temperature T is also 4 in the label distribution registration.

The stochastic gradient descent method is used for both datasets.

C. Results

The TDMD was compared with other current methods. All experiments were implemented on the Caffe framework. Recently proposed methods (you only look at once (YOLOv3) [39], LO-Det 608 [41], box boundary-aware vectors (BBAVectors) [51], and R3Det [52]) are selected for the comparison.

1) DOTA Result: Table II reports the mAP of the TDMD and other target detection models. The bold values denote the best performances. The first row lists the target classes, detection accuracy, and speed. The first column lists the name of the target detection model. The TDMD model achieves 75.47% mAP and 34.5 fps. Its detection speed ranks second after the LO-Det 608, but its detection accuracy is 9.3% higher. The detection speed of the TDMD is 5.3 fps faster, and its mAP is 1.47% higher than that of Pelee, a real-time detection model. The SCRDET++ achieves the highest mAP of 76.81%, but its detection speed is only 13 fps, much slower than that of the TDMD model. The proposed TDMD provides the best classification results for the RA, HA, SP, and HC classes. The mAP of the TDMD is only 1% lower for the PL, BD, and TC classes. The mAP of the TDMD without knowledge distillation (TDMD-no KD) is 2.24% lower than that of the TDMD with knowledge distillation.

TABLE III
DETECTION ACCURACY AND MODEL SIZE OF TDMD AND OTHER METHOD ON DOTA DATASET

Method	Model Size	mAP
HSD-Res-9-256[54]	8.5MB	65.5
TDMD	11.8MB	75.47
Pelee[49]	26.5MB	74.00
LO-Det 608[41]	26.9MB	66.17
Yolo v3 tiny_618[39]	35.0MB	65.9
SSD512[34]	99.8MB	20.8
R^2CNN[40]	170MB	60.67
Yolo v2[36]	192MB	21.39
Yolo v3[39]	235MB	60.0
ROI Trans[43]	273MB	67.74
SCRDet[47]	338MB	72.61
Fast R-CNN VGG16[55]	538MB	59.2

Table III lists the detection accuracy and model size of the TDMD and other models on the DOTA dataset. The first row shows the model size and mAP. The first column lists the name of the target detection model. The TDMD exhibits a 9.3% mAP improvement, and the model size is 15.1 MB lower compared with the recently proposed lightweight LO-Det 608 detector.

