Sampling Equivariant Self-Attention Networks for Object Detection in Aerial Images

Guo-Ye Yang, Xiang-Li Li, Zi-Kai Xiao, Tai-Jiang Mu, and Shi-Min Hu, Senior Member, IEEE

Abstract—Objects in aerial images show greater variations in scale and orientation than in other images, making them harder to detect using vanilla deep convolutional neural networks. Networks with sampling equivariance can adapt sampling from input feature maps to object transformation, allowing a convolutional kernel to extract effective object features under different transformations. However, methods such as deformable convolutional networks can only provide sampling equivariance under certain circumstances, as they sample by location. We propose sampling equivariant self-attention networks, which treat self-attention restricted to a local image patch as convolution sampling by masks instead of locations, and a transformation embedding module to improve the equivariant sampling further. We further propose a novel randomized normalization module to enhance network generalization and a quantitative evaluation metric to fairly evaluate the ability of sampling equivariance of different models. Experiments show that our model provides significantly better sampling equivariance than existing methods without additional supervision and can thus extract more effective image features. Our model achieves state-of-the-art results on the DOTA-v1.0, DOTA-v1.5, and HRSC2016 datasets without additional computations or parameters.

Index Terms—Sampling equivariance, aerial images, object detection, self-attention.

I. INTRODUCTION

Aerial image object detection is the basis of aerial image analysis and processing, with a wide range of applications in natural disaster prevention, ecological protection, and city management. Aerial images have particular characteristics due to their capture mode: objects in aerial images have greater rotational and scale variations than in most images. This requires the detection approach to handle a range of geometric transformations flexibly. Currently, convolutional neural networks (CNNs) are widely used for object detection, typically using multiple frequency- and orientation-specific kernels to identify objects subject to a range of transformations [1]. This usually increases the number of parameters and computations. Furthermore, acquiring aerial images of objects under a range of conditions for training models is costly.

In this paper, we propose the sampling equivariance property that benefits the CNNs for feature extraction, especially for images with objects under various transformations, such as aerial images. Convolutions are calculated in two steps: (i) sample features from the input feature map, and (ii) generate output features by matrix multiplication of the sampled features and convolutional kernels. A layer is considered sampling equivariant if the layer can adjust the sampling locations according to the underlying object transformation. Sampling equivariant layers can thus extract object features independently of transformations using the same set of convolutional kernels. Therefore, sampling equivariant networks have better feature extraction capabilities for the same number of parameters and achieve better accuracy.

Deformable convolutional networks (DCNs) [2] attempt to reach sampling equivariance by adding a regressive offset to each regular grid-sampling location of a convolution. However, they only achieve sampling equivariance for scaling...
transformations, but not rotation, reflection, and skew, due to the two drawbacks of sampling by location. First, ambiguities may arise under different transformations if some image areas have similar appearances: in Fig. 1(a), samples 1 and 2 are located on the left and right wings of the aircraft, respectively; when the image is rotated by 180 degrees to Fig. 1(b) if the model identifies the nose of the aircraft as its tail and vice versa, this will lead to errors in the regression of the left- and right-wing positions. Second, sampling by location can result in indistinguishable representation of different parts of the object. For example, as Fig. 1(e) shows, we want samples 1 and 2 to represent the rotor and cabin of the helicopter, respectively; however, the sampled features might be similar since the centers of the two parts are very close.

We propose a novel sampling equivariant self-attention layer, the SES-layer, to achieve general sampling equivariance. An SES-layer samples features via masks instead of locations, using a local self-attention mechanism to fulfill a position-independent feature sampling. As shown in Fig. 2, using masks for sampling can overcome the above ambiguities by representing similar areas with one mask and can represent different parts of an object better. Equivariant sampling by masks is illustrated in Fig. 2. Most self-attention-based models require a positional encoding layer to provide spatial context for feature learning. However, positional encoding affects sampling mask generation, resulting in poor sampling equivariance, which we call prior sampling error. Furthermore, sampled features may lose information about where and how the sampling was performed, which is important in the object detection task, e.g., how many wings there are, where the wings are, and what the wing shapes are. We thus propose a transformation embedding module to encode the sampling masks, providing the spatial context and sampling information for output features while avoiding prior sampling error. Additionally, models based on self-attention are prone to overfitting when given relatively small training datasets. We propose a randomized normalization module (RNM) to enhance the generalization ability of the SES-layer.

In the evaluation, the sampling equivariance ability of different models is hard to measure, as they use different methods to sample features. We thus propose a quantitative evaluation metric, i.e., the average earth mover’s distance (AEMD), to fairly evaluate the ability of sampling equivariance of different models. Quantitative experiments show that our model provides much better sampling equivariance without additional supervision and thus obtains higher accuracy without increasing the number of computations or model parameters. Our model also achieves new state-of-the-art results on DOTA-v1.0 [4], DOTA-v1.5 and HRSC2016 [5].

In summary, our contributions are as follows:

• A sampling equivariant self-attention layer (SES-layer), which significantly improves sampling equivariance within the network.
• An RNM, which can improve the generalization ability of the network.
• A new quantitative evaluation metric, i.e., the AEMD, which is used for evaluating the sampling equivariance ability of different models.

II. RELATED WORK

A. Transformation Equivariant and Invariant Features

Earlier work on extracting transformation-invariant features from 2D images mostly used manually designed feature extractors [6], [7]. Recently, researchers have used learnable neural networks for this purpose, and these methods can be divided into two categories. The first kind of method obtains equivariant/invariant features by constraining the equivariance/invariance of the learned features through training objectives [8]. Wang et al. [9] propose learning discriminative and equivariant query embeddings for query-based instance segmenters by supervising the uniqueness and equivariance of the segmentation results. Yang et al. [10] constrain the matching between image representations and object queries to be rotation equivariant for better fisheye object detection. Such methods usually require image pairs before and after transformation for training, and thus, their equivariance relies on the transformations used during data augmentation.

The second kind of method usually designs a specialized network structure to extract equivariant/invariant features [11]. Some methods [12], [13] propose structures that are equivariant to group transformations, and some methods [14], [15] achieve rotation-equivariance by pooling the multiple rotated convolutional kernel outputs. However, these methods can only obtain equivariance on several manually set discrete transformations. For better flexibility, other methods employ learnable deformation modules and extract transformation-invariant features by indirectly supervising visual tasks. Jaderberg et al. [16] propose spatial transformer networks (STNs) to perform a learnable global transformation on the feature map to obtain invariant features. Dai et al. [2] propose DCNs to transform local features by adding a regressed offset to each regular
grid-sampling location of the convolution. This method is further improved by follow-ups [17], [18], [19]. Such methods obtain transformation invariance for local features by changing sampling locations or weights, which requires the sampling locations to have transformation equivariance. However, our experiments showed that this only holds for the scaling transformation. We believe this is due to sampling by location. We instead use masks for sampling, which overcomes ambiguities and better represents irregular areas, yielding better sampling equivariance.

