Efficient Object Detection in Optical Remote Sensing Imagery via Attention-Based Feature Distillation

Pourya Shamsolmoali, Jocelyn Chanussot, Huiyu Zhou, Yue Lu

Efficient Object Detection in Optical Remote Sensing Imagery via Attention-based Feature Distillation

Pourya Shamsolmoali, *Member, IEEE,* Jocelyn Chanussot, *Fellow, IEEE,* Huiyu Zhou, and Yue Lu, *Senior Member, IEEE*

***Abstract*—Efficient object detection methods have recently received great attention in remote sensing. Although deep convolutional networks often have excellent detection accuracy, their deployment on resource-limited edge devices is difficult. Knowledge distillation (KD) is a strategy for addressing this issue since it makes models lightweight while maintaining accuracy. However, existing KD methods for object detection have encountered two constraints. First, they discard potentially important background information and only distill nearby foreground regions. Second, they only rely on the global context, which limits the student detector's ability to acquire local information from the teacher detector. To address the aforementioned challenges, we propose Attention-based Feature Distillation (AFD), a new KD approach that distills both local and global information from the teacher detector. To enhance local distillation, we introduce a multi-instance attention mechanism that effectively distinguishes between background and foreground elements. This approach prompts the student detector to focus on the pertinent channels and pixels, as identified by the teacher detector. Local distillation lacks global information, thus attention global distillation is proposed to reconstruct the relationship between various pixels and pass it from teacher to student detector. The performance of AFD is evaluated on two public aerial image benchmarks, and the evaluation results demonstrate that AFD in object detection can attain the performance of other state-of-the-art models while being efficient.**

***Index Terms*—Deep neural network, object detection, knowledge distillation, remote sensing images.**

I. Introduction

RECENTLY, due to the advancement of deep convolution neural networks (CNNs), significant progress has been made in object detection in remote sensing images [1]–[4]. Nevertheless, most of cutting-edge CNNs, require a large amount of processing power, preventing them from being used on mobile phones and embedded systems. Knowledge Distillation (KD) [5], Weight pruning [6], and model quantization [7], are a few examples of the model compression strategies developed to address this problem. KD in particular has gained popularity as a method for both model compression and model accuracy improvement because of its simplicity and efficacy. In the KD [5], [8], [9], a heavyweight teacher network's prediction logits are used to train a smaller, more manageable student network. Therefore, the teacher network's soft labels can assist the student network in making decisions like the teacher network, leading to better performance despite the student network's relatively few parameters.

The detection of objects and classification of object types in remote sensing images is complicated due to the presence of multiple objects distributed across various locations. This results in vagueness and imbalance in the details of detection. The representations of different positions, such as background, foreground, centers, or borders, may have varying contributions, making the task of KD challenging. The conventional KD approaches [10]–[12] were established for the classification tasks (see Fig. 1(a)), due to a lack of localization performance, cannot be used for the detection tasks. For example, hint learning [13] is suggested to distill the transitional feature maps, but it does not pass the localization and classification knowledge of the teacher detector to the student detector. In order to address this concern, [8] introduces a new approach to object detection that improves feature extraction, information localization, and classification. Still, because of the disparity between the background and foreground, [8] is not able to efficiently extract the teacher's knowledge. In [14], a feature distillation method is developed, which uses ground truth to filter background regions in order to only perform distillation from the efficient foreground regions. However, this solution does not solve the issue of assigning equal weights to different target regions. Consequently, in [15], the authors suggest applying mechanism of attention to global features in order to build soft weighted masks, whereby these masks facilitate the access of information from certain and highly important locations. However, we have noticed two main problems that arise when relying only on global feature contexts, potentially resulting in the loss of important information within the teacher's features. Firstly, there is a tendency to primarily concentrate on foreground areas while disregarding the background. Neglecting the background is unfavorable for accurate object detection in remote sensing images [16], [17] as it contains valuable information that should not be overlooked. Therefore, efficiently balancing and using all information from both the foreground and the background is the key to boosting distillation performance in object detection. Second, some significant local features that

Manuscript received 23 January 2023; accepted 28 October 2023.
P. Shamsolmoali and Y. Lu (*Corresponding author*) are with the School of Communication and Electronic Engineering, East China Normal University, Shanghai 200241, China. (Emails: (pshams, ylu)@cee.ecnu.edu.cn).
J. Chanussot is with the GIPSA-lab, Université Grenoble Alpes, CNRS, Grenoble INP 38000, France. (Email: jocelyn.chanussot@grenoble-inp.fr).
H. Zhou is with the School of Computing and Mathematical Sciences, University of Leicester, Leicester LE1 7RH, United Kingdom. (Email: hz143@leicester.ac.uk).

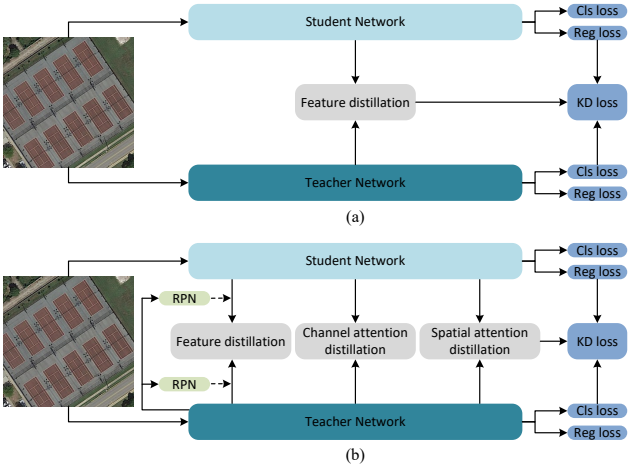

Fig. 1: KD detection pipelines. (Top) Conventional approaches. (Bottom) Our AFD method. The AFD focuses on obtaining information from the teacher network on both local and global basis.

are uniformly distributed in all regions might be overlooked given that the global mask-based approaches just pay attention to the features' global contexts. Applying the softmax function to the global region, would produce an enhanced mask that supplies considerable attention to a foreground object while ignoring the other objects and background areas [15].

To detect and classify objects in remote sensing images, we propose Attention-based Feature Distillation (AFD) to address the above constraints, as illustrated in Fig. 1(b). In AFD, we propose a new multi-instance attention strategy that is based on the detector's local and global context features. AFD applies an attention mechanism to local and global features to generate attention masks. In this procedure, the model estimates the attention of various channels and pixels of the teacher's feature map, enabling the student detector to more focus to the teacher's most significant channels and pixels. It also distills the relationship between various pixels from the teacher network and feeds it to the student network. To further extract the teacher's information, the created mask is applied on the extracted features, the Region Proposal Network (RPN) features, the classification output, and the regression output.

Additionally, we incorporate a feature map normalization technique and minimize the MSE loss between the normalized features. This approach aims to mitigate the adverse impact of magnitude disparities between the teacher and student detectors, as well as variations between different Feature Pyramid Network (FPN) layers and channels. In our AFD model, all loss functions exclusively operate on features, allowing for direct integration with different one/two-stage detectors. To evaluate AFD's performance, on two challenging benchmark aerial datasets (DOTA [18], NWPU VHR-10 [19]), a comprehensive set of experiments were conducted. The results show that AFD outperforms state-of-the-art KD approaches in object detection. The following is a summary of this paper's significant inventions and contributions:

- We introduce an attention-based model for distilling both local and global information from the teacher detector. As a result, student detector focuses more to the foreground objects and less to the background pixels.
- We introduce local and global distillation to enhance the student detector's attention to important teacher channels and pixels, while also fostering an understanding of pixel relationships.
- Comprehensive experiments conducted on two challenging benchmark datasets to thoroughly evaluate our approach. The results demonstrate impressive improvements over other detectors. To illustrate the impact of each module on our propose model's performance, we also performed a comprehensive ablation study.

The rest of this paper is structured as follows. Section II dedicated to the brief review of CNN and KD-based object detection methods in natural and remote sensing images. Section III describes the proposed AFD model. The dataset details, experimental and evaluation results are given in Section IV. Section V concludes the paper.