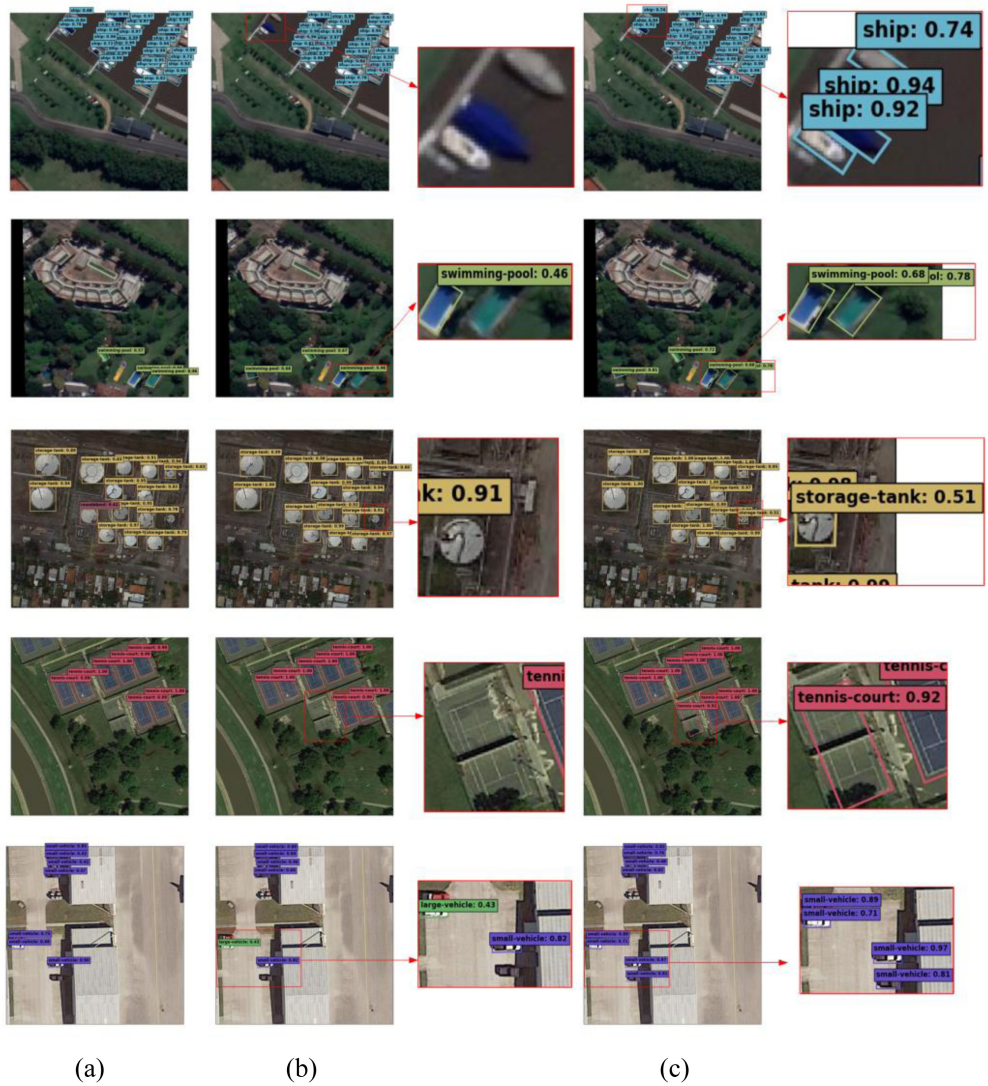

Fig. 4. Visualization results on the DOTA dataset. From left to right, the detection results of (a) Pelee, (b) TDMD-no KD, and (c) TDMD.

TABLE IV
DETECTION RESULTS OF NWPU VHR-10 DATASET

Method	PL	SH	ST	BD	TC	BC	GT	HA	BR	VH	mAP	Fps
EDAI-16[56]	76.9	72.8	60.3	69.5	59.6	64.7	78.4	67.1	71.4	68.5	68.9	13.3
RICNN[57]	88.35	77.34	85.27	88.12	40.83	58.45	86.73	68.6	61.51	71.1	72.63	-
COPD[29]	89.11	81.73	97.32	89.38	73.27	73.41	82.99	73.39	62.86	83.3	80.68	-
Faster R-CNN(R50)[55]	94.6	82.3	65.3	95.5	81.9	89.7	92.4	72.4	57.5	77.8	80.9	19.9
RICAOD[29]	99.70	90.80	90.61	92.91	90.29	80.31	90.81	80.29	68.53	87.14	87.12	-
HyperNet[58]	99.4	89.7	**98.6**	90.9	90.6	90.3	89.2	80.3	68.9	88.6	88.7	-
Yolov3[39]	90.91	90.91	90.81	**99.13**	**90.86**	90.91	**99.47**	90.05	**90.91**	90.35	92.43	7
Pelee[49]	99.52	**93.46**	90.88	97.25	90.72	96.3	95.95	88.91	88.9	90.75	93.26	29.2
TDMD-no KD	99.56	90.42	90.80	90.50	90.62	98.33	95.63	90.38	89.75	97.80	93.39	**34.5**
TDMD	**99.70**	90.98	91.60	90.78	90.69	**98.67**	96.52	**90.39**	90.86	**97.91**	**93.81**	**34.5**

Compared with the HSD-Res-9-256 detector, the TDMD's mAP shows a 9.97% mAP improvement, and the model size is 3.3 MB higher.

