Structured Adversarial Self-Supervised Learning for Robust Object Detection in Remote Sensing Images

Cong Zhang, *Graduate Student Member, IEEE*, Kin-Man Lam, *Senior Member, IEEE*, Tianshan Liu, Yui-Lam Chan, *Member, IEEE*, and Qi Wang, *Senior Member, IEEE*

***Abstract*— Object detection plays a crucial role in scene understanding and has extensive practical applications. In the field of remote sensing object detection, both detection accuracy and robustness are of significant concern. Existing methods heavily rely on sophisticated adversarial training strategies that tend to improve robustness at the expense of accuracy. However, detection robustness is not always indicative of improved accuracy. Therefore, in this article, we research how to enhance robustness, while still preserving high accuracy, or even improve both simultaneously, with simple vanilla adversarial training or even in the absence thereof. In pursuit of a solution, we first conduct an exploratory investigation by shifting our attention from adversarial training, referred to as adversarial fine-tuning, to adversarial pretraining. Specifically, we propose a novel pretraining paradigm, namely, structured adversarial self-supervised (SASS) pretraining, to strengthen both clean accuracy and adversarial robustness for object detection in remote sensing images. At a high level, SASS pretraining aims to unify adversarial learning and self-supervised learning into pretraining and encode structured knowledge into pretrained representations for powerful transferability to downstream detection. Moreover, to fully explore the inherent robustness of vision Transformers and facilitate their pretraining efficiency, by leveraging the recent masked image modeling (MIM) as the pretext task, we further instantiate SASS pretraining into a concise end-to-end framework, named structured adversarial MIM (SA-MIM). SA-MIM consists of two pivotal components: structured adversarial attack and structured MIM (S-MIM). The former establishes structured adversaries for the context of adversarial pretraining, while the latter introduces a structured local-sampling global-masking strategy to adapt to hierarchical encoder architectures. Comprehensive experiments on three different datasets have demonstrated the significant superiority of the proposed pretraining paradigm over previous counterparts for remote sensing object detection. More importantly, regardless of with or without adversarial fine-tuning, it enables simultaneous improvements in detection accuracy and robustness as expected, promisingly alleviating the dependence on complicated adversarial fine-tuning.**

Manuscript received 28 August 2023; revised 1 January 2024 and 29 January 2024; accepted 23 February 2024. Date of publication 14 March 2024; date of current version 21 March 2024. *(Corresponding author: Cong Zhang.)*

Cong Zhang, Kin-Man Lam, and Yui-Lam Chan are with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong (e-mail: cong-clarence.zhang@connect.polyu.hk).

Tianshan Liu is with the School of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu 210049, China.

Qi Wang is with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, Shaanxi 710071, China.

Digital Object Identifier 10.1109/TGRS.2024.3375398

***Index Terms*— Adversarial learning, remote sensing object detection, self-supervised pretraining (SPT), structured knowledge, vision Transformers (ViTs).**

I. Introduction

OBJECT detection is a fundamental task in the field of remote sensing scene understanding, with a wide range of real-world applications, including environmental monitoring, intelligent transportation, and military deployment [1], [2], [3], [4], [5], [6]. Thanks to the easy availability of large-scale remote sensing data, deep learning has recently dominated the field [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]. Specifically, convolutional neural networks (CNNs) and vision Transformers (ViTs) [21], [22], [23] have successively demonstrated their exceptional representational abilities, while the latter, as a more promising alternative to CNNs, has received increasing popularity, which reflects their significant potential for future applications in remote sensing object detection [24], [25], [26].

However, deep learning is a double-edged sword, especially in the context of remote sensing. On the one hand, defeating most traditional algorithms based on handcrafted features, deep learning-based methods can proficiently extract highly discriminative features and extensively achieve milestone performance across different tasks. On the other hand, it has been established that deep learning-based models, including both CNNs and Transformers, are highly vulnerable to adversarial attacks [27], [28], [29], which can be easily performed by adding imperceptible perturbations to the original legitimate (clean) images. ViTs, while demonstrated to be more robust against adversarial attacks than CNNs [30], [31], [32], still fail to provide satisfactory results. This deficiency is certainly a problem in the safety-critical field of remote sensing object detection, where the vulnerability of deep models to adversarial images is even more serious [33]. As shown in the left half of Fig. 1(a), a well-trained object detector can accurately detect an airplane (APL) in a clean image [top left of Fig. 1(a)], but it can be completely fooled to incorrectly identify the same APL as a dam in an adversarial sample, with only subtle perturbations [middle left of Fig. 1(a)]. To defend against such adversarial attacks, *adversarial training* has become an effective and popular approach [29], [34], which can promote an object detector to become more robust, enabling it

to localize and classify objects even in adversarial images [bottom left of Fig. 1(a)]. Adversarial robustness is commonly used to measure detection performance in such scenarios. To further enhance this robustness, recent detection methods [35], [36], [37], [38] tend to develop more advanced yet complicated adversarial training strategies or extra specialized and sophisticated detector components. However, adversarial training always comes at a cost: it significantly limits the final performance upper bound and forces a tradeoff between clean accuracy and adversarial robustness. As depicted in the first set of comparison results in Fig. 1(b), after time- and resource-consuming adversarial training, the clean accuracy rate is unintentionally reduced from 60.6% to 58.6%. Then, two crucial questions naturally arise.

1) With very simple or vanilla adversarial training instead of complicated training, can the adversarial robustness of object detectors be generally boosted while still maintaining their clean accuracy (i.e., free from the tradeoff)?
2) Even without adversarial training and any additional components, can the clean accuracy and adversarial robustness of object detectors be simultaneously improved?

This work is different from existing robust detectors dedicated to developing various sophisticated adversarial training strategies, as we mainly answer the above questions from a novel perspective, namely, SASS pretraining.

Existing robust object detectors [35], [36], [37], [38] typically follow a two-step paradigm: upstream natural pretraining on clean images and downstream adversarial fine-tuning on perturbed images. Natural pretraining here is based on the fully-supervised ImageNet classification [39], while adversarial fine-tuning actually refers to the aforementioned adversarial training. However, in such a setting, pretraining and fine-tuning are treated completely isolated, and we argue that ImageNet supervised pretraining (SPT) is unsuitable or suboptimal for robust object detection in remote sensing. Specifically, ImageNet SPT may lead to models encoding less valuable directional knowledge that can be inherited and transferred for fine-tuning, thereby compromising the model generalizability and reducing the clean accuracy of a robust model [40], [41], [42]. Inspired by this, instead of developing sophisticated adversarial fine-tuning approaches as in the literature, this work explores how to make the upstream pretrained models more attentive and transferable to downstream robust detection. In other words, we shift our focus from fine-tuning to pretraining and provide an advanced solution to the above questions. Ideally, a pretraining paradigm should not only facilitate generalizable features to improve clean accuracy but also promote robust representations against potential attacks [43], with the following three desiderata.

A. Adversarial Learning

Robust pretrained models generated by adversarial pretraining, which incorporate adversarial learning into pretraining on the input or feature levels, have witnessed superior performance over naturally (nonadversarially) pretrained models [44], [45], [46]. Such *adversarial pretraining* schemes should generally consist of two crucial and intertwined components: an informative adversary to perturb legitimate images accounting for enhancing robustness and an efficient pretext task that should be interrelated with the adversary. We thoroughly explore their relationships and interactions, and posit that both components are of equal significance in generating robust pretrained models.

B. Unsupervised Pretext

Various pretext tasks enable adversarial pretraining, yet some of which, despite adversaries, still lead to pretrained representations that are highly susceptible to unpredictable downstream attacks [47]. For example, it is uncultured to simply introduce adversaries to ImageNet SPT since supervised adversarial pretraining suffers from label leakage [34], overfitting pretrained models to particular perturbations, and hindering their generalization to clean examples or unseen adversaries. In contrast, unsupervised pretraining can alleviate this issue [43] and exhibit desirable label efficiency, free from prohibitively expensive large-scale annotations in the context of remote sensing. Therefore, self-SPT (SSPT), specifically masked image modeling (MIM), is engaged with adversarial learning as the pretext task to facilitate our *adversarial SSPT*.

C. Structured Knowledge

As described in both questions, the dual goals of our pretraining are always to boost clean accuracy (related to generalization) and adversarial robustness for downstream object detection, regardless of whether adversarial fine-tuning is applied or not. However, downstream tasks usually have different learning objectives than pretraining. This motivates us to ponder: how to ensure that the merits in generalizability and robustness from pretrained representations can be successfully inherited by downstream detection after fine-tuning? In other words, how to guarantee the powerful transferability of pretrained models? This work will demonstrate the significant importance of modeling structured knowledge. Based on the proposed two structured strategies, i.e., structured adversarial attacks and structured MIM (S-MIM), structured knowledge can be embedded into adversarial SSPT.

The above three essential desiderata for pretraining are not independent but mutually interdependent and harmoniously complementary to one another. By tightly incorporating them, this article proposes a novel and versatile pretraining paradigm, namely, SASS pretraining, for robust object detection in remote sensing images. In theory, it suggests a general guideline for designing robust pretraining strategies that are not limited to a specific pretext task. For better generalization and downstream performance, we instantiate the proposed SASS pretraining on the state-of-the-art pretext task for Transformers, i.e., MIM, to form structured adversarial MIM (SA-MIM). Fig. 2 compares different MIM-based pretraining paradigms. Our SA-MIM comprises two critical components: structured adversarial attacks and S-MIM. The former generates perturbed samples in an unsupervised manner, facilitating the pretrained models to build robust intrasample relationships through multigrained adversarial attacks. The latter masks the adversarial tokens and then reconstructs their corresponding

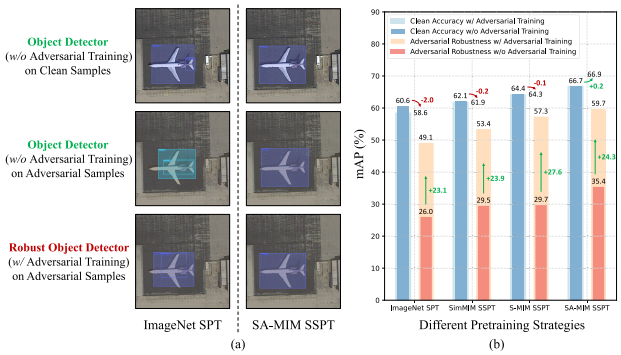

Fig. 1. Performance comparison of different pretraining strategies on downstream remote sensing object detection. SPT and SSPT denote supervised pretraining and self-supervised pretraining, respectively, while adversarial training refers to adversarial fine-tuning here. (a) Qualitative comparison of ImageNet SPT and the proposed SA-MIM SSPT under different training (fine-tuning) settings. Typically, an object detector is naturally trained, while a robust object detector is adversarially trained. It can be observed that without adversarial training, the ImageNet pretrained detector misidentifies the APL as a dam on the adversarial example, while the SA-MIM pretrained detector correctly detects this APL. With adversarial training, the SA-MIM pretrained robust detector localizes the APL more precisely. (b) Quantitative comparison of four different pretraining strategies on downstream detection. The performance of each pretraining strategy is characterized by a set of four values related to clean accuracy or adversarial robustness with or without adversarial training (fine-tuning). The proposed SA-MIM SSPT remarkably improves the generalization and robustness of the pretrained models, achieving the best detection performance across all varied configurations.

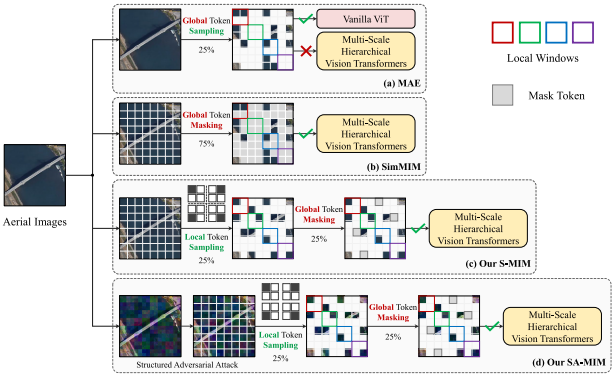

Fig. 2. Comparison of four different MIM-based pretraining paradigms. "Sampling" involves randomly selecting a certain proportion of tokens and discarding the rest, with only the sampled tokens visible to the Transformer for pretraining. "Masking," in contrast, leverages learnable mask tokens to mask a certain proportion of the original tokens, and both are fed to the Transformers for pretraining. Typically, sampling stands for higher computational efficiency than masking. (a) MAE can only be applied to the vanilla ViT [21] and cannot construct structured hierarchical representations due to the inconsistency in the number of tokens within each local window. (b) SimMIM can support multiscale Transformers, yet it suffers from heavy computational overheads brought by abundant mask tokens. (c) Our proposed S-MIM enables structured hierarchical representations locally and globally with high efficiency. (d) Our SA-MIM fully respects the proposed SASS pretraining strategy. In contrast to the above three SSPT paradigms, SA-MIM enjoys both higher robustness and generalizability.

legitimate tokens, which can efficiently model both local and global semantic information in a structured hierarchy and function as a problematic regularization in the context of adversarial pretraining [42]. Moreover, for high flexibility, S-MIM is purposely devised to be decoupled from adversarial pretraining, as shown in Fig. 2(c), which also independently serves as an SSPT paradigm on clean examples. Experiments demonstrate the effectiveness of our method for both object detection and robust object detection. For instance, in Fig. 1(a), even without adversarial fine-tuning, the detector pretrained by our SA-MIM successfully identifies the aircraft in the adversarial sample. If adversarial fine-tuning is adopted, the precision will be further improved with higher confidence. Fig. 1(b) quantitatively presents that, with or without adversarial fine-tuning, our method outperforms other supervised or unsupervised counterparts, including ImageNet SPT and SimMIM SSPT, in terms of both clean accuracy and adversarial robustness. Our contributions can be summarized as follows.

1) This article provides affirmative answers to the aforementioned two questions and verifies two critical scientific facts. First, if an object detector is initialized with a robust pretrained model, instead of a natural one, after downstream adversarial fine-tuning, its detection robustness can be considerably enhanced, while the clean accuracy is still kept high. It is suggested that adversarial pretraining can work synergistically with adversarial fine-tuning to facilitate high-quality detection. Second, even without adversarial fine-tuning, empowering a simple detector with a robust pretrained model can also improve detection performance in both clean and adversarial examples.
2) A novel and advanced pretraining paradigm, namely, SASS pretraining, is systematically proposed for Transformers. To the best of our knowledge, this is the first time that adversarial learning and self-supervision have been structurally integrated into pretraining to boost robust object detection in remote sensing. SASS is further instantiated into the SA-MIM SSPT framework, which equips pretrained Transformers with strong generalizability, robustness, and transferability for downstream detection.
3) To enable adversarial learning of SA-MIM, the structured adversarial attack is devised by applying learnable multigrained perturbations to unlabeled images in a structured manner, resulting in an effective input-level adversary.
4) S-MIM, which refers to S-MIM, is tailor-made for a self-supervised pretext task, to construct a feature-level hierarchy for multiscale Transformers, with both local and global structural constraints. Not only can it be flexibly adapted to SA-MIM but it can also function as a standalone pretraining paradigm with high computational efficiency.

The rest of this article is organized as follows. First, the related work is reviewed in Section II. Then, some preliminaries and the proposed method are elaborated on in Sections III and IV, respectively. Section V focuses on experiments and analyses. Finally, Section VI draws the conclusion.

II. Related Work

Three topics are highly relevant to this work: remote sensing object detection, adversarial learning in remote sensing, and MIM. This section provides a review of these three topics.

A. Remote Sensing Object Detection

Recent remote sensing object detectors [48], [49], [50], [51], [52] are mainly based on CNNs or ViTs. For example,

MidNet [48] proposed an anchor-and-angle-free detector with a novel oriented object representation for ship detection, which adopted a pretrained symmetric CNN architecture as the backbone. EMO2-DETR [50] developed an efficient matching paradigm to alleviate the relative redundancy issue for Transformer-based remote sensing object detection. However, these methods rely on pretraining CNNs or Transformers in natural scenes while fine-tuning in remote sensing scenarios, which may result in domain gap and then performance degradation. Moreover, previous literature rarely considers detection robustness in remote sensing. In contrast, this work aims to simultaneously enhance the standard accuracy and adversarial robustness of Transformer-based detectors by SSPT in the context of remote sensing.