B. Self-Attention in Computer Vision

The successful attention mechanism has been applied in computer vision [20], [21], [22], [23], a popular self-attention approach with powerful representational capabilities is the transformer [24]. ViT [3] treats different patches of an image as words and uses the transformer structure to improve image recognition results. DETR [25] provides a transformer-based end-to-end object detection pipeline. Some recent methods [26], [27] apply self-attention to a local image patch. Hu et al. [28] propose a local relation layer for image recognition. Zhao et al. [29] propose self-attention networks (SANs) with two forms of self-attention, pairwise and patchwise. We regard self-attention restricted to a local patch as a convolution sampled by masks instead of locations. SAN uses patchwise weights to regress the masks in different sampling locations, leading to inequalities and making it difficult to obtain sampling equivariance. Thus, our model is implemented on top of SAN using pairwise. Our model has significantly better sampling equivariance than existing local self-attention methods, and we further provide an RNM to solve the overfitting problem of self-attention models trained on relatively small datasets.

C. Object Detection in Aerial Images

Object detection has been investigated for decades [30], [31]; aerial images are particularly challenging with respect to varying scale and orientation. Researchers have proposed methods to improve detection results by considering how to learn effectively in a rotation-invariant way [32], [33], [34], using special network structures for oriented object detection [35], [36], [37], [38], [39], [40], or designing specialized oriented object representations [41], [42], [43] and losses [44], [45]. Cheng et al. [32], [33] propose to train a rotation-invariant feature extractor by rotating the data and imposing the rotation invariance of extracted features. Ding et al. [46] propose a region of interest (RoI) transformer that can extract features in an oriented bounding box. Han et al. [47] propose S²-Net by aligning deep features adaptively and reducing inconsistency between localivation and classification. ReDet [34] gives a rotation equivariant detector to extract rotation equivariant features. Xie et al. [48] present an effective two-stage oriented object detection framework, Oriented R-CNN, with an oriented region proposal network and an oriented R-CNN head. GSDet [38] proposes to consider the physical size of objects when regressing the bounding box and the category by predicting the height of the image being captured. Xu et al. [49] propose a novel representation for an oriented bounding box with four vertices ‘gliding’ on the sides of an axis-aligned bounding box; it achieves state-of-the-art performance. This network is concise and achieves good results; therefore, we use it as a baseline for measuring the effectiveness of our proposed model. Most of these methods obtain transformation-invariant features with backbones lacking general sampling equivariance, which is inefficient. SES-Net has sampling equivariance for diverse transformations; therefore, it extracts more informative features for better results.

III. METHODOLOGY

A. Overview

Our network for object detection in aerial images, the sampling equivariant self-attention network (SES-Net), is shown in Fig. 3(a). It consists of two main components: an SES-layer and an RNM. The SES-layer considers self-attention restricted to a local patch of the input feature map and extracts features with a sliding window as a convolution sampled by a mask rather than location. The RNM enhances the generalization ability of the SES-layer when trained on relatively small datasets.

In this section, we first formally define sampling equivariance (Sec. III-B) and analyze why existing methods fail to achieve generic sampling equivariance. Then, we detail the design of our SES-layer (Sec. III-C) and the RNM (Sec. III-D) of our SES-Net. Finally, we introduce the AEMD metric to evaluate the sampling equivariance ability of different models quantitatively (Sec. III-E).

B. Sampling Equivariance

A simplified vanilla 2D-convolution is given by:

$$G(x) = \sum_{i=1}^{K_h} \sum_{j=1}^{K_w} w_{i,j} \cdot F(x + s_{i,j}^\text{Conv})$$

where $F$ and $G$ are the input and the output feature maps, respectively. $w$ is the convolutional kernel, and $(K_h, K_w)$ is the kernel size. $s_{i,j}^\text{Conv} = (i - (K_h + 1)/2, j - (K_w + 1)/2)$ determines the fixed sampling location offsets for the convolution.

This structure with fixed sampling locations is relatively inflexible and requires a large number of parameters to accommodate a range of transformations. The sampling offsets $s$ of a sampling equivariant convolutional layer can adapt to the transformations of objects in $F$, and the layer can thus extract the features of the object under different transformations using the same kernel $w$. Formally, we say a layer is sampling equivariant under transformation group $T$, if:

$$T \cdot x + s_{i,j} (T \circ F, T \cdot x) = T \cdot (x + s_{i,j}(F, x))$$

for all $i \in \{1, \ldots, K_h\}, j \in \{1, \ldots, K_w\}$, and $T \in \mathbb{T}$, where $s_{i,j}(F, x)$ is the $(i, j)$-th sampling offset for center $x$ with input feature map $F$, $T \cdot x$ applies transform $T$ to $x$, and $T \circ F$ applies transform $T$ to $F$, which satisfies $(T \circ F)(x) = F(T^{-1} \cdot x)$. If $T$ is a linear transformation, Eq.2 can be simplified to:

$$s_{i,j}(T \circ F, T \cdot x) = T \cdot s_{i,j}(F, x)$$

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
C. Sampling Equivariant Self-Attention Layer

DCNs [2], [17], [18] suggest learning the sampling offsets, giving the model a certain degree of flexibility. A DCN layer replaces the sampling location offsets \( s_{ij}^{\text{Conv}} \) in Eq. 1 by \( s_{ij}^{\text{DCN}}(F, x) = g_{ij}^{\text{DCN}}(F)(x) \), which regresses the \((i, j)\)-th offset using a two-channel convolution \( g_{ij}^{\text{DCN}} \). However, sampling by location may cause ambiguities when some image areas have similar appearances under different transformations, leading to a poor sampling equivariance. In particular, if the object in \( x \) is self-similar under transformation \( T \) (e.g., for axisymmetric objects under the reflection transformation), the features around \( x \) in \( F \) will be similar to those around \( T \cdot x \) in \( T \circ F \), making \( g_{ij}^{\text{DCN}}(F)(x) \approx g_{ij}^{\text{DCN}}(T \circ F)(T \cdot x) \rightarrow s_{ij}^{\text{DCN}}(F, x) \approx s_{ij}^{\text{DCN}}(T \circ F, T \cdot x) \rightarrow s_{ij}^{\text{DCN}}(F, x) \approx T \cdot s_{ij}^{\text{DCN}}(F, x) \). This implies that if DCNs can achieve sampling equivariance, they can only place sampling locations at the invariant points under transformation \( T \). This greatly limits the sampling range of DCNs and consequently leads to lower precision and increased loss. The training process will drive DCNs to perform sampling over a broader range, which also makes it hard to achieve sampling equivariance. Our experiments show that DCNs only provide good sampling equivariance under scaling transformations, in which ambiguities rarely appear, allowing DCNs to adjust their receptive field to object size, as described in [2]. However, achieving complete sampling equivariance for other transformations for the reasons described above is difficult.

C. Sampling Equivariant Self-Attention Layer

Local self-attention models [26], [28], [29] apply self-attention to local image patches and use sliding windows to extract features, showing promising performance on feature learning. However, they do not consider local self-attention in terms of sampling by masks. Thus their models have drawbacks such as prior sampling error and lack of sampling information, making it difficult to achieve sampling equivariance.

Instead of sampling a single location at each offset position, our SES-Net uses masks for sampling, giving greater flexibility and avoiding ambiguities. A simplified SES-layer is given by:

\[
G(x) = \sum_{i=1}^{N} w_i \cdot F(x, s_i^{\text{SES}}(F, x)),
\]

\[
F(x, s) = \sum_{k=1}^{M} \sum_{l=1}^{M} F(x + (k - \bar{M}, l - \bar{M})) \cdot s_{k,l}
\]

(4)

In contrast to Eq. 1, which takes \( K_h \times K_w \) sampling locations, the SES-layer takes \( N \) sampling masks. \( M \) is the size of the sampling masks, \( \bar{M} = (M + 1)/2 \), and \( s_i^{\text{SES}}(F, x) = M \times M \) sampling mask centered at position \( x \), given the input feature map \( F \), subject to \( \sum_{l=1}^{M} \sum_{k=1}^{M} s_i^{\text{SES}}(F, x)_{k,l} = 1 \). Similar to the definition in Eq. 3, an SES-layer is sampling equivariant under a linear transformation \( T \), if:

\[
s_i^{\text{SES}}(T \circ F, T \cdot x) = T \circ s_i^{\text{SES}}(F, x)
\]

(5)

for all \( i \in \{1, \ldots, N\} \), and \( T \in \mathbb{T} \).