Fig. 4 shows the visualization results of the Pelee, TDMD-no KD and TDMD. In general, the proposed TDMD model detects more targets than the other models.

2) NWPU VHR-10 Result: Table IV list the results for the NWPU VHR-10 dataset. The first row shows the target classes, detection accuracy, and speed. The first column lists the name of the target detection model. The TDMD model achieves 93.81% mAP and 34.5 fps, the optimum performance. The detection speed of the TDMD is 5.3 fps faster, and the mAP is 0.55% higher

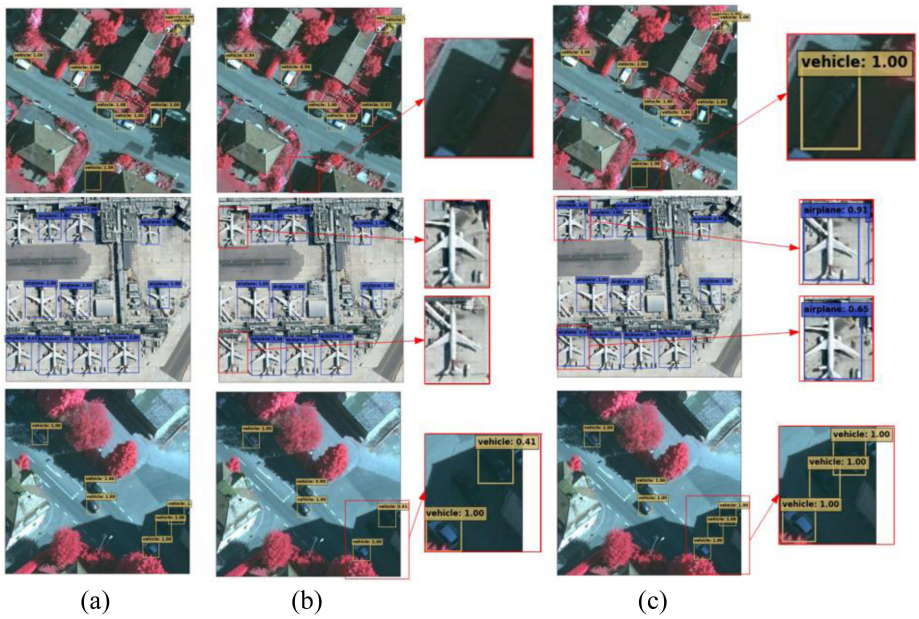

Fig. 5. Visualization results on the NWPU VHR-10 dataset. From left to right, the detection results of (a) Pelee, (b) TDMD-no KD, and (c) TDMD.

TABLE V
DETECTION ACCURACY AND MODEL SIZE OF TDMD AND OTHER METHOD ON NWPU VHR-10 DATASET

Method	Model Size	mAP
TDMD	11.6MB	93.81
Pelee[49]	26.4MB	93.26
EDAI-16[56]	90MB	68.9
Faster R-CNN[55]	161.2MB	80.9
Yolov3[39]	246MB	92.43

than that of Pelee. The TDMD model achieves the best results for the PL, BC, UA, and VH classes. The TDMD has a 0.42% higher mAP after implementing multiscale feature transition and label distribution registration.

Table V reports the detection accuracy and model size on the NWPU VHR-10 dataset. The first row shows the model size and mAP. The first column lists the name of the target detection model. The TDMD model achieves the highest mAP of 93.81% mAP with a model size of only 11.6 MB.

Fig. 5 shows the visualization results of the Pelee, TDMD-no KD and TDMD methods on the NWPU VHR-10 dataset. The TDMD method detects more vehicles and planes than the other methods.

IV. DISCUSSION

Additional comparison experiments are conducted on the DOTA and NWPU VHR-10 datasets to evaluate the performance of multiscale feature transition and label distribution registration. The TDMD is tested on Nvidia Titan Xp and Jetson TX2, as shown in Fig. 6.