II. RELATED WORK

Given wide references on object detection models, we focus only on the most recent and closely relevant studies including CNN-based object detection and KD methods.

A. Object Detection

Current CNN-based object detection models, whether one-stage [20]–[22] or two-stage [23]–[25], need considerable processing resources to achieve desired performance, making them impractical for use on embedded devices with limited computation power. These detectors often have a strong backbone, such as VGGs [26] and ResNets [27]. Consequently, some researches focus on creating lightweight backbone. MobileNet [28] is a lightweight deep neural network that using depth-separable convolutions with a complementing search strategie. Single Shot multibox Detector (SSD) [20], MobileNetV2-SSD [29] and MobileNetV3 [30] are three examples of lightweight detectors created by combining MobileNet with one-stage detectors.

Existing object detection approaches often rely on adapting image classification frameworks [26], [27] to tackle detection tasks. But since classification and detection tasks are so distinct from one another, a lightweight backbone is not ideal for direct deployment. Hence, some lightweight detectors like Tiny-deeply supervised object detection [31] and Pelee [32] have developed specific backbones. To accomplish effective real-time detection, ThunderNet [33] proposes integrating a compacted backbone with a RPN.

These lightweight detectors often do not provide good detection results when used for remote sensing imagery due to its complex background and multiscale objects. In light of this need, several deep learning-based detectors have been proposed for remote sensing images. To better focus on tiny objects, the authors of [34] propose adopting the atrous spatial feature pyramid component and integrating multiscale context information through a loss weighted by region. Merging attention with deformable convolution for object detection is proposed in [35], via context-based deformable module on the basis of contextual information. Spatial misalignment between

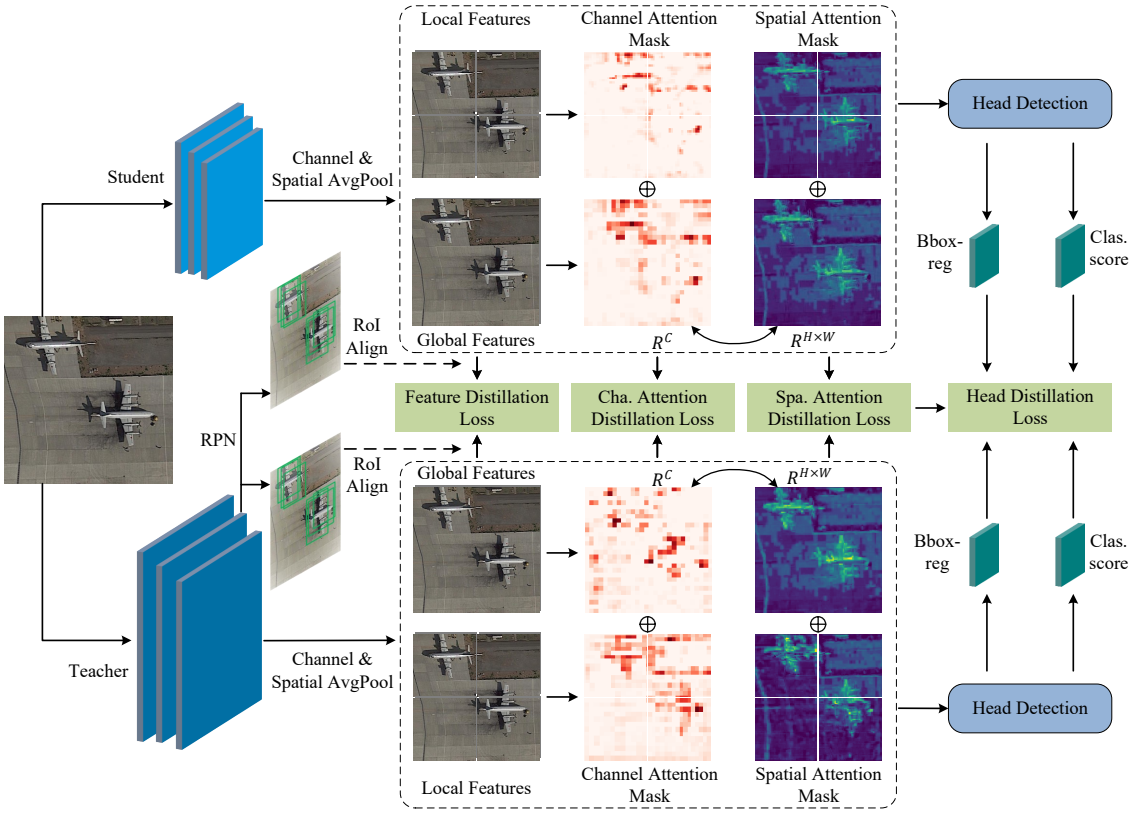

Fig. 2: Architecture of our KD method. The enhancement of AFD is based on three points. (a) Our new KD approach distils both local and global information from the teacher network. (b) For local distillation, a multi-instance attention mechanism is proposed to identify the background from the foreground. (c) Attention local and global distillation is proposed to reconstruct the relationship between various pixels and pass it from teacher to student detector.

anchors and ground truth is one of the problems of object detection in remote sensing images. To address this issue, in [36] a unique pseudo-anchor proposal module is introduced. To efficiently address the problem of rotated objects, [2] proposes a method that learns rotation invariants and trains the network by applying additional constraints. In addition, to boost geospatial recognition accuracy, [37] proposes a network based on local-contextual feature fusion and [1] introduce a new model to enhance feature maps quality for better object detection. In [38], a survey of object detection and tracking methods for remote sensing images is gathered that provides thoughts for models further development. In [39], a pyramid single-shot detector is proposed for small object detection in remote sensing images. To further enhance small object detection, in [40], an interactive U-Net architecture is proposed which has higher feature learning by utilising object's global context information. In [41], a detector is proposed that uses spatial-frequency channel features by incorporating both rotation-invariant channel features and original spatial channel features which enhances the system's robustness, and accuracy. However, these detectors are difficult to deploy on devices with limited storage and computational power.

B. KD for Object Detection

To develop precise and lightweight detectors for natural scenes, researchers have extensively utilize KD in recent years [8]. The application of KD in this particular task focuses on distilling the distinct locations of the detector. However, during the process of imitating the feature maps, the imbalanced distribution of foreground and background pixels is often disregarded, leading to inferior performance. To solve this problem, the authors of [14] suggest an imitation technique for fine-grained features, which focuses the detector's attention on the objects. To identify the central foreground pixels, a 2-dimension Gaussian mask is applied in the ground-truth regions for feature distillation in [42]. This strategy decreases the imbalance at the expense of eliminating the backgrounds. On the other hand, recent analysis [43], [44] has shown that the background areas contain important information. Specially, remote sensing objects are often connected to their environments. During the distillation process, it is important to pay attention to the areas around the objects and the background. In [45], Focal and Global Distillation is proposed that consists of, focal distillation for foreground-background separation, and global distillation for pixel relationship restoration.

In [8], the authors integrate the boundary regression loss of teacher detector with the regression elements and the transmission of unbounded regression data, which lacks distinctions between objects with varied levels of difficulty in regression. In [42], a successful classification and regression model is designed using the ℓ_1 and binary cross-entropy losses. These distillation approaches continue to underestimate the importance of background, which result in loss of contextual information around the objects. [44] introduces a distillation mask-based method that focuses on discriminative patches by

determining the differences between the teacher and student results. Moreover, in order to restrict the feature maps, a multiscale feature transition on the output of the FPN is applied. Similarly, in [15], a soft mask-based method is developed which extracts feature attention from its backbone. Conversely, current global attention masks generally overlook other significant regions as networks primarily have attention to small objects or regions. In order to further boost detection performance, we propose creating attention masks for all the local patches. These masks would direct attention to other important patches that contain local information.

III. METHODOLOGY

In this section, we outline the details of our local and global attention mask to accurately represent the characteristics of features. Then describe how the feature distillation and head distillation are accomplished. Fig. 2 represents the overview of our proposed AFD method.