To adapt the sampling mask to the object under different transformations, we learn \( s_i^{\text{SES}}(F, x) \) with a self-attention mechanism [24]. A typical self-attention process for generating a sampling mask can be written as:

\[
s_i^{\text{SES}}(F, x)_{k,l} = \psi_i(\text{PE}(k, l) + \text{Query}(F(x)) - \text{Key}(F(x + (k - \bar{M}, l - \bar{M}))))
\]

(6)

where Query and Key are linear layers, PE denotes positional encoding, and \( \psi_i \) is the mask regression module composed of two lots of BatchNorm-ReLU-Linear. In practice, PE destroys the positional independence of the self-attention structure, causing sampling mask generation to be affected by their relative position, which we call prior sampling error. This means that equivariant masks are difficult to generate, so we
remove the PE term in Eq. 6, having:
\[
\begin{align*}
S_{i}^{\text{SES}}(T \circ F, T \cdot x)_{k,i} &= \psi_{i}(\text{Query}((T \circ F)(T \cdot x)) \\
& - \text{Key}(T \circ F)(T \cdot (x + T^{-1} \cdot (k - \tilde{M} \cdot l - \tilde{M})))) \\
& = \psi_{i}(\text{Query}(F(x)) - \text{Key}(F(x + T^{-1} \cdot (k - \tilde{M} \cdot l - \tilde{M})))) \\
& = s_{i}^{\text{SES}}(F, x)_{T^{-1} \cdot (k, l)} \\
& = (T \circ s_{i}^{\text{SES}}(F, x))_{k,i}
\end{align*}
\]
for all \( k, l \in \{1, \ldots, M\} \). This equation satisfies Eq. 5, meaning that the SES-layer is theoretically sampling equivariant under linear transformations. However, due to the limited sampling range and the discretization error in sampling, in practice, the SES-layer achieves only approximate sampling equivariance.

Previous works \cite{28, 29} required features to be positionally encoded to provide spatial context for the sampled features. In contrast, we discard PE in exchange for better sampling equivariance. Furthermore, the sampled features do not contain information about how sampling was performed, which is important in identifying object shapes and positions. However, it has not been presented in previous works. To solve these two problems, we propose a transformation embedding module, which embeds the sampling information into the sampled features with a grouped linear layer. It also provides spatial context for the sampled features without destroying the position independence of the self-attention structure. Compared to PE, our transformation embedding module does not result in an increased number of parameters or computational workload.

The architecture of our SES-layer with local self-attention mechanism is illustrated in Fig. 3(b). \( F' \) and \( F'' \) are feature maps with and without random perturbations, respectively, sharing the same shape of \([C, H, W]\) for the number of channels, height, and width, respectively. \( F' \) and \( F'' \) are first fed into three linear layers to obtain the three output tensors \( \text{value} \), \( \text{query} \) and \( \text{key} \) with shapes of \([C/r_{1}, H, W]\), \([C/r_{2}, H, W]\), and \([C/r_{2}, H, W]\), respectively; \( r_{1}, r_{2} \) are the bottleneck dimension reduction factors. The sampling mask \( s \) with shape \([N, C/r_{1}, H, W]\) is given by \( s = \text{softmax}(\psi(\eta(Q) - \delta(K))) \). \( \eta \) and \( \delta \) are a repeat and an unfolding operator with a kernel size \( M \), respectively, sharing the same output shape of \([C/r_{2}, C/r_{2}, H, W]\), \( \eta(Q) \) repeats \( Q \) for \( M^{2} \) times on the second dimension, and \( \delta(K) \) unfolds the tensor \( K \), such that \( \delta(K)(c, (i - 1) \ast M + j, h, w) = K(c, h + i \ast M, j \ast M) \), for \( c \in \{1, \ldots, C/r_{2}\}, i, j \in \{1, \ldots, M\}, h \in \{1, \ldots, H\}, \) and \( w \in \{1, \ldots, W\} \). \( \psi \) is the mask regression module composed of BatchNorm-ReLU-Linear\(_{C/r_{2}}\)-BatchNorm-ReLU-Linear\(_{C/r_{2}}\), which operates on the first dimension. The sampled features \( V' \) with shape \([C/r_{1}, H, W]\) are given by a sampling operation on \( V \) with \( s \), modified following Eq. 4 to reduce the number of calculations; \( V'(c, h, w) = \sum_{l=1}^{M} \sum_{j=1}^{M} V(c, h + k \ast M, w + l \ast M) \cdot s \left( \frac{c - (1 - 1) \ast M + j, h, w}{C/r_{1}/N} \right) + 1, (k - 1) \ast M + l, h, w) \), for \( c \in \{1, \ldots, C/r_{1}\}, h \in \{1, \ldots, H\}, \) and \( w \in \{1, \ldots, W\} \). We then embed the sampling mask \( s \) as the sampling transformation information into the sampled features \( V' \) with a transformation embedding module. In this module, we reshape \( V' \) into \([N, C/r_{1}/N, H, W]\], concatenate \( s \) and \( V' \) on the second dimension and flatten the first two dimensions, then pass the tensor into a linear layer with \( N \) groups to obtain output features with embedded transformation information. Linear grouping is used as we believe different sampled features can embed their corresponding transformation information independently. This makes training easier and reduces the number of parameters.

Experiments in Tab. 1 show that the proposed structure achieves significantly better sampling equivariance for a wider range of transformations, including rotation, reflection, skew, and scaling. Better recognition results are also obtained on various datasets. The SES-layer can simply replace any convolution layer with kernel size greater than 1 in a CNN to enable sampling equivariant and efficient feature extraction.

\section{D. Randomized Normalization Module}

A network based on self-attention requires massive training data to generalize well \cite{3}. However, the training sets for some aerial image object detection datasets \cite{4, 5} are relatively small. We enhance the generalization ability of our model by adding random noise to the SES-layer input. However, directly adding noise to the input feature map leads to data distribution inconsistencies between training and testing, further degrading the performance. We thus propose an RNM based on batch normalization (BN) and explore how to combine it with the SES-layer. The RNM architecture is shown in Fig. 3(c).

For an input feature map \( F \), the RNM produces two outputs during training. The first is the typical BN output \( F' \):
\[
F' = \gamma \frac{F - \mu_{F}}{\sigma_{F}^{2} + \epsilon} + \beta
\]
where \( \mu_{F} \) is the mean of \( F \) across channels, \( \sigma_{F}^{2} \) is its variance, \( \gamma \) and \( \beta \) are learnable weights for BN, and \( \epsilon \) is 0.00001. The second is a randomized output \( F'' \) computed as:
\[
F'' = \gamma \frac{\hat{F} - \mu_{\hat{F}}}{\sqrt{\sigma_{\hat{F}}^{2} + \epsilon}} + \beta, \quad \hat{F} = F + \mathcal{N}(0, r)
\]
where \( \mathcal{N}(0, r) \) is a normal distribution noise with mean 0 and variance (random size) \( r \).

