remote sensing

MDPI

Article

Co-ECL: Covariant Network with Equivariant Contrastive Learning for Oriented Object Detection in Remote Sensing Images

Yunsheng Zhang [1,2], Zijing Ren [1], Zichen Ding [3], Hong Qian [4], Haiqiang Li [4] and Chao Tao [1,*]

[1] School of Geosciences and Info-Physics, Central South University, Changsha 410083, China; zhangys@csu.edu.cn (Y.Z.); 215012151@csu.edu.cn (Z.R.)
[2] Xiangjiang Laboratory, Changsha 410205, China
[3] Intellectual Property Protection Center of Inner Mongolia Autonomous Region, Hohhot 010000, China; dzichen75@gmail.com
[4] Inner Mongolia Tongdao Yao Digital Technology Co., Ltd., Hohhot 010000, China; qh19880710@gmail.com (H.Q.); lhaiqiang08@gmail.com (H.L.)
* Correspondence: kingtaochao@csu.edu.cn

Abstract: Contrastive learning allows us to learn general features for downstream tasks without the need for labeled data by leveraging intrinsic signals within remote sensing images. Existing contrastive learning methods encourage invariant feature learning by bringing positive samples defined by random transformations in feature spaces closer, where transformed samples of the same image at different intensities are considered equivalent. However, remote sensing images differ from natural images in their top-down perspective results in the arbitrary orientation of objects and in that the images contain rich in-plane rotation information. Maintaining invariance to rotation transformations can lead to the loss of rotation information in features, thereby affecting angle information predictions for differently rotated samples in downstream tasks. Therefore, we believe that contrastive learning should not focus only on strict invariance but encourage features to be equivariant to rotation while maintaining invariance to other transformations. To achieve this goal, we propose an invariant–equivariant covariant network (Co-ECL) based on collaborative and reverse mechanisms. The collaborative mechanism encourages rotation equivariance by predicting the rotation transformations of input images and combines invariant and equivariant learning tasks to jointly supervise the feature learning process to achieve collaborative learning. The reverse mechanism introduces a reverse rotation module in the feature learning stage, applying reverse rotation transformations with equal intensity to features in invariant learning tasks as in the data transformation stage, thereby ensuring their independent realization. In experiments conducted on three publicly available oriented object detection datasets of remote sensing images, our method consistently demonstrated the best performance. Additionally, these experiments on multi-angle datasets demonstrated that our method has good robustness on rotation-related tasks.

Keywords: oriented object detection; remote sensing (RS); contrastive learning; equivariant learning

Citation: Zhang, Y.; Ren, Z.; Ding, Z.; Qian, H.; Li, H.; Tao, C. Co-ECL: Covariant Network with Equivariant Contrastive Learning for Oriented Object Detection in Remote Sensing Images. *Remote Sens.* **2024**, *16*, 516. https://doi.org/10.3390/rs16030516

Academic Editor: Eufemia Tarantino

Received: 1 November 2023
Revised: 13 January 2024
Accepted: 20 January 2024
Published: 29 January 2024

1. Introduction

Object detection is a fundamental remote sensing task that has important applications in urban planning, land management, and disaster monitoring. Compared to natural images, remote sensing images are taken from a top-down perspective, so the problem of orientation is more prominent. Different objects of the same class are often distributed with arbitrary orientations, coupled with the characteristic of dense object distributions. Maintaining horizontal bounding boxes (HBBs) [1–7] for detection causes a mismatch between the bounding boxes and the objects. Therefore, object detection in remote sensing images achieves better object localization by using oriented bounding boxes (OBBs) [8–12].

Currently, the success of oriented object detection in the supervised learning paradigm is usually achieved by relying on the strong learning ability of networks to fit the rotation information of the data, but this is limited by the quantity of high-quality labeled data. Even if rotation transformations are applied to data, networks can fit only the rotation information in a single dataset, that is, the angle distribution of the oriented object. Rotation-equivariant features cannot be learned in this way.

Therefore, to address the limitations of the oriented object detection task caused by the lack of high-quality labeled data, we leverage the rich intrinsic signals of remote sensing images with the help of self-supervised learning paradigms to learn good representations that can be transferred to downstream tasks. Currently, self-supervised learning achieves feature learning by constructing pretext tasks [13–16] to generate pseudo-labels; the most classical method is contrastive learning [17–22]. For a given set of images, contrastive learning first applies multiple random transformations to the images, where different transformation samples of the same image are positive samples of each other and transformation samples of other images become negative samples. Then, the distances between positive samples are decreased during feature learning while the distances between negative samples are increased. This means that the different transformation samples of the same image in the feature space are equal, so contrastive learning learns the invariant features of the transformations used through this process, which we call strict invariant learning [23–25]. The choice of different transformations in invariant learning has a significant impact [26,27] on feature learning. Typical transformations, such as random crop and color distortion, have been shown to significantly improve performance [22] and are widely used in contrastive learning.

When rotation transformations are added to invariant learning, transformed samples from the same image with different rotation intensities are forced to be close together in the feature space. This rotation-invariant assumption leads networks learning the same features for differently rotated samples. However, this assumption leads to the loss of rotation information in learned features, limiting the expressiveness of representations in samples with different rotations. In particular, objects in remote sensing images have arbitrary orientations. If a network loses the rich rotation information of these objects, then the features of differently rotated samples are difficult to distinguish, thus making it difficult to achieve the accurate localization of OBBs. Therefore, we consider maintaining invariant learning while encouraging features to remain equally transformed following the rotation transformations applied to images, which we call rotation-equivariant learning. This addresses the contrastive learning limitations when transferred to oriented object detection tasks. Rotation-equivariant learning first appeared in self-supervised learning in the RotNet [16] method proposed by Gidaris et al. A network is trained by predicting four independent rotation transformations of images. Some methods [27–30] conduct rotation-equivariant or near-rotation-equivariant learning by introducing rotation angle prediction tasks or relaxing exact alignments in rotated views in the self-supervised learning paradigm; however, these methods do not achieve the combination of invariant learning and equivariant learning. Some other methods [31–33] learn both invariance and equivariance and apply different learning properties for different transformations; however, the addition of transformations destroys the structure of contrastive learning and increases the number of samples to be learned. This increased number of samples can help networks to learn better features, confusing final accuracy improvements.

To address the above problems, we propose an invariant–equivariant covariant network (Co-ECL) based on collaborative and reverse mechanisms to realize invariant learning tasks and equivariant learning tasks. The invariant learning task encourages features to remain invariant to random transformations by decreasing the distances between differently transformed samples of the same input in the feature space. The equivariant learning task encourages features to remain equivariant to rotation transformations by recognizing rotation-transformed samples of the same input with different intensities. The collaborative mechanism combines these two tasks, jointly supervising the feature learning process to

achieve collaborative learning. The reverse mechanism introduces a reverse rotation module, which applies reverse rotation transformations to the features involved in the invariant learning task, ensuring independent learning. Our method extends strict invariant learning into the collaborative learning of invariance and equivariance and addresses the challenge of achieving better rotation equivariance through supervised learning by employing a self-supervised paradigm. Experiments on three datasets demonstrated the effectiveness of our proposed method.