A. Local and Global Attention-based Mask

A fundamental part of the proposed AFD is the local and global attention-based masks (LGAM), which we discuss below. LGAM incorporates channel attention M_{cha} with spatial attention M_{spa} methods. To get the channel attention masks, a softmax is applied to the channel dimension, as the average weight of the feature components $|x_{i,j}|$ over the channel dimension. The proposals $\mathcal{P}_x^T$ generated by the RPN of teacher are shared with the teacher and the student detector in order to obtain the same candidates for loss computation between the detectors. The RPN module has a positive impact on proposal quality, localization accuracy, and efficiency. It allows the student detector to benefit from the teacher's knowledge, leading to improved object detection performance.

$$M_{cha}(x) = HW.\rho\ (\frac{\frac{1}{HW}\cdot\sum_{i=1}^{H}\sum_{j=1}^{W}(|x_{i,j}, \mathcal{P}_x^T|)}{\mathcal{T}}), \quad (1)$$

in which $\rho(.)$ is the softmax operation and $\mathcal{T}$ denotes the temperature parameter. For an input feature, H and W denote its height and width. Consequently, the channel-wise feature components $|x_k|$ is utilized in the operations of softmax with the H and W dimensions to generate the spatial attention masks as written below:

$$M_{spa}(x) = C.\rho\ (\frac{\frac{1}{C}\cdot\sum_{c=1}^{C}(|x_c, \mathcal{P}_x^T|)}{\mathcal{T}}), \quad (2)$$

in which C denotes the feature's channel of input. To create LGAM that incorporate both local and global perspectives, we divide each FPN output feature into P local features $f_p \in \mathbb{R}^{I\times I\times C}$, in which I is the predetermined instance size and $p \in \{1, 2, ..., P\}$. Consequently, we can write the local channel and spatial masks (L_{ch}, L_{sp}) as:

$$L_{ch}, p = M_{cha}(f_p^T) + M_{cha}(f_p^S),\ L_{ch} = \otimes(L_{ch,1}, L_{ch,2}, ..., L_{ch,P},) \quad (3)$$

$$L_{sp}, p = M_{spa}(f_p^T) + M_{spa}(f_p^S),\ L_{sp} = \otimes(L_{sp,1}, L_{sp,2}, ..., L_{sp,P},) \quad (4)$$

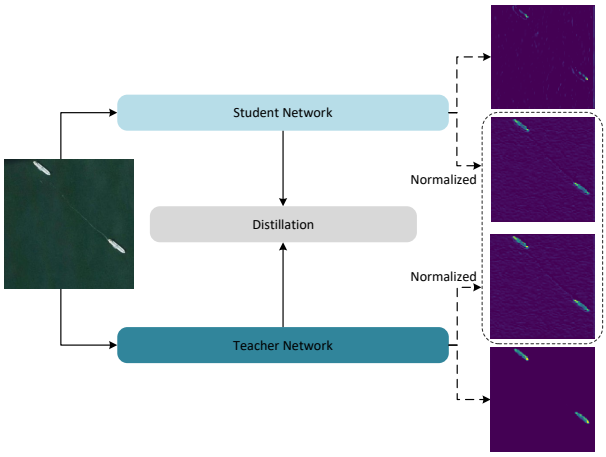

Fig. 3: Activations for the input image before and after normalization. This process fills the gap between the patterns of the teacher detector and the student detector, providing a more effective and smoother transfer of knowledge.

in which T stands for teacher, S for student, and $\otimes$ for the concatenation operator. An effective feature distillation approach must considers magnitude difference while generating pairs for imitation. In addition, by analyzing the activation patterns, we observe that the dominating FPN layers and channels may directly interact with the student's training phase and lead to sub-optimal performance, which is ignored by previous studies. To overcome this problem and optimize the learning process as shown in Fig. 3, we suggest first normalize the teacher's features and the student ones. This involves transforming the features to have a zero mean and a unit variance. Once the normalization is complete, the next step is to minimize the mean squared error (MSE) between the normalized features. It is also important that the normalization follow the convolution property, to ensure that features are normalized uniformly at different regions of the feature map. Let $\mathbb{V}$ represent the whole set of feature map values that include the components of mini-batch and its spatial locations. Therefore, for an u-size mini-batch and $h \times w$-size feature maps we take the functional mini-batch of $m = ||\mathbb{V}|| = u - size.hw$. Let $s^{(c)} \in \mathbb{R}^m$ be the c^{th} channel in a batch of FPN outputs, therefore, we can obtain the normalized values from the teacher T and the student S detector. Consequently, in the same way that local features are normalized, the global feature $\mathcal{F} \in \mathbb{R}^{H\times W\times C}$ can be normalized. Therefore, the global channel and spatial masks (G_ch, G_sp) can be written as:

$$\begin{aligned} G_{ch} &= M_{cha}(\mathcal{F}^T) + M_{cha}(\mathcal{F}^S), \\ G_{sp} &= M_{spa}(\mathcal{F}^T) + M_{spa}(\mathcal{F}^S). \end{aligned} \quad (5)$$

In order to build our final channel attention masks LG_{ch} and spatial attention masks LG_{sp}, we integrate the local and global masks as illustrated below:

$$LG_{ch} = \frac{1}{2}.(L_{ch} + G_{ch}),\ LG_{sp} = \frac{1}{2}.(L_{sp} + G_{sp}). \quad (6)$$

In our model, the feature maps normalization of both student and teacher detectors helps to align the magnitudes, improve

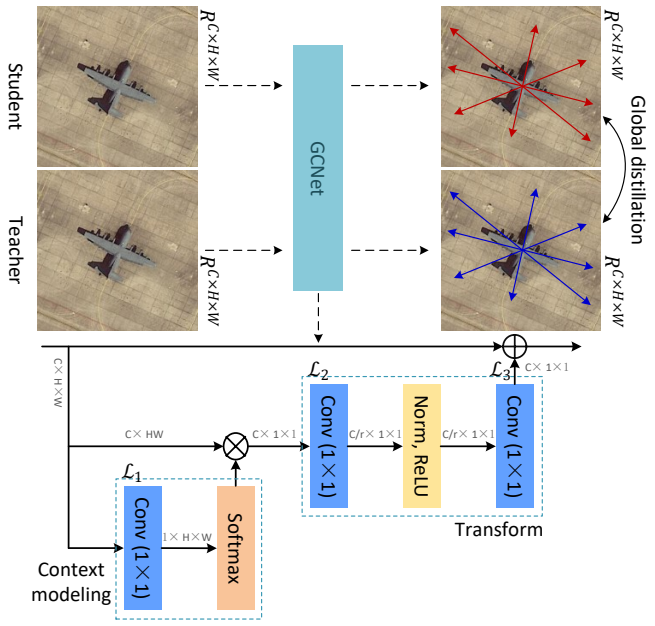

Fig. 4: The framework of global distillation. The feature maps from the student and teacher detectors used as the inputs.

knowledge transfer, and enhance stability. These effects collectively contribute to more effective and efficient distillation and result in improved student detector performance.

B. Feature-based Distillation

In KD, the features of teacher detector generally contain more information than the features of student detector. Hence, we distill the FPN's intermediate features to boost students' performance. To carefully distill the region of interest, all layers' features are combined with the spatial and channel attention masks. We define the loss of feature distillation as:

$$\ell_{fd} = \sum_{l=1}^{L}(\sum_{c=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W}(\mathcal{F}_{lcij}^{T} - \phi_{adj}(\mathcal{F}_{lcij}^{S}))^2 . LG_{sp,l} . LG_{ch,l})^{\frac{1}{2}}, \tag{7}$$

in which depth of the FPN is shown by L, l stands for the l^{th} FPN layer, while i and j are the locations of the feature map with the corresponding H and W. ϕ_{adj} is the 1×1 convolution layer used to adjust the student's features to those of the teacher. In addition, $LG_{ch,l}$ and $LG_{sp,l}$ are the mean channel and the spatial masks of the l^{th} layer, respectively. Further, the attention features are distilled in order to support the student to generate better LGAM. So, we can write the procedure of channel attention feature and spatial attention feature extraction as $AF_{ch}(x) = \frac{1}{C} . \sum_{c=1}^{C} x_c$ and $AF_{sp}(x) = \frac{1}{HW} . \sum_{i=1}^{H}\sum_{j=1}^{W} x_{ij}$. Through the distillation of local and global channel attention features, the channel attention loss is computed. Furthermore, the features obtained from local and global spatial attention are equivalent, while the local features are derived through splitting the global features in the spatial domain. Spatial attention loss, unlike channel attention loss, mainly uses global spatial attention features.