B. Adversarial Learning in Remote Sensing

Adversarial learning is an emerging technique with enormous practical potential in the field of remote sensing scene understanding. Most of the existing works tend to develop preferable adversarial attacks on remote sensing images to degrade accuracy in different downstream tasks, including scene classification [53], [54], [55] and object detection [56], [57], [58]. On the contrary, the others [33], [59], [60] emphasize advanced adversarial defense techniques to improve the robustness of remote sensing scenarios, but all of them heavily rely on complicated downstream adversarial fine-tuning. Moreover, recent studies [44], [45] have revealed that attaining adversarial robustness through pretraining, instead of fine-tuning, can remarkably enhance model representational capabilities and contribute to downstream performance gains in terms of both robustness and accuracy. This motivates us to introduce adversarial learning into remote sensing pretraining. Nevertheless, previous adversarial pretraining methods are limited to CNNs and cannot be directly applied to more advanced and robust Transformers. Consequently, how to efficiently and adversarially pretrain Transformers, especially in the context of remote sensing, remains to be explored. In this work, we propose a novel adversarial pretraining scheme, through which pretrained Transformer models are both generalizable and robust for downstream remote sensing object detection.

C. Masked Image Modeling

As an advanced SSPT paradigm, MIM aims to obtain discriminative unsupervised representations by predicting masked regions while only being conditioned on visible context [61], [62], [63], [64], [65]. Compared to previous contrastive learning-based SSPT [66], [67], MIM-based methods are more tightly coupled to the Transformer family and have widely demonstrated superior performance on various downstream tasks, including object detection. Currently, MAE [62] and SimMIM [63] are the two most popular MIM methods. MAE accelerates the pretraining procedure by only taking visible tokens as input, while SimMIM attempts to integrate MIM with multiscale Transformers, such as Swin Transformer [22]. However, both methods still have their fatal limitations, as depicted in Fig. 2. In addition, recent research [45], [68] has also validated the benefits of learned unsupervised representations for adversarial robustness. Concretely, it is suggested that self-supervision can be intentionally incorporated into adversarial pretraining as a pretext task through contrastive learning. However, unlike contrastive learning, which inherently generates pairs of positive and adversarial negative samples through different augmented views, MIM does not explicitly introduce adversaries. This provides unprecedented research opportunities to adapt adversarial pretraining to MIM for the increasingly popular Transformers. Thus, this work will further demonstrate that, by devising appropriate perturbations on unlabeled samples, MIM-based self-supervised representation learning can be efficiently extended and well aligned with adversarial pretraining to shape adversarial SSPT.

III. Preliminaries: From Pretraining to Fine-Tuning for Object Detection and Robust Object Detection

Given a clean image x as input, a well-trained object detector f parameterized by $\boldsymbol{\theta}$ predicts a varying number of K objects, i.e., $f(x;\boldsymbol{\theta}) \rightarrow \{(\hat{\mathbf{b}}_k, \hat{\mathbf{c}}_k)\}_{k=1}^{K}$, where $\hat{\mathbf{b}}_k = [d_k^x, d_k^y, w_k, h_k]$ denotes a 4-D bounding box and $\hat{\mathbf{c}}_k = [\hat{c}_k^0, \hat{c}_k^1, \ldots, \hat{c}_k^C]$ represents its class probabilities over $C+1$ categories (including background $\hat{c}_k^0$). d_k^x and d_k^y are the center-point coordinates of $\hat{\mathbf{b}}_k$, while w_k and h_k are the width and height of $\hat{\mathbf{b}}_k$. In the context of robust object detection, the same detector f yet parameterized by $\tilde{\boldsymbol{\theta}}$ is expected to make accurate predictions on both clean example x and its adversarial counterpart $\tilde{x}$, i.e., $f(x;\tilde{\boldsymbol{\theta}}) \rightarrow \{(\hat{\mathbf{b}}_k, \hat{\mathbf{c}}_k)\}_{k=1}^{K}$ and $f(\tilde{x};\tilde{\boldsymbol{\theta}}) \rightarrow \{(\hat{\mathbf{b}}'_k, \hat{\mathbf{c}}'_k)\}_{k=1}^{K}$. It is worth noting that the correctness and consistency of both predictions, $(\hat{\mathbf{b}}_k, \hat{\mathbf{c}}_k)$ and $(\hat{\mathbf{b}}'_k, \hat{\mathbf{c}}'_k)$, are important [36]. "Correctness" directly determines the accuracy and robustness, while "consistency" indicates the practical value and potential of a robust detector in real-world complex remote sensing scenarios. Consequently, our objectives are always to enhance both clean accuracy and adversarial robustness and reduce the gap between them.

A. Object Detection

To obtain $f(\cdot;\boldsymbol{\theta})$, most object detection methods [69], [70], [71], [72] generally involve two stages: upstream pretraining and downstream detection fine-tuning. Concretely, for a pretext task $\mathcal{T}_p$, the pretraining stage trains the model p from scratch on an upstream pretraining dataset $\mathcal{D}_p$ and saves the pretrained parameters $\boldsymbol{\theta}^p$. Based on the object detection task $\mathcal{T}_{\text{det}}$, the fine-tuning stage first initializes the detector f with $\boldsymbol{\theta}^p$ to form $f(\cdot;\boldsymbol{\theta}^p)$ and then fine-tunes it on a downstream detection dataset $\mathcal{D}_{\text{det}}$, formulated as $f(\cdot;\boldsymbol{\theta}^p) \rightarrow f(\cdot;\boldsymbol{\theta})$. It should be noted that $\boldsymbol{\theta}^p$ may be only a part of $\boldsymbol{\theta}$. For instance, $\boldsymbol{\theta}^p$ typically refers to the pretrained weights in the backbone, while $\boldsymbol{\theta}$ includes all parameters in the entire detector. Since we consider a compact one-stage detector whose parameters are mainly determined by its backbone, this issue will be omitted below for simplicity. Generally, previous object detection algorithms take the supervised image classification on ImageNet [39] as the pretraining task $\mathcal{T}_p$ (i.e., $\mathcal{T}_p$ is ImageNet

SPT), whose objective function can be formulated as follows:

$$\min_{\boldsymbol{\theta}^p} \mathbb{E}_{(x_p,y_p)\sim\mathcal{D}_p}\mathcal{L}_{pt}(x_p, y_p; \boldsymbol{\theta}^p) \tag{1}$$

$$\mathcal{L}_{pt}(x_p, y_p; \boldsymbol{\theta}^p) = \mathcal{L}_{\text{CE}}\big(p(x_p; \boldsymbol{\theta}^p), y_p\big) \tag{2}$$

where x_p denotes the pretraining image sampled from $\mathcal{D}_p$ and y_p represents its class ground truth. $\mathcal{L}_{pt}$ is the pretraining loss function that measures the distance between the output of $p(\cdot; \boldsymbol{\theta}^p)$ and the ground truth, thus minimizing it to result in an appropriate estimation of $\boldsymbol{\theta}^p$ in (1). In practice, ImageNet SPT utilizes the cross-entropy loss $\mathcal{L}_{\text{CE}}$ to instantiate $\mathcal{L}_{pt}$. After pretraining with the initialized detector $f(\cdot; \boldsymbol{\theta}^p)$, fine-tuning can be conducted to further update $\boldsymbol{\theta}^p$ to $\boldsymbol{\theta}$ on $\mathcal{D}_{\text{det}}$, formulated as follows:

$$\min_{\boldsymbol{\theta}\leftarrow\boldsymbol{\theta}^p} \mathbb{E}_{(x_t,\mathbf{b}_t,\mathbf{c}_t)\sim\mathcal{D}_{\text{det}}}\mathcal{L}_{\text{det}}(x_t, \mathbf{b}_t, \mathbf{c}_t; \boldsymbol{\theta}) \tag{3}$$

$$\mathcal{L}_{\text{det}}(x_t, \mathbf{b}_t, \mathbf{c}_t; \boldsymbol{\theta}) = \mathcal{L}_{\text{loc}}\big(f(x_t; \boldsymbol{\theta}), \mathbf{b}_t\big) + \lambda\mathcal{L}_{\text{cls}}\big(f(x_t; \boldsymbol{\theta}), \mathbf{c}_t\big) \tag{4}$$

where x_t is the clean training sample from $\mathcal{D}_{\text{det}}$, while $\mathbf{b}_t$ and $\mathbf{c}_t$ denote the bounding box labels and class labels of all objects in x_t, respectively. It can be observed that the update of $\boldsymbol{\theta}$ is achieved by minimizing the detection loss $\mathcal{L}_{\text{det}}$. λ represents the weight to balance the localization loss $\mathcal{L}_{\text{loc}}$ (e.g., smooth-L1 loss [73]) and the classification loss $\mathcal{L}_{\text{cls}}$ (e.g., focal loss [69]).

B. Robust Object Detection

Akin to object detection, existing robust object detection methods [35], [36], [37] usually adopt the same ImageNet SPT as the pretraining task $\mathcal{T}_p$ to obtain $f(\cdot; \boldsymbol{\theta}^p)$, i.e., (1) and (2) remain unchanged. However, a robust object detector $f(\cdot, \tilde{\boldsymbol{\theta}})$ is supposed to outperform $f(\cdot, \boldsymbol{\theta})$ against potential attacks on adversarial examples. To this end, a common practice lies in the implementation of robust detection-oriented adversarial fine-tuning on $\mathcal{D}_{\text{det}}$, formulated as

$$\min_{\tilde{\boldsymbol{\theta}}\leftarrow\boldsymbol{\theta}^p} \mathbb{E}_{(x_t,\mathbf{b}_t,\mathbf{c}_t)\sim\mathcal{D}_{\text{det}}}\mathcal{L}_{\text{det}}(x_t, \mathbf{b}_t, \mathbf{c}_t; \tilde{\boldsymbol{\theta}}) + \mathcal{L}_{\text{det}}(\tilde{x}_t, \mathbf{b}_t, \mathbf{c}_t; \tilde{\boldsymbol{\theta}}). \tag{5}$$

$\tilde{x}_t \in \mathcal{A}_{\text{det}}$ is the adversarial counterpart of x_t generated by attacking the overall detection loss $\mathcal{L}_{\text{det}}$ and

$$\mathcal{A}_{\text{det}} \triangleq \{\tilde{x}_t \mid \arg\max_{\tilde{x}_t\in\mathcal{X}_t} \mathcal{L}_{\text{det}}(\tilde{x}_t, \mathbf{b}_t, \mathbf{c}_t; \tilde{\boldsymbol{\theta}})\} \tag{6}$$

where

$$\mathcal{X}_t = \{\tilde{x}_t \cap [0, 255]^n \big| \|\tilde{x}_t - x_t\|_\infty \le \epsilon\}. \tag{7}$$

$\mathcal{X}_t$ is defined as the adversarial sample space centered on the clean image x_t with the perturbation budget ϵ. In this way, even with the pretrained weights $\boldsymbol{\theta}^p$ as an initialization, the detector parameters can be estimated by using (6), instead of (3), to form a robust object detector $f(\cdot, \tilde{\boldsymbol{\theta}})$ after fine-tuning.

IV. SASS Pretraining

This article introduces a novel pretraining strategy, namely, SASS pretraining, which is characterized by an adversarial unsupervised (self-supervised) learning paradigm under the guidance of carefully designed structured knowledge and instantiated as SA-MIM. In this section, we first interpret the adversarial SSPT and then present the two crucial components of the proposed SA-MIM, structured adversarial attack and S-MIM, respectively.

A. Adversarial SSPT

Our object detection framework still respects the aforementioned dual-stream training paradigm, i.e., upstream pretraining and downstream fine-tuning. Distinctively, this work proposes to replace the previous full SPT on ImageNet with adversarial SSPT in the remote sensing domain, as illustrated in Fig. 3. Focusing only on pretraining can provide high flexibility for detection-oriented fine-tuning downstream. Moreover, the benefits of adversarial training and self-supervised learning can be combined into a single unified pretraining paradigm that is generalizable to both detection- and robust detection-based downstream tasks. We interpret this pretraining and highlight its characteristics in this section.

1) Problem Definition: Ideally, during adversarial SSPT, the capability of the pretrained models to remove adversarial patterns and encode contextual information should be gradually strengthened with high label efficiency. Thus, this pretraining paradigm is characterized by its "adversary" and "self-supervision." We start with the definition of adversarial pretraining (instead of adversarial fine-tuning), which strives to introduce a proper adversary to the pretraining sample x_p in (1). To find an adversary for x_p, the associated pretraining loss is first maximized over a constrained perturbation δ, belonging to the perturbation distribution Ψ_δ, as follows:

$$\tilde{\delta} = \arg\max_{\delta\in\Psi_\delta} \mathcal{L}_{pt}\big(\mathcal{P}(x_p, \delta), y_p; \boldsymbol{\theta}^p\big) \tag{8}$$

where $\mathcal{P}(x_p, \delta)$ represents an adversarial mapping function that perturbs the clean pretraining example x_p by δ. Then, with the optimized pixelwise perturbations $\tilde{\delta}$, the robust pretrained weights $\tilde{\boldsymbol{\theta}}^p$ can be learned as follows:

$$\tilde{\boldsymbol{\theta}}^p = \arg\min_{\boldsymbol{\theta}^p} \mathbb{E}_{(x_p,y_p)\sim\mathcal{D}_p}\mathcal{L}_{pt}\big(\mathcal{P}(x_p, \tilde{\delta}), y_p; \boldsymbol{\theta}^p\big). \tag{9}$$

The pretrained model, parameterized by $\tilde{\boldsymbol{\theta}}^p$, has a certain degree of resistance against subtle attacks, yet it is sensitive to the pretext task. This is because an image-label pair (x_p, y_p) is required to compute adversaries and update the pretrained weights, leading to the patterns still resembling the conventional fully SPT (e.g., ImageNet SPT taking image classification as a pretext). Instead, to align with unsupervised representation learning without access to manual annotations, adversarial SSPT allows the transformation of the benign input as self-supervision, denoted as $\mathcal{Y}(x_p)$, to replace y_p in (8) and (9), rewritten as follows:

$$\tilde{\delta} = \arg\max_{\delta\in\Psi_\delta} \mathcal{L}_{pt}\big(\mathcal{P}(x_p, \delta), \mathcal{Y}(x_p); \boldsymbol{\theta}^p\big) \tag{10}$$

$$\tilde{\boldsymbol{\theta}}^p = \arg\min_{\boldsymbol{\theta}^p} \mathbb{E}_{x_p\sim\mathcal{D}_p}\mathcal{L}_{pt}\big(\mathcal{P}(x_p, \tilde{\delta}), \mathcal{Y}(x_p); \boldsymbol{\theta}^p\big). \tag{11}$$

It can be observed that adversarial SSPT is defined as a two-step min–max optimization process, and for clarity of demonstration, (10) and (11) can be jointly expressed in a

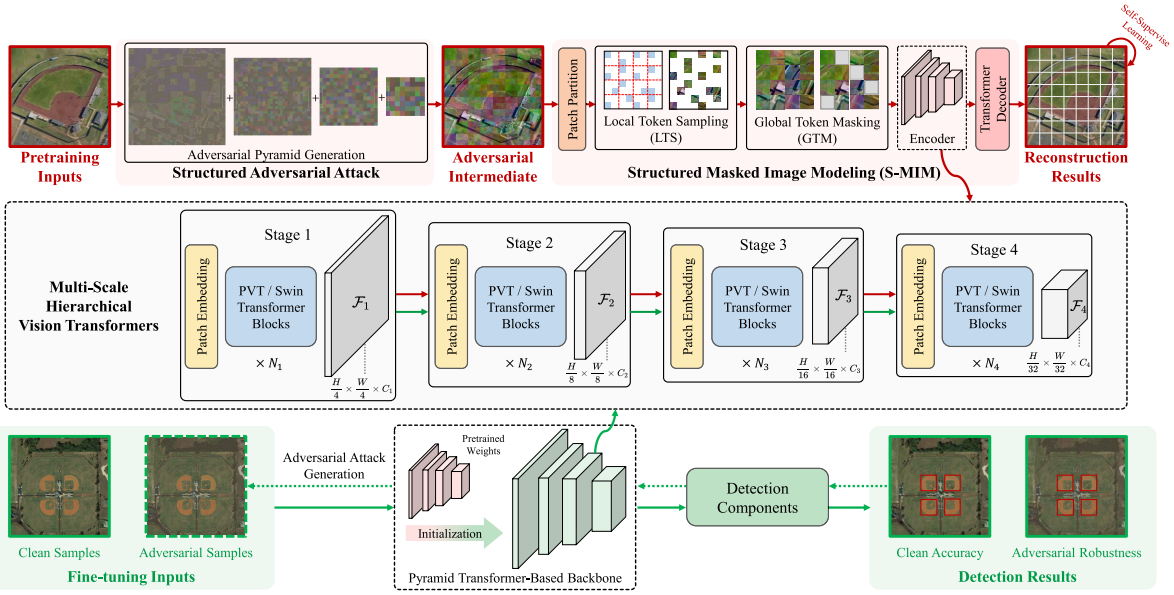

Fig. 3. Overview of the pretraining and fine-tuning paradigm proposed in this work for robust object detection in remote sensing images, where the red and green arrows indicate the data flows in the pretraining and fine-tuning stages, respectively. (Top) Workflow of our SASS pretraining based on the proposed SA-MIM, consisting of structured adversarial attack and S-MIM. (Bottom) Fine-tuning pipeline for robust object detection, leveraging standard adversarial training. The middle shows the hierarchical ViT shared by our pretraining and fine-tuning stages, which can handle multiscale geospatial instances in remote sensing scenarios.

generic form similar to (1) as follows:

$$\min_{\boldsymbol{\theta}^p} \mathbb{E}_{x_p \sim \mathcal{D}_p} \left[\max_{\delta \in \Psi_\delta} \mathcal{L}_{pt}(x_p, \delta; \boldsymbol{\theta}^p) \right] \tag{12}$$

where only clean images x_p are employed as both input and supervisory signals, without external labels. Equation (12) finds pixelwise perturbations by maximizing the pretraining loss and optimizes the network parameters by enhancing the worst case performance of the pretrained model. Accordingly, the generated optimal robust pretrained weights $\tilde{\boldsymbol{\theta}}^p$ can seamlessly substitute $\boldsymbol{\theta}^p$ in (3) and (5), as initialization for natural/adversarial fine-tuning. It is worth noting that the pretext task $\mathcal{T}_{pt}$ and the corresponding objective function $\mathcal{L}_{pt}$ also need to be redesigned in an unsupervised manner to enable (12), which will be introduced in Section IV-C.