The main contributions of this paper are as follows:

(1) We propose an invariant–equivariant covariant network based on collaborative and reverse mechanisms to achieve invariant and equivariant learning tasks;
(2) The collaborative mechanism combines invariant and equivariant learning tasks, jointly supervising the feature learning process to achieve collaborative learning, while the reverse mechanism introduces a reverse rotation module in the feature learning stage, applying reverse rotations of equal intensity to features in invariant learning tasks as in the data transformation stage, thereby ensuring their independent realization;
(3) Experiments on three oriented object detection datasets of remote sensing images demonstrated the effectiveness of the proposed Co-ECL method.

2. Materials and Methods

2.1. Related Work

2.1.1. Oriented Object Detection in Remote Sensing Images

Due to the arbitrary orientation of objects in remote sensing images, oriented object detection in remote sensing images uses OBBs for more accurate object localization. Currently, oriented object detection in remote sensing images is improved in two ways: OBB definition and feature extraction. Common OBB definitions comprise two types: five-parameter [8] and eight-parameter OBBs [34]. The five-parameter OBBs add angle parameters to horizontal boxes, indicating that the HBBs are rotated by an angle around the center point. The eight-parameter OBBs contain the coordinates of the four corner points of the horizontal external boxes and the offset values of the corresponding corner points, which indicate the positions of the corresponding corner point coordinates of the OBBs. Further methods are dedicated to extracting rotation-invariant features. Cheng et al., designed a rotation-invariant layer and introduced a rotation-invariant regular constraint [35], finding that the features approximated the rotation invariance but needed a large quantity of data for training. Ding et al. studied the transformation from HRoIs to RRoIs in the RoI Transformer [8] method and extracted rotation-invariant features using rotated position-sensitive RoI Align. However, since CNNs cannot learn good rotation-equivariant features under the supervised learning paradigm, rotation-invariant features cannot be efficiently extracted on this basis, even with the addition of sub-modules. That is, the success of rotated object detection in feature learning is usually achieved by relying on the strong learning ability of backbone networks to fit rotation information.

Recently, Han et al. introduced rotation-equivariant networks [36] to learn rotation-equivariant features in oriented object detection and proposed the use of RiRoI Align to obtain features that are rotation-invariant in space and orientation. Although this method learns rotation-equivariant features, it increases the complexity of network structures and introduces many nonessential computations. In contrast, our method makes full use of intrinsic signals in remote sensing images by introducing a self-supervised learning paradigm, which learns rotation-equivariant features in pretraining and does not introduce additional computations.

2.1.2. Rotation in Self-Supervised Learning

Rotation transformations first appeared in self-supervised learning in the RotNet [16] method proposed by Gidaris et al. This method applies four independent rotation transformations to input images and uses the rotation information as pseudo labels, thus allowing

networks to learn strong semantic rotation information. Feng et al. [28] decoupled representations using a rotation prediction task and an instance discrimination task. The method was proposed to have both rotation-related and rotation-unrelated features, but it is too simplistic in the way that it is divided. This method was the first to divide features into rotation-related and rotation-unrelated features; however, the division is too simple. Xiao et al. [27] captured varying and invariant factors for visual representations by constructing separate embedding spaces, each of which is invariant to all but one augmentation. Wang et al. [29] relaxed the exact alignment in multi-view methods by introducing an adaptive residual vector to encode both pretext-invariant and pretext-aware representations. In remote sensing, Wen et al. [30] proposed a new encoding scheme to define and characterize the rotation patterns of ordered sequences among intra-class objects and made their network learn to predict the rotation patterns of a baseline sequence. While the above methods conduct rotation-equivariant or near-rotation-equivariant learning in the self-supervised learning paradigm, our method combines invariant learning and equivariant learning using a collaborative mechanism to learn better representations.

Recently, Dangovski et al. [31] proposed that better features can be learned by maintaining invariance for some transformations while maintaining equivariance for others. Xie et al. [32] had the same point of view and proposed a framework for determining the augmentation types to which feature representations should be invariant and equivariant during pretraining. In remote sensing, Ji et al. [33] solved the few-shot scene classification problem by combining three tasks in training, including rotation prediction, contrastive prediction, and semantic class prediction. While the above methods apply different learning properties for different transformations, the addition of transformation branches destroys the contrastive learning structure and increases the number of samples to be learned. In contrast, our method introduces rotation-equivariant learning without affecting contrastive invariance learning using a reverse mechanism to ensure the independence of both. It also maintains the dual-branch structure of contrastive learning, avoiding any confusion in accuracy improvements due to increased numbers of samples.

2.2. Method

This section introduces our proposed Co-ECL method. The overall network is presented in Section 2.2.1. The collaborative mechanism is introduced in Section 2.2.2 and the reverse mechanism is introduced in Section 2.2.3.

2.2.1. Overview

Our objective is to learn features that remain equivariant to rotation transformations while remaining invariant to other transformations. To achieve this goal, our method consists of the equivariant learning task and the invariant learning task. The network structure contains data transformation, feature learning, and two task branches, which are shown in Figure 1. Specifically, two views are generated during the data transformation stage. After the feature learning stage, features are sent to two separate branches. The invariant learning branch encourages features to remain invariant to random transformations by decreasing the distances between differently transformed samples of the same input in the feature space. The equivariant learning branch encourages features to remain equivariant to rotation transformations by recognizing rotation-transformed samples of the same input with different intensities.

(1) Data transformation. For a given input image x, two different views are first generated by applying a random data transformation τ (without rotation transformation) to image x. The two views include $q = t(x), t \in \tau$ and $k_+ = t'(x), t' \in \tau$. We then define the set of N rotation transformations as $R = \{r(\cdot|y)\}_{y=1}^{N}$, where $r(\cdot|y)$ is the operator that is applied to view the rotation transformation with label y. The set of rotation transformations R is used to generate N transformation samples for both views. We set the view q to maintain a rotation angle of 0°, that is, $q^r = r(q|y = 1)$. For view k_+, $N - 1$ rotation transformations are performed to generate $N - 1$ samples

$k_+^{r_i} = r(k_+|y_i), i \in 2,3,\ldots,N$ with different rotation transformation intensities. Here, we adopt eight rotation transformations at 45° intervals, meaning that the rotation transformation set is denoted as $R = \{r(\cdot|y)\}_{y=1}^{8}$.

(2) Feature learning. We define $f(\cdot;\theta)$ as an encoder with learnable parameters θ and $p_e(\cdot;\phi_{p_e})$ and $p_i(\cdot;\phi_{p_i})$ as projection heads with learnable parameters ϕ_{p_e} and ϕ_{p_i}, respectively. Let $r^{-1}(\cdot|y)$ denote the reverse rotation transformation applied to the features, with an intensity corresponding to the intensity of the rotation transformation during data transformation. For views q^r and $k_+^{r_i}, i \in 2,3,\ldots,N$, after passing through the encoder, features $h_q^r = f(q^r;\theta)$ and $h_{k_+}^{r_i} = f(k_+^{r_i};\theta), i \in 2,3,\ldots,N$ are generated, collectively referred to as h^r. Feature h^r is passed through $p_e(\cdot;\phi_{p_e})$ to generate the representation $z^r = p_e(h^r;\phi_{p_e})$, which is then sent to the equivariant learning branch. Simultaneously, after applying the reverse rotation transformation to feature h^r and computing the average feature h_{k_+} of $h_{ki_+}, i \in 2,3,\ldots,N$, we obtain feature $h = \{h_q, h_{k_+}\} = r^{-1}(h^r|y)$ without any rotation transformation information. h is then passed through $p_i(\cdot;\phi_{p_i})$ to generate the representation $z = \{z_q, z_{k_+}\} = p_i(h;\phi_{p_i})$, which is sent to the invariant learning branch.