After applying ReLU to \( F'' \) and \( F' \), we take \( F' \) as the input to \( V \), and \( F'' \) as the inputs to \( Q \) and \( K \) in the SES-layer. Overfitting the \text{query} and \text{key} in the self-attention structure makes it difficult to generate more generalized sampling masks during inference, so a random perturbation should be added to the \text{query} and \text{key}. However, adding a random perturbation to the \text{value} introduces noise to the features, which is not conducive to training. Our experiments show that this randomization approach best improves the generalization ability of the model (see Tab. VI). RNM behaves like BN during inference, so it does not require additional computations or network parameters.

\section{E. Metric for Sampling Equivariance}

We design a new unified metric to objectively measure the capability of the sampling equivariance of different sampling
methods. Both our SES-Net and SAN [29] use masks to sample feature maps. In contrast, DCNs use locations and sampling by bilinear interpolation. Since each sample feature is calculated by the weighted average of cells in a feature map, we can construct a sampling graph of sampling locations and masks using the weights. We use the earth mover’s distance (EMD) [50] to quantitatively evaluate the equivariance of sampling locations and masks before and after a transformation $T$ is applied.

Specifically, given $N$ images $\{I_i\}_{i=1}^N$, we apply a random transformation $T_i$ on $I_i$ to obtain the transformed image $\hat{I}_i$. During feature extraction, there are $M_d$ feature maps with stride $d$. We first randomly select a stride $d_i$, then feed $I_i$ and $\hat{I}_i$ into the network to obtain the $j$-th feature map $g_{i,j,k}$ and $\hat{g}_{i,j,k}$. Finally, we use $T_i$ to transform $g_{i,j,k}$ to obtain an ideally equivariant sampling graph $\hat{g}_{i,j,k}$. Average EMD (AEMD) is then used to measure the distance between $\hat{g}_{i,j,k}$ and $\hat{g}_{i,j,k}$, representing the capability of sampling equivariance:

$$ \text{AEMD} = \alpha \frac{1}{N} \sum_{i=1}^{N} \frac{1}{M_d} \sum_{j=1}^{M_d} \frac{1}{\mathcal{O}_{d_i,j}} \sum_{k=1}^{\mathcal{O}_{d_i,j}} \text{EMD}(\hat{g}_{i,j,k}, \hat{g}_{i,j,k}) $$

(10)

where $\alpha$ normalizes AEMD to a maximum value of 1, and $\mathcal{O}_{d_i,j}$ indicates the number of samples at a sampling center on the $j$-th feature map of stride $d_i$. A smaller value of AEMD indicates better sampling equivariance.

IV. EXPERIMENTS

A. Datasets

DOTA is a popular aerial image dataset for oriented object detection. We conduct experiments on both DOTA-v1.0 [4] and DOTA-v1.5. DOTA-v1.0 has 1,411 training images, 458 validation images, and 937 testing images of sizes ranging from 800 $\times$ 800 to 4000 $\times$ 4000 captured by different sensors and platforms, containing 188,282 instances in 15 common categories: plane (PL), ship (SH), storage tank (ST), baseball diamond (BD), tennis court (TC), basketball court (BC), ground track field (GTF), harbor (HA), bridge (BR), large vehicle (LV), small vehicle (SV), helicopter (HC), roundabout (RA), soccer-ball field (SBF) and swimming pool (SP). DOTA-v1.5 is annotated with an extra category, i.e., container crane (CC) and additional small objects, resulting in 402,089 object instances and making it more challenging.

HRSC2016 [5] is another widely used aerial image dataset of ships, containing 436 training images, 181 validation images, and 444 testing images of sizes ranging from 300 $\times$ 300 to 1500 $\times$ 900.

Following the standard protocol, we use the training and validation sets for training and the testing set for testing and adopt mean average precision (mAP) to measure the object detection accuracy.

1https://captain-whu.github.io/DOAI2019/

B. Implementation Details

Most previous work [32], [35], [36], [46], [47], [51] designed special modules to handle rotation and scaling, so it is difficult to measure the effectiveness of SES-Net using these works. Thus, we used Gliding [49] as a baseline model; it has a relatively simple design and excellent performance. We also incorporated our SES-layer into the state-of-the-art oriented R-CNN (ORCNN) [48] to test the versatility of our method.

We applied our SES-layer and RNM to Gliding and ORCNN with ResNet as the backbone by simply replacing [BN, ReLU, Conv3 $\times$ 3] by [RNM, ReLU, SES-layer] in the backbone bottleneck; these are denoted SES-G and SES-O respectively. We followed the default configurations, setting the Gliding and SES-G backbones to ResNet101, the ORCNN and SES-O backbones to ResNet50 on DOTA, and ResNet101 on HRSC2016. We set the sampling mask size to $M = 7$, the dimension reduction factor $r_1 = 1$, $r_2 = 16$, and the number of samples $N$ to 1/8 of the input feature map channel size $C$ for all SES-layers. We set the random size $r = 0.01$ for SES-O, $r = 0.005$ for SES-G in DOTA, and $r = 0.002$ for SES-G in HRSC2016. Due to the limited training data, we replaced bottlenecks only in the last stage of ResNet to ease training. We used parameters pretrained on ImageNet as initial weights. The learning rate was decayed by a cosine annealing schedule, initially set to 0.015, 0.04, 0.005, and 0.01 for SES-G on DOTA and HRSC2016, and SES-O on DOTA and HRSC2016, respectively. Other hyperparameters followed the baselines. Our model was implemented on Jittor [52], an efficient deep learning framework. To achieve higher calculation accuracy in AEMD, we upsampled the 7 $\times$ 7 sampling graph by a factor of 8 to 56 $\times$ 56. As a result, the maximum value of EMD($\cdot$, $\cdot$) became 56+56 = 112, and we thus set the AEMD normalization parameter $\alpha = 1/112$.

C. Sampling Equivariance

Quantitative and qualitative experiments on sampling equivariance included comparisons of SES-Net with a deformable convolutional network DCNv2 [18], and a local self-attention model SAN [29] with a pairwise SA-layer (denoted SAN-pairwise) and with a patchwise SA-layer (denoted SAN-patchwise), where SA-layer is the SAN’s self-attention layer. For fairness, we conducted this experiment on ImageNet [53], a widely used dataset. Our SES-Net10, SES-Net15, and SES-Net19 were based on SAN10, SAN15, and SAN19, respectively, by replacing SA-layers with SES-layers. To more accurately evaluate sampling equivariance of the SES-layer, we omitted the RNM module here. We used SAN [29] and DCNv2 [18] with ResNet-50 as backbones for our baseline models, trained SES-Net, SAN and DCNv2 from scratch on ImageNet, and followed the hyperparameters of SAN for classification. DCNv2 used for classification is denoted DCNv2-C. We also compared DCNv2 with weights provided by mmdetecion [54] trained on MSCOCO [55].

We considered four transformations: rotation, reflection, skewing, and scaling. We randomly selected $N = 5,000$ images for each experiment from the ImageNet validation set. We used AEMD (see Eq. 10) to evaluate the sampling
equivariance; results are shown in Tab. I. The AEMD distance for DCNv2 is 3.8–8.7 times that of SES-Net19 for rotation, reflection, and skewing transformations, which is significant. DCNv2 is closer to our model for scaling. The reason why DCNv2 fails to achieve complete sampling equivariance is that it uses locations for sampling. Our models also outperform SAN-pairwise for each task, with 1.1–2.1 times smaller AEMD: our structure can eliminate prior sampling error while retaining spatial context and incorporating sampling transformation information into the output features. SAN-pairwise has better sampling equivariance than SAN-patchwise, as the structure of the latter leads to inequality of different sampling locations.