2) Self-Supervision in Remote Sensing: The supervisory signal for adversarial SSPT should be independent of the label space. The intuition behind this mechanism is to effectively exploit discrimination entangled in the image or feature space to form automatic supervision. In the context of remote sensing, raw unlabeled data are usually abundant, yet well-labeled data are scarce due to the vast expense of expert annotation [74], which properly harmonizes with the adversarial SSPT scheme. Thus, it is desirable and feasible to explore self-supervision from plentiful uncurated satellite and aerial images during the pretraining phase to benefit extensive downstream remote sensing tasks limited by a lack of annotations [75], [76]. Our goal is to align the proposed adversarial SSPT with the single remote sensing domain, eliminating the dependence on natural scene training samples such as ImageNet to avoid the domain gap from upstream to downstream. In this work, we leverage a very large-scale dataset named Million-AID [77] for remote sensing in-domain pretraining, i.e., Million-AID is utilized as $\mathcal{D}_p$. It contains more than one million nonoverlapping remote sensing scenes with impressive data diversity, richness, and scalability, making it the best choice as our pretraining dataset in practice. On the one hand, unlike its competitive counterparts that mainly involve multispectral imagery [78], [79], all examples in Million-AID are in the RGB format, which is more suitable for pretraining deep vision models and also consistent with the downstream datasets for remote sensing object detection. On the other hand, it covers almost all scenes and categories that will be encountered during fine-tuning, significantly alleviating the difficulty of transfer learning. It is worth noting that although this dataset contains partial annotations, only the images are employed during pretraining, by discarding all labels, to respect an unsupervised routine. Specifically, a raw image, its augmented variants, or the corresponding features can serve as the supervisory signal $\mathcal{Y}(x_p)$ in (10) and (11). We will demonstrate that with the framework using our SASS pretraining, each clean uncurated remote sensing image can be regarded as an ideal prototype for self-supervision, i.e., $\mathcal{Y}(x_p) = x_p$. The above strategy of exploring self-supervision has two favorable merits: 1) internal self-supervision, instead of external supervision, has been proven to boost the model's robustness against uncertainty and outliers [80] and 2) constraining $\mathcal{D}_p$ (for pretraining) and $\mathcal{D}_{\text{det}}$ (for detection fine-tuning) to the same remote sensing context can also address the domain gaps.

3) Multiscale Hierarchical ViTs: For various remote sensing tasks, ViTs have extensively outperformed CNNs due to their proficiency in modeling global dependencies [81]. Meanwhile, recent studies [30], [31], [32] have also demonstrated that Transformers can generalize better on perturbed examples than CNNs with stronger adversarial robustness. Therefore, Transformers appear to be a preferred architecture

for modern detectors or robust detectors, which will be adopted by this work. Moreover, considering the large variation in the scale of remote sensing instances, it is necessary to construct hierarchical Transformer representations during both upstream pretraining and downstream detection fine-tuning, similar to pyramid features in CNNs, rather than simply using the vanilla plain Transformers, such as ViT [21]. This implies that an identical hierarchical Transformer model should be pretrained with respect to the scheme of adversarial SSPT to encode consistent structured knowledge. PVT [23] and Swin Transformer [22] are the two most representative ViTs with pyramid structures, both of which are examined in this work. As illustrated in the middle of Fig. 3, typically, hierarchical representations can be structured by four successive stages, each containing N_i Transformer blocks, where $i \in \{1, 2, 3, 4\}$ denotes the stage index. Given an input pretraining/fine-tuning image $x_p, x_t \in \mathbb{R}^{H\times W\times 3}$, the four output hierarchical representations can be denoted as $\{\mathcal{F}_i\}$, with different feature dimensions C_i. The corresponding spatial scales are fixed to $(H/4) \times (W/4)$, $(H/8) \times (W/8)$, $(H/16) \times (W/16)$, and $(H/32) \times (W/32)$, respectively. This shared Transformer represents the most important architecture to be trained for both pretraining and fine-tuning, and also accounts for most of the computational overhead. It is worth noting that although hierarchical Transformers are off the shelf, the remaining key challenge lies in how to instantiate adversarial SSPT (rather than traditional supervised classification pretraining) to pretrain them for both improved generalization and robustness. In light of this, we propose an efficient pretraining framework, namely, SA-MIM, whose two components will be detailed in Sections IV-B and IV-C.

4) From Adversarial SSPT to SASS Pretraining: As emphasized above, modeling structured knowledge plays an important role in enhancing the transferability of pretrained models, which naturally strengthens the connections between the upstream pretext and downstream target tasks [82]. Essentially, there are two underlying influencing factors, learning difficulty and hierarchical representation consistency, corresponding to the input space and feature space, respectively. Specifically, sufficient pretraining difficulty prevents Transformers from learning ambiguous and less discriminative low-level representations that usually indicate trivial solutions that bypass the designated pretext task, leading to poor downstream transferability. We address this problem by designing and performing structured adversarial attacks on unlabeled inputs and updating them when adversarial SSPT proceeds (see Section IV-B). Meanwhile, hierarchical Transformer representations embody critical feature-level structured information [65], [83], and thus, it is necessary to keep this pretrained structured hierarchy consistent for detection fine-tuning. However, as illustrated in Fig. 2(a) and (b), previous MIM-style pretraining methods either fail to build such hierarchical representations (such as MAE [62]) or suffer from discrepancies between pretraining and fine-tuning (such as SimMIM [63]). To overcome these obstacles, we propose an efficient S-MIM framework (see Section IV-C), successfully constructing a representational hierarchy with structured information.

B. Structured Adversarial Attack

1) Motivations: Adversarial SSPT is characterized by min–max optimization, as defined in (10)–(12), wherein the initial step involves establishing the context for adversarial learning. To achieve this, for Transformers, the most straightforward way is to directly introduce a pixelwise adversarial perturbation δ onto the clean pretraining input to form its perturbed version, formulated as follows:

$$\mathcal{P}(x_p, \delta) = x_p + \delta, \quad \text{s.t. } \delta \in \Psi_\delta. \tag{13}$$

Beneficially, this simple input-level adversary has rendered adversarial pretraining viable, which effectively mitigates the tradeoff between generalization (responsible for clean accuracy) and adversarial robustness [43]. However, it still suffers from two fatal weaknesses. First, (13) imposes insufficient pretraining difficulty for (10) and (11), leading the network to bias toward locally low-level and trivial details while disregarding more discriminative contextual information. Second, pretrained models are prone to overfitting to such a simple type of perturbation and can only inherit limited in-distribution robustness after fine-tuning. The above deficiencies motivate us to propose a more advanced pattern of input-level adversarial attack to replace (13), which should yield more informative adversaries to distort both local details and global context [84], so that the pretrained Transformers are forced to investigate high-level structured knowledge and semantic correlations. Furthermore, it is worth remembering that our ultimate goal is always to enhance both the generalization and robustness of downstream detectors, regardless of whether adversarial fine-tuning is involved. Therefore, this adversarial attack should result in significant image differences yet still preserving the object identity and discrimination to be learned. As illustrated in Fig. 3, we propose a structured adversarial attack that constructs an adversarial pyramid to synthesize adversarial intermediates in the latent space for the subsequent self-supervised pretext task.

2) Adversarial Pyramid Generation: Concretely, our proposed structured adversarial attack directly operates on massive raw remote sensing pretraining images without manual annotations. As depicted in Fig. 3, taking a legitimate example x_p as input, structured multigrained perturbations δ_g are generated and added to x_p [instead of using only one single-grained perturbation δ in (13)] to form the adversarial intermediate $x_p^{(a)}$, formulated as follows:

$$x_p^{(a)} \leftarrow \mathcal{P}(x_p, \delta_g) = \mathcal{C}\left(x_p + \sum_{g\in\mathcal{G}} m_g \cdot \mathcal{B}_g(\delta_g)\right), \quad \text{s.t. } \delta_g \in \Psi_{\delta_g} \tag{14}$$

where $\mathcal{C}(\cdot)$ is a clipping function to keep the generated adversarial intermediates within the normal range and $\mathcal{G} = \{1, 8, 16, 32\}$ is a set of granularities. For instance, $g = 1$ means the value of the pixelwise perturbation in (13). As shown in Fig. 3, different granularities typically indicate perturbing the clean input at different scales, making this attack more flexible and structured. We construct a four-level adversarial pyramid consisting of four different scales, while

the perturbations are constrained at each scale. In particular, $\mathcal{B}_g(\cdot)$ represents a reflection, which aligns the scales of multigrained perturbations with that of x_p to facilitate elementwise addition, and m_g denotes the multiplicative constant that controls the perturbation strength for granularity g. In practice, developing stronger perturbations for coarser granularities is more effective than equal strengths across all granularities, so $m_g = [1, 8, 16, 32]$ by default. In (14), δ_g is the learned multigrained adversarial perturbation, and the generation and optimization for each δ_g follow an iterative gradient-based attack approach, i.e., project gradient descent (PGD) [29], formulated as follows:

$$\delta_g \leftarrow \prod_{\Psi_{\delta_g}} \left(\delta_g + \tau \cdot \operatorname{sign}\left(\nabla_{\delta_g} \mathcal{L}_{pt}(x_p, \delta_g; \boldsymbol{\theta}^p) \right) \right) \tag{15}$$

where ∇_{δ_g} calculates the gradients of the pretraining objective function with respect to the perturbation δ_g and $\operatorname{sign}(\cdot)$ denotes the sign function. τ is the learning rate, and the constraint set for multigrained perturbations is empirically defined as $\Psi_{\delta_g} = \{\delta_g \mid \|\delta_g\|_\infty \leq \varepsilon\}$, based on the l_∞-norm, which confines each perturbation with a specified sphere with the magnitude ε. Notably, ε is fixed to $8/255$ for all levels (or granularities) of the adversarial pyramid. It can be observed that this adversarial pyramid allows for more flexible and higher magnitude perturbations, as shown in Fig. 3, generally suggesting stronger attacks than those for downstream object detection. This structured adversary, encapsulated in (14) and (15), not only establishes the context for adversarial pretraining but also presents increased learning difficulties for better generalizability of the learned representations.

C. Structured MIM

Our proposed structured adversarial attack has successfully produced an adversarial intermediate, followed by a pretext task, to jointly perform adversarial SSPT according to (12). In fact, both contrastive learning [66], [67] and recent MIM [62], [63], [85], [86] can be potentially adopted as the unsupervised pretext task, but the latter stands out in this work due to its two advantages. First, it has been proven [87] that MIM-based pretraining can contribute more to downstream dense prediction tasks, including detection and segmentation, with stronger generalization, compared to contrastive pretraining, which is sensitive to the selection of positive-negative pairs and data augmentation. Second, and more importantly, MIM is tightly coupled with image-patch tokenization in Transformers and synergizes with our structured attack, which also operates on multigrained image patches, enabling more flexible and discriminative representation learning.

1) Motivations: As a self-supervised pretext task, MIM randomly masks some patches of the input and then forces Transformer architectures to make masked predictions only based on the visible patches. Currently, MAE [62] and SimMIM [63] are considered the two most representative MIM-based pretraining strategies. However, neither of them holds the optimal choice for our adversarial SSPT. On the one hand, as illustrated in Fig. 2(a), MAE globally and randomly samples patch tokens with a low sampling ratio (typically 25%), and only these sampled patches are visible and fed into the Transformer encoder. Such design significantly reduces the complexity of pretraining yet adversely destroys the spatial structure of the input image. Therefore, MAE is inapplicable to multiscale hierarchical Transformers. On the other hand, as shown in Fig. 2(b), SimMIM globally selects some patch tokens and replaces them with mask tokens at a high masking ratio (typically 75%). All the informative tokens and mask tokens are fed into the encoder for pretraining. Despite preserving the image structure to accommodate hierarchical Transformers, this approach sacrifices the substantial computational burden of handling uninformative mask symbols. In contrast, an ideal MIM strategy should not only maintain the structure for hierarchical Transformers but also achieve high computational efficiency. We argue that the "global" property of MAE and SimMIM undermines such structured information required by hierarchical Transformers, such as PVT and Swin Transformer, which usually rely on local window-based self-attention. Therefore, to address this issue, inspired by [85] and [86], as depicted in Fig. 2(c), our proposed S-MIM introduces a novel *two-step* strategy that incorporates both local and global operations, ensuring structural integrity and high efficiency. Moreover, S-MIM can be integrated with structured adversarial attacks into a unified end-to-end framework, i.e., SA-MIM in Fig. 2(d), to jointly boost structured knowledge embedding during pretraining.

2) Overview: As illustrated in Fig. 3, S-MIM still adopts an efficient asymmetric encoder–decoder design, similar to MAE, where the hierarchical Transformer is pretrained to serve as the encoder that will be transferred to downstream detection. It follows the "first-uniform-then-random" guideline with two core steps, namely, local token sampling (LTS) and global token masking (GTM). LTS strictly samples a *uniform* number of patches from each local window, preserving the local structure distribution yet sacrificing randomness. Consequently, GTM is further developed to *randomly* and globally masks some already sampled regions as learnable mask tokens. They can complement each other to efficiently convert dense tokens into sparse tokens while ideally modeling structural information.