TABLE I: OUR AFD'S ABILITY FOR GENERALIZATION IN DIFFERENT OBJECT DETECTION MODELS ON THE DOTA AND NWPU. WE COMPARE OUR MODEL'S PERFORMANCE BY mAP (%), FPS (f/s), AND NUMBER OF PARAMETERS (M).

Method	DOTA			NWPU		
	mAP	FPS	Params	mAP	FPS	Params
FR-CNN (T)	72.18	18	92.65	90.91	22	73.62
FR-CNN (S)	65.27	32	60.17	85.93	39	41.15
AFD	70.54	32	60.17	89.76	38	41.15
Cascade (T)	77.39	16	120.24	93.14	19	101.22
Cascade (S)	70.47	30	87.80	88.42	36	68.77
AFD	76.91	30	87.80	91.85	35	68.77
RetinaNet (T)	72.94	19	87.74	90.86	23	68.87
RetinaNet (S)	64.47	32	55.36	85.91	41	36.28
AFD	73.08	32	55.36	90.93	40	36.28
ATSS (T)	74.18	18	55.45	92.90	22	51.44
ATSS (S)	67.42	33	18.97	86.52	39	18.95
AFD	72.94	33	18.97	92.67	39	18.95
FCOS (T)	72.56	19	67.98	91.84	22	64.33
FCOS (S)	67.71	33	31.56	87.21	41	31.85
AFD	71.82	33	31.56	90.63	40	31.85

Thus, we can write our channel attention and spatial attention losses (ℓ_{cha}, ℓ_{spa}) as:

$$\ell_{cha} = \frac{1}{2} . (\parallel AF_{ch}(\mathcal{F}^S) - AF_{ch}(\mathcal{F}^T) \parallel_2 + \frac{1}{N} . \sum_{p=1}^{P} \parallel AF_{ch}(f_p^S) - AF_{ch}(f_p^T) \parallel_2), \tag{8}$$

$$\ell_{spa} = \parallel AF_{sp}(\mathcal{F}^S) - AF_{sp}(\mathcal{F}^T) \parallel_2. \tag{9}$$

Now we can write the feature attention loss ℓ_{fa} by combining the channel attention loss and spatial attention loss

$$\ell_{fa} = \ell_{cha} + \ell_{spa}. \tag{10}$$

C. Global Distillation

The relationship [46], [47] between different pixels contains useful information that is used to boost detection task performance. In addition to feature attention, which aims to sever the relationship between background and foreground, a global distillation approach is also proposed. This approach facilitates the transfer of key knowledge from the teacher detector to the student detector by leveraging the global relationships between neighboring pixels in the feature maps. To compel the student detector to acquire knowledge about the pixels' relationship from the teacher detector, we use GcNet [46] for the purpose of extracting the global relation information from an image, as demonstrated in Fig. 4. Consequently, we can write the global loss as:

$$\ell_{glob} = \Lambda . \sum (B(\mathcal{F}^T) - B(\mathcal{F}^S))^2, \tag{11}$$

in which $B(\mathcal{F}) = \mathcal{F} + \mathcal{L}_3(\mathcal{N}(ReLU(\mathcal{L}_2(\sum_{j=1}^{n^p} \frac{e^{\mathcal{L}_1 \mathcal{F}_j}}{e^{\mathcal{L}_1 \mathcal{F}_M}} \mathcal{F}_j))))$, $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ are layers of CNN. $\mathcal{N}$ is the normalizing layer, n^p denotes the number of pixels and Λ represents the loss-balancing hyper-parameter. This hyper-parameter controls the trade-off between performance gains and knowledge transfer in the student detector. It allows fine-tuning the amount of knowledge transferred from the teacher detector while ensuring that the student detector learns effectively from its own training data.

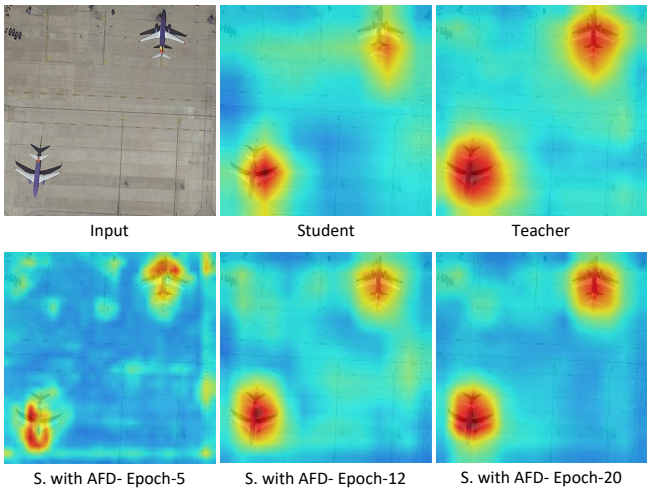

Fig. 5: Visualization of the attention maps produced by the student detector, teacher detector, and different training phases of student detector using AFD. Red denotes the highest level of attention, whereas other colors denote lower.

D. Head Distillation

By directing attention towards the outputs of the students, the distillation process stimulates them to attain performance levels comparable to that of the teacher. Nonetheless, in the case of remote sensing images, where there exists a substantial imbalance between background and foreground, directly distilling the outputs from the teacher's head may adversely affect the detection performance of the student. That's why spatial attention masks are used to ensure that the response-based distillation is as accurate as possible. In particular, from the FPN we take the spatial attention masks (see Eq. (6)) to preform masked head distillation and we can write the classification head loss ℓ_{cls-h} as:

$$\ell_{cls-h} = \sum_{l=1}^{L}\sum_{c=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W} \ell_{ce}(o^S_{lcij}, o^T_{lcij}).LG_{sp,l}, \tag{12}$$

in which o^S and o^T denote the classification head's outputs for both student and teacher detectors, and ℓ_{ce} is the cross-entropy loss. As stated in [8], the student model receives inappropriate information from unbounded teacher outputs. To address this problem, IoU loss is adopted to distill the localization head and we can define it's loss as:

$$\ell_{loc-h} = \sum_{l=1}^{L}\sum_{c=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W} \ell_{IoU}(k^S_{lcij}, k^T_{lcij}).LG_{sp,l}, \tag{13}$$

in which k denotes the output of the localization head.

By incorporating the outputs generated by the modules of the detector with the Faster R-CNN [25], we properly set the distillation losses and then compute the total loss by aggregating the ℓ_{cls-h} and ℓ_{loc-h} for object detection as,

$$\ell_{total} = \nu\ell_{fd} + \upsilon\ell_{fa} + \ell_{glob} + \beta(\ell_{cls-h} + \ell_{loc-h}) + \ell_{rpn}, \tag{14}$$

in which ν, υ and β represent the balancing-parameters for the different losses. The ℓ_{rpn} denotes the loss of RPN [25] in two-stage detector which written as:

$$\begin{aligned}\ell_{rpn} = \lambda_1 \frac{1}{\mathcal{N}_{cls-h}} \sum_i \ell_{cls-h}(p_i, p_i^*) \\ + \lambda_2 \frac{1}{\mathcal{N}_{reg}} \sum_i p_i^* \ell_{reg}(t_i, t_i^*)\end{aligned} \tag{15}$$

in which i is the index of a bounding box (BB), p_i is the probability of the i^{th} anchor predicted as an object, p_i^* is the ground-truth type appointed to the i^{th} anchor (0 if the box is negative and 1 for the positive one), ℓ_{reg} is the smooth-ℓ_1 loss, t_i represents the detected regression offset for i^{th} anchor and t_i^* denotes the target BB regression offset for the i^{th} positive anchor. The hyper-parameters λ_1 and λ_2 denote the balancing factors for losses, which we adjusted to 1 in our experiments for simplicity. $\mathcal{N}_{cls-h}, \mathcal{N}_{reg}$ denote normalization parameters that help to reduce the effect of various object scales, resulting in more effective training.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we encompass a detailed description of the datasets used, the evaluation metrics employed, and the experiments undertaken to assess the effectiveness and efficiency of our KD approach. Furthermore, to determine the effect of each module on the overall performance of the proposed architecture, a comprehensive ablation study is conducted.