Fig. 4 shows sampling examples from the ImageNet validation set for DCNv2-C, SAN10-pairwise, and SES-Net10 under rotation and reflection. Fig. 4(b) shows results for DCNv2-C, numbered circles representing different sampling locations for DCNv2. Comparing locations 2, 4, and 7 in images in the first column shows that these locations have not rotated as the image rotates, so are not sampling equivariant. Rows (a1, a2) show sampling masks for SAN10-pairwise. Due to prior sampling error and lack of transformation information, they do not always provide good sampling equivariance: see row a2, columns 3–4. Rows (a3, a4) show sampling masks for our method, with good sampling equivariance under different image transformations. Our model provides significantly better sampling equivariance without special supervision.

D. Ablation Studies

All our ablation studies were conducted on the DOTA-v1.0 object oriented bounding box (OBB) task. To verify the effectiveness of the SES-layer, we replaced the SES-layer in SES-G with the following layers:

- DCNv2, which aims to obtain sampling equivariance by regressing the sampling offsets, denoted as “G+DCNv2”.
- E2CNN [13] with cyclic group $C_8$, which can obtain equivariant features under discrete rotations by multiples of 45°, denoted as “G+E2CNN”.
- a pairwise SA-layer, denoted as “G+SA-pair”.
- a patchwise SA-layer, denoted as “G+SA-patch”.

We also conducted experiments on G+SA-pair and G+SA-patch with RNM because they were compatible. The results are shown in Tab. II; the numbers of parameters and FLOPS for the backbones of different models (denoted as B-Params and B-FLOPs, respectively) are also reported. Our SES-G surpassed others by 1% – 2.02% mAP, while the number of parameters and FLOPs remained similar or smaller. Because compared to previous models, which achieved equivariance only under restricted transformations, such as scaling for DCNv2 and rotation for E2CNN, our SES-layer obtained better and more comprehensive sampling equivariance, thus
can extract features from objects undergoing various transformations more efficiently.

We conducted experiments with the PE and the TEM to validate the SES-layer design. We evaluated the AEMD for SES-Net10 with the positional encoding added SES-layer (denoted as SES-Net10 + PE). The results in Tab. I show that the PE led to worse sampling equivariance, which we believe is because the PE breaks the position independence of self-attention and causes prior sampling error. We also tested SES-G with SES-layer with and without the PE and the TEM; the results are shown in Tab. III. Using PE provides the spatial context for the model and thus achieves better results (see rows 1 and 2). Our TEM retains the function of providing spatial context and provides the sampling transformation information for the features so it achieves even better results (see rows 1 and 3). Since the TEM already provides the spatial context, adding PE to the model with the TEM will not bring additional benefits. Instead, the additional PE will destroy positional independence, resulting in worse results (see rows 3 and 4).

We also tested the TEM using different methods to embed the sampling transformation information. With $N$ sampling masks as the sampling transformation information and $N$ sampled features for an SES-layer, we tested embedding all sampling masks in each sampled feature with $N$ different linear layers (denoted SES-G+all), embedding each sampling mask into the corresponding sampled feature with $N$ different linear layers (denoted SES-G+corr.), and embedding each sampling mask into the corresponding sampled features with a shared linear layer (denoted SES-G+share). As Tab. II shows, SES-G+corr. achieved the best results since different sampled features can embed their corresponding sampling information independently, which reduces the number of parameters and makes training easier (see rows 8 and 9). Differently sampled features may contain different semantic information and are, therefore, difficult to embed with a shared linear layer (see rows 3 and 4).

### Table II

Ablation Study on SES-layer and the Setting of the Transformation Embedding Module Using the DOTA-v1.0 Dataset

| Method          | mAP   | B-Params | B-FLOPs |
|-----------------|-------|----------|---------|
| Gliding         | 74.82 | 44.5M    | 7.8G    |
| G+DCNv2         | 75.26 | 44.9M    | 7.8G    |
| G+E2CNN         | 75.16 | 39.4M    | 7.8G    |
| G+SA-patch      | 75.68 | 39.3M    | 7.6G    |
| G+SA-patch+RNM  | 75.84 | 39.3M    | 7.6G    |
| G+SA-pair       | 75.81 | 38.4M    | 7.6G    |
| G+SA-pair+RNM   | 75.84 | 38.4M    | 7.6G    |
| SES-G (Ours)    | 76.84 | 38.5M    | 7.6G    |
| SES-G+linear    | 76.34 | 44.0M    | 7.9G    |
| SES-G+corr.     | 76.84 | 38.5M    | 7.6G    |
| SES-G+share     | 75.97 | 38.4M    | 7.6G    |

### Table III

Ablation Study on the PE and the Transformation Embedding Module (TEM) of SES-layer Using the DOTA-v1.0 Dataset

| Method | PE | TEM | mAP   | B-Params | B-FLOPs |
|--------|----|-----|-------|----------|---------|
| SES-G  | x  | x   | 75.44 | 38.4M    | 7.6G    |
| SES-G  | ✓  | x   | 75.84 | 38.4M    | 7.6G    |
| SES-G  | x  | ✓   | 76.84 | 38.5M    | 7.6G    |
| SES-G  | ✓  | ✓   | 75.60 | 38.5M    | 7.6G    |

### Table IV

Experiments on Different Architectures with Varying Sampling Ranges Using the DOTA-v1.0 Dataset.

| Method   | mAP   | B-Params | B-FLOPs |
|----------|-------|----------|---------|
| G+KS-3   | 74.82 | 44.5M    | 7.8G    |
| G+KS-5   | 75.04 | 57.1M    | 8.5G    |
| G+KS-7   | 75.03 | 76.0M    | 9.4G    |
| G+KS-9   | 75.04 | 101.2M   | 10.6G   |
| SES-G+SR-3 | 75.68 | 38.4M    | 7.6G    |
| SES-G+SR-5 | 75.93 | 38.4M    | 7.6G    |
| SES-G+SR-7 | 76.84 | 38.5M    | 7.6G    |
| SES-G+SR-9 | 75.85 | 38.5M    | 7.6G    |
| G+KS-7   | 75.03 | 76.0M    | 9.4G    |
| G+SA-patch+SR-7 | 75.68 | 39.3M | 7.6G |
| G+SA-pair+SR-7 | 75.81 | 38.4M | 7.6G |
| G+DCNv2+SR-21* | 75.26 | 44.9M | 7.8G |
| SES-G+SR-7 (Ours) | 76.84 | 38.5M | 7.6G |

### Table V

Ablation Study on RNM Using the DOTA-v1.0 Dataset

| Method | RNM | mAP   |
|--------|-----|-------|
| SES-G  | x   | 76.10 |
| SES-G  | ✓   | 75.84 |
| SES-O  | x   | 80.69 |
| SES-O  | ✓   | 81.44 |