3) Local Token Sampling: In the proposed SA-MIM framework shown in Fig. 3, the adversarial intermediate $x^{(a)}$ is partitioned into nonoverlapped image patches of size $\mathcal{P} \times \mathcal{P}$ as input. Then, LTS respects a structured sampling constraint: uniform sampling with a ratio of α for each nonoverlapped 2×2 local window in the whole image space. The window size is $\mathcal{W} \times \mathcal{W}$ and $\mathcal{W} := 2\mathcal{P}$. The sampling positions within each window are randomly selected, resulting in a binary sampling map, denoted as $\mathcal{S}_\alpha \in \{0, 1\}^{H \times W}$, where 0 means that the corresponding pixel is sampled out, and vice versa. Accordingly, the sampled patches can be represented as follows:

$$x_p^{(a)} = x_p^{(a)} \odot \mathcal{S}_\alpha \tag{16}$$

where $\odot$ indicates the elementwise multiplication. The total number of visible (sampled) patches in $x_p^{(a)}$ is $\alpha(HW/\mathcal{P}^2)$. The remaining unsampled patches are dropped without participating in the subsequent encoding process to improve computational efficiency. In this way, the number of manipulative

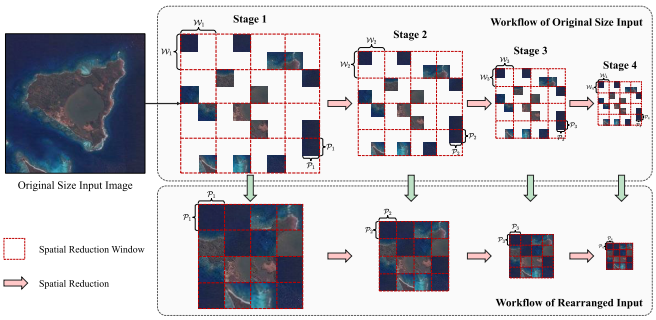

Fig. 4. Illustration of the compatibility of LTS with PVT [23], showcasing the workflow for sampled tokens across four different encoding stages in PVT. Due to the spatial reduction in PVT, the patch size $\mathcal{P}_i$ and the window size $\mathcal{W}_i$ gradually decrease with consecutive stages, subject to $\mathcal{W}_i := 2\mathcal{P}_i$. Specifically, (top) presents the workflow for the original-sized input with blank placeholders, while (bottom) demonstrates the workflow of the rearranged compact input. Since the effective elements in each spatial reduction window are identical across the two pipelines, they are essentially equivalent, and the lower one will be adopted in practice.

elements in each window is consistently equal, thereby enabling information propagation across different stages of the hierarchical Transformer. In practice, as depicted in Fig. 3, α is set to 0.25, i.e., one and only one token will be sampled within each local window. Moreover, LTS can generate uniformly sparse patches compatible with different hierarchical Transformer architectures, including PVT [23] and Swin Transformer [22], with structured locality. Taking PVT in Fig. 4 as an example, the proposed LTS can ideally align with its spatial reduction over nonoverlapped local windows, which is introduced in PVT to construct hierarchical representations and diminish computational complexity. Furthermore, by removing blank placeholders and rearranging the sampled patches, the workflow with compact tokens is directly compatible with PVT.

4) Global Token Masking: Although LTS retains local structured knowledge and is compatible with hierarchical Transformers, its uniformly distributed sampling implicitly provides shortcuts for pixel reconstruction and reduces the difficulty of the self-supervised pretext task [62]. To alleviate this degradation, GTM is proposed to enhance global randomness, which further benefits the discrimination and robustness of the learned latent representations. Concretely, as illustrated in Fig. 3, given the rearranged compact patches $x_p^{(a)}$, GTM performs globally random masking with a masking ratio β, formulated as follows:

$$\tilde{x}_p^{(a)} = x_p^{(a)} \odot \mathcal{M}_\beta + s_{\text{mask}} \odot (1 - \mathcal{M}_\beta) \tag{17}$$

where $\mathcal{M}_\beta$ is a binary random mask with similar characteristics to $\mathcal{S}_\alpha$, while s_{mask} represents the learnable mask symbol. Noticeably, GTS performs masking rather than sampling. Different from LTS, which completely discards the unsampled patches, GTM preserves the structure by replacing the selected patches with shared mask tokens. This simple yet efficient operation makes the reconstruction pretext task more challenging and focuses the encoder to be pretrained on modeling high-quality semantic correlations.

5) Pretraining Target: It is worth noting that in the context of our proposed adversarial SSPT, only the latent adversarial intermediates $x_p^{(a)}$, instead of the original clean pretraining images x_p, are visible to S-MIM. However, the optimization objective is still based on its original version. Specifically, as shown in Fig. 3, an asymmetric encoder–decoder design, similar to MAE [62], is adopted, where the pretrained parameters of the encoder will be inherited by downstream detection, while the lightweight decoder is only utilized during pretraining, aimed at predicting pixel values for the reconstruction pretask. The pretraining loss is defined by the mean square error (mse) between x_p and the reconstructed patches, i.e., the output of the pretraining model $p(x_p^{(a)}; \boldsymbol{\theta}^p)$

$$\mathcal{L}_{pt} = \left\| \left(x_p - p(x_p^{(a)}; \boldsymbol{\theta}^p)\right) \odot \left(1 - \mathcal{S}_\alpha \odot \mathcal{M}_\beta\right) \right\|_2^2. \tag{18}$$

As can be observed from the term $(1 - \mathcal{S}_\alpha \odot \mathcal{M}_\beta)$, the pretraining loss is computed only on the unsampled and masked regions. Moreover, such adversarial reconstruction, based on perturbed patches, implicitly learns adversary elimination, which can encourage the pretraining model to leverage the unmasked structured context to perform imputation through complex semantic reasoning, rather than relying on simple low-level details.

V. Experiments and Analysis

In this section, extensive experiments on both object detection and robust object detection in remote sensing images were conducted to evaluate the effectiveness and superiority of the proposed SASS pretraining.

A. Dataset Description

To comprehensively evaluate the proposed SASS pretraining paradigm, three remote sensing object detection datasets, namely, DIOR [8], NWPU VHR-10 [88], and DOTA [89], are utilized in the experiments.

1) DIOR Dataset: DIOR [8] is currently the largest dataset for object detection in remote sensing images, consisting of 23 463 images with a total of 190 288 instances. It includes 20 geospatial categories: APL, airport (APO), baseball field (BF), basketball court (BC), bridge (BR), chimney (CH), dam (DAM), expressway toll station (ETS), expressway service area (ESA), golf field (GF), ground track field (GTF), harbor (HA), overpass (OP), ship (SH), stadium (STA), storage tank (STO), tennis court (TC), train station (TS), vehicle (VE), and windmill (WM). In the experiments, 1/3, 1/6, and 1/2 of the original images are randomly selected for training, validation, and testing, respectively.

2) NWPU VHR-10 Dataset: NWPU VHR-10 [88] contains a total of 800 very-high-resolution (VHR) optical remote sensing images, of which 650 are positive samples, while the remaining are negatives. This dataset involves ten different geospatial categories, namely, APs, SHs, STs, baseball diamonds (BD), TCs, BCs, GTFs, HAs, BRs, and VEs. In our experiments, positive images are randomly selected for training and testing, with a ratio of 3:1.

3) DOTA Dataset: DOTA [89] is another representative large-scale remote sensing object detection dataset, consisting of 2806 remote sensing images and 188 282 instances in total. Images are categorized into 15 classes, including plane

(PL), BD, BR, ground field track (GFT), small VE (SV), large VE (LV), SH, TC, BC, STO, soccer ball field (SBF), roundabout (RA), HA, swimming pool (SP), and helicopter (HC). In the experiments, this dataset is partitioned into the training set and the test set, with a ratio of 3:1. In addition, considering the extreme range of spatial resolutions of images in this dataset, from less than 800×800 pixels to more than 4000×4000 pixels, we split all images into 600×600 patches with 150-pixel overlap to improve training efficiency.

B. Evaluation Metrics and Implementation Setup

In this work, we are actually concerned with both object detection and robust object detection. The former is evaluated by *clean accuracy*, while the latter is measured by *adversarial robustness*. Specifically, average precision (AP) is utilized as the metric, which is defined as the area under the precision–recall (PR) curve for each category. mAP represents the mean of APs across all categories, with a higher mAP generally indicating better detection performance. To measure the localization precision of different detectors, we also calculate mAPs with diverse intersection-over-union (IoU) thresholds, i.e., mAP_{50}, mAP_{75}, and $\text{mAP}_{50:95}$. It is worth noting that mAP can refer to both clean accuracy and adversarial robustness, depending on whether it is computed on clean samples or adversarial samples. Since our goal is to simultaneously enhance both clean accuracy and adversarial robustness, instead of improving one at the expense of the other, we not only care about their absolute values but also their relative relationship. Therefore, relative rPC [36], which is defined as the ratio between adversarial robustness and clean accuracy, is introduced as an evaluation metric. A higher relative rPC usually suggests a better balance between the two objectives. In the experiments, unless otherwise specified, RetinaNet [69] with PVT-Tiny [23] as the backbone (i.e., the encoder to be pretrained and fine-tuned) is employed as the default remote sensing object detector. During *pretraining*, the models are pretrained for 100 epochs by default, with AdamW [90] as the optimizer, an initial learning rate of 1×10^{-4} and the weight decay of 0.05. During *fine-tuning*, both natural fine-tuning and adversarial fine-tuning will be examined, and all the detectors are fine-tuning for 12 epochs, also optimized by AdamW with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. In addition, for adversarial fine-tuning, we exploit PGD [29] with a budget $\epsilon = 50/255$ to generate adversarial examples. All experiments are conducted on 16 NVIDIA Tesla V100 GPUs.

C. Comparison of Different Pretraining Paradigms

As mentioned above, the proposed SASS pretraining is instantiated as the SA-MIM SSPT framework, which will be quantitatively and qualitatively compared with its counterparts, including the SPT and SSPT frameworks, on the DIOR validation set.

1) Quantitative Results With Natural and Adversarial Fine-Tuning: As stated in the two initial questions, our goal is to improve both the clean accuracy and adversarial robustness of remote sensing object detectors, instead of a tradeoff, whether using natural fine-tuning or adversarial fine-tuning. Tables I and II tabulate the quantitative detection results obtained by natural or adversarial fine-tuning of different pretrained models, respectively. It is worth noting that since our primary objective is to compare the effects of different pretraining paradigms for the same fine-tuned detector, for the sake of simplicity and generality, only vanilla adversarial fine-tuning is employed here. From Tables I and II, it can be that the proposed SA-MIM can extensively outperform other pretraining strategies, in terms of both clean accuracy and adversarial robustness across different evaluation criteria. In particular, with adversarial fine-tuning in Table II, SA-MIM exhibits the highest relative rPC of 89.2%, showcasing a promising balance between the detection generalization and robustness. When considering Tables I and II together, it can be further inferred that previous pretraining frameworks usually achieve better robustness at the expense of clean accuracy, while our method can better align with adversarial fine-tuning and attains simultaneous improvement in both aspects. Moreover, even performing natural fine-tuning on a detector pretrained by SA-MIM can result in comparable detection robustness to adversarial fine-tuning the same detector from scratch (e.g., 18.5% versus 18.2% mAP_{75} and 19.2% versus 19.6% $\text{mAP}_{50:95}$). This implies an advantage of our SASS pretraining for real-world detectors, as natural fine-tuning consumes less computational resources and is, therefore, preferred in practice. In addition, some representative PR curves corresponding to the detection results in Table II are plotted in Fig. 5, where our SA-MIM SSPT consistently exhibits the best detection accuracy and robustness across various categories. It is worth noting that, in some cases, the robustness of our method is even better than the clean accuracy achieved by other methods. Overall, the above quantitative experimental results have validated the effectiveness and superiority of the proposed pretraining paradigm over competitors, including the most popular ImageNet SPT and the recent advanced SimMIM SSPT.

2) Qualitative Comparison: In addition to quantitative analysis, we also qualitatively compare four pretraining paradigms, i.e., ImageNet SPT, SimMIM SSPT, our S-MIM SSPT, and our SA-MIM SSPT, as shown in Figs. 6 and 7. Concretely, Fig. 6 depicts the average attention, represented by heat maps, across all pretraining images. It can be clearly observed that compared to ImageNet SPT that fails to establish impactful attention and instead forms trivial solutions biased toward the image edges, the three SSPT paradigms focus more on valuable and discriminative regions. Moreover, SA-MIM tends to distribute attention more evenly across the whole image to explore structured knowledge during pretraining. This differs from previous typical MIM methods, such as SimMIM [63], which performs global masking and mainly extracts semantics from the central regions of patches. However, pretrained models are expected to encode more generalizable knowledge, rather than overfitting the perceived objects at the center. By introducing locality with structural constraints, our S-MIM mitigates this centralization bias, and SA-MIM further incorporates adversarial learning to disrupt internal correlations and guide the pretrained models to learn more robust and

TABLE I

PERFORMANCE EVALUATION AND COMPARISON OF FOUR DIFFERENT PRETRAINING PARADIGMS WITH NATURAL FINE-TUNING, WHERE THE VALUES BEFORE AND AFTER THE "/" SYMBOL REFER TO CLEAN ACCURACY AND ADVERSARIAL PERFORMANCE, RESPECTIVELY

Pretraining Paradigms	APL GTF	APO HA	BF OP	BC SH	BR STA	CH STO	DAM TC	ESA TS	ETS VE	GF WM	mAP_{50} (ΔmAP_{50})	mAP_{75} (ΔmAP_{75})	$mAP_{50:95}$ ($\Delta mAP_{50:95}$)	Relative rPC
From Scratch	84.3/27.4 61.3/44.8	18.1/3.2 23.6/12.1	94.5/81.3 31.8/ 11.0	48.3/12.2 65.8/19.8	9.2/3.2 88.1/69.1	69.3/13.1 65.4/31.4	21.1/2.1 82.4/57.9	62.0/31.9 13.6/4.0	42.2/7.5 41.9/9.1	31.5/4.9 46.7/0.9	50.1/22.3 (—/—)	26.6/12.8 (—/—)	27.2/12.5 (—/—)	44.5%
ImageNet SPT	88.2/42.3 80.3/48.6	36.6/11.0 28.3/11.9	95.9/82.8 31.1/10.2	57.6/25.3 71.8/30.8	14.5/8.4 95.5/40.3	84.9/3.6 76.7/56.0	36.9/4.4 90.8/76.1	72.7/3.4 26.0/9.3	50.8/13.1 49.0/8.8	60.4/22.5 63.6/**10.5**	60.6/26.0 (+10.5/+3.7)	31.5/13.8 (+4.9/+1.0)	32.4/14.1 (+5.2/+1.6)	42.9%
SimMIM SSPT	87.7/28.1 79.2/**56.1**	41.3/5.6 37.4/16.1	96.1/**87.5** 41.2/21.6	57.4/24.2 74.2/31.0	20.2/8.5 93.2/78.8	82.0/15.3 74.9/51.4	42.3/11.4 90.1/68.4	75.7/30.4 23.9/9.4	53.2/15.3 50.8/9.3	53.9/17.7 66.7/3.6	62.1/29.5 (+12.0/+7.2)	34.6/16.7 (+8.0/+3.9)	34.8/16.7 (+7.6/+4.2)	47.5%
S-MIM SSPT (Ours)	88.2/27.0 80.1/55.2	46.8/7.1 **39.4/19.8**	**96.2**/86.1 **44.3**/19.3	59.3/22.3 **76.2**/33.1	23.6/**9.8** **95.9/81.0**	83.8/13.7 76.3/51.3	46.3/**12.6** **90.9**/68.7	**76.6**/26.0 25.2/10.2	56.6/12.3 **52.1**/8.2	59.1/25.8 70.8/4.3	64.4/29.7 (+14.3/+7.4)	37.5/17.1 (+10.9/+4.3)	36.7/16.8 (+9.5/+4.3)	46.1%
SA-MIM SSPT (Ours)	**88.8/55.6** **82.9**/55.9	**64.6/33.1** 25.1/14.8	95.6/83.1 40.1/**26.0**	**68.7/36.5** 73.5/**42.4**	**23.7**/9.2 94.8/77.6	**88.1/20.7** **77.7/57.6**	**54.9**/4.9 90.0/**79.4**	73.4/**27.4** **34.9/15.7**	**58.7/16.0** 48.8/**16.3**	**77.6/33.2** **71.3**/3.7	**66.7/35.4** (**+16.6/+13.1**)	**38.4/18.5** (**+11.8/+5.7**)	**37.4/19.2** (**+10.2/+6.7**)	**53.1**%

TABLE II

PERFORMANCE EVALUATION AND COMPARISON OF DIFFERENT PRETRAINING PARADIGMS WITH ADVERSARIAL FINE-TUNING, WHERE THE VALUES BEFORE AND AFTER THE "/" SYMBOL REFER TO CLEAN ACCURACY AND ADVERSARIAL PERFORMANCE, RESPECTIVELY