A. Datasets and Evaluation Metrics

DOTA [18] is a remote sensing image dataset for object detection that contains 2806 images of various sizes. It consists of 15 types of objects with various dimensions and orientations.

NWPU VHR-10 [19] is a dataset that contains 650 remote sensing images of different sizes. It consists of 10 types of objects.

For the DOTA dataset, the images are cropped into the 800×800 pixels patches with 200 pixels overlap with the neighboring patches. For the NWPU dataset there are not enough images for training. For expanding the training dataset, we performed rescaling, rotation and flipping.

The metrics of *mean average precision* (mAP), *frames per second* (FPS), and *number of parameters* (Params) are used to assess the performance of AFD. Following is how mAP is calculated:

$$mAP = \int_0^1 P(R)\, dR, \tag{16}$$

where the predicted rates for accuracy and recall are P and R, respectively, and d denotes the coordinates of the estimated center point. In addition, to more accurately assess a method's ability for localisation and classification, the metrics of *Localization Error* and *Confusions with Background* [48], [49] are used.

TABLE II: COMPARISON WITH STATE-OF-THE-ART OBJECT DETECTION KD METHODS ON THE DOTA USING RETINANET.

Methods	Plane	BD	Bridge	GFT	SV	LV	Ship	TC	BC	ST	SBF	RA	Harbor	SP	HC	mAP
Teacher	88.96	83.18	52.68	62.67	72.51	73.84	79.42	90.57	81.98	79.82	54.77	67.82	75.49	71.84	58.51	72.94
Student	86.67	72.63	40.58	51.69	71.82	65.61	77.46	89.64	68.22	72.78	41.59	56.42	64.87	69.05	37.98	64.47
FGFI [14]	88.45	75.96	44.51	56.30	72.89	63.58	74.96	90.78	76.81	72.19	48.43	61.63	70.06	69.11	45.51	67.41
TAR [42]	89.16	76.52	46.55	59.98	73.90	66.24	78.56	90.78	78.57	75.67	43.71	66.93	71.40	72.16	45.32	69.03
DKD [16]	89.48	81.48	46.38	60.52	76.25	64.18	78.36	90.79	78.60	78.31	53.12	65.06	73.05	**74.11**	59.98	71.31
FGD [45]	89.60	81.55	**47.63**	60.34	76.19	64.26	78.26	90.46	78.43	78.45	52.81	65.20	73.32	73.67	60.14	71.36
LD [9]	89.64	81.54	46.57	**60.73**	76.42	64.15	78.51	**90.84**	78.72	78.41	53.26	65.23	73.18	73.91	60.21	71.43
AFD (ours)	**89.91**	**82.63**	47.59	60.58	**77.30**	**65.37**	**79.14**	90.75	**79.22**	**78.84**	**54.56**	**65.94**	**74.49**	74.06	**62.48**	**73.08**

TABLE III: COMPARISONS OF DETECTION RATE AND SPEED FOR OUR MODEL WITH CASCADE AGAINST OTHER DETECTORS ON THE DOTA DATASET FOR HBB TASK. THE BEST RESULTS ARE HIGHLIGHTED.

Methods	Plane	BD	Bridge	GFT	SV	LV	Ship	TC	BC	ST	SBF	RA	Harbor	SP	HC	mAP	FPS
RICA [37]	86.97	80.93	46.68	67.47	66.19	71.56	74.33	86.43	80.37	71.42	51.76	64.78	71.35	76.84	56.11	70.21	24
DRN [50]	89.63	82.71	47.25	64.05	76.20	74.33	85.76	90.53	86.15	84.82	57.77	61.95	69.34	69.72	58.46	73.25	9
FMSSD [34]	89.15	83.51	49.23	69.84	69.32	74.57	77.83	90.64	83.62	75.28	55.37	67.42	75.31	80.72	60.36	73.48	17
Pelee [32]	87.61	73.84	52.93	73.88	72.32	78.15	76.30	90.16	79.24	76.13	44.89	68.20	72.63	78.81	79.36	73.62	28
BBAVectors [51]	88.65	84.07	52.14	69.58	78.24	80.37	**88.03**	90.82	87.16	86.41	56.07	65.71	67.02	71.94	63.97	75.34	11
R3Det [52]	89.84	83.79	48.23	66.85	78.71	**83.36**	87.90	**90.86**	85.44	85.46	65.74	62.75	67.49	78.83	72.64	76.53	10
Scrdet++ [53]	**90.12**	**85.23**	55.61	**74.17**	76.48	73.28	86.11	90.53	**87.30**	**87.24**	**69.73**	68.81	73.38	72.65	67.43	**77.20**	14
AFD (ours)	89.81	77.68	**56.17**	70.65	**78.94**	81.62	84.28	90.35	75.23	76.90	51.65	**75.24**	**75.92**	**82.54**	**86.67**	76.91	**30**

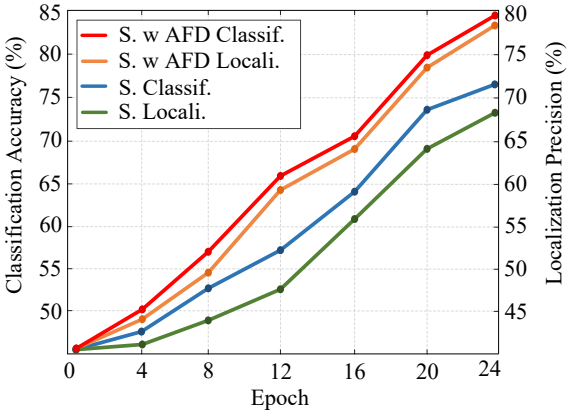

Fig. 6: Classification accuracy and localization precision of the student detector on FR-CNN with and without our AFD module on the DOTA dataset during training.

B. Implementation Details

While the labels on DOTA objects are in a quadrilateral form, those on NWPU are in the common axis-aligned bounding boxes (BBs). In order to have both options, our AFD method provides both oriented and horizontal BBs (HBB, OBB), in which HBB:$\{x_{min}, y_{min}, x_{max}, y_{max}\}$, OBB:$\{x_{center}, y_{center}, W, H, \theta\}$, while W is width, H is height and θ is between $[0, 90^{\circ})$ for all objects. In training, a set of rotating rectangles that appropriately overlap with the given quadrilateral labels provide the OBB ground truth. AFD only generates HBB results for the NWPU, due to the datasets' lack of OBB ground truth. AFD, on the other hand, generates both OBB and HBB outputs for the DOTA.

We compare the performance of our model with those of previous KD methods using a variety of object detection strategies [14]–[17], [43] in order to show the efficacy of our approach. Our implementation is on the basis of MMDetection [54] and on ImageNet the backbone networks are pretrained. The model is implemented in Pytorch and we use four GeForce RTX3090 GPUs for training with the batch size of 16.

The network is trains with 24 epochs using stochastic gradient descent (SGD). The initial learning rate is 0.02 for FR-CNN and 0.01 for the other modules, which are decrease in the 16^{th} and 22^{nd} epochs by a factor of 10. The momentum and weight decay are set to 0.9 and 1e-4, respectively. The hyper-parameters of losses are set to ($\kappa = 5 \times 10^{-4}$, $\upsilon = 2 \times 10^{-2}$, $\beta = 1 \times 10^{-1}$, and $\mathcal{T} = 1 \times 10^{-1}$) in the case of one-stage detectors ($\kappa = 6 \times 10^{-5}$, $\upsilon = 4 \times 10^{-3}$, $\beta = 1 \times 10^{-1}$, and $\mathcal{T} = 4 \times 10^{-1}$) in the case of two-stage detectors.