(3) Invariant learning branch. Invariant learning refers to bringing the features of differently transformed views of the same image closer while pushing away the transformed views of other images. This means bringing the features z_q and z_{k_+} from different views generated from the same input image x closer. During training, the network is optimized by minimizing the cosine similarity between positive samples using the InfoNCE loss function.

(4) Equivariant learning branch. Equivariant learning refers to predicting the rotation intensity label y for the input image x after applying the rotation transformation $r(x|y)$, that is, predicting the rotation intensity labels of views q^r and $k_+^{r_i}, i \in 2,3,\ldots,N$. During training, this task is optimized by minimizing the cross-entropy loss of predicting rotation transformation labels.

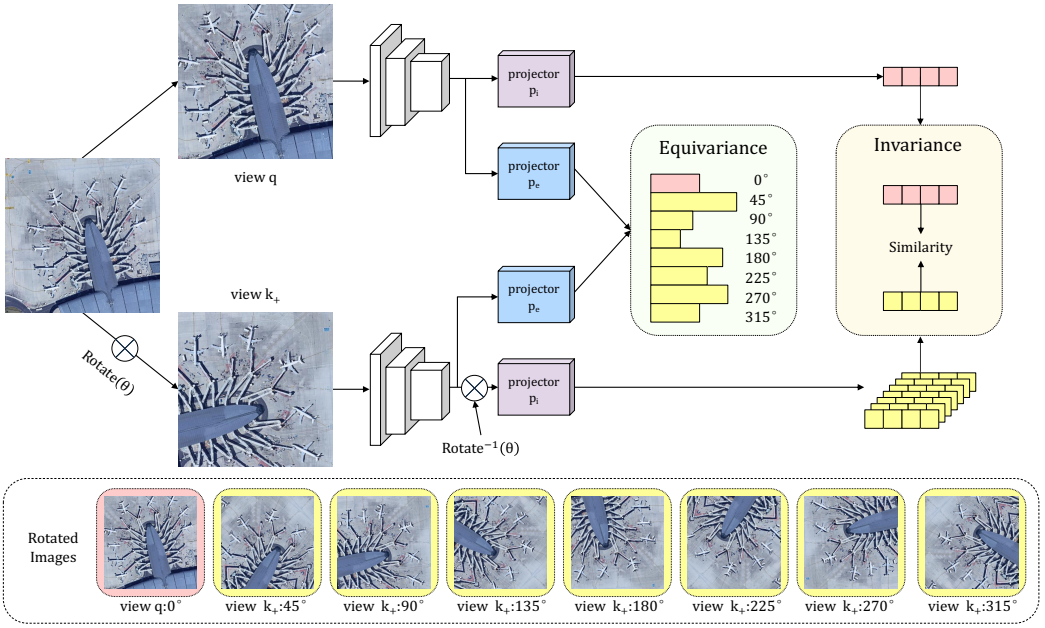

Figure 1. Overall structure of Co-ECL: data transformation, feature learning, invariant learning branch, and equivariant learning branch.

2.2.2. Collaborative Mechanism

The collaborative mechanism of covariant networks refers to the joint supervision of the feature learning stage by the invariant and equivariant learning tasks. Next, we introduce each of these two tasks.

(1) Invariant Learning Task

We define x as the input image, f as the encoder network that performs feature learning, R as a set of transformations, and $r(x)$, r as the function that applies the transformation r to the input image x. The encoder f learns the invariance of the transformation R when any input image x satisfies Equation (1), as shown in Figure 2a.

$$f(x) = f(r(x)), r \in R \tag{1}$$

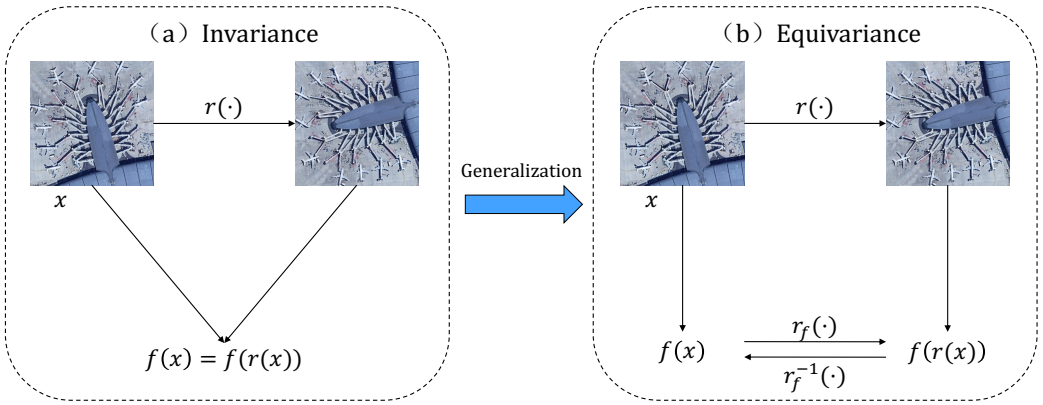

Figure 2. Conceptual diagram of invariance and equivariance. Invariance is described by Equation (1), while equivariance is described by Equation (4) and also Equation (7).

Contrastive learning reduces the distances between different samples of the same image in the feature space after applying random transformations during the learning process. This means that in the invariant learning task of contrastive learning, if the features corresponding to the input image x after different transformations r and r' are the same, then encoder f has learned the invariance of transformation R, as described in Equation (2).

$$f(r(x)) = f\left(r'(x)\right), r \in R \tag{2}$$

Invariant learning tasks use the InfoNCE loss function to minimize the similarity distances between differently transformed views of the same image. Specifically, when the similarity distances between the features z_q and z_{k_+} from different views generated from the same input image x are closest, the loss function reaches its minimum. Here, dot products are used to calculate the similarity. The invariant loss function is as follows:

$$L_{in} = -log\frac{exp\left(z_q \cdot z_{k_+}/\tau\right)}{\sum_{i=0}^{K} exp\left(z_q \cdot z_{k_i}/\tau\right)} \tag{3}$$

where τ is the temperature hyperparameter and K is the number of negative samples.