### Table VI

Alternative Configurations Using the DOTA-v1.0 Dataset

| Method          | mAP   |
|-----------------|-------|
| SES-G+Rand-QKV  | 75.23 |
| SES-G+Rand-QK   | 76.84 |
| SES-G+Rand-QV   | 74.02 |
| SES-G+Rand-KV   | 75.21 |
| SES-G+Rand-Q    | 75.78 |
| SES-G+Rand-K    | 75.87 |
| SES-G+Rand-V    | 75.74 |
| SES-G+$\tau$ = 0.001 | 75.86 |
| SES-G+$\tau$ = 0.005 | 76.84 |
| SES-G+$\tau$ = 0.02 | 76.45 |
| SES-G+$\tau$ = 0.08 | 75.72 |
| SES-G+$\tau$ = 0.32 | 75.80 |

### Table VII

Experiments on Training Under Different Data Augmentations Using the DOTA-v1.0 Dataset

| Scaling | Reflection | Rotation | mAP   |
|---------|------------|----------|-------|
| ×       | ×          | ×        | 67.85 | 68.79 (+0.94) |
| ✓       | ×          | ×        | 74.55 | 75.49 (+0.94) |
| ×       | ✓          | ×        | 70.78 | 71.76 (+0.98) |
| ×       | ×          | ✓        | 71.51 | 73.35 (+1.84) |
| ✓       | ✓          | ✓        | 76.21 | 78.71 (+2.50) |
rows 9 and 10). We thus chose SES-G+corr. to build the TEM and implemented it with a group linear layer.

In Tab. IV, we evaluated the effect of sampling ranges on network accuracy. We tested SES-G with an SES-layer with different sizes of sampling masks (denoted as SES-G+SR-x). We replaced the SES-layer in SES-G by convolution with different kernel sizes (denoted as G+KS-x). Larger convolution kernel sizes require more kernels to identify objects in different transformations, making it difficult to obtain better performance from larger convolution kernels with the same number of kernels (see rows 1–4). However, our sampling equivariant network can extract object features under different transformations using the same set of kernels, so it can handle a larger sampling range, thus boosting existing models (see rows 5–8). The optimal sampling range of SES-G is 7, which we thus use as the sampling mask size. We also compared SES-G+SR-7 with G+SA-patch, G+SA-pair, and G+DCNv2 with a competitive sampling range (denoted G+SA-patch+SR-7, G+SA-pair+SR-7, and G+DCNv2-SR-21* respectively). Although the sampling range of DCNv2 is theoretically infinite, we statistically found that the actual sampling range of DCNv2 in G+DCNv2 is about 21. When the sampling ranges of different models are similar, SES-G+SR-7 achieved the best results, which shows that its superiority not only comes from a larger sampling range but also the enhanced feature extraction capability due to the sampling equivariant network.

To verify the effectiveness of our RNM module, we replaced it in SES-G and SES-O by BN. The results in Tab. V show that RNM allows the model to generate more general sampling masks, enhancing the generalization ability of the model and providing significantly improved results on a relatively small-scale dataset.

We conducted further experiments with different configurations of RNM, using various combinations of input features with or without random perturbation to Q, K, and V in the SES-layer: e.g. we tried random perturbation of $F''$ for $Q$ but no perturbation of $F'$ for $K$ and $V$. This model is denoted SES-G+Rand-Q; other combinations are denoted similarly. Results in Tab. VI show that SES-G+Rand-QK works best, verifying our hypothesis that the query and key of the self-attention structure can easily overfit the training data. Thus, random perturbations should be added to the query and key input features, while doing so to the value increases noise in those features, which is not conducive to training.

We also considered different random sizes from $r = 0.001$ to $r = 0.32$ for SES-G using DOTA-v1.0. Results in Tab. VI show the optimal value is $r = 0.005$ in this case, but may differ for other datasets and models and must be determined experimentally.

To evaluate the improvement of our method compared to the baseline model under different data augmentations, we tested training Gliding and SES-G with and without different data augmentations: (1) Scaling: randomly scale the training image by a factor of 0.5, 1, or 1.5. (2) Reflection: randomly horizontally and vertically reflect the training image. (3) Rotation: randomly rotate the training image by 0, 90, 180, or 270 degrees. The results in Tab. VII show that our model can significantly improve Gliding under different data augmentation conditions. Moreover, the improvements are particularly pronounced under more complex data augmentation strategies (see the last row). This is because our model with sampling equivariance can learn objects under different deformations using the same set of convolutional kernels. In contrast, models without sampling equivariance can only learn different transformations through multiple convolutional kernels, whose capability is limited to the number of kernels.

### E. Generalizability and Speed

To test the generality of our models, we compared SES-G and SES-O with baseline models Gliding and ORCNN on DOTA-v1.0, DOTA-v1.5, and HRSC2016 datasets. The results in Tab. VIII show that, compared to the baselines, our models have significantly better mAP in each case, using fewer parameters and FLOPs. ORCNN uses an oriented region proposal network (RPN) to handle object rotation and scaling, so it partially duplicates the functionality of the SES-layer. However, SES-O still improves the results using fewer parameters and FLOPs: our method is robust and better handles various transformations.

We also evaluated the inference and training speed of our model and the baseline model. We measured the frames per second (FPS) for inference and training of SES-O and ORCNN using an NVIDIA Tesla A100 GPU. As Tab. IX shows, our model demonstrates a comparable inference speed to the baseline model. However, the training speed is lower due to the extra computational overhead imposed by the RNM during training.

---

**TABLE VIII**

**PERFORMANCE OF THE PROPOSED MODELS ON DIFFERENT DATASETS. EXPERIMENTS ON HRSC2016 WERE EVALUATED UNDER VOC2012 METRICS**

| Method    | DOTA-v1.0 mAP | DOTA-v1.5 mAP | HRSC2016 mAP | Backbone | Params | FLOPs |
|-----------|---------------|---------------|--------------|----------|--------|-------|
| Gliding   | 74.82         | 66.21         | 94.67        | R-101    |        |       |
| SES-G     | 76.84 (+2.02) | 67.89 (+1.68) | 95.35 (±0.68)| R-101+SES|        |       |
| ORCNN     | 80.78         | 76.04         | 97.54        | R-50     |        |       |
| SES-O     | 81.44 (+0.66) | 77.62 (+1.58) | 98.14 (±0.60)| R-50+SES|        |       |

**TABLE IX**

**INFEERENCE AND TRAINING FPS OF THE PROPOSED AND BASELINE MODELS WITH DIFFERENT BACKBONES**

| Method | Backbone | Inference | Training |
|--------|----------|-----------|----------|
| ORCNN  | R-101    | 24.1      | 16.6     |
| SES-O  | R-101+SES| 24.1      | 13.8     |
| ORCNN  | R-50     | 32.6      | 20.3     |
| SES-O  | R-50+SES| 32.7      | 16.6     |

---

*Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.*
TABLE X
DETECTION ACCURACY ON DIFFERENT OBJECTS (AP) AND OVERALL PERFORMANCE (mAP) EVALUATION ON THE DOTA-v1.0 AND DOTA-v1.5 TEST SETS. THE RED AND BLUE RESULTS REPRESENT THE BEST AND THE SECOND-BEST, RESPECTIVELY.