An ablation study is conducted as follows.

1) Lightweight rate: The lightweight attention network has variable sizes. The percentage is used to represent different rates. The effect of the number of initial pretrained parameters of the teacher network is also evaluated. We use suffix (-no param) to represent the TDMD trained without pretrained parameters.
2) Feature transition: We use the suffix (-mft) to indicate the TDMD model with the multiscale feature transition.
3) Label registration: To evaluate the effect of the label distribution registration, we use the suffix (-ldr) to represent the TDMD model with label distribution registration.

Fig. 6. Inference devices. (a) Nvidia Titan Xp. (b) Jetson TX2.

Table IV lists the results of the ablation study on the DOTA dataset. The TDMD models with initial pretrained parameters have substantially higher detection accuracies. The detection performance is higher for the larger model size. TDMD (75%)-mft&ldr, TDMD (50%)-mft&ldr, and TDMD (25%)-mft&ldr achieve mAP values of 77.28%, 75.47%, and 66.62%, respectively. The multiscale feature transition and label distribution registration improve the detection performance. For example, for the 50% model size, the TDMD (50%)-ldr. TDMD (50%)-mft, and TDMD (50%)-mft&ldr exhibit improvements in the mAP values of 1.44%, 0.72%, and 2.24% compared to model TDMD (50%)(with param). Then, the detection speeds are also reported in the following table and the smaller model size corresponds to the faster detection speed.

TABLE VI
ABLATION EXPERIMENTAL RESULTS ON DOTA DATASET

Settings	mAP	Size	Params	Titan	TX2
TDMD (25%)(no param)	55.00	8.7 MB	2.03M	35.5 fps	7.2 fps
TDMD (25%)(with param)	63.56				
TDMD (25%)-ldr	66.02				
TDMD (25%)-mft	64.43				
TDMD (25%)-mft&ldr	66.62				
TDMD (50%)(no param)	68.25	11.8 MB	2.84M	34.5 fps	7.0 fps
TDMD (50%)(with param)	73.23				
TDMD (50%)-ldr	74.67				
TDMD (50%)-mft	73.95				
TDMD (50%)- mft&ldr	75.47				
TDMD (75%)(no param)	72.66	16.8 MB	3.95M	32.8 fps	6.7 fps
TDMD (75%)(with param)	76.23				
TDMD (75%)-ldr	76.64				
TDMD (75%)-mft	76.42				
TDMD (75%)-mft&ldr	77.28				

TABLE VII
ABLATION EXPERIMENTAL RESULTS ON NWPU VHR-10 DATASET

Settings	mAP	Size	Params	Titan	TX2
TDMD (25%)(no param)	88.65	8.5 MB	1.99M	35.5 fps	7.2 fps
TDMD (25%)(with param)	90.20				
TDMD (25%)-ldr	91.60				
TDMD (25%)-mft	91.35				
TDMD (25%)-mft&ldr	92.01				
TDMD (50%)(no param)	91.24	11.6 MB	2.79M	34.5 fps	7.0 fps
TDMD (50%)(with param)	93.39				
TDMD (50%)-ldr	93.75				
TDMD (50%)-mft	93.52				
TDMD (50%)- mft&ldr	93.81				
TDMD (75%)(no param)	92.32	16.6 MB	3.91M	32.8 fps	6.7 fps
TDMD (75%)(with param)	93.55				
TDMD (75%)-ldr	94.06				
TDMD (75%)-mft	93.74				
TDMD (75%)-mft&ldr	94.22				

Table VII reports the results of the ablation study on the NWPU VHR-10 dataset. The results are similar to that of the DOTA dataset. The TDMD (75%)-mft&ldr, TDMD (50%)-mft&ldr, and TDMD (25%)-mft&ldr achieve mAP values of 94.22%, 93.81, and 92.01%, respectively. Moreover, because the number of target classes is lower in the NWPU VHR-10 dataset than in the DOTA dataset, the number of classification and location parameters is lower, resulting in a smaller model size for the NWPU VHR-10.