(2) Equivariant Learning Task

For input image x, applying the transformation r to the image x is the same as the equivalent transformation r_f of features, meaning that equivariance encourages the transformations applied to features to have the same intensity as those applied to the input image, as shown in Figure 2b. By extending invariance to Equation (4), the encoder f learns the equivariance of transformation R. In other words, equivariance is the generalization of invariance.

$$r_f(f(x)) = f(r(x)), r \in R \tag{4}$$

We define a set of N rotation transformations as $R = \{r(x|y)\}_{y=1}^{N}$, where $r(x|y)$ is the rotation transformation applied to image x with an intensity label defined as y. The rotation-equivariant task trains a network to recognize the intensity of the rotation

transformation applied to the input image. Here, we optimize the network by minimizing the cross-entropy loss between the predicted probability distribution and the rotation label y. The equivariance loss function is as follows:

$$L(x_i, y) = -\frac{1}{N}\sum_{y=1}^{N} log\left(p_e\left(f(r(x_i|y);\theta);\phi_{p_e}\right)\right) \quad (5)$$

Thus, given a set of training images $\{x_i\}_{i=1}^{B}$, the overall loss function during training is

$$L_{equi} = \frac{1}{B}\sum_{i=1}^{B} L(x_i, y) \quad (6)$$

(3) Collaborative Mechanism

We combine the invariant learning task and the equivariant learning task using a dual-branch network to jointly supervise the feature learning process, as shown in Figure 1. The overall loss function is the weighted sum of the invariant loss L_{in} and the equivariant loss L_{equi}, with λ assumed to be 0.1 by default.

$$L = L_{in} + \lambda L_{equi} \quad (7)$$

2.2.3. Reverse Mechanism

The reverse mechanism of the covariant network applies reverse rotation transformations of equal intensity to features in the invariant learning task as in the data transformation stage, which ensures that the features in that task do not contain rotation transformations.

In Equation (4), we define the concept of equivariance. In this equation, equivariance is achieved by encouraging the transformations applied to features to have the same intensity as those applied to the input image. We define the reverse transformation of r as r^{-1} and the reverse transformation follows the equation $x = r^{-1}(r(x))$. Similarly, the reverse transformation of the transformation r_f applied to the features is denoted as r_f^{-1}. Therefore, we can transform Equation (4) into Equation (8), where the transformed features after the reverse transformation are the same as the original features, as shown in Figure 2b. When specifying a counterclockwise rotation transformation, the reverse rotation transformation corresponds to a clockwise rotation.

$$f(x) = r_f^{-1}(f(r(x))), r \in R \quad (8)$$

In the data transformation stage of the covariant network, we apply rotation transformations to both views q and k_+. To ensure that the rotation transformations do not interfere with the invariant learning process, we refer to Equation (8) and introduce a feature reverse rotation module. This module applies reverse rotation transformations of the same intensity as the rotation transformations in the data transformation stage to the features in the invariant learning task, ensuring that the features in the invariant task do not contain rotation information. This ensures independence between the invariant learning task and the equivariant learning task while combining them.

The reverse rotation module is implemented using a spatial transformer [37] structure. This module can perform spatial transformations on features and be inserted at any position within the CNN. In the covariant network, we insert it after the encoder f and before the projection head p_i. Since the rotation transformation intensity is the same as that of the input image, the parameters of the spatial transformer module remain fixed during network training.

3. Experiments

We evaluated our method on three oriented object detection datasets of remote sensing images. The datasets and parameter settings are described in Section 3.1. The main experimental results with state-of-the-art comparisons are presented in Section 3.2. The ability of

the model to learn rotation-equivariant features is evaluated in Section 3.3. The effect of different numbers of rotations on network learning is evaluated in Section 3.4. In Section 3.5, we present some ablation studies.

3.1. Datasets and Implementation Details

We trained our method using the FAIR1M [38] dataset and validated it on three oriented object detection datasets: DOTA-v1.0 [9], DOTA-v1.5, and DIOR-R [39]. For pretraining, all images from the FAIR1M dataset were used. For the DOTA-v1.0 and DOTA-v1.5 datasets, 80% of the original training data were randomly divided into a training dataset, then 10% of the divided training dataset was randomly selected for fine-tuning, and the original validation data were used as the test dataset. For the DIOR-R dataset, 10% of the original training data were randomly selected for fine-tuning, and the original validation data were used as the evaluation dataset.The datasets are shown in detail in Table 1.

Table 1. Detailed information about the FAIR1M, DOTA-v1.0, DOTA-v1.5, and DIOR-R datasets. All datasets are publicly available.

Dataset	FAIR1M [38]	DOTA-v1.0 [9]	DOTA-v1.5	DIOR-R [39]
Categories	37	15	16	20
Instances	>1 million	188,282	403,318	192,518
Image Width	1000–10,000	800–4000	800–4000	800
Crop Width	1000	1024	1024	800
Annotation	OBB	OBB	OBB	OBB
Usage	Pretraining	Fine-tuning	Fine-tuning	Fine-tuning

FAIR1M is a remote sensing dataset for fine-grained oriented object detection. The images in the FAIR1M dataset were collected from different sensors and platforms, with spatial resolutions ranging from 0.3 m to 0.8 m. There are more than 1 million instances and more than 16,488 images in this dataset. All objects in the FAIR1M dataset are annotated with respect to 5 categories and 37 sub-categories according to oriented bounding boxes (OBBs). No labels from the dataset were used in the pretraining.

The DOTA-v1.0 images were collected from Google Earth, GF-2, and JL-1 satellites provided by the China Centre for Resources Satellite Data and Application. It is one of the largest datasets for oriented object detection in aerial images and contains 15 common categories, 2806 images, and 188,282 instances. Each image is from 800 × 800 to 40,000 × 40,000 pixels in size. The object categories in DOTA-v1.0 are as follows: plane (PL), baseball diamond (BD), bridge (BR), ground track field (GTF), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer field (SBF), roundabout (RA), harbor (HA), swimming pool (SP), and helicopter (HC). For training and testing, we cropped the original images into a series of 1024×1024 patches with a stride of 824.

DOTA-v1.5 uses the same images as DOTA-v1.0, but the extremely small instances (less than 10 pixels) are also annotated. Moreover, a new category was added: container crane. It contains 403,318 instances in total. Compared to DOTA-v1.0, DOTA-v1.5 is more challenging but also more stable during training. For training and testing, we cropped the original images into a series of 1024 × 1024 patches with a stride of 824.

The angle distribution of objects for each category in the cropped DOTA-v1.0 dataset is shown in Figure 3. The horizontal axis represents the categories contained in the DOTA-v1.0 dataset, while the vertical axis represents the proportion of the number of objects in each angle interval to the total number of objects in that category. The different colors of the legend represent data from different angle intervals. It shows that in this remote sensing dataset, there are two main categories, as shown in Figure 4: (a) ground-nonfixed objects and (b) ground-fixed objects. Ground-nonfixed objects include vehicles, planes, ships, etc. This type of target has an arbitrary orientation and can be distributed in the image at any angle. Ground-fixed objects mainly include basketball courts, soccer fields,

and ground track fields. This type of target has a regular orientation and can be distributed in the image at one or several fixed angles. Remote sensing object detection datasets are typically collected from several fixed remote sensing satellites. Due to the limitations of satellite trajectory and shooting angle, the angles of ground-fixed objects are concentrated in a few angle intervals. The angle distribution of ground-nonfixed objects is more uniform over the intervals due to the arbitrary object distribution.