C. Evaluations using Various Detection Architectures

We evaluate our AFD's generalization ability on several detection frameworks, including RetinaNet [21] (one-stage detector), FR-CNN [25] and Cascade [56] (two-stage detectors), and ATSS [22] and FCOS [57] (anchor-free detectors). We use ResNet18 as student backbones and ResNet101 as teacher backbones for all detectors. As reported in Table I, our AFD shows substantial improvements in mAP across various types of detectors. Specifically, when applied to the DOTA dataset, our approach achieves an impressive average mAP enhancement of 5.8 points, surpassing the performance of standard two-stage detectors. Among the various detectors, our model shows the most significant performance improvement when applied to RetinaNet, enhancing its mAP from an initial 64.47 to 73.08. Moreover, when AFD used on the NWPU dataset, it achieves remarkable mAP values of 91.85, 92.67, and 90.93 for Cascade, ATSS, and RetinaNet, respectively. The achieved mAP values not only demonstrate competitive performance but also show instances where the student detectors outperform their respective teacher detectors. Notably, in the case of RetinaNet, the incorporation of our AFD yields a marginal yet discernible improvement in student detector performance, attributed to the efficacy of our attention module. These results show that our AFD is adaptable and can be effectively integrated into a wide range of detector architectures, yielding significant performance improvements.

Fig. 5 shows the visualization of attention maps derived from both the student and teacher detectors, as well as from

Fig. 7: Sample detection results of the student using Cascade detector trained with AFD on the DOTA.

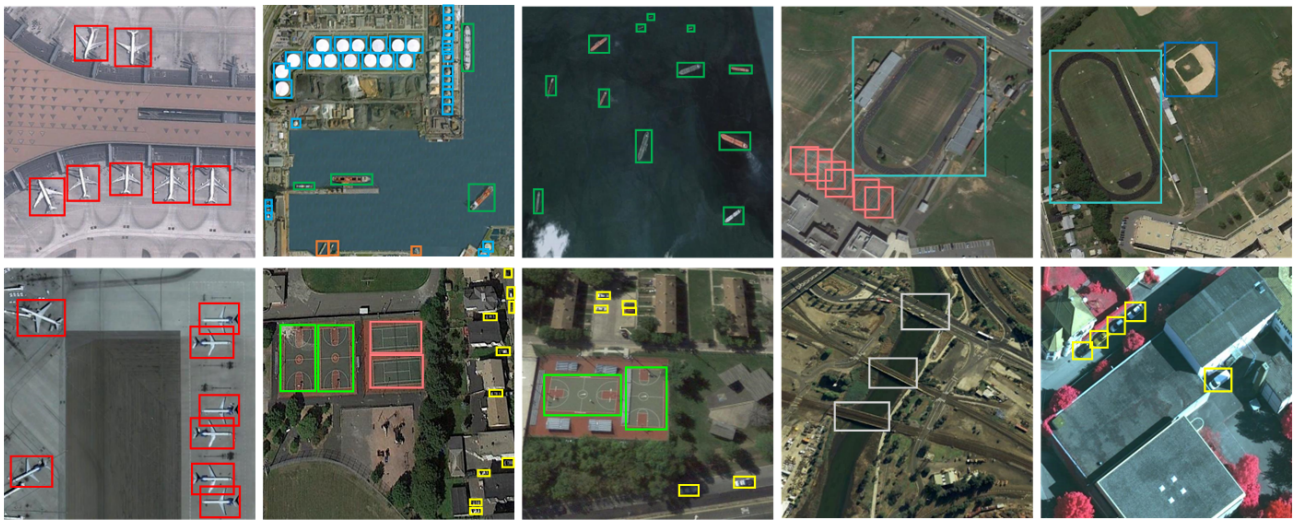

Fig. 8: Sample detection results of the student using Cascade detector trained with AFD on the NWPU.

TABLE IV: PERFORMANCE COMPARISONS BETWEEN OUR KD DETECTOR WITH CASCADE AND STATE-OF-THE-ART OBJECT DETECTION MODELS ON NWPU DATASET.

Methods	Plane	SH	ST	BD	TC	BC	GTF	Harbor	Bridge	Vehicle	mAP	FPS
EDAI [55]	77.12	72.86	60.34	70.08	59.48	64.82	78.61	67.59	71.94	68.84	69.17	14
RICA [37]	96.43	85.69	89.37	91.48	85.66	78.35	89.30	75.46	69.89	74.96	83.66	31
FMSSD [34]	**99.62**	88.71	89.54	97.23	84.65	**95.28**	98.56	73.76	79.43	87.54	89.43	24
Pelee [32]	99.45	**91.15**	96.21	**97.82**	88.79	90.34	98.30	**86.63**	**86.92**	87.85	92.34	29
AFD (ours)	99.51	90.88	**97.13**	97.36	**89.45**	94.67	**98.95**	85.94	86.34	**88.19**	**92.85**	**35**

different stages of the student detector using our AFD method. By comparing these attention maps at various training stages, we can observe the student's progressive learning process and its attempt to align with the teacher's guidance. Fig. 5 demonstrates that the teacher detector shows more accurate focus on the airplanes in the image compared to the student detector, which has undergone only five epochs of training. However, as our AFD progresses, we observe a gradual convergence of attention between the student and teacher. This observation potentially explains why our smaller distillation method even outperforms the teacher model in certain instances (using RetinaNet as detector).

Moreover, Fig. 6 illustrates the progression of classification and localization accuracy during the training process for student detectors using FR-CNN, both with and without the inclusion of our AFD. The data clearly indicates that the incorporation of AFD has a substantial and positive influence on enhancing the performance of the student detector. Notably, the classification accuracy increases from 76.8% without AFD to an impressive 84.3% when AFD is applied.

D. Comparative Analysis with Cutting-edge KD Approaches

On the DOTA dataset we evaluate our model together with recent KD methods using RetinaNet to compare the results

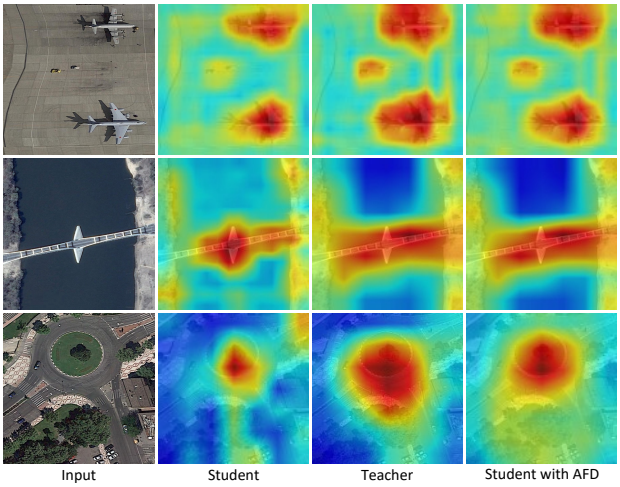

Fig. 9: Attention maps visualization from different detectors.

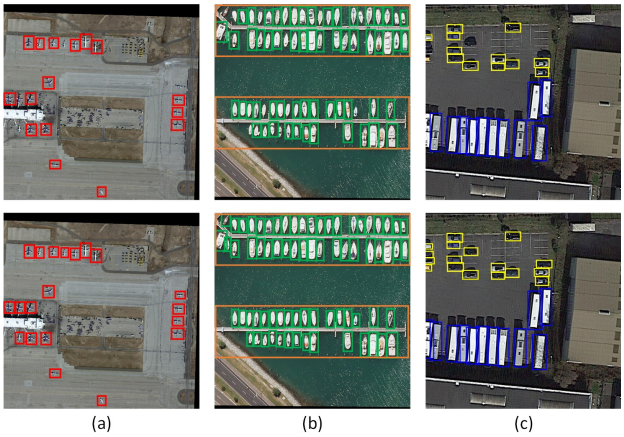

Fig. 10: The qualitative evaluation of improvement in distillation learning on the DOTA. The top row is the results of student detector without distillation learning and the bottom row is the student that learned by AFD. (a) Plane, (b) ship and harbor, (c) large and small vehicles.

with those of other KD approaches. The teacher detector is RetinaNet with ResNet-101, while the student detector is RetinaNet with ResNet-18. We conducted a thorough performance evaluation of AFD, comparing it to other state-of-the-art KD models. The results provide clear evidence of AFD's superior performance. As shown in Table II, our AFD surpasses all state-of-the-art KD approaches in distillation performance. To be more specific, our model achieved a remarkable improvement, surpassing the dynamic global distillation method [16] by an impressive margin of 1.77 mAP. This result highlights the superiority of our approach in KD. Moreover, our model achieved an impressive 73.08 mAP, surpassing the recently developed distillation methods FGD [45] and LD [9], which achieved 71.36 and 71.43 mAP, respectively. These results highlight the significant impact of our method's superior extraction of local and global knowledge, resulting in a substantial enhancement in distillation performance.