| Method          | Backbone | PL   | BD   | BR   | GTF  | SV   | LV   | SH   | TC   | RC   | ST   | SBF  | RA   | HA   | SP   | HC   | CC   | mAP |
|-----------------|----------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| Azimi et al. [56] | R-101    | 81.36| 74.30| 47.70| 70.32| 64.89| 67.82| 69.98| 90.76| 79.06| 78.20| 53.64| 62.90| 67.02| 64.17| 50.23| -    | 68.16|
| RoI Trans. [46] | R-101    | 88.64| 75.22| 63.44| 75.92| 68.81| 73.68| 83.59| 90.74| 77.27| 81.46| 58.39| 53.54| 62.83| 58.93| 47.67| -    | 69.56|
| R$^3$Det [36]   | R-101    | 99.54| 81.99| 48.46| 62.52| 70.48| 74.29| 77.54| 90.80| 81.39| 83.54| 61.97| 59.82| 65.44| 67.46| 60.05| -    | 71.69|
| SCRDet [35]     | R-101    | 99.98| 80.65| 52.09| 68.36| 66.36| 60.32| 72.41| 90.85| 87.94| 86.86| 65.02| 66.68| 66.25| 68.24| 65.21| -    | 72.61|
| Li et al. [57]  | R-101    | 90.41| 85.21| 55.00| 78.27| 76.19| 72.19| 82.14| 90.70| 78.22| 86.67| 66.62| 68.43| 75.43| 72.70| 57.99| -    | 76.36|
| DRN [37]        | H-104    | 88.91| 80.22| 43.52| 63.35| 73.48| 70.69| 84.94| 90.14| 83.85| 84.11| 50.12| 58.41| 67.62| 68.60| 52.50| -    | 70.70|
| GSDet [38]      | R-101    | 81.12| 76.78| 40.78| 75.89| 64.50| 58.37| 74.21| 89.92| 79.40| 78.83| 64.54| 63.67| 66.04| 58.01| 52.13| -    | 68.28|
| ACE [43]        | DLA-34   | 89.5 | 75.91| 51.60| 60.00| 72.78| 77.81| 86.5 | 90.89| 79.58| 85.7 | 49.07| 59.46| 65.7 | 57.1 | 63.9 | -    | 71.7 |

TABLE XI
EXPERIMENTS ON THE HRSC2016 TEST SET; * INDICATES THAT MAP IS EVALUATED UNDER VOC2012 METRICS; OTHERS ARE EVALUATED UNDER VOC2007 METRICS.

| Method          | Backbone | PL   | BD   | BR   | GTF  | SV   | LV   | SH   | TC   | RC   | ST   | SBF  | RA   | HA   | SP   | HC   | CC   | mAP |
|-----------------|----------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| Azimi et al. [56] | R-101    | 81.36| 74.30| 47.70| 70.32| 64.89| 67.82| 69.98| 90.76| 79.06| 78.20| 53.64| 62.90| 67.02| 64.17| 50.23| -    | 68.16|
| RoI Trans. [46] | R-101    | 88.64| 75.22| 63.44| 75.92| 68.81| 73.68| 83.59| 90.74| 77.27| 81.46| 58.39| 53.54| 62.83| 58.93| 47.67| -    | 69.56|
| R$^3$Det [36]   | R-101    | 99.54| 81.99| 48.46| 62.52| 70.48| 74.29| 77.54| 90.80| 81.39| 83.54| 61.97| 59.82| 65.44| 67.46| 60.05| -    | 71.69|
| SCRDet [35]     | R-101    | 99.98| 80.65| 52.09| 68.36| 66.36| 60.32| 72.41| 90.85| 87.94| 86.86| 65.02| 66.68| 66.25| 68.24| 65.21| -    | 72.61|
| Li et al. [57]  | R-101    | 90.41| 85.21| 55.00| 78.27| 76.19| 72.19| 82.14| 90.70| 78.22| 86.67| 66.62| 68.43| 75.43| 72.70| 57.99| -    | 76.36|
| DRN [37]        | H-104    | 88.91| 80.22| 43.52| 63.35| 73.48| 70.69| 84.94| 90.14| 83.85| 84.11| 50.12| 58.41| 67.62| 68.60| 52.50| -    | 70.70|
| GSDet [38]      | R-101    | 81.12| 76.78| 40.78| 75.89| 64.50| 58.37| 74.21| 89.92| 79.40| 78.83| 64.54| 63.67| 66.04| 58.01| 52.13| -    | 68.28|
| ACE [43]        | DLA-34   | 89.5 | 75.91| 51.60| 60.00| 72.78| 77.81| 86.5 | 90.89| 79.58| 85.7 | 49.07| 59.46| 65.7 | 57.1 | 63.9 | -    | 71.7 |

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Gliding [49], Oriented R-CNN (ORCNN) [48], Oriented Faster R-CNN (OFRCNN) [30], Oriented Mask R-CNN (OMRCNN) [31], RC2 [41], RRPN [59], R2PN [60], RRD [61], CSL [42], and R2CNN [62]. These methods choose different backbones, including: ResNet-50 and ResNet-101 (R-50, and R-101) [63], Hourglass-104 (H-104) [64], DLA-34 [65], DarkNet-53 [66], and ReR50 [34]. We first conduct experiments for the OBB task on the DOTA-v1.0 and DOTA-v1.5 test sets. The results are shown in Tab. X. Note that the results of ACE and Li et al. were directly drawn from their paper since they did not provide the implementation code; we use the results of Li et al. without a model ensemble for a fair comparison. Our method outperforms the others with the best mAP, and 11 out of 15 top-2 category results on DOTA-v1.0 and 12 out of 16 top-2 category results on DOTA-v1.5. Various detection results on DOTA-v1.0 using our SES-G and the baseline method are shown in Fig. 5: our method both has a higher detection rate and provides more accurate bounding boxes.

Results of other experiments on the HRSC2016 test set are shown in Tab. XI, again showing that our method outperforms all others under both VOC2007 and VOC2012 metrics.

V. CONCLUSION
Sampling equivariance is an important property, which can enhance the representation ability of a model without increasing model parameters. We analyzed why existing methods do not have complete sampling equivariance and proposed a new solution, SES-Net. It has significantly better sampling equivariance than existing methods without increasing the number of model parameters or computational effort. We also proposed an RNM module to enhance the generalization ability of the model by adding random perturbations to part of the data without changing the data distribution.

Object detection in aerial images has greater sampling equivariance requirements due to the overhead shooting characteristics. We applied our proposed SES-layer to existing aerial image object detection methods to provide better sampling equivariance. RNM further enhances the generalization ability of our proposed network, and it works well with the SES-layer, achieving new state-of-the-art performance on various benchmarks.

A. LIMITATIONS AND FUTURE WORK
Although the proposed SES-Net offers improved sampling equivariance and higher accuracy, it is prone to overfitting when trained on relatively small datasets due to its reliance on the self-attention structure. To address this issue, we introduced RNM, which enhances generalization on smaller training sets. However, this comes at the cost of a slower network training speed. In the future, we will focus on exploring network architectures that exhibit both sampling equivariance and superior generalization capabilities. Furthermore, we believe that sampling equivariance has the potential to enhance feature extraction and can be applied to a wider range of network modules. For instance, incorporating sampling equivariance into RoI Align may yield a more effective extraction of RoI features. We consider this a promising avenue for further exploration in the future.

ACKNOWLEDGMENT
The authors would like to thank Yuan-Chen Guo for helpful discussions and help in writing, and Dr. Dun Liang for helpful discussions.