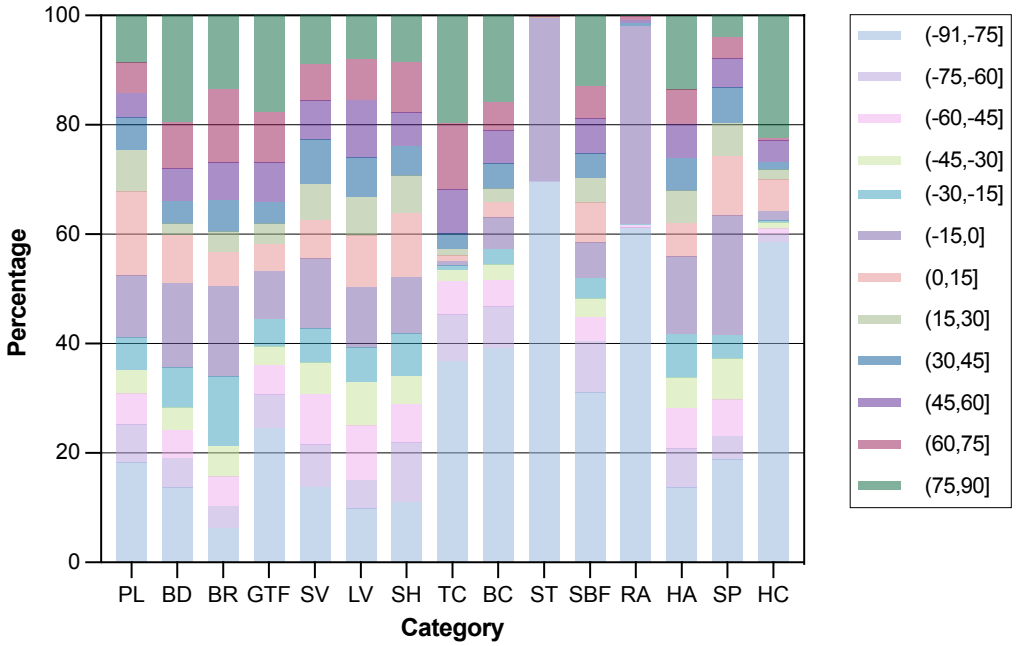

Figure 3. Angle distribution of objects in each category of the DOTA-v1.0 dataset.

Figure 4. Example images of ground-nonfixed and ground-fixed objects in the DOTA-v1.0 dataset: (**a**) ground-nonfixed objects; (**b**) ground-fixed objects.

DIOR-R is a challenging remote sensing dataset for oriented object detection, which shares the same images as the DIOR [40] dataset labeled with horizontal annotations. There is a total of 23,463 images and 192,518 instances, covering 20 classes. The size of each image

is 800 × 800 and the spatial resolutions range from 0.5m to 30m. The object categories in DIOR-R are as follows: airplane (APL), airport (APO), baseball Field (BF), basketball Court (BC), bridge (BR), chimney (CH), dam (DAM), expressway toll station (ETS), expressway service area (ESA), golf course (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). The original image size was used for both training and testing.

Pretraining. Our method is based on MoCo-v2 [21]. We optimized the model using synchronized SGD with a weight decay of 0.0001 and a momentum of 0.9. We used a batch size of 128 in a GPU. The optimization took 200 epochs with an initial learning rate of 0.015 and a cosine learning rate schedule. The learning rate was multiplied by 0.1 at 120 and 160 epochs. The backbone network was ResNet50 [41] and the temperature hyperparameter was set to 0.07.

Fine-tuning. The detector was fine-tuned using the RoI Transformer [8] method implemented in the MMRotate-v0.3.4 [42]. The backbone network weights obtained from pretraining were use and frozen during training. We optimized the model using synchronized SGD with a weight decay of 0.0001 and a momentum of 0.9. The optimization took 200 epochs with an initial learning rate of 0.005 and a batch size of 4. We adopted a learning rate warm-up of 500 iterations and the learning rate was divided by 10 at 133 and 183 epochs.

3.2. Main Results

We compared our method to other current state-of-the-art methods on three separate remote sensing image datasets for oriented object detection. Tables 2–4 show the experimental results on DOTA-v1.0, DOTA-v1.5, and DIOR-R, respectively. Our method showed the best performance compared to the baseline methods on all three datasets, with improvements of 0.39%, 0.33%, and 0.32%, respectively. Combining the results from the DOTA-v1.0 dataset in Tables 2 and 3 and the angle distribution for each category in Figure 3, we found that compared to the baseline method, ground-fixed objects, such as GTF, BC, and SBF, achieved an average accuracy improvement of 2.56% on DOTA-v1.0 and DOTA-v1.5, which was 1.50% higher than ground-nonfixed objects, such as PL, BR, SV, and LV. Our method showed more significant improvements on ground-fixed objects where the angle information provided was more unbalanced. This demonstrates that the Co-ECL method does not rely on the network fitting the angle distribution of the oriented object for feature learning, but really learns better rotation equivariance through the covariant network. However, our method performed poorly on two circular objects: ST and RA. We found that this was because circular objects do not have obvious angle information. Circular objects also do not have different angle information for the network, but this type of object has different angle annotations. This is contradictory for network training, so the accuracy was reduced.

Table 2. Comparisons to state-of-the-art methods on DOTA-v1.0. The results in red indicate the best results in each column.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
RotNet [16]	56.06	19.72	12.53	15.51	29.98	45.58	62.10	59.34	4.01	43.70	9.11	7.20	9.48	37.94	9.49	28.12
SimCLR [22]	57.94	25.34	11.52	16.26	27.78	42.62	52.67	56.41	4.09	52.07	22.67	12.68	10.21	43.10	6.79	29.48
Barlowtwins [43]	59.41	27.93	12.43	18.24	28.75	45.33	59.49	61.97	4.17	50.41	16.75	17.76	11.92	40.91	7.11	30.84
SimSiam [24]	61.39	30.17	7.67	18.32	31.26	45.24	61.34	54.14	4.22	52.59	17.74	14.82	11.59	50.44	6.85	31.19
ESSL [31]	61.17	28.61	13.38	20.05	30.80	45.93	60.68	61.65	4.79	47.69	21.28	12.39	10.90	46.67	8.21	31.61
MoCo-v2 [21]	61.72	33.69	14.98	17.51	33.31	47.50	62.19	65.81	4.43	47.02	23.36	12.28	15.85	37.66	11.47	32.58
Co-ECL	62.46	31.98	15.80	20.87	36.16	47.99	61.81	64.19	5.92	47.08	25.08	11.98	18.61	35.83	8.87	32.97

Table 3. Comparisons to state-of-the-art methods on DOTA-v1.5. The results in red indicate the best results in each column.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	CC	mAP
RotNet [16]	49.76	18.56	6.68	9.85	49.40	47.41	57.22	61.69	9.14	46.71	0.81	15.74	21.22	19.45	3.19	0.98	26.11
SimCLR [22]	58.99	31.07	9.61	12.99	47.75	46.68	50.29	60.13	10.27	50.53	6.64	17.08	15.87	28.95	4.33	0.54	28.23
Barlowtwins [43]	55.81	32.28	8.67	13.55	46.43	47.36	54.58	67.47	12.35	46.39	14.70	19.88	22.07	27.65	4.55	1.25	29.69
SimSiam [24]	57.62	34.88	8.17	12.65	50.20	47.03	55.00	59.90	9.24	52.94	7.68	18.72	16.13	43.00	1.34	1.24	29.73
ESSL [31]	57.65	33.33	9.79	13.77	48.52	47.73	56.29	62.71	10.42	48.95	9.31	19.90	18.34	35.17	2.27	2.14	29.77
MoCo-v2 [21]	57.78	31.74	11.61	15.92	45.99	49.62	57.67	71.52	15.17	44.81	11.51	20.13	22.41	28.15	4.16	0.97	30.57
Co-ECL	58.29	29.26	11.98	18.32	47.65	50.64	55.25	69.91	16.50	39.04	16.56	16.05	23.87	28.98	9.65	2.36	30.90

Table 4. Comparisons to state-of-the-art methods on DIOR-R. The results in red indicate the best results in each column.