E. Comparative Analysis with CNN-based Detectors

We evaluate our KD method using the Cascade detector and compare it to other CNN-based object detection models.

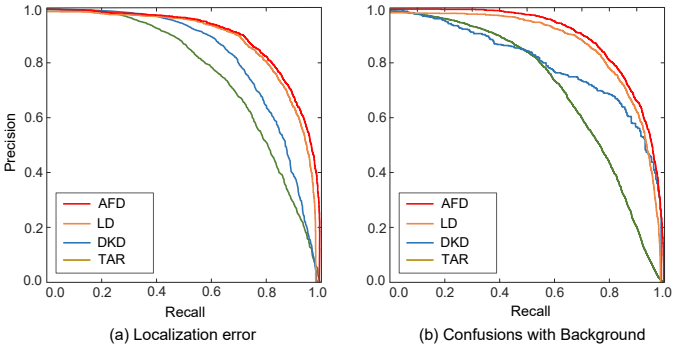

Fig. 11: Performance analysis of the KD detectors on the DOTA dataset. (a) mean Localization error. (b) Confusions with Background.

In Tables III on the DOTA and Table IV on the NWPU datasets, we respectively evaluate the results of our KD detector compared to the recent object detection models. On the DOTA, FMSSD detects at a rate of 73.48 mAP and processing 17 FPS. On the other hand, BBAVectors [51] and R3Det [52] detect at a rate of 75.34 and 76.53 mAP while processing at 11 and 10 FPS, respectively. Although SCRDET++ [53] obtains the highest mAP of 77.20, its detection speed is just 14 FPS, which notably lower when compared to the performance of our AFD model. Empirically, we find our AFD obtains a better detection/speed trade-off compared to other detectors (76.91 mAP / 30 FPS). Some sample detection results of AFD are shown in Fig. 7. AFD has the most accurate classification results for the Bridge, Small vehicle, Roundabout, Harbor, Swimming Pool, and Helicopter classes.

As reported in Table IV, on the NWPU, AFD obtains state-of-the-art results. Some sample detection results of our model are shown in Fig. 8. Our detector achieves 92.85% mAP with detection speed of 35 FPS which shows the superiority of AFD compare to other state-of-the-art models for object detection in remote sensing images. AFD detects 6 and 11 FPS faster and has a 0.51% and 3.42% higher mAP than Pelee [32] and FMSSD [34], respectively. AFD performs best in the Storage tank, Tennis court, Ground track field, and Vehicle classes.

Fig. 9 illustrates the attention maps produced by the student detector, teacher detector, and student detector with our AFD method using Cascade. Upon observing the attention maps of both the teacher and student detectors, noticeable disparities in pixel distribution become apparent prior to applying the distillation process. However, following training with AFD, the student detector has a similar pixel distribution to the teacher detector, indicating that the student relies on the same regions as the teacher. This shows how AFD improves the performance of the student detector.

The first row in Fig. 10 shows the baseline visualization results (student detector without KD learning) and the next row represents the results of the student detector learned with our AFD. AFD has a stronger feature extraction ability than the baseline, and the objects are more accurately detected. For example, the detected small vehicles at the bottom of Fig. 10(c) show that AFD produces more accurate regression results than the baseline. To conduct a more comprehensive evaluation of the AFD's performance, in Fig. 11 the local-

TABLE V: ABLATION STUDY FOR EACH MODULE'S CONTRIBUTION TO AFD. ℓ_{fd}, ℓ_{rpn}, ℓ_{cls-h}, and ℓ_{loc-h} DENOTE DIFFERENT DISTILLATION LOSSES OF OUR METHOD. "LGAM" IS OUR MASK FOR DISTILLATION LOSSES.

ℓ_{fd}	ℓ_{rpn}	LGAM	ℓ_{cls-h}	ℓ_{loc-h}	mAP
×	×	×	×	×	71.82
✓	×	×	×	×	73.41
✓	✓	×	×	×	74.63
✓	✓	✓	×	×	76.23
✓	✓	✓	✓	×	76.56
✓	✓	✓	×	✓	76.48
✓	✓	✓	✓	✓	76.91

TABLE VI: EXPERIMENTS ON HOW PERFORMANCE CHANGES WITH DIFFERENT ATTENTION MASKS AND WITHOUT NORMALIZATION.

Local	Global	Normalization	mAP
×	×	×	75.54
×	✓	×	75.82
✓	×	✓	76.65
✓	✓	✓	76.91

ization error and confusions with Background curves on the DOTA dataset is shown. As it shows, our model outperforms the other baselines in terms of localisation and classification accuracy. This performance is due to our proposed local and global distillation approach which enables the student detector to understand the relationship between pixels. The following factors have significantly contributed to this progress.

1) The proposed attention feature distillation approach enhances the students' learning of foreground objects while suppressing students' learning of background pixels.
2) The proposed local and global feature distillation allows the student detector to not only focus on the important pixels and channels of the teacher, but also to recognize the connection between different pixels.

F. Ablation Study

In this section, a comprehensive ablation experiments is conducted to assess the importance of the proposed modules of our framework.

1) AFD Modules: We compare the AFD detection performance with and without different modules of Eq. (14) to evaluate the impact of each of them. Table V reports some results of our ablation experiment. On the classification and regression heads, AFD improves mAP by 0.33 and 0.25, respectively. When the distillation process is applied jointly to both the classification and regression heads, our model shows an improvement of 0.68 mAP. These results provide clear evidence that each component of our total distillation loss significantly contributes to the overall enhancement.

2) Distillation Effect of Local and Global Attention Masks: Since the global attention mask is not distributed equally,

TABLE VII: EFFECT OF SPATIAL AND CHANNEL ATTENTION.

Spatial attention	Channel attention	mAP
×	×	73.44
×	✓	76.73
✓	×	76.08
✓	✓	76.91

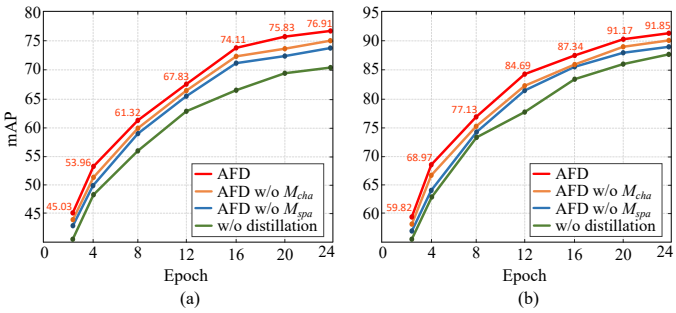

Fig. 12: The mAP during training. (a) Performance evaluation on DOTA. (b) Performance evaluation on NWPU. Channel and spatial masks can improve mAP in training stages.

TABLE VIII: THE RESULTS OF VARIOUS LOSS FUNCTIONS UTILIZED IN ℓ_{loc-h}.

Loss	ℓ_1	$smooth-\ell_1$	ℓ_{MSE}	ℓ_{IoU}
mAP	76.84	76.69	76.72	76.91

KD only with the global feature attention module can extract information from large objects. Alternatively, we believe that small objects can be retrieved using local feature attention. We conduct different distillations using local and global attention masks to determine which is more effective. Based on the data in Table VI, it is clear that local attention leads to a higher mAP than global attention. Despite that, the global attention results are inferior to the local ones, however, detection performance further improves when both local and global attention are used together. Following these observations, AFD's local attention mask should work with the global attention mask to boost performance further. Furthermore, by using normalization process, a large uniform value of loss weight can be obtained between the teacher and student detectors to maintain the balance of the detection and distillation losses. This balance is important to ensure that both the detection loss and the distillation loss have appropriate contributions to the overall learning process. Without normalization, one of the losses (detection or distillation) might dominate the training process due to large magnitude differences. This can lead to an imbalance in learning objectives and hinder the effectiveness of KD. Normalization mitigates this issue by ensuring that both losses have comparable magnitudes and influence the training process in a more balanced manner.