V. CONCLUSION

This article proposed a lightweight target detection model for remote sensing images called TDMD based on knowledge distillation. A lightweight attention network was designed. The multiscale feature transition method learns knowledge from the MSCCA using a feature pyramid and Euclidean distance constraint. Label distribution registration employs a soft label that controls the smoothness of the probability distribution. The results showed that the proposed TDMD achieved 75.47% and 93.81% mAP on the DOTA and NWPU VHR-10 datasets. The model sizes were only 11.8 MB and 11.6 MB on these datasets, 43% smaller than that of the predecessor model. In a future study, we will investigate nonstructured pruning of the lightweight attention network to obtain a smaller model size and lower computational cost.

REFERENCES

[1] B. Zhang *et al.*, "Progress and challenges in intelligent remote sensing satellite systems," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1814–1822, 2022.

[2] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, Art. no. 25.

[4] H. Ghanbari, M. Mahdianpari, S. Homayouni, and F. Mohammadimanesh, "A meta-analysis of convolutional neural networks for remote sensing applications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3602–3613, 2021.

[5] Q. Ran, Q. Wang, B. Zhao, Y. Wu, S. Pu, and Z. Li, "Lightweight oriented object detection using multiscale context and enhanced channel attention in remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5786–5795, 2021.

[6] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 MB model size," 2016, *arXiv:1602.07360*.

[7] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[8] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6848–6856.

[9] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1–9.

[10] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," in *Proc. Int. Conf. Learn. Representat.*, 2014, pp. 1–13.

[11] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan, "Knowledge adaptation for efficient semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 578–587.

[12] Y. Han, C. Deng, B. Zhao, and B. Zhao, "Spatial-temporal context-aware tracking," *IEEE Signal Process. Lett.*, vol. 26, no. 3, pp. 500–504, Mar. 2019.

[13] Y. Liu, C. Shu, J. Wang, and C. Shen, "Structured knowledge distillation for dense prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2020.3001940.

[14] Y. Han, C. Deng, B. Zhao, and D. Tao, "State-aware anti-drift object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4075–4086, Aug. 2019.

[15] C. Deng, S. He, Y. Han, and B. Zhao, "Learning dynamic spatial-temporal regularization for UAV object tracking," *IEEE Signal Process. Lett.*, vol. 28, pp. 1230–1234, 2021.

[16] C. Deng, D. Jing, Y. Han, S. Wang, and H. Wang, "FAR-Net: Fast anchor refining for arbitrary-oriented object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6505805.

[17] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–10.

[18] T. Wang, L. Yuan, X. Zhang, and J. Feng, "Distilling object detectors with fine-grained feature imitation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4933–4942.

[19] L. Zhang and K. Ma, "Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–14.

[20] C. Deng, Y. Han, and B. Zhao, "High-performance visual tracking with extreme learning machine framework," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2781–2792, Jun. 2019.