Method	APL	APO	BF	BC	BR	CH	DAM	ETS	ESA	GF	GTF	HA	OP	SH	STA	STO	TC	TS	VE	WM	mAP
RotNet [16]	76.78	6.25	79.08	17.56	7.96	63.15	1.29	21.48	15.92	28.39	30.02	11.45	8.61	73.49	44.61	59.00	70.78	4.86	40.86	17.97	33.98
SimSiam [24]	65.73	2.79	79.18	22.12	9.09	64.12	4.24	22.59	16.24	33.77	38.63	9.12	10.99	73.35	64.38	60.91	68.26	7.83	43.71	16.84	35.69
SimCLR [22]	71.93	7.58	79.49	25.55	7.20	68.16	3.70	26.11	16.01	32.20	38.34	11.22	11.89	71.55	65.28	61.29	71.26	7.78	40.08	15.98	36.63
MoCo-v2 [21]	65.95	6.82	81.06	37.48	8.52	72.34	1.99	27.32	16.68	29.19	38.45	14.81	10.68	79.44	60.23	61.75	72.15	10.60	41.88	18.18	37.78
ESSL [31]	72.65	6.06	80.85	34.75	8.52	68.45	3.88	31.13	17.31	31.04	38.30	14.08	11.78	78.91	59.38	61.32	72.08	7.15	42.63	17.77	37.90
Barlowtwins [43]	68.02	6.14	81.08	44.30	5.76	76.07	1.51	25.10	18.36	32.37	39.40	12.28	13.61	76.96	56.97	62.05	72.24	8.34	41.21	18.11	37.99
Co-ECL	74.31	5.00	80.84	34.42	7.51	68.60	3.47	30.22	17.94	35.50	38.00	13.44	12.19	78.96	60.86	61.15	72.25	9.12	43.68	18.77	38.31

3.3. Rotation Equivariance Measurement

To verify whether our method could learn better rotation equivariance, we applied rotation transformations to the data and validated the method on multi-angle datasets. The validated data came from the test datasets of DOTA-v1.0, DOTA-v1.5, and DIOR-R. We applied rotation transformations to generate data under four angles: 0°, 90°, 180°, and 270°.

3.3.1. Important Regions

To verify whether the model could learn better rotation equivariance after pretraining, we visualized the important regions for predicting the concept on multi-angle images. We used the Grad-CAM [44] method for the visualizations. The Grad-CAM method is a popular CNN visualization approach, which uses the global averages of gradients to compute the weights in feature maps, producing coarse localization maps that highlight important regions in images for predicting the concept. Specifically, we used the pretraining backbone network weights to obtain visualization results on the DOTA-v1.0 dataset and compared them to the MoCo-v2. Figure 5 shows that MoCo-v2 could only focus on the object region at certain angles, while our method could focus on the object region on all different angle images. This proves that our method can focus on objects with different angles and learn better rotation equivariance.

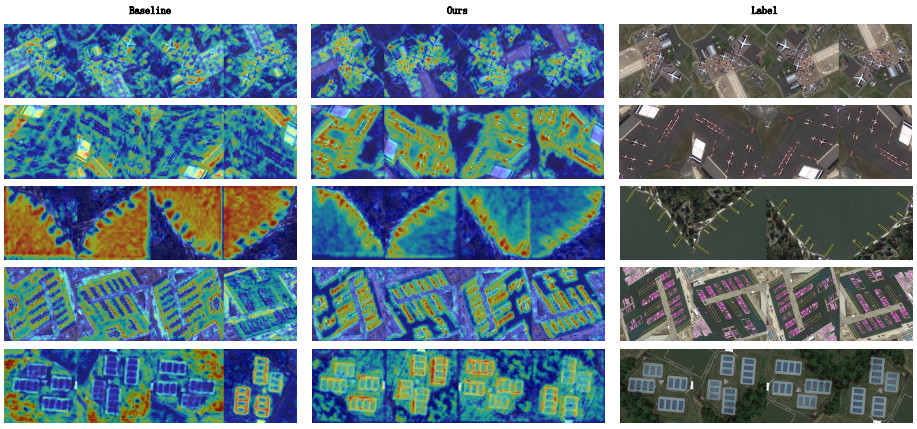

Figure 5. Visualizations of coarse localization maps on multi-angle images from the DOTA-v1.0 dataset. The hottest areas in the figures indicate important regions for predicting the concept.

3.3.2. Detection Accuracy

Since the angle distributions of oriented objects are fixed in a fixed remote sensing dataset, testing on these datasets can only reflect the detection results of oriented objects within that fixed angle distribution. The angle distributions of the oriented objects in the training and test datasets were similar, so it is possible that the network performed angle fitting via its strong learning ability to achieve better detection results. To more accurately measure the rotation equivariance learned by the network, we used multi-angle datasets to compare the overall levels and degrees of deviation. Specifically, we tested the multi-angle datasets separately to obtain the detection results of oriented objects within each angle distribution. We then took the averages of the accuracy $\overline{AP}$ and the coefficient of variation c_v for each angle distribution. The average accuracy $\overline{AP}$ for each angle distribution represented the overall detection accuracy. A larger average $\overline{AP}$ represented better overall detection results on multi-angle datasets. The coefficient of variation c_v is the ratio of the standard deviation to the mean and is used to compare the dispersion of data when means are not equal. Here, the coefficient of variation c_v represented the degree of detection deviation. A smaller coefficient of variation c_v represented less variation in detection on multi-angle datasets.

Tables 5–7 show that our method had the highest average accuracy $\overline{AP}$ on all three datasets, with improvements of 0.75%, 0.59%, and 0.12% compared to the best-performing method on each dataset. The improvement was most obvious on the DOTA-v1.0 dataset and poorest on the DIOR-R dataset. We found that this was because some categories are forcefully annotated by horizontal boxes, although the objects are not exactly horizontal. This confuses the information about the angles that the network needed to learn and thus, could have affected the network training and detection results.The coefficient of variation c_v also decreased compared to the MoCo-v2 method by 0.48%, 2.5%, and 0.26% on the three datasets. The average decrease was 1.08%. The decrease was most obvious on the DOTA-v1.5 dataset and poorest on the DIOR-R dataset. The decrease in the degree of deviation was most obvious and higher on the DOTA-v1.5 dataset than the other two datasets. It is possible that the DOTA-v1.5 dataset has more detailed annotations, providing a greater amount of quantifiable rotation information during testing. On datasets with higher difficulty levels in rotation-related tasks, the improvement of Co-ECL compared to the baseline methods became more obvious.