3) Analysis of Different Attentions and Distillation Losses: As shown in Table VII, spatial and channel attentions boost mAP by 3.29 and 2.64, respectively. On the other hand, combining the two attentions result in 3.47 mAP improvements. The findings indicate the valuable contributions of both channel attention and spatial attention, highlighting their effective combination as a means to enhance overall performance. Indeed, the teacher detector effectively guides the student detector's focus towards important components by using a spatial and channel attention mask. We analyze the impact of different masks in Fig. 12. Each attention mask increases our model efficiency, particularly the spatial attention mask. However, the best result is obtained by combining both the masks. Additionally, we evaluate the effects of the various loss functions of Eq. (13). In order to evaluate the regression

loss in Eq. (13), we analyze the result of several losses such as ℓ_{IoU}, ℓ_{MSE}, ℓ_1, and $smooth-\ell_1$. The ℓ_{IoU} has the best performance compared to the others, as reported in Table VIII.

V. Conclusion

This paper introduced AFD, a new mask-based KD approach for target detection in remote sensing images that efficiently uses local and global attention methods to obtain local features and background information. To extract local features, we split the feature maps of input image into patches and apply attention methods. Our method enhances distillation performance by extracting both fine-grained features and more important background information from a range of objects. We showed that AFD outperforms other KD techniques when combined with the different detection systems. The detection results demonstrate that AFD surpasses state-of-the-art models performance and can be adopted in various detectors such as single-stage, two-stage, and even anchor-free. Moreover, we conducted an ablation experiment and analysis, demonstrating the importance of distilling local information from multiple regions for object detection. We believe our work represents a turning point for traditional KD approaches that only rely on global information, into more effective model that incorporate both local and global information.

References

[1] P. Shamsolmoali, J. Chanussot, M. Zareapoor, H. Zhou, and J. Yang, "Multi-patch feature pyramid network for weakly supervised object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, 2022.

[2] P. Shamsolmoali, M. Zareapoor, J. Chanussot, H. Zhou, and J. Yang, "Rotation equivariant feature image pyramid network for object detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, 2022.

[3] Y. Liu, Q. Li, Y. Yuan, Q. Du, and Q. Wang, "Abnet: Adaptive balanced network for multiscale object detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[4] X. Wu, D. Hong, and J. Chanussot, "Uiu-net: U-net in u-net for infrared small object detection," *IEEE Trans. Image Process.*, 2022.

[5] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[6] Y. Guo, A. Yao, and Y. Chen, "Dynamic network surgery for efficient dnns," *Proc. Adv. Neural Inform. Process. Syst.*, vol. 29, 2016.

[7] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2704–2713.

[8] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," *Proc. Adv. Neural Inform. Process. Syst.*, vol. 30, 2017.

[9] Z. Zheng, R. Ye, Q. Hou, D. Ren, P. Wang, W. Zuo, and M.-M. Cheng, "Localization distillation for object detection," *arXiv preprint arXiv:2204.05957*, 2022.

[10] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3967–3976.

[11] L. Chen, D. Wang, Z. Gan, J. Liu, R. Henao, and L. Carin, "Wasserstein contrastive representation distillation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16 296–16 305.

[12] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1365–1374.

[13] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

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

[15] L. Zhang and K. Ma, "Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[16] Y. Zhang, Z. Yan, X. Sun, W. Diao, K. Fu, and L. Wang, "Learning efficient and accurate detectors with dynamic knowledge distillation in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022.

[17] Y. Yang, X. Sun, W. Diao, H. Li, Y. Wu, X. Li, and K. Fu, "Adaptive knowledge distillation for lightweight remote sensing object detectors optimizing," *IEEE Trans. Geosci. Remote Sens.*, 2022.

[18] 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 Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.

[19] R. Dong, D. Xu, J. Zhao, L. Jiao, and J. An, "Sig-nms-based faster r-cnn combining transfer learning for small target detection in vhr optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8534–8545, 2019.

[20] 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.*, 2016, pp. 21–37.

[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2980–2988.

[22] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9759–9768.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, 2015.

[24] R. Girshick, "Fast r-cnn," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2016.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

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

[28] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[29] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[30] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proc.IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.

[31] Y. Li, J. Li, W. Lin, and J. Li, "Tiny-dsod: Lightweight object detection for resource-restricted usages," *arXiv preprint arXiv:1807.11013*, 2018.

[32] R. J. Wang, X. Li, and C. X. Ling, "Pelee: A real-time object detection system on mobile devices," *Proc. Adv. Neural Inform. Process. Syst.*, vol. 31, 2018.

[33] Z. Qin, Z. Li, Z. Zhang, Y. Bao, G. Yu, Y. Peng, and J. Sun, "Thundernet: Towards real-time generic object detection on mobile devices," in *Proc.IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6718–6727.

[34] 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, 2019.

[35] X. Sun, Y. Liu, Z. Yan, P. Wang, W. Diao, and K. Fu, "Sraf-net: Shape robust anchor-free network for garbage dumps in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6154–6168, 2020.

[36] T. Xu, X. Sun, W. Diao, L. Zhao, K. Fu, and H. Wang, "Assd: Feature aligned single-shot detection for multiscale objects in aerial imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2021.

[37] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2337–2348, 2017.

[38] X. Wu, W. Li, D. Hong, R. Tao, and Q. Du, "Deep learning for unmanned aerial vehicle-based object detection and tracking: a survey," *IEEE Geosci. Remote Sens. Magazine*, vol. 10, no. 1, pp. 91–124, 2021.

[39] P. Shamsolmoali, M. Zareapoor, J. Yang, E. Granger, and J. Chanussot, "Enhanced single-shot detector for small object detection in remote sensing images," in *Proc. Int. Geosci. Remote Sens. Symposium*, 2022, pp. 1716–1719.
[40] X. Wu, D. Hong, Z. Huang, and J. Chanussot, "Infrared small object detection using deep interactive u-net," *IEEE Geosci. Remote Sens. Letters*, vol. 19, pp. 1–5, 2022.
[41] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao, "Orsim detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5146–5158, 2019.
[42] R. Sun, F. Tang, X. Zhang, H. Xiong, and Q. Tian, "Distilling object detectors with task adaptive regularization," *arXiv preprint arXiv:2006.13108*, 2020.
[43] J. Guo, K. Han, Y. Wang, H. Wu, X. Chen, C. Xu, and C. Xu, "Distilling object detectors via decoupled features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2154–2164.
[44] X. Dai, Z. Jiang, Z. Wu, Y. Bao, Z. Wang, S. Liu, and E. Zhou, "General instance distillation for object detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7842–7851.
[45] Z. Yang, Z. Li, X. Jiang, Y. Gong, Z. Yuan, D. Zhao, and C. Yuan, "Focal and global knowledge distillation for detectors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4643–4652.
[46] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc.IEEE Int. Conf. Comput. Vis. workshops*, 2019.
[47] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.
[48] 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, 2020.
[49] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, and X. Sun, "Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 294–308, 2020.
[50] X. Pan, Y. Ren, K. Sheng, W. Dong, H. Yuan, X. Guo, C. Ma, and C. Xu, "Dynamic refinement network for oriented and densely packed object detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11 207–11 216.
[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 Winter Conf. App. of 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. Aartif. Intell.*, vol. 35, no. 4, 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.
[54] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "Mmdetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.
[55] L. Li, G. Cao, J. Liu, and Y. Tong, "Efficient detection in aerial images for resource-limited satellites," *IEEE Geosci. Remote Sens. Letters*, vol. 19, pp. 1–5, 2021.
[56] Z. Cai and N. Vasconcelos, "Cascade r-cnn: high quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1483–1498, 2019.
[57] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proc.IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.