Pretraining Paradigms	APL GTF	APO HA	BF OP	BC SH	BR STA	CH STO	DAM TC	ESA TS	ETS VE	GF WM	mAP_{50} (ΔmAP_{50})	mAP_{75} (ΔmAP_{75})	$mAP_{50:95}$ ($\Delta mAP_{50:95}$)	Relative rPC
From Scratch	83.0/59.2 60.3/52.2	12.7/5.7 26.8/21.4	93.8/90.7 25.8/22.6	41.1/36.4 63.3/49.0	11.5/8.6 88.9/86.4	68.6/52.1 67.4/54.9	20.2/8.6 83.9/78.7	53.7/39.7 11.2/9.3	31.7/19.1 40.7/27.7	43.0/32.9 32.0/6.6	48.0/38.1 (—/—)	23.0/18.2 (—/—)	24.9/19.6 (—/—)	79.4%
ImageNet SPT	86.9/79.9 75.3/68.1	36.3/21.7 32.6/30.0	94.7/93.3 36.0/34.2	49.4/39.9 71.7/59.5	15.8/11.6 90.8/85.9	78.5/63.3 70.5/61.9	36.1/25.5 87.4/83.5	73.0/50.7 26.8/24.4	51.4/32.3 48.5/34.5	44.7/44.1 65.9/37.5	58.6/49.1 (+10.6/+11.0)	30.2/23.0 (+7.2/+4.8)	31.4/25.4 (+6.5/+5.8)	83.8%
SimMIM SSPT	88.1/82.5 74.2/67.4	36.4/28.2 35.0/28.4	95.6/93.2 42.8/39.4	59.3/51.3 72.6/58.6	23.5/17.6 93.3/90.1	82.9/74.7 75.9/68.7	45.7/31.5 90.0/85.2	75.3/63.4 21.1/18.1	58.2/46.6 50.0/35.6	48.8/47.6 70.2/40.8	61.9/53.4 (+13.9/+15.3)	33.5/26.5 (+10.5/+8.3)	34.0/28.4 (+9.1/+8.8)	86.3%
S-MIM SSPT (Ours)	88.2/83.5 79.3/73.0	**48.0**/38.8 35.3/35.9	95.6/94.3 43.3/41.3	61.3/54.0 73.9/61.4	24.8/20.1 93.2/90.9	82.9/74.7 73.8/67.5	**50.9/39.7** 89.9/86.5	79.3/**71.9** 27.8/22.2	57.0/**49.5** 52.2/40.0	55.5/52.1 73.6/48.1	64.3/57.3 (+16.3/+19.2)	36.0/30.2 (+13.0/+12.0)	35.8/31.2 (+10.9/+11.6)	89.1%
SA-MIM SSPT (Ours)	**88.6/85.3** **82.0/76.6**	46.4/**41.2** **43.6/38.7**	**96.4/94.8** **46.7/46.6**	**61.5/54.9** **76.6/62.9**	**27.6/24.4** **96.1/93.0**	**84.9/78.6** **78.3/71.2**	50.6/38.8 **92.1/88.7**	**82.3**/67.3 **32.0/26.6**	**60.5**/48.5 **53.9/43.0**	**58.6/54.2** **78.6/58.0**	**66.9/59.7** (**+18.9/+21.6**)	**39.4/32.9** (**+16.4/+14.7**)	**38.4/33.1** (**+13.5/+13.5**)	**89.2**%

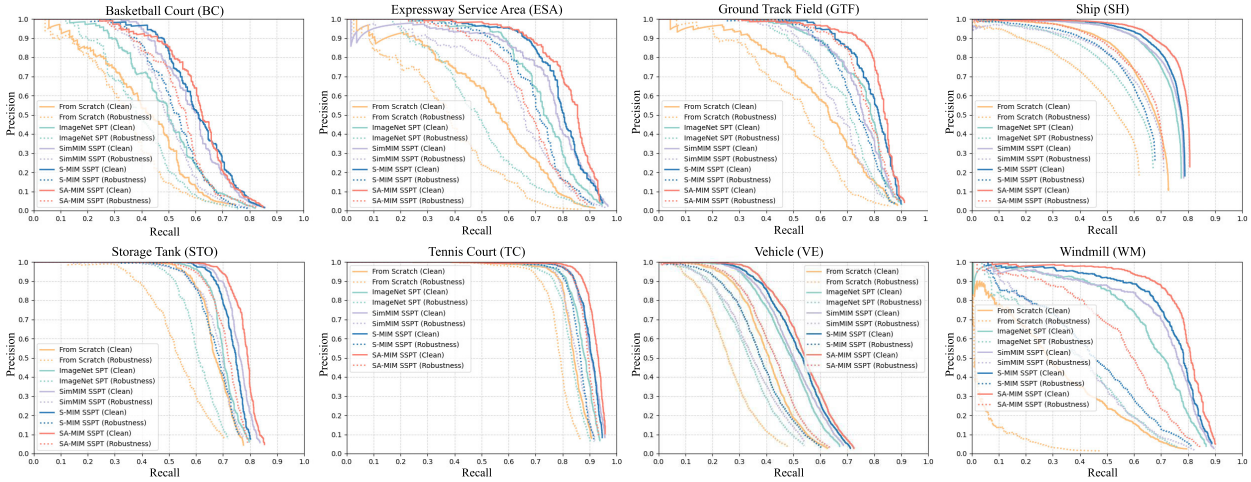

Fig. 5. PR curves of eight representative categories for adversarially fine-tuning the same detector pretrained by five different methods, i.e., from scratch, ImageNet SPT, SimMIM SSPT, S-MIM SSPT, and SA-MIM SSPT. "Clean" and "robustness" refer to clean accuracy and adversarial robustness.

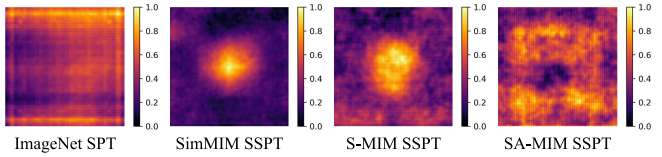

Fig. 6. Normalized average attention of four different pretraining paradigms. The proposed S-MIM SSPT and SA-MIM SSPT can attend to more regions than SimMIM SSPT, which only focuses on the center of images.

generalizable semantics. Fig. 7 illustrates the attention flow of the same adversarially fine-tuned detector with four different pretrained models. As shown in Fig. 7(a)–(e), on clean samples, although all four methods are able to focus the most of the attention on or around potential objects, the three SSPT methods generally outperform ImageNet SPT. Furthermore, our proposed SA-MIM SSPT achieves the most accurate attention distribution, minimizing background distractions. As presented in Fig. 7(f)–(j), on adversarial examples, SSPT appears to perform better against adversarial attacks than ImageNet SPT, which cannot perceive perturbed objects. However, even with adversarial fine-tuning, the detectors pretrained by SimMIM and S-MIM still struggle with imperceptible perturbations, as shown in Fig. 7(h) and (i). In contrast, in Fig. 7(j), SA-MIM showcases the strongest robustness, which demonstrates the superiority of our proposed SASS pretraining.

D. Ablation Studies

We conduct comprehensive ablation studies on the DIOR validation set to further verify the effectiveness of each component or step in the proposed SA-MIM pretraining framework.

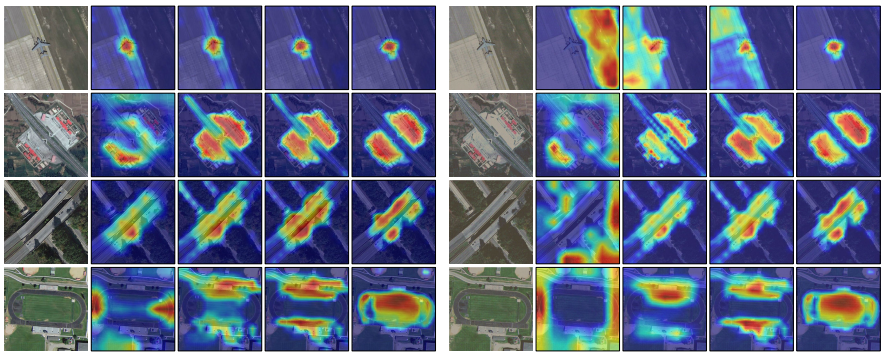

Fig. 7. Visualization of attention flow for different pretrained models after adversarial fine-tuning. (a) and (f) Clean samples and their corresponding adversarial versions, respectively. (b)–(e) Attention maps generated on the clean samples in (a) four different pretraining paradigms, i.e., ImageNet SPT, SimMIM SSPT, S-MIM SSPT, and SA-MIM SSPT, respectively. (g)–(j) Attention maps generated by the four different pretraining paradigms on the adversarial samples in (f).

TABLE III

ABLATION EXPERIMENTS FOR THE PROPOSED STRUCTURED ADVERSARIAL ATTACK ON THE DIOR DATASET, WHERE "AA" IS THE ABBREVIATION OF ADVERSARIAL ATTACKS, WHILE "PIXEL-AA" AND "PATCH-AA" REFER TO THE FINEST-LEVEL AA (I.E., PIXELWISE AA) AND THE COARSET-LEVEL AA

Pretraining Paradigms	Adversarial Attack	Structured Adversaries	Clean Accuracy			Adversarial Robustness			Relative rPC
			mAP_{50}	mAP_{75}	$mAP_{50:95}$	mAP_{50}	mAP_{75}	$mAP_{50:95}$	
ImageNet SPT	×	×	58.6	30.2	31.4	49.1	23.0	25.4	83.8%
S-MIM SSPT	×	×	64.3 (+5.7)	36.0 (+5.8)	35.8 (+4.4)	57.3 (+8.2)	30.2 (+7.2)	31.2 (+5.8)	89.1%
Pixel-AA + S-MIM SSPT	✓	×	65.2 (+6.6)	36.6 (+6.4)	36.5 (+5.10)	57.6 (+8.5)	31.1 (+8.1)	31.6 (+6.2)	88.3%
Patch-AA + S-MIM SSPT	✓	×	64.9 (+6.3)	35.8 (+5.6)	36.1 (+4.7)	56.5 (+7.4)	29.1 (+6.1)	30.5 (+5.1)	87.1%
SA-MIM SSPT (2-Level Pyramid)	✓	✓	65.8 (+7.2)	38.3 (+8.1)	37.2 (+5.8)	58.1 (+9.0)	31.2 (+8.2)	31.8 (+6.4)	88.3%
SA-MIM SSPT (3-Level Pyramid)	✓	✓	66.5 (+7.9)	38.3 (+8.1	37.4 (+6.0)	58.8 (+9.7)	31.4 (+8.4)	32.1 (+6.7)	88.4%
SA-MIM SSPT (4-Level Pyramid)	✓	✓	**66.9 (+8.3)**	**39.4 (+9.2)**	**38.4 (+7.0)**	**59.7 (+10.6)**	**32.9 (+9.9)**	**33.1 (+7.7)**	**89.2%**

TABLE IV

ABLATION EXPERIMENTS FOR THE PROPOSED STRUCTURED ADVERSARIAL ATTACK ON THE OTHER TWO DATASETS: NWPU VHR-10 AND DOTA. "AA" IS THE ABBREVIATION OF ADVERSARIAL ATTACKS, WHILE "PIXEL-AA" AND "PATCH-AA" REFER TO THE FINEST-LEVEL AA AND THE COARSET-LEVEL AA. "CA" AND "AR" REPRESENT "CLEAN ACCURACY" AND "ADVERSARIAL ROBUSTNESS," RESPECTIVELY

Pretraining Paradigms	Adversarial Attack	Structured Adversaries	NWPU VHR-10 Dataset			DOTA Dataset		
			CA mAP_{50}	AR mAP_{50}	Relative rPC	CA mAP_{50}	AR mAP_{50}	Relative rPC
ImageNet SPT	×	×	88.0	63.5	72.2%	40.4	18.8	46.5%
S-MIM SSPT	×	×	91.5 (+3.5)	67.5 (+4.0)	73.8%	42.6 (+2.2)	24.6 (+5.8)	57.7%
Pixel-AA + S-MIM SSPT	✓	×	92.4 (+4.4)	71.3 (+7.8)	77.2%	43.4 (+3.0)	29.2 (+10.4)	67.3%
Patch-AA + S-MIM SSPT	✓	×	92.2 (+4.2)	70.8 (+7.3)	76.8%	43.1 (+2.7)	28.1 (+9.3)	65.2%
SA-MIM SSPT (2-Level Pyramid)	✓	✓	93.3 (+5.3)	72.7 (+9.2)	77.9%	44.2 (+3.8)	31.6 (+12.8)	71.5%
SA-MIM SSPT (3-Level Pyramid)	✓	✓	94.0 (+6.0)	75.5 (+12.0)	80.3%	**44.9 (+4.5)**	33.4 (+14.6)	74.4%
SA-MIM SSPT (4-Level Pyramid)	✓	✓	**94.1 (+6.1)**	**76.4 (+12.9)**	**81.2%**	**44.9 (+4.5)**	**34.6 (+15.8)**	**77.1%**

1) Effect of Structured Adversarial Attack: As depicted in Fig. 3, the structured adversarial attack is the first component of SA-MIM, which constructs an adversarial pyramid to generate adversarial intermediates in the latent space. We ablate its effect on the DIOR dataset in Table III and the NWPU VHR-10 and DOTA datasets in Table IV. As can be seen from Tables III and IV, with single-level adversarial attacks only, i.e., without establishing structured perturbations, the detectors pretrained by "Pixel-AA + S-MIM SSPT" and "Patch-AA + S-MIM SSPT" outperform ImageNet SPT and S-MIM SSPT, neither of which perform adversarial attacks, in terms of clean accuracy and adversarial robustness. This validates the benefits of incorporating adversarial learning into pretraining for downstream detection fine-tuning. As introduced in Section IV-B, the proposed structured adversarial attack constructs an adversarial pyramid through multigrained perturbations. Specifically, it can be concluded that when leveraging the four-level structured adversaries, we can obtain the best detection performance across different evaluation metrics, e.g., in Table III, an accuracy gain of +8.3% mAP_{50} and a robustness gain of +10.6% mAP_{50}. Similar benefits can also be observed from Table IV. This verifies the superiority of introducing structured adversaries over simply using a single-level adversary.

Visualization of Multigrained Perturbations: We include some exemplar visualizations of structured adversarial attacks against legitimate pretraining images in Fig. 8. There are four different levels of adversarial perturbation, from coarse-grained to fine-grained, with the first being a regular pixelwise perturbation. Perturbations of different granularities serve distinct purposes. Specifically, as illustrated in Fig. 8, the finest-grained pixelwise perturbation captures and disrupts

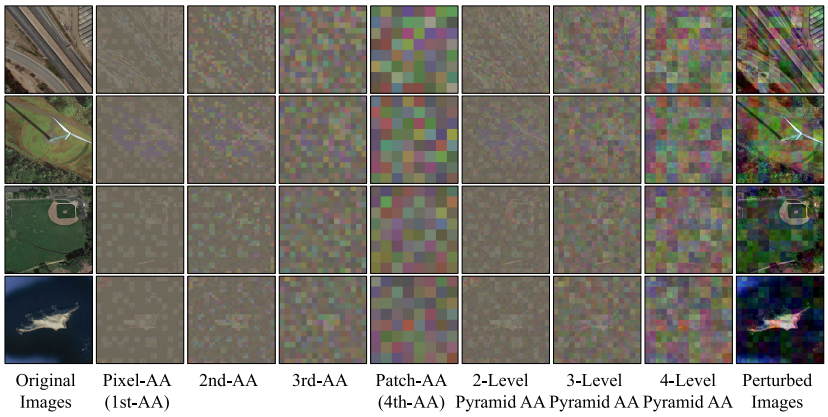

Fig. 8. Visualizations of the proposed structured adversarial attack on the original pretraining images. "AA" refers to the abbreviation of adversarial attacks, and there are four different levels of AA that can be developed to construct an adversarial pyramid. Typically, the four-level pyramid AA is utilized to generate perturbed images.

texture structure information, while coarser granularity typically allows higher perturbation thresholds, enabling stronger attacks to perturb high-level semantics and provide more challenging pretraining objectives. Furthermore, it is worth noting that, thanks to the characteristics of gradient-based updates, such adversarial attacks only affect the sampled regions to avoid invalid and redundant computations on invisible patches. By visually comparing the original images with the corresponding perturbed images in Fig. 8, we can reveal that integrating multigrained perturbations (the four-level pyramid is constructed by default) noticeably disrupts low-level details and high-level correlations, facilitating the pretrained models to explore more robust and generalizable semantic knowledge for the reconstruction pretask.

2) Effect of Principal Components in S-MIM: As mentioned earlier, S-MIM can be separated from SA-MIM to independently function as a self-supervised pretext task. Table V ablates its two core steps or components, i.e., LTS and GTM, and their corresponding hyperparameters. ImageNet SPT can be regarded as the baseline without performing MIM during pretraining, while both MAE [62] and SimMIM [63] have been introduced as recent MIM-based SSPT counterparts. Since the MAE random sampling strategy is not applicable to hierarchical Transformers, we can only adopt its variant, namely, gridwise sampling (GS), which is also proposed in MAE [62], for comparison in Table V. It can be observed that leveraging only LTS with the sampling ratio $\alpha = 0.25$, S-MIM outperforms MAE in terms of both clean accuracy and adversarial robustness and presents comparable accuracy to SimMIM yet inferior robustness. This phenomenon highlights the benefit of preserving structured locality by LTS yet suggests the potential hazards of overreliance on such local cues for downstream predictions on adversarial samples. The local cues may be misleading or even destroyed, consequently hindering the detection robustness. As tabulated in the last three rows of Table V, further introducing GTM can alleviate this issue and simultaneously improve accuracy and robustness. When the masking ratio $\beta = 0.25$, S-MIM consistently achieves the best performance across all metrics. However, if β is set too high, such as 0.40, less discriminative information can be preserved. This will adversely affect the reasoning capability of the pretrained models and lead to the degradation of fine-tuning performance. The above results validate the effectiveness of the proposed LTS and GTM in S-MIM.