The optimization of these two metrics showed that the Co-ECL method could achieve good detection results on test data with different angle distributions by only relying on the single angle distribution data. This means that the model produces accurate predictions even when the angle distribution of the test set is different from that of the training set. This demonstrates that the Co-ECL method truly learns good rotation-equivariant features of objects and improves the robustness of the model on rotation-related tasks.

Table 5. Comparisons to state-of-the-art methods on DOTA-v1.0 in terms of the averages of the accuracy $\overline{AP}$ and coefficient of variation c_v.

Method	0°	90°	180°	270°	$\overline{AP}$	c_v
RotNet [16]	28.12	24.32	28.70	24.96	26.52	7.19
SimCLR [22]	29.48	24.84	29.11	25.08	27.13	8.00
Barlowtwins [43]	30.84	26.11	30.58	26.40	28.48	7.83
ESSL [31]	31.61	26.82	32.07	27.24	29.43	8.21
SimSiam [24]	31.19	27.11	32.15	27.76	29.55	7.30
MoCo-v2 [21]	32.58	27.98	32.20	28.09	30.21	7.23
Co-ECL	32.97	29.12	33.11	28.64	30.96	6.75

Table 6. Comparisons to state-of-the-art methods on DOTA-v1.5 in terms of the averages of the accuracy $\overline{AP}$ and coefficient of variation c_v.

Method	0°	90°	180°	270°	$\overline{AP}$	c_v
RotNet [16]	26.11	20.29	25.87	20.18	23.11	12.45
SimCLR [22]	28.23	23.06	27.55	23.12	25.49	9.47
Barlowtwins [43]	29.69	22.45	28.73	22.67	25.88	12.92
MoCo-v2 [21]	30.57	23.32	29.26	23.37	26.63	12.45
ESSL [31]	29.77	23.77	28.91	24.29	26.69	10.04
SimSiam [24]	29.73	24.13	29.15	24.33	26.83	9.74
Co-ECL	30.90	24.68	29.28	24.82	27.42	9.95

Table 7. Comparisons to state-of-the-art methods on DIOR-R in terms of the averages of the accuracy $\overline{AP}$ and coefficient of variation c_v.

Method	0°	90°	180°	270°	$\overline{AP}$	c_v
RotNet [16]	33.98	27.39	34.05	27.53	30.74	9.20
SimSiam [24]	35.69	30.64	35.42	30.86	33.15	6.25
SimCLR [22]	36.63	31.22	36.50	31.69	34.01	6.41
Barlowtwins [43]	37.99	32.12	37.87	31.62	34.90	7.79
MoCo-v2 [21]	37.78	31.98	37.71	32.42	34.97	6.81
ESSL [31]	37.90	32.60	38.15	32.97	35.40	6.66
Co-ECL	38.31	32.79	38.10	32.88	35.52	6.55

3.4. Effect of Different Numbers of Rotations

To explore how the number of rotation transformations affected the ability of our method to learn good representations, we defined three tasks with different numbers of rotations. The three tasks used images with 4, 8, and 12 rotations, where the rotations were all spaced at 45° intervals. To avoid black border areas appearing after the rotations and thus causing the leakage of GT angle information, we used a center crop after rotation to obtain images of the same size as the original and then used reflection padding to fill the black areas [45], as shown in Figure 6. Based on the results of the experiments on the three datasets in Table 8, we found that we achieved the highest accuracy after applying eight rotation transformations. This proves that for the oriented object detection task, choosing eight rotation transformations during pretraining is the optimal rotation transformation scheme.We analyzed this as a result of a more detailed angle division of oriented objects in remote sensing. Compared to other downstream tasks, oriented object detection is more difficult in the task of recognizing the rotation intensity. Therefore, more rotation transformations need to be applied during pretraining to learn better representations of rotation equivariance.

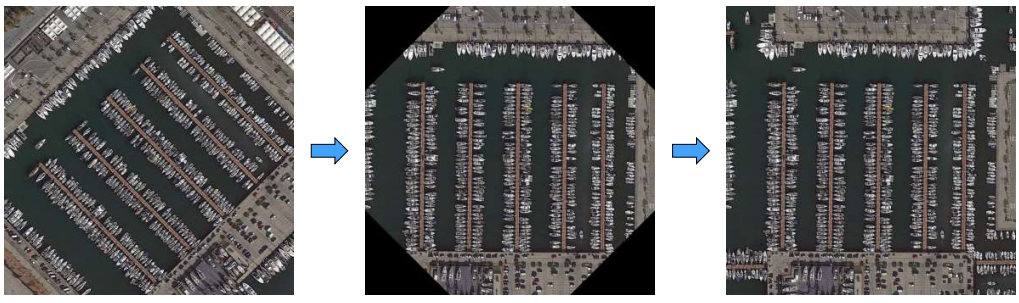

Figure 6. Reflection padding of rotated images.

Table 8. Exploring how the number of rotation transformations affects the quality of learned features. Comparisons of the three rotation transformation schemes on each of the three datasets.

Number	Rotations	DOTA-v1.0	DOTA-v1.5	DIOR-R
4	0°, 90°, 180°, 270°	32.44	30.16	38.02
8	0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°	32.97	30.90	38.31
12	0°, 30°, 60°, 90°, 120°, 150°, 180°, 210°, 240°, 270°, 300°, 330°	31.72	30.74	37.23

3.5. Ablation Studies

In this section, we present a series of ablation experiments conducted on three datasets to evaluate the effectiveness of our proposed method.

3.5.1. Collaborative Mechanism

To evaluate the effectiveness of the collaborative mechanism, we compared the Co-ECL method to a method that performs strict invariant learning (including rotation transformations). The method that performs strict invariant learning (including rotation transformations) is achieved by adding rotation transformations that are equal to those in Co-ECL to the baseline method MoCo-v2. This means that this method simply adds rotation-invariant learning to contrastive learning. As shown in Table 9, the experimental results for the method that performs strict invariant learning (including rotation transformations) were not only poorer than those of the Co-ECL method, but even worse than those of MoCo-v2. This demonstrates that rotation transformations are more suitable for equivariant learning in contrastive learning. This implies that downstream oriented object detection in remote sensing images needs to learn rotation-equivariant features more. It also demonstrates that conducting the collaborative learning of invariance and rotation equivariance is more helpful for feature learning.

Table 9. Ablation studies with and without the collaborative mechanism on three datasets.

Method	Rotation	DOTA-v1.0	DOTA-v1.5	DIOR-R
MoCo-v2	-	32.58	30.57	37.78
	Invariance	31.66	30.35	37.65
Co-ECL	Equivariance	32.97	30.90	38.31

3.5.2. Reverse Mechanism

To evaluate the effectiveness of the reverse mechanism, we compared the Co-ECL method to a method that does not include the reverse rotation module. The learning of rotation equivariance and invariance was conducted using this method. Table 10 shows that the network structure without the reverse rotation module performed extremely poorly on all three datasets. This shows that the transformations applied to an image can be counteracted by applying reverse transformations of equal intensity at the feature level, verifying the equivariant reversibility. It also demonstrates that the reverse mechanism played an important role in ensuring the independence of invariant and equivariant learning.