REFERENCES
[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), vol. 25, Dec. 2012, pp. 1097–1105.
[2] J. Dai et al., “Deformable convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 764–773.
[3] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” 2020, arXiv:2010.11929.
[4] 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.
[5] Z. Liu, L. Yuan, L. Weng, and Y. Yang, “A high resolution optical satellite image dataset for ship recognition and some new baselines,” in Proc. 9th Int. Conf. Pattern Recognit. Appl. Methods, 2017, pp. 324–331.
[6] D. G. Lowe, “Object recognition from local scale-invariant features,” in Proc. 7th IEEE Int. Conf. Comput. Vis., 1999, pp. 1150–1157.
[7] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, Nov. 2004.
[8] G.-J. Qi, L. Zhang, F. Lin, and X. Wang, “Learning generalized transformation equivariant representations via AutoEncoding transformations,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 4, pp. 2045–2057, Apr. 2022.
[9] W. Wang, J. Liang, and D. Liu, “Learning equivariant segmentation with instance-unique querying,” in Proc. Adv. Neural Inf. Process. Syst., vol. 35, 2022, pp. 12826–12840.
[10] L. Yang, L. Li, X. Xin, Y. Sun, Q. Song, and W. Wang, “Large-scale person detection and localization using overhead fisheye cameras,” 2023, arXiv:2307.08252.
[11] J. Bruna and S. Mallat, “Invariant scattering convolution networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1872–1886, Aug. 2013.
[12] T. Cohen and M. Welling, “Group equivariant convolutional networks,” in Proc. Int. Conf. Mach. Learn., 2016, pp. 2990–2999.
[13] M. Weiler and G. Cesa, “General E(2)-equivariant steerable CNNs,” in Proc. Conf. Neural Inf. Process. Syst. (NeurIPS), vol. 32, 2019, pp. 14334–14345.
[14] P. Shamsolmoali, M. Zareapoor, J. Chausoot, H. Zhou, and J. Yang, “Rotation equivariant feature image pyramid network for object detection in optical remote sensing imagery,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 5608614.
[15] K. Zhu et al., “Multi-oriented rotation-equivariant network for object detection on remote sensing images,” IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2022.
[16] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, “Spatial transformer networks,” 2015, arXiv:1506.02025.
[17] H. Gao, X. Zhu, S. Lin, and J. Dai, “Deformable kernels: Adapting effective receptive fields for object deformation,” 2019, arXiv:1910.02940.
[18] X. Zhu, H. Hu, S. Lin, and J. Dai, “Deformable ConvNets v2: More deformable, better results,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2019, pp. 9308–9316.
[19] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable DETR: Deformable transformers for end-to-end object detection,” 2020, arXiv:2010.04159.
[20] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.
[21] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7794–7803.
[22] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, “Beyond self-attention: External attention using two linear layers for visual tasks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 5, pp. 5436–5447, May 2023.
[23] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, “SegNeXt: Rethinking convolutional attention design for semantic segmentation,” in Proc. Adv. Neural Inf. Process. Syst., vol. 35, 2022, pp. 1140–1156.
[24] A. Vaswani et al., “Attention is all you need,” 2017, arXiv:1706.03762.
X. Yang, J. Yan, Z. Feng, and T. He, “End-to-end object detection with transformers,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2020, pp. 213–229.

P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, “Stand-alone self-attention in vision models,” 2019, arXiv:1906.05909.

J. Liu et al., “Ewin transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 9992–10002.

H. Hu, Z. Zhang, Z. Xie, and S. Lin, “Local relation networks for image recognition,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 3463–3472.

H. Zhao, J. Jia, and V. Koltun, “Exploring self-attention for image recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10073–10082.

S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” 2015, arXiv:1506.01497.

K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2980–2988.

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.

G. Cheng, J. Han, P. Zhou, and D. Xu, “Learning rotation-invariant and Fisher discriminative convolutional neural networks for object detection,” IEEE Trans. Image Process., vol. 28, no. 1, pp. 265–278, Jan. 2019.

J. Han, J. Ding, N. Xue, and G.-S. Xia, “ReDet: A rotation-equivariant detector for aerial object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 2785–2794.

X. Yang et al., “SCRDet: Towards more robust detection for small, cluttered and rotated objects,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8231–8240.

X. Yang, J. Yan, Z. Feng, and T. He, “RDet: Refined single-stage detection with feature refinement for rotating object,” 2019, arXiv:1908.05612.

X. Pan et al., “Dynamic refinement network for oriented and densely packed object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 11204–11213.

W. Li, W. Wei, and L. Zhang, “GSDet: Object detection in aerial images based on scale reasoning,” IEEE Trans. Image Process., vol. 30, pp. 4599–4609, 2021.

L. Hou, K. Lu, and J. Xue, “Refined one-stage oriented object detection method for remote-sensing images,” IEEE Trans. Image Process., vol. 31, pp. 1545–1558, 2022.

Z. Huang, W. Li, X.-G. Xia, and R. Tao, “A general Gaussian heatmap label assignment for arbitrary-oriented object detection,” IEEE Trans. Image Process., vol. 31, pp. 1895–1910, 2022.

L. Liu, Z. Pan, and B. Lei, “Learning a rotation invariant detector with rotatable bounding box,” 2017, arXiv:1711.09405.

X. Yang and J. Yan, “Arbitrary-oriented object detection with circular smooth label,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2020, pp. 677–694.

P. Dai, S. Yao, Z. Li, S. Zhang, and X. Cao, “ACE: Anchor-free corner evolution for real-time arbitrarily-oriented object detection,” IEEE Trans. Image Process., vol. 31, pp. 4076–4089, 2022.

W. Qian, X. Yang, S. Peng, Y. Guo, and J. Yan, “Learning modulated loss for rotated object detection,” 2019, arXiv:1911.08299.

X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, “Rethinking rotated object detection with Gaussian Wasserstein distance loss,” 2021, arXiv:2101.11952.

J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, “Learning RoI transformer for oriented object detection in aerial images,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 2844–2853.

J. Han, J. Ding, J. Li, and G.-S. Xia, “Align deep features for oriented object detection,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 5602511.

X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, “Oriented R-CNN for object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 3500–3509.

Y. Xu et al., “Gliding vertex on the horizontal bounding box for multi-oriented object detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 4, pp. 1452–1459, Apr. 2021.
Xiang-Li Li received the B.S. degree from the Dalian University of Technology in 2019. He is currently pursuing the Ph.D. degree with Tsinghua University. His research interests include computer graphics, image analysis, and computer vision.

Zi-Kai Xiao received the B.S. degree from Tsinghua University in 2023, where he is currently pursuing the Ph.D. degree. His research interests include computer graphics, computer vision, and photonic computing.

Tai-Jiang Mu received the bachelor’s and Ph.D. degrees in computer science and technology from Tsinghua University, Beijing, China, in 2011 and 2016, respectively. He is currently an Assistant Researcher with the Department of Computer Science and Technology, Tsinghua University. His research interests include visual media learning, computer graphics, and image processing.

Ralph R. Martin received the Ph.D. degree from Cambridge University in 1983. He is currently an emeritus Professor with Cardiff University. He has authored over 300 papers and 14 books, covering such topics, such as geometric modeling, computer vision, and computer graphics. He is currently the Associate Editor-in-Chief of Computational Visual Media.

Shi-Min Hu (Senior Member, IEEE) received the Ph.D. degree from Zhejiang University in 1996. He is currently a Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing. He has published over 100 articles in journals and refereed conferences. His research interests include digital geometry processing, video processing, rendering, computer animation, and computer-aided geometric design. He is the Editor-in-Chief of Computational Visual Media. He is on the editorial board of several journals, including Computer-Aided Design and Computers and Graphics.