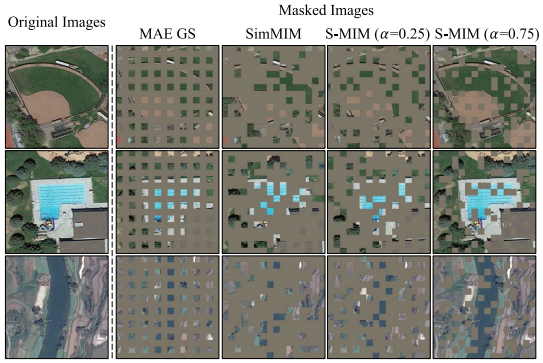

Fig. 9. Visualization of different token sampling or token masking strategies. MAE [62] and SimMIM [63] adopt global gridwise sampling and global random masking, respectively, while our proposed LTS performs local sampling under structured local constraints.

Visualization of Different Sampling/Masking Strategies: Fig. 9 visualizes three different token sampling or masking strategies, including MAE global GS [62], SimMIM global random masking [63], and our proposed LTS. Token sampling and masking strategies can determine the pretext task difficulty, further influencing the quality of the pretrained representations. The grid layout in MAE significantly reduces sampling randomness and simplifies reconstruction. Global random masking in SimMIM introduces the highest randomness yet lacks stability, misleading the pretrained encoders to focus on irrelevant details and hindering its downstream transferability. The proposed LTS strikes a balance between randomness and stability by incorporating structured locality, with adequate learning difficulty. However, it can be seen that setting the sampling ratio too high, such as 0.75 in Fig. 9, will noticeably undermine the reconstruction difficulty, and thus, it is set to 0.25 by default.

3) Effect of Longer Pretraining Schedules: Table VI investigates the benefit of adopting longer pretraining schedules for fine-tuning. By default, SA-MIM is pretrained for only

TABLE V

ABLATION STUDIES OF THE PROPOSED S-MIM AND ITS TWO KEY HYPERPARAMETERS. "MAE GS" REPRESENTS THE GRIDWISE SAMPLING STRATEGY PROPOSED IN MAE [62]. MAE [62] AND SIMMIM [63] INVOLVE EITHER TOKEN SAMPLING OR TOKEN MASKING, WHILE OUR S-MIM INCORPORATES BOTH IN LTS AND GTM SEQUENTIALLY WITH TWO HYPERPARAMETERS α AND β

Pretraining Paradigms	Sampling Ratio α	Masking Ratio β	Clean Accuracy			Adversarial Robustness			Relative rPC
			mAP_{50}	mAP_{75}	$mAP_{50:95}$	mAP_{50}	mAP_{75}	$mAP_{50:95}$	
ImageNet SPT	–	–	58.6	30.2	31.4	49.1	23.0	25.4	83.8%
MAE GS SSPT [62]	0.25	–	57.9 (-0.7)	29.9 (-0.3)	31.1 (-0.3)	46.2 (-2.9)	22.0 (-1.0)	24.0 (-1.4)	79.8%
SimMIM SSPT [63]	–	0.75	61.9 (+3.3)	33.5 (+3.3)	34.0 (+2.6)	53.4 (+4.3)	26.5 (+3.5)	28.4 (+3.0)	86.3%
S-MIM (*w/* LTS only) SSPT	0.25	–	59.0 (+0.4)	31.0 (+0.8)	32.1 (+0.7)	47.7 (-1.4)	22.9 (-0.1)	25.0 (-0.4)	80.8%
S-MIM (*w/* LTS only) SSPT	0.75	–	56.4 (-2.2)	28.4 (-1.8)	29.9 (-1.5)	45.4 (-3.7)	21.5 (-1.5)	23.5 (-1.9)	80.5%
S-MIM (*w/* LTS + GTM) SSPT	0.25	0.10	63.3 (+4.7)	34.2 (+4.0)	34.8 (+3.4)	56.0 (+6.9)	28.8 (+5.8)	30.1 (+4.7)	88.5%
S-MIM (*w/* LTS + GTM) SSPT	0.25	0.25	**64.3 (+5.7)**	**36.0 (+5.8)**	**35.8 (+4.4)**	**57.3 (+8.3)**	**30.2 (+7.2)**	**31.2 (+5.8)**	**89.1%**
S-MIM (*w/* LTS + GTM) SSPT	0.25	0.40	61.3 (+2.7)	32.3 (+2.1)	33.3 (+1.9)	52.3 (+3.2)	25.4 (+2.4)	27.6 (+2.2)	85.3%

TABLE VI

EFFECT OF DIFFERENT PRETRAINING EPOCHS. THE DEFAULT NUMBER OF PRETRAINING EPOCHS IS FIXED TO 100, WHILE LONGER PRETRAINING SCHEDULES CAN BRING MORE IMPROVEMENTS IN DETECTION FINE-TUNING PERFORMANCE

Pretraining Schedules	Clean Accuracy			Adversarial Robustness			Relative rPC
	mAP_{50}	mAP_{75}	$mAP_{50:95}$	mAP_{50}	mAP_{75}	$mAP_{50:95}$	
SA-MIM SSPT 100-Epochs	66.9	39.4	38.4	59.7	32.9	33.1	89.2%
SA-MIM SSPT 200-Epochs	68.8 (+1.9)	41.7 (+2.3)	39.9 (+1.5)	63.8 (+4.1)	36.6 (+3.7)	36.2 (+3.1)	92.7%
SA-MIM SSPT 300-Epochs	69.0 (+2.1)	41.2 (+1.8)	39.8 (+1.4)	65.3 (+5.6)	37.6 (+4.7)	36.8 (+3.7)	94.6%
SA-MIM SSPT 400-Epochs	**70.5 (+3.6)**	**43.5 (+4.1)**	**41.7 (+3.3)**	65.8 (+6.1)	38.6 (+5.7)	37.8 (+4.7)	93.3%
SA-MIM SSPT 500-Epochs	69.5 (+2.6)	42.7 (+3.3)	40.8 (+2.4)	65.7 (+6.0)	38.5 (+5.6)	37.7 (+4.6)	94.5%
SA-MIM SSPT 600-Epochs	70.3 (+3.4)	43.5 (+4.1)	41.6 (+3.2)	**66.8 (+7.1)**	**39.8 (+6.9)**	**38.6 (+5.5)**	**95.0%**

TABLE VII

COMPARISON OF DETECTION PERFORMANCE OF DIFFERENT MULTISCALE HIERARCHICAL TRANSFORMERS WITH DIFFERENT MODEL SIZES AS THE ENCODERS FOR PRETRAINING AND FINE-TUNING

Encoders (Backbones)	Pretraining Paradigms	Clean Accuracy			Adversarial Robustness			Relative rPC
		mAP_{50}	mAP_{75}	$mAP_{50:95}$	mAP_{50}	mAP_{75}	$mAP_{50:95}$	
PVT-Tiny	ImageNet SPT	58.6	30.2	31.4	49.1	23.0	25.4	83.8%
	SA-MIM SSPT	66.9 (+8.3)	39.4 **(+9.2)**	38.4 **(+7.0)**	59.7 (+10.6)	32.9 **(+9.9)**	33.1 **(+7.7)**	89.2%
PVT-Small	ImageNet SPT	60.3	33.5	33.5	50.7	27.3	27.8	84.1%
	SA-MIM SSPT	68.7 **(+8.4)**	41.8 (+8.3)	40.0 (+6.5)	60.6 (+9.9)	34.2 (+6.9)	34.2 (+6.4)	88.2%
PVT-Medium	ImageNet SPT	62.9	35.1	35.1	54.0	29.1	29.7	85.9%
	SA-MIM SSPT	**69.7** (+6.8)	**41.9** (+6.8)	**40.5** (+5.4)	**64.8 (+10.8)**	37.2 (+8.1)	36.6 (+6.9)	93.0%
PVT-Large	ImageNet SPT	61.7	34.5	34.2	51.3	27.8	28.1	83.1%
	SA-MIM SSPT	66.9 (+5.2)	40.6 (+6.1)	38.7 (+4.5)	61.8 (+10.5)	36.4 (+8.6)	35.3 (+7.2)	92.4%
Swin-Tiny	ImageNet SPT	61.4	35.4	34.8	55.8	30.7	31.0	90.9%
	SA-MIM SSPT	66.7 (+5.3)	40.8 (+5.4)	39.2 (+4.4)	63.8 (+8.0)	37.6 (+6.9)	36.6 (+5.6)	95.7%
Swin-Small	ImageNet SPT	61.1	34.3	33.9	56.2	31.0	30.8	92.0%
	SA-MIM SSPT	68.0 (+6.9)	41.7 (+7.4)	39.9 (+6.0)	64.5 (+8.3)	**38.9** (+7.9)	**37.4** (+6.6)	**94.9%**

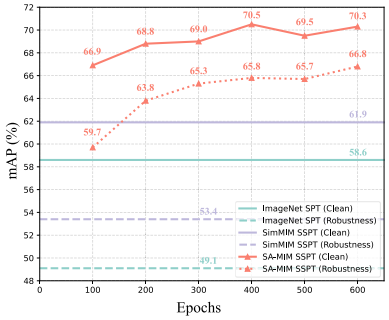

Fig. 10. Comparison of SA-MIM at different pretraining epochs and with two competitors in terms of detection performance after adversarial fine-tuning.

100 epochs to save computational resources and time. The experimental results demonstrate that appropriately increasing the number of pretraining epochs can lead to better detection performance, in terms of accuracy and robustness. Moreover, significant improvements in terms of relative rPC can be observed in Table VI, which indicates that our proposed SASS pretraining not only effectively enhances detection accuracy and robustness simultaneously but also gradually minimizes their discrepancy. This is highly desirable for practical applications. Fig. 10 shows that SA-MIM consistently outperforms its competitors, including ImageNet SPT and SimMIM SSPT, regardless of the number of pretraining epochs.

4) Effect of Different Pretrained Encoders: For experimental efficiency, we mainly leverage PVT-Tiny [23] as the encoder to be pretrained and fine-tuned in ablation studies. Table VII explores whether the proposed SA-MIM can be scaled to larger models and different hierarchical Transformer architectures. Experimental results demonstrate that the proposed pretraining paradigm can consistently improve the detection performance of PVT series at various scales, achieving a maximum accuracy gain of +8.4 mAP_{50} and a maximum robustness gain of +10.8 mAP_{50}. As illustrated in Table VII, SA-MIM is also compatible with Swin Transformer [22], outperforming ImageNet SPT by a large margin.

5) Visualization of Reconstructions: We visualize the SA-MIM-based pretraining process in Fig. 11, along with

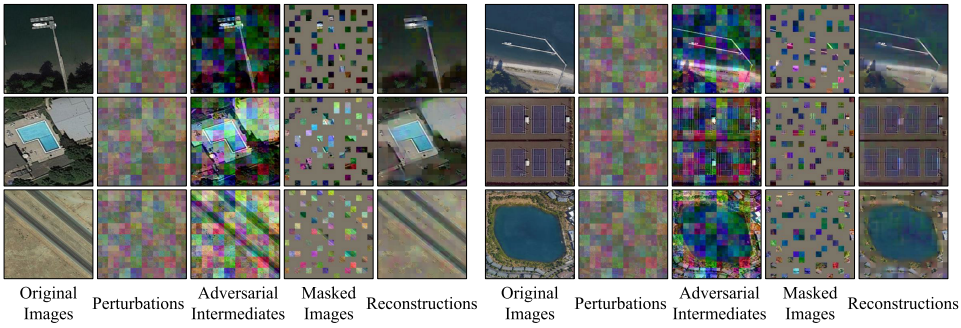

Fig. 11. Reconstructions of the proposed SA-MIM pretraining. For better visualizations, the sampled visible patches from the original legitimate images are displayed at the corresponding positions of the final reconstruction results.

TABLE VIII

COMPARISON OF DETECTION PERFORMANCE OF IMAGENET SPT AND THE PROPOSED SA-MIM SSPT WITH DIFFERENT FINE-TUNING STRATEGIES

Pretraining Paradigms	Fine-tuning Strategies	Clean Accuracy			Adversarial Robustness			Relative rPC
		mAP_{50}	mAP_{75}	$mAP_{50:95}$	mAP_{50}	mAP_{75}	$mAP_{50:95}$	
ImageNet SPT	*w/o* Adversarial Fine-tuning	60.6	31.5	32.4	26.0	13.8	14.1	42.9%
	Vanilla Adversarial Fine-tuning	58.6 (-2.0)	30.2 (-1.3)	31.4 (-1.0)	49.1 (+23.1)	23.0 (+9.2)	25.4 (+11.3)	83.8%
	Structured Adversarial Fine-tuning	58.4 (-2.2)	29.4 (-2.1)	30.7 (-1.7)	51.4 (+25.4)	24.6 (+10.8)	26.4 (+12.3)	88.0%
SA-MIM SSPT	*w/o* Adversarial Fine-tuning	66.7	38.4	37.4	35.4	18.5	19.2	53.1%
	Vanilla Adversarial Fine-tuning	**66.9 (+0.2)**	**39.4 (+1.0)**	**38.4 (+1.0)**	59.7 (+24.3)	32.9 (+14.4)	33.1 (+13.9)	89.2%
	Structured Adversarial Fine-tuning	66.5 (-0.2)	37.0 (-1.4)	37.0 (-0.4)	**62.2 (+26.8)**	**33.3 (+14.8)**	**33.9 (+14.7)**	**93.5%**

some exemplar intermediate representations and final reconstruction results. It can be observed that, under multigrained perturbations, the adversarial intermediates have presented dramatic differences from their original versions. Then, the low-level details and high-level contextual relations are further destroyed, as shown in the derived masked adversarial images. Interestingly, the foreground geospatial regions of interest can still be roughly recovered after pretraining, showcasing the ability of SA-MIM to plausibly explore discriminative semantics with strong generalization. More importantly, in Fig. 11, the reconstructions depict a notable absence of adversaries, suggesting that the pretrained models have a certain level of resilience against potential adversarial attacks, i.e., improved robustness.

6) Effect of Complex Adversarial Fine-Tuning Strategies: As mentioned above, by default, we only exploit the simplest vanilla adversarial fine-tuning in the experiments for efficiency. Table VIII further investigates the efficacy of adapting the proposed structured adversarial attacks as a more complex adversarial fine-tuning strategy, namely, structured adversarial fine-tuning, for downstream detection fine-tuning. When utilizing ImageNet SPT as the pretraining paradigm, the two adversarial fine-tuning strategies can improve robustness by sacrificing detection accuracy. With SA-MIM SSPT, using vanilla adversarial fine-tuning can enhance both clean accuracy and adversarial robustness, whereas structured adversarial fine-tuning leads to a slight decrease in accuracy, albeit providing more robustness improvements. This phenomenon demonstrates the effectiveness of structured attacks in boosting adversarial robustness, both for pretraining and fine-tuning. However, it also indicates that more complex adversarial fine-tuning may unavoidably degrade clean accuracy, which is an undesirable compromise in remote sensing scenarios. Therefore, adversarial pretraining is more effective and advantageous to adversarial fine-tuning in the context of accurate and robust remote sensing object detection.

7) Generalization on Oriented Object Detection: The above experiments focus on horizontal object detection, while oriented object detection is also prevalent in the field of remote sensing object detection. Thus, it is highly instructive to evaluate the generalization of the proposed pretraining paradigm on this task. Table IX compares our SA-MIM SSPT with the traditional ImageNet SPT based on three recent representative oriented object detectors, i.e., R^3Det [91], S^2A-Net [92], and ReDet [93]. It can be observed that replacing ImageNet SPT with our SA-MIM SSPT can consistently lead to performance improvements in terms of clean accuracy and adversarial robustness. Specifically, empowered by SA-MIM SSPT, S^2A-Net attains an accuracy of 64.9% mAP_{50} and a robustness of 58.2% mAP_{50}, remarkably outperforming the original ImageNet SPT-based version. These results validate the effectiveness of our pretraining paradigm on the task of remote sensing-oriented object detection.