[21] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee, "Multiresolution knowledge distillation for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14902–14912.
[22] X. Dai *et al.*, "General instance distillation for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7842–7851.
[23] J. Guo *et al.*, "Distilling object detectors via decoupled features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2154–2164.
[24] Z. Huang and N. Wang, "Like what you like: Knowledge distill via neuron selectivity transfer," 2017, *arXiv:1707.01219*.
[25] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1921–1930.
[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representat.*, 2015, pp. 1–14.
[28] G.-S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.
[29] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogrammetry Remote Sens.*, vol. 98, pp. 119–132, 2014.
[30] D. Hong, J. Yao, D. Meng, Z. Xu, and J. Chanussot, "Multimodal GANs: Toward crossmodal hyperspectral–multispectral image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5103–5113, Jun. 2020.
[31] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, "X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, pp. 12–23, 2020.
[32] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, "Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model," *ISPRS J. Photogrammetry Remote Sens.*, vol. 178, pp. 68–80, 2021.
[33] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 761–769.
[34] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[35] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, 1964.
[36] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7263–7271.
[37] 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.
[38] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, Art. no. 29.
[39] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, *arXiv:1804.02767*.
[40] Y. Jiang *et al.*, "R2CNN: Rotational region CNN for orientation robust scene text detection," 2017, *arXiv:1706.09579*.
[41] Z. Huang, W. Li, X. G. Xia, H. Wang, and R. Tao, "LO-Det: Lightweight oriented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603515.
[42] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "Dssd: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*.
[43] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2849–2858.
[44] O. Acatay, L. Sommer, A. Schumann, and J. Beyerer, "Comprehensive evaluation of deep learning based detection methods for vehicle detection in aerial imagery," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, 2018, pp. 1–6.
[45] 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.
[46] P. Wang, X. Sun, W. Diao, and K. Fu, "FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, May 2019.
[47] X. Yang *et al.*, "Scrdet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8232–8241.
[48] X. Pan *et al.*, "Dynamic refinement network for oriented and densely packed object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11207–11216.
[49] R. J. Wang, X. Li, and C. X. Ling, "Pelee: A real-time object detection system on mobile devices," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, Art. no. 31.
[50] K. Fu, Z. Chang, Y. Zhang, and X. Sun, "Point-based estimator for arbitrary-oriented object detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4370–4387, May 2020.
[51] J. Yi, P. Wu, B. Liu, Q. Huang, H. Qu, and D. Metaxas, "Oriented object detection in aerial images with box boundary-aware vectors," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 2150–2159.
[52] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3163–3171.
[53] X. Yang, J. Yan, W. Liao, X. Yang, J. Tang, and T. He, "Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022, to be published, doi: 10.1109/TPAMI.2022.3166956.
[54] R. Zhang, X. Jiang, J. An, and T. Cui, "A light anchor-free detection network for remote sensing images via heatmap-saliency distillation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6503905.
[55] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, Art. no. 28.
[56] L. Li, G. Cao, J. Liu, and Y. Tong, "Efficient detection in aerial images for resource-limited satellites," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 6001605.
[57] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.
[58] T. Kong, A. Yao, Y. Chen, and F. Sun, "Hypernet: Towards accurate region proposal generation and joint object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 845–853.

Boya Zhao was born in 1990. He received the B.S. degree in electronic and information engineering from the Hebei University of Technology, Tianjin, China, in 2013 and the Ph.D. degree in information and communication engineering from Beijing Institute of Technology, Beijing, China, in 2019.

He is currently an Assistant Professor with the Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing. His research interests include the object detection in complex background and on-board real-time information processing.

Qing Wang received the B.S. degree in software engineering from Qingdao Institute of Technology, China, in 2017. He is currently working toward the M.S. degree in computer technology from College of Information Science and Technology at Beijing University of Chemical Technology, Beijing, China.

Yuanfeng Wu (Senior Member, IEEE) received the B.S. and M.S. degrees in computer science from China University of Mining and Technology, Beijing, China, in 2004 and 2007, respectively, and the Ph.D. degree in cartography and geographical information system from the Graduate University of Chinese Academy of Sciences, Beijing, in 2010.

He is currently an Associate Professor with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include the onboard real-time algorithms, high-performance computing implementation and hyperspectral image processing.

Qingqing Cao received the B.S. degree in communication engineering from the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China, in 2022. She is currently working toward the M.S. degree in electronic and information engineering with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

Her research interests include hyperspectral image processing and object detection.

Qiong Ran received the Ph.D. degree in cartography and geographic information system from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2009.

She is currently with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing. She has authored and coauthored more than 10 papers in China and abroad. Her research interests include, hyperspectral image analysis, and applications.