E. Generalization of Detection Robustness on Other Datasets

We conducted experiments on two other remote sensing object detection datasets for generalization validation and tabulated the results in Table X. It can be seen that, on the NWPU VHR-10 dataset, SA-MIM SSPT consistently outperforms the traditional ImageNet SPT. Due to the limited training samples in this dataset, adversarial fine-tuning, which can be considered a form of data augmentation, plays an important role in enhancing both accuracy and robustness. Thus, its absence results in a drastic performance decline. However, meanwhile, a distinct and noteworthy observation comes on the DOTA dataset that contains sufficient training samples. Surprisingly, even without any adversarial fine-tuning (i.e., with only natural fine-tuning), our SA-MIM SSPT paradigm achieves superior or comparable accuracy and robustness

TABLE IX

PERFORMANCE COMPARISON OF IMAGENET SPT AND THE PROPOSED SA-MIM SSPT FOR THE TASK OF ORIENTED OBJECT DETECTION

Oriented Object Detectors	Pretraining Paradigms	Clean Accuracy			Adversarial Robustness			Relative rPC
		mAP_{50}	mAP_{75}	$mAP_{50:95}$	mAP_{50}	mAP_{75}	$mAP_{50:95}$	
R^3Det [91]	ImageNet SPT	55.5	25.7	28.0	40.9	15.1	19.3	73.7%
	SA-MIM SSPT	60.2 **(+4.7)**	30.4 **(+4.7)**	31.6 (+3.6)	51.3 (+10.4)	25.0 **(+9.9)**	26.9 **(+7.6)**	85.2%
S^2A-Net [92]	ImageNet SPT	60.7	25.9	30.2	47.6	17.7	22.2	78.4%
	SA-MIM SSPT	**64.9** (+4.2)	29.0 (+3.1)	33.5 (+3.3)	**58.2** (+10.6)	24.1 (+6.4)	29.0 (+6.8)	**89.7%**
ReDet [93]	ImageNet SPT	60.3	36.0	34.4	45.8	27.1	26.3	76.0%
	SA-MIM SSPT	64.6 (+4.3)	**40.1** (+4.1)	**38.1 (+3.7)**	56.5 **(+10.7)**	**34.3** (+7.2)	**33.1** (+6.8)	87.5%

TABLE X

DETECTION PERFORMANCE COMPARISON OF IMAGENET SPT AND THE PROPOSED SA-MIM SSPT ON THE OTHER TWO DATASETS

Pretraining Paradigms	Fine-tuning Strategies	NWPU VHR-10 Dataset					DOTA Dataset				
		Clean Accuracy		Adversarial Robustness		Relative rPC	Clean Accuracy		Adversarial Robustness		Relative rPC
		mAP_{50}	$mAP_{50:95}$	mAP_{50}	$mAP_{50:95}$		mAP_{50}	$mAP_{50:95}$	mAP_{50}	$mAP_{50:95}$	
ImageNet SPT	Vanilla Adversarial Fine-tuning	88.0	53.5	63.5	34.1	72.2%	40.4	20.6	18.8	9.3	46.5%
SA-MIM SSPT	*w/o* Adversarial Fine-tuning	79.1 (-8.9)	45.3 (-8.2)	33.2 (-30.3)	16.0 (-18.1)	42.0%	44.0 (+3.6)	21.9 (+1.3)	19.4 (+0.6)	9.3 (+0.0)	44.1%
	Vanilla Adversarial Fine-tuning	**94.1 (+6.1)**	**60.5 (+7.0)**	76.4 (+12.9)	42.0 (+7.9)	81.2%	**44.9 (+4.5)**	**23.6 (+3.0)**	34.6 (+15.8)	18.2 (+8.9)	77.1%
	Structured Adversarial Fine-tuning	93.1 (+5.1)	58.5 (+5.0)	**88.0 (+24.5)**	**51.2 (+17.1)**	**94.5%**	44.6 (+4.2)	23.0 (+2.4)	**37.3 (+18.5)**	**18.9 (+9.6)**	**83.6%**

TABLE XI

DETECTION ACCURACY COMPARISON OF VARIOUS DETECTORS ON THE DIOR DATASET. OTHER DETECTORS ARE GENERALLY BASED ON IMAGENET SPT, WHILE OURS LEVERAGES THE PROPOSED SA-MIM SSPT. † REPRESENTS THAT PVT-SMALL IS ADOPTED AS THE DETECTION BACKBONE. THE TOP TWO BEST RESULTS IN EACH COLUMN ARE REPRESENTED IN BOLD FONT

Detectors	Pretraining Paradigms	APL	APO	BF	BC	BR	CH	DAM	ESA	ETS	GF	GTF	HA	OP	SH	STA	STO	TC	TS	VE	WM	mAP_{50}
HRBM [94]	ImageNet SPT	42.2	69.7	62.0	79.0	27.7	68.9	50.1	60.5	49.3	64.4	65.3	42.3	46.8	11.7	53.5	24.5	70.3	53.3	20.4	56.2	50.9
SSD [95]	ImageNet SPT	59.5	72.7	72.4	75.7	29.7	65.8	56.6	63.5	53.1	65.3	68.6	49.4	48.1	59.2	61.0	46.6	76.3	55.1	27.4	65.7	58.6
CornerNet [96]	ImageNet SPT	58.8	84.2	72.0	80.8	**46.4**	75.3	64.3	81.6	**76.3**	79.5	79.5	26.1	60.6	37.6	70.7	45.2	84.0	57.1	43.0	75.9	64.9
PANet [97]	ImageNet SPT	60.2	72.0	70.6	80.5	43.6	72.3	61.4	72.1	66.7	72.0	73.4	45.3	56.9	71.7	70.4	62.0	80.9	57.0	47.2	84.5	66.1
CF2PN [98]	ImageNet SPT	78.3	78.3	76.5	88.4	37.0	71.0	59.9	71.2	51.2	75.6	77.1	**56.8**	58.7	**76.1**	70.6	55.5	**88.8**	50.8	36.9	86.4	67.3
CSFF [99]	ImageNet SPT	57.2	79.6	70.1	87.4	46.1	76.6	62.7	82.6	73.2	78.2	81.6	50.7	59.5	73.3	63.4	58.5	85.9	61.9	42.9	**86.9**	68.0
O^2-DNet [100]	ImageNet SPT	61.2	80.1	73.7	81.4	45.2	75.8	64.8	81.2	**76.5**	79.5	79.7	47.2	59.3	72.6	70.5	53.7	82.6	55.9	**49.1**	77.8	68.4
SCRDet++ [101]	ImageNet SPT	64.3	79.0	73.2	85.7	45.8	76.0	68.4	79.3	68.9	77.7	77.9	**56.7**	**62.2**	70.4	67.7	60.4	80.9	63.7	44.4	84.6	69.4
MSFC-Net [102]	ImageNet SPT	**85.8**	76.2	74.4	**90.1**	44.2	78.1	55.5	60.9	59.5	76.9	73.7	49.6	57.2	**89.6**	69.2	**76.5**	86.7	51.8	**55.2**	84.3	70.1
SB-MSN [103]	ImageNet SPT	**79.6**	82.2	76.4	**89.8**	45.6	78.2	64.8	58.9	59.3	79.2	**82.4**	51.8	60.8	74.4	**79.7**	**66.4**	85.6	65.4	45.1	79.9	70.3
GLNet [104]	ImageNet SPT	62.9	83.2	72.0	81.1	**50.5**	79.3	67.4	86.2	70.9	81.8	**83.0**	51.8	**62.6**	72.0	75.3	53.7	81.3	**65.5**	43.4	**89.2**	70.7
RetinaNet (Ours)	SA-MIM SSPT	70.4	**85.3**	**78.8**	86.7	39.5	**80.4**	**71.1**	**86.8**	60.3	**84.6**	80.7	49.1	59.0	70.1	77.1	56.6	88.6	64.4	40.9	84.6	**70.8**
RetinaNet† (Ours)	SA-MIM SSPT	79.0	**86.4**	**83.4**	87.0	38.5	**79.7**	**72.3**	**87.1**	58.8	**85.7**	82.1	50.6	59.2	70.7	**81.0**	55.8	**89.8**	**65.7**	40.9	84.8	**71.9**

TABLE XII

DETECTION ACCURACY COMPARISON OF VARIOUS DETECTORS ON THE NWPU VHR-10 DATASET. OTHER DETECTORS ARE GENERALLY BASED ON IMAGENET SPT, WHILE OURS LEVERAGES THE PROPOSED SA-MIM SSPT. THE TOP TWO BEST RESULTS IN EACH COLUMN ARE REPRESENTED IN BOLD FONT

Detectors	Pretraining Paradigms	AP	SH	ST	BD	TC	BC	GTF	HA	BR	VE	mAP_{50}
RICNN [88]	ImageNet SPT	88.4	77.3	85.3	88.1	40.8	58.5	86.7	68.6	61.5	71.1	72.6
SSD [95]	ImageNet SPT	90.6	83.7	77.4	97.4	87.6	69.3	**100.0**	88.2	**98.2**	38.4	83.1
MSCA [105]	ImageNet SPT	99.5	80.0	90.4	90.5	90.6	77.3	**100.0**	76.1	65.9	80.6	85.1
HRBM [94]	ImageNet SPT	**99.7**	90.8	**90.6**	92.9	90.3	80.1	90.8	80.3	68.5	87.1	87.1
SAPNet [106]	ImageNet SPT	97.8	87.6	67.2	94.8	**99.5**	**99.5**	95.9	**96.8**	68.0	85.1	89.2
FMSSD [107]	ImageNet SPT	**99.7**	89.9	90.3	98.2	86.0	**96.8**	99.6	75.6	80.1	88.2	89.2
CAD-Net [108]	ImageNet SPT	97.0	77.9	**95.6**	93.6	87.6	87.1	99.6	**100.0**	86.2	**89.9**	91.5
MEDNet [109]	ImageNet SPT	99.2	**94.4**	82.2	98.5	95.4	95.2	98.3	88.1	75.1	89.3	**91.6**
RetinaNet (Ours)	SA-MIM SSPT	**99.8**	90.3	87.4	**99.9**	95.0	93.7	**99.7**	75.2	**94.7**	80.1	**91.6**
Faster R-CNN (Ours)	SA-MIM SSPT	**99.7**	**94.0**	**90.6**	**99.7**	**98.8**	96.2	**100.0**	74.2	87.5	**89.8**	**93.1**

TABLE XIII

DETECTION ACCURACY COMPARISON OF VARIOUS DETECTORS ON THE DOTA DATASET. OTHER DETECTORS ARE GENERALLY BASED ON IMAGENET SPT, WHILE OURS LEVERAGES THE PROPOSED SA-MIM SSPT. THE TOP TWO BEST RESULTS IN EACH COLUMN ARE REPRESENTED IN BOLD FONT

Detectors	Pretraining Paradigms	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP_{50}
MSCA [105]	ImageNet SPT	78.1	67.7	28.3	42.2	24.0	62.2	48.3	82.6	45.0	38.4	40.5	36.4	69.2	38.5	36.1	49.2
AF-SSD [110]	ImageNet SPT	**88.0**	61.6	31.2	43.6	59.4	32.8	54.8	90.7	62.4	74.0	29.8	54.0	33.0	49.8	23.9	52.6
R-FCN [111]	ImageNet SPT	81.0	59.0	31.6	59.0	49.8	45.0	49.3	69.0	52.1	67.4	41.8	51.4	45.2	53.3	33.9	52.6
FR-H [89]	ImageNet SPT	80.3	**77.6**	32.9	**68.1**	53.7	52.5	50.0	90.4	**75.1**	59.6	57.0	49.8	61.7	56.5	41.9	60.5
HSF-Net [112]	ImageNet SPT	80.0	69.9	37.7	58.0	**66.8**	64.2	71.8	87.9	**69.4**	61.8	47.5	52.8	66.8	**59.2**	41.8	62.4
RFB-Net [113]	ImageNet SPT	87.8	49.8	35.7	36.4	**78.0**	**77.5**	**90.1**	**93.8**	63.8	**89.5**	33.6	**67.9**	63.2	**72.8**	20.5	64.0
GLNet [104]	ImageNet SPT	**89.4**	71.5	**43.1**	**68.4**	31.9	52.5	53.9	79.9	66.9	**77.2**	**74.1**	**64.6**	**73.1**	**59.2**	**74.0**	**65.3**
RetinaNet (Ours)	SA-MIM SSPT	81.0	75.1	**41.4**	59.1	57.1	**80.2**	75.9	90.8	64.3	64.4	58.6	51.7	69.1	50.0	56.0	65.0
Faster R-CNN (Ours)	SA-MIM SSPT	82.2	**76.1**	**43.1**	62.5	60.3	**80.2**	**76.3**	**91.3**	68.8	65.0	**59.8**	55.5	**70.2**	53.5	**61.8**	**67.1**

to ImageNet SPT with adversarial fine-tuning. The detection performance of our method can be further elevated when coupled with adversarial fine-tuning. These experimental results demonstrate the generalization of the proposed pretraining paradigm and also foreshadow its promising potential to replace computationally expensive adversarial fine-tuning.

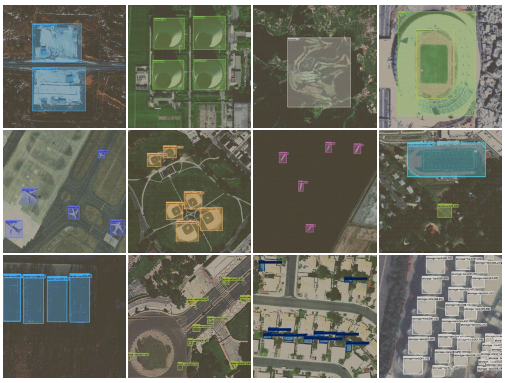

Fig. 12. Some qualitative robust object detection results generated by our SA-MIM pretraining with vanilla adversarial fine-tuning. The adversarial samples in the three rows originate from three different datasets, namely, DIOR, NWPU VHR-10, and DOTA.

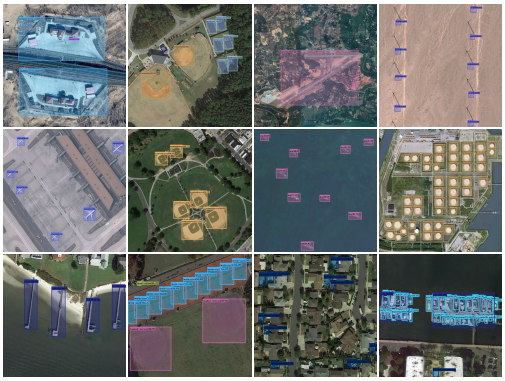

Fig. 13. Some qualitative object detection results generated by our SA-MIM pretraining with standard natural fine-tuning. The detection examples in the three rows are from DIOR, NWPU VHR-10, and DOTA.

Fig. 12 visualizes some robust detection results on different datasets.

F. Comparison of Standard Detection Accuracy With Recent State-of-the-Art Remote Sensing Object Detectors

This work focuses on a novel pretraining paradigm for robust object detection with high generalizability and robustness. However, the pretrained weights are also expected to go beyond robust detection and serve as a general-purpose initialization to extensively boost object detection. Therefore, we strictly train two fundamental detectors, i.e., RetinaNet and Faster R-CNN, in the context of remote sensing object detection, which are initialized based on our pretraining method. Then, we compare their detection performance with other state-of-the-art detectors on the three datasets in Tables XI–XIII. It can be observed from Table XI that a simpler detector, e.g., RetinaNet, when empowered with SA-MIM SSPT, can achieve comparable or superior detection accuracy over all other specially designed detectors that may introduce additional sophisticated components yet still follow ImageNet SPT. Meanwhile, in Tables XII and XIII, our RetinaNet and Faster R-CNN, respectively, also achieve competitive detection results, closely approaching and outperforming the recent methods, such as GLNet [104] and MEDNet [109], which, however, are computationally intensive with lower inference efficiency. These experimental results demonstrate that robust and generalizable structured knowledge learned during pretraining can generally benefit remote sensing object detection. Some qualitative detection results on these three datasets are visualized in Fig. 13.

VI. Conclusion

This article proposes a novel pretraining paradigm, namely, SASS pretraining, for robust object detection in remote sensing imagery. It can be characterized by three design principles: adversarial learning, unsupervised pretext, and structured knowledge guidance. To fully exploit the superior robustness inherent in ViTs over CNNs, SASS pretraining is instantiated as an MIM-style pretraining framework, called SA-MIM, which consists of two crucial components: structured adversarial attacks and S-MIM. The structured adversarial attack generates multigrained perturbations on the remote sensing in-domain pretraining samples to establish the context for adversarial learning. With the adversarial intermediates in the latent space, S-MIM introduces structured local and global constraints to accommodate hierarchical Transformers required by downstream detection. This adversarial reconstruction pretask in SA-MIM not only acquires generalizable representations beneficial for clean accuracy but also enhances the adversarial robustness of the pretrained models. Comprehensive experiments have been conducted on three public datasets for remote sensing object detection: DIOR, NWPU VHR-10, and DOTA. The quantitative and qualitative results consistently reveal that initializing the detector with our adversarially pretrained model by SA-MIM can simultaneously benefit detection accuracy and robustness, instead of sacrificing one for the other, like previous methods. The effectiveness and superiority of the proposed pretraining paradigm have also been demonstrated for both object detection and robust object detection in remote sensing images.

References

[1] D. Yu and S. Ji, "A new spatial-oriented object detection framework for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4407416.

[2] Y. Wu, K. Zhang, J. Wang, Y. Wang, Q. Wang, and X. Li, "GCWNet: A global context-weaving network for object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5619912.

[3] Q. Wang, Y. Liu, Z. Xiong, and Y. Yuan, "Hybrid feature aligned network for salient object detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5624915.

[4] C. Zhang, K.-M. Lam, and Q. Wang, "CoF-Net: A progressive coarse-to-fine framework for object detection in remote-sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5600617.

[5] Q. Lin, J. Zhao, G. Fu, and Z. Yuan, "CRPN-SFNet: A high-performance object detector on large-scale remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 416–429, Jan. 2022.

[6] B. Liu, C. Xu, Z. Cui, and J. Yang, "Progressive context-dependent inference for object detection in remote sensing imagery," *IEEE Trans. Image Process.*, vol. 32, pp. 580–590, 2023.

[7] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.

[8] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[9] J. Ding et al., "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7778–7796, Nov. 2022.

[10] Z. Li et al., "Deep learning-based object detection techniques for remote sensing images: A survey," *Remote Sens.*, vol. 14, no. 10, p. 2385, May 2022.

[11] X. Yang and J. Yan, "On the arbitrary-oriented object detection: Classification based approaches revisited," *Int. J. Comput. Vis.*, vol. 130, no. 5, pp. 1340–1365, May 2022.

[12] X. Yang et al., "Detecting rotated objects as Gaussian distributions and its 3-D generalization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4335–4354, Apr. 2023.

[13] J. Shen, C. Zhang, Y. Yuan, and Q. Wang, "Enhancing prospective consistency for semi-supervised object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5619312.

[14] Y. Liu, Z. Xiong, Y. Yuan, and Q. Wang, "Distilling knowledge from super resolution for efficient remote sensing salient object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5609116.

[15] C. Zhang, T. Liu, J. Xiao, K.-M. Lam, and Q. Wang, "Boosting object detectors via strong-classification weak-localization pretraining in remote sensing imagery," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–20, 2023.

[16] Y. Ju, M. Jian, C. Wang, C. Zhang, J. Dong, and K.-M. Lam, "Estimating high-resolution surface normals via low-resolution photometric stereo images," *IEEE Trans. Circuits Syst. Video Technol.*, early access, 2024, doi: 10.1109/TCSVT.2023.3301930.

[17] 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, 2021.

[18] C. Zhang, J. Su, Y. Ju, K.-M. Lam, and Q. Wang, "Efficient inductive vision transformer for oriented object detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5616320.

[19] T. Zhang et al., "Foreground refinement network for rotated object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5610013.

[20] Y. Ju, B. Shi, M. Jian, L. Qi, J. Dong, and K.-M. Lam, "NormAttention-PSN: A high-frequency region enhanced photometric stereo network with normalized attention," *Int. J. Comput. Vis.*, vol. 130, no. 12, pp. 3014–3034, Dec. 2022.

[21] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[22] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[23] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.

[24] Y. Zhou, S. Chen, J. Zhao, R. Yao, Y. Xue, and A. E. Saddik, "CLT-Det: Correlation learning based on transformer for detecting dense objects in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4708915.

[25] J. Xue, D. He, M. Liu, and Q. Shi, "Dual network structure with interweaved global-local feature hierarchy for transformer-based object detection in remote sensing image," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6856–6866, 2022.

[26] W. Lu et al., "A CNN-transformer hybrid model based on CSWin transformer for UAV image object detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1211–1231, 2023.

[27] C. Szegedy et al., "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2014.

[28] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[29] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[30] Y. Mo, D. Wu, Y. Wang, Y. Guo, and Y. Wang, "When adversarial training meets vision transformers: Recipes from training to architecture," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 18599–18611.

[31] K. Mahmood, R. Mahmood, and M. van Dijk, "On the robustness of vision transformers to adversarial examples," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7838–7847.

[32] P. Benz, S. Ham, C. Zhang, A. Karjauv, and I. S. Kweon, "Adversarial robustness comparison of vision transformer and MLP-mixer to CNNs," in *Proc. Brit. Mach. Vis. Conf.*, 2021.

[33] Y. Xu, B. Du, and L. Zhang, "Assessing the threat of adversarial examples on deep neural networks for remote sensing scene classification: Attacks and defenses," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1604–1617, Feb. 2021.

[34] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[35] H. Zhang and J. Wang, "Towards adversarially robust object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 421–430.

[36] X. Chen, C. Xie, M. Tan, L. Zhang, C.-J. Hsieh, and B. Gong, "Robust and accurate object detection via adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16617–16626.

[37] P. Chen, B. Kung, and J. Chen, "Class-aware robust adversarial training for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 10420–10429.

[38] Z. Dong, P. Wei, and L. Lin, "Adversarially-aware robust object detector," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 297–313.

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[40] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, "A self-supervised approach for adversarial robustness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 262–271.

[41] H. Zhang and J. Wang, "Defense against adversarial attacks using feature scattering-based adversarial training," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1829–1839.

[42] X. Mao et al., "Enhance the visual representation via discrete adversarial training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 7520–7533.

[43] G.-J. Qi and M. Shah, "Adversarial pretraining of self-supervised deep networks: Past, present and future," 2022, *arXiv:2210.13463*.

[44] T. Chen, S. Liu, S. Chang, Y. Cheng, L. Amini, and Z. Wang, "Adversarial robustness: From self-supervised pre-training to fine-tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 696–705.

[45] Z. Jiang, T. Chen, T. Chen, and Z. Wang, "Robust pre-training by adversarial contrastive learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 16199–16210.

[46] L. Fan, S. Liu, P.-Y. Chen, G. Zhang, and C. Gan, "When does contrastive learning preserve adversarial robustness from pretraining to finetuning?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 21480–21492.

[47] R. Gupta, N. Akhtar, A. Mian, and M. Shah, "Contrastive self-supervised learning leads to higher adversarial susceptibility," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 14838–14846.

[48] Y. Liang, J. Feng, X. Zhang, J. Zhang, and L. Jiao, "MidNet: An anchor-and-angle-free detector for oriented ship detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5612113.

[49] J. Feng, Y. Liang, X. Zhang, J. Zhang, and L. Jiao, "SDANet: Semantic-embedded density adaptive network for moving vehicle detection in satellite videos," *IEEE Trans. Image Process.*, vol. 32, pp. 1788–1801, 2023.

[50] Z. Hu et al., "EMO2-DETR: Efficient-matching oriented object detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5616814.

[51] Y. Tian, M. Zhang, J. Li, Y. Li, H. Yang, and W. Li, "FPNFormer: Rethink the method of processing the rotation-invariance and rotation-equivariance on arbitrary-oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5605610.

[52] X. Yang, S. Zhang, S. Duan, and W. Yang, "An effective and lightweight hybrid network for object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5600711.

[53] L. Chen, Z. Xu, Q. Li, J. Peng, S. Wang, and H. Li, "An empirical study of adversarial examples on remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7419–7433, Sep. 2021.

[54] C. Shi et al., "Multifeature collaborative adversarial attack in multimodal remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5631815.

[55] Y. Xu and P. Ghamisi, "Universal adversarial examples in remote sensing: Methodology and benchmark," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5619815.

[56] X. Sun, G. Cheng, L. Pei, H. Li, and J. Han, "Threatening patch attacks on object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5609210.

[57] J. Lian, S. Mei, S. Zhang, and M. Ma, "Benchmarking adversarial patch against aerial detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5634616.

[58] Y. Zhang et al., "Adversarial patch attack on multi-scale object detection for UAV remote sensing images," *Remote Sens.*, vol. 14, no. 21, p. 5298, Oct. 2022.

[59] Y. Su, G. Zhang, S. Mei, J. Lian, Y. Wang, and S. Wan, "Reconstruction-assisted and distance-optimized adversarial training: A defense framework for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5624613.

[60] G. Cheng, X. Sun, K. Li, L. Guo, and J. Han, "Perturbation-seeking generative adversarial networks: A defense framework for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5605111.

[61] H. Bao, L. Dong, S. Piao, and F. Wei, "BEit: BERT pre-training of image transformers," in *Proc. Int. Conf. Learn. Represent.*, 2022.

[62] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16000–16009.

[63] Z. Xie et al., "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9653–9663.

[64] L. Huang, S. You, M. Zheng, F. Wang, C. Qian, and T. Yamasaki, "Green hierarchical vision transformer for masked image modeling," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 19997–20010.

[65] X. Zhang et al., "HiViT: A simpler and more efficient design of hierarchical vision transformer," in *Proc. Int. Conf. Learn. Represent.*, 2023.

[66] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[67] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9640–9649.

[68] Q. Hu, X. Wang, W. Hu, and G.-J. Qi, "AdCo: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1074–1083.

[69] 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.*, Oct. 2017, pp. 2980–2988.

[70] 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, Jun. 2017.

[71] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[72] L. Liu et al., "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, Feb. 2020.

[73] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[74] K. Ayush et al., "Geography-aware self-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10181–10190.

[75] O. Manas, A. Lacoste, X. Giro-i-Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9414–9423.

[76] P. Akiva, M. Purri, and M. Leotta, "Self-supervised material and texture representation learning for remote sensing tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8203–8215.

[77] Y. Long et al., "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-AID," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4205–4230, 2021.

[78] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6172–6180.

[79] X. Sun et al., "RingMo: A remote sensing foundation model with masked image modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2022, Art. no. 5612822.

[80] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 15637–15648.

[81] A. A. Aleissaee et al., "Transformers in remote sensing: A survey," *Remote Sens.*, vol. 15, no. 7, p. 1860, Mar. 2023.

[82] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3533–3545.

[83] Y. Tian et al., "Integrally pre-trained transformer pyramid networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1–13.

[84] C. Herrmann et al., "Pyramid adversarial training improves ViT performance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13409–13419.

[85] X. Li, W. Wang, L. Yang, and J. Yang, "Uniform masking: Enabling MAE pre-training for pyramid-based vision transformers with locality," 2022, *arXiv:2205.10063*.

[86] J. Liu, X. Huang, J. Zheng, Y. Liu, and H. Li, "MixMAE: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 6252–6261.

[87] X. Chen et al., "Context autoencoder for self-supervised representation learning," 2022, *arXiv:2202.03026*.

[88] 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.

[89] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

[90] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2019.

[91] 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.*, vol. 35, no. 4, May 2021, pp. 3163–3171.

[92] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5602511.

[93] J. Han, J. Ding, N. Xue, and G. Xia, "ReDet: A rotation-equivariant detector for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2786–2795.

[94] 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, Apr. 2017.

[95] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 21–37.

[96] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.

[97] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[98] W. Huang, G. Li, Q. Chen, M. Ju, and J. Qu, "CF2PN: A cross-scale feature fusion pyramid network based remote sensing target detection," *Remote Sens.*, vol. 13, no. 5, p. 847, Feb. 2021.

[99] G. Cheng, Y. Si, H. Hong, X. Yao, and L. Guo, "Cross-scale feature fusion for object detection in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 431–435, Mar. 2020.

[100] H. Wei, Y. Zhang, Z. Chang, H. Li, H. Wang, and X. Sun, "Oriented objects as pairs of middle lines," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 268–279, Nov. 2020.

[101] 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.*, vol. 45, no. 2, pp. 2384–2399, Feb. 2022.

[102] T. Zhang, Y. Zhuang, G. Wang, S. Dong, H. Chen, and L. Li, "Multiscale semantic fusion-guided fractal convolutional object detection network for optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5608720.

[103] W. Han et al., "Improving training instance quality in aerial image object detection with a sampling-balance-based multistage network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10575–10589, Dec. 2021.

[104] Z. Teng, Y. Duan, Y. Liu, B. Zhang, and J. Fan, "Global to local: Clip-LSTM-Based object detection from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603113.

[105] J. Chen, L. Wan, J. Zhu, G. Xu, and M. Deng, "Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 681–685, Apr. 2019.

[106] S. Zhang, G. He, H.-B. Chen, N. Jing, and Q. Wang, "Scale adaptive proposal network for object detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 864–868, Oct. 2019.

[107] 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, Dec. 2020.

[108] G. Zhang, S. Lu, and W. Zhang, "CAD-Net: A context-aware detection network for objects in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10015–10024, Dec. 2019.

[109] Q. Lin, J. Zhao, B. Du, G. Fu, and Z. Yuan, "MEDNet: Multiexpert detection network with unsupervised clustering of training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022.

[110] X. Lu, J. Ji, Z. Xing, and Q. Miao, "Attention and feature fusion SSD for remote sensing object detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.

[111] 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.*, vol. 29, 2016, pp. 379–387.

[112] Q. Li, L. Mou, Q. Liu, Y. Wang, and X. X. Zhu, "HSF-Net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7147–7161, Dec. 2018.

[113] S. Liu and D. Huang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 385–400.

Cong Zhang (Graduate Student Member, IEEE) received the B.E. degree from the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China, in 2018, and the M.E. degree from the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, in 2021. He is currently pursuing the Ph.D. degree with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong.

His research interests include remote sensing, computer vision, and machine learning.

Kin-Man Lam (Senior Member, IEEE) received the M.Sc. degree in communication engineering from the Department of Electrical Engineering, Imperial College of Science, Technology and Medicine, London, U.K., in 1987, and the Ph.D. degree from the Department of Electrical Engineering, The University of Sydney, Camperdown, NSW, Australia, in 1996.

In 1996, he joined the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, again as an Assistant Professor, where he became an Associate Professor in 1999 and has been a Professor since 2010. He is currently an Associate Dean of the Faculty of Engineering, The Hong Kong Polytechnic University. His research interests include image processing, computer vision, and human face analysis and recognition.

Prof. Lam is currently a Member-at-Large of the Asia-Pacific Signal and Information Processing Association (APSIPA). He was the VP-Member Relations and Development and VP-Publications of APSIPA from 2014 to 2017 and from 2017 to 2021, respectively. He was the General Co-Chair of the APSIPA Annual and Summit 2015 and IEEE International Conference on Multimedia and Expo (ICME) 2017 and the Technical Chair of IEEE Visual Communications and Image Processing (VCIP) 2020. He was the Director-Student Services and the Director-Membership Services of the IEEE Signal Processing Society from 2012 to 2014 and from 2015 to 2017, respectively. He was an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING from 2009 to 2014 and *Digital Signal Processing* from 2014 and 2018. He was an Editor of HKIE TRANSACTIONS from 2013 to 2018 and an Area Editor of the *IEEE Signal Processing Magazine* from 2015 to 2017.

Tianshan Liu received the B.Sc. and M.Sc. degrees from the School of Internet of Things Engineering, Jiangnan University, Wuxi, China, in 2016 and 2019, respectively, and the Ph.D. degree from the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, in 2023.

He is currently with the Faculty of the School of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include computer vision, multimedia computing, and video understanding.

Yui-Lam Chan (Member, IEEE) received the B.Eng. (Hons.) and Ph.D. degrees from The Hong Kong Polytechnic University, Hong Kong, in 1993 and 1997, respectively.

In 1997, he joined The Hong Kong Polytechnic University, where he is currently an Associate Professor with the Department of Electronic and Information Engineering. He is actively involved in professional activities and has authored over 140 research papers in various international journals and conferences. His research interests include multimedia technologies, signal processing, and image and video compression.

Dr. Chan was the Secretary of the 2010 IEEE International Conference on Image Processing. He was also the Publication Chair of the IEEE International Conference on Multimedia and Expo. He has served as an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China, where he is also with the Key Laboratory of Intelligent Interaction and Applications, Ministry of Industry and Information Technology. His research interests include computer vision, pattern recognition, and remote sensing.