Table 10. Ablation studies with and without the reverse mechanism on three datasets. The $\surd$ and $\times$ indicate the method with and without the reverse rotation module respectively.

Method	Reverse Rotation Module	DOTA-v1.0	DOTA-v1.5	DIOR-R
Co-ECL	$\times$	28.08	27.43	34.21
	$\surd$	32.97	30.90	38.31

4. Discussion

To address the limitations of contrastive learning when applied to downstream rotation-related tasks, we propose collaborative and reverse mechanisms based on con-

trastive learning. These mechanisms extend strict invariant learning into the collaborative learning of invariance and equivariance, learning features that remain equivariant to rotation transformations while maintaining invariance to other transformations. We conducted sufficient experiments on three datasets to validate the effectiveness of our method.

Compared to methods [21,22] that perform strict invariant learning, our method based on the collaborative mechanism can learn features that are better suited to rotation-related tasks. In experiments on multi-angle images, Co-ECL could focus on more comprehensive object regions in multi-angle images compared to the MoCo-v2 method. In the baseline experiments, the improvement in accuracy for ground-fixed objects was significantly higher than ground-nonfixed objects. This demonstrates that our method does not rely on the network fitting the angle distributions of oriented objects for feature learning as it truly learns the rich rotation angle information of the objects. The ablation experiments on the collaborative mechanism also demonstrated that conducting the collaborative learning of invariance and equivariance is more advantageous for the network to acquire good representations.

Compared to methods [31–33] that also perform collaborative learning, the Co-ECL method improved the robustness of the model while the reverse mechanism ensured that the dual-branch structure of contrastive learning was not destroyed. In multi-angle detection experiments, Co-ECL outperformed the ESSL method in both the average accuracy $\overline{AP}$ and the coefficient of variation c_v. This demonstrates that our method alleviates the dependence of the model on high-quality labeled data and improves its robustness on rotation-related tasks. Furthermore, the ablation experiments on the reverse mechanism demonstrated that this mechanism ensures the independent learning of invariance and equivariance. It also enables Co-ECL to maintain the dual-branch structure of contrastive learning, avoiding confusion in accuracy improvements due to increased numbers of samples.

At the same time, it should be pointed out that the network also has limitations in rotation information learning. In the rotation equivariance measurement experiments, the performance of our method in terms of the coefficient of variation was reduced compared to the baseline method. However, there are still methods with lower coefficients of variation c_v. The reason for this may be that the enhancement of our method was not completely balanced across all angles, resulting in large differences. In order to focus the network on more comprehensive angle information during the learning process, we propose that an equivariant task could be designed to better align with oriented objects or more comprehensive rotation transformations could be performed. These issues are all worthy of future study.

5. Conclusions

In this study, we constructed a covariant network based on contrastive learning for oriented object detection in remote sensing images. To achieve the accurate localization of OBBs, we proposed a collaborative mechanism that combines invariant learning and rotation-equivariant learning so that rotation-equivariant features can be truly learned. We also introduced a reverse mechanism, ensuring that the network retains a dual-branch structure while maintaining the independent learning of invariance and equivariance. The combination of these two mechanisms extends strict invariant learning into the collaborative learning of invariance and equivariance, addressing the limitations of contrastive learning when applied to downstream rotation-related tasks. In experiments conducted on three publicly available oriented object detection datasets of remote sensing images, our method consistently demonstrated the best performance. This demonstrates that our proposed method effectively leverages the rich intrinsic signals in remote sensing images, learns better rotation-equivariant features, enhances the robustness of models in rotation-related tasks, and introduces a novel solution to the challenge of learning better rotation equivariance under the supervised learning paradigm. However, in the rotation equivariance measurement experiments, our method showed unbalanced improvements

across different angles and some performance gaps. This suggests that there is still room for improvement in the design of rotation-equivariant learning tasks. In the future, we will conduct further research on how to create more comprehensive rotation-equivariant learning tasks.

Author Contributions: Conceptualization, methodology, software, and writing—original draft preparation, Z.R.; supervision, Y.Z. and C.T.; software and validation, Z.D.; data curation, H.Q. and H.L.; writing—review and editing, Y.Z. and C.T. All authors have read and agreed to the published version of the manuscript.

Funding: The authors gratefully acknowledge the financial support provided by the Major Program Project of Xiangjiang Laboratory (No. 22XJ01010) and the Natural Science Foundation of Hunan for Distinguished Young Scholars under Grant (2022JJ10072).

Data Availability Statement: Data associated with this research are available online. The FAIR1M dataset is available at https://www.gaofen-challenge.com/benchmark (accessed on 5 January 2022). The DOTA-v1.0 and DOTA-v1.5 datasets are available at https://captain-whu.github.io/DOTA/index.html (accessed on 5 March 2019). The DIOR-R dataset is available at https://gcheng-nwpu.github.io#Datasets (accessed on 14 June 2022).

Conflicts of Interest: Authors Hong Qian and HaiQiang Li were employed by the Inner Mongolia Tongdao Yao Digital Technology Co., Ltd. The remaining authors declare no conflicts of interest.

References

1. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep learning for generic object detection: A survey. *Int. J. Comput. Vis.* **2020**, *128*, 261–318. [CrossRef]
2. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
3. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
4. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
5. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
6. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
7. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
8. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI transformer for oriented object detection in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2849–2858.
9. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.
10. Yang, X.; Yan, J. Arbitrary-oriented object detection with circular smooth label. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part VIII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 677–694.
11. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8232–8241.
12. Zhang, G.; Lu, S.; Zhang, W. CAD-Net: A context-aware detection network for objects in remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10015–10024. [CrossRef]
13. Doersch, C.; Gupta, A.; Efros, A.A. Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1422–1430.
14. Zhang, R.; Isola, P.; Efros, A.A. Colorful image colorization. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part III 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 649–666.

15. Noroozi, M.; Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 69–84.
16. Gidaris, S.; Singh, P.; Komodakis, N. Unsupervised representation learning by predicting image rotations. *arXiv* **2018**, arXiv:1803.07728.
17. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality reduction by learning an invariant mapping. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; IEEE: New York, NY, USA, 2006; Volume 2, pp. 1735–1742.
18. Dosovitskiy, A.; Springenberg, J.T.; Riedmiller, M.; Brox, T. Discriminative unsupervised feature learning with convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2014**, *27*. [CrossRef]
19. Oord, A.V.D.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
20. Bachman, P.; Hjelm, R.D.; Buchwalter, W. Learning representations by maximizing mutual information across views. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
21. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
22. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 1597–1607.
23. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9650–9660.
24. Chen, X.; He, K. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 15750–15758.
25. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap your own latent—A new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21271–21284.
26. Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; Isola, P. What makes for good views for contrastive learning? *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6827–6839.
27. Xiao, T.; Wang, X.; Efros, A.A.; Darrell, T. What should not be contrastive in contrastive learning. *arXiv* **2020**, arXiv:2008.05659.
28. Feng, Z.; Xu, C.; Tao, D. Self-supervised representation learning by rotation feature decoupling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10364–10374.
29. Wang, Y.; Geng, Z.; Jiang, F.; Li, C.; Wang, Y.; Yang, J.; Lin, Z. Residual relaxation for multi-view representation learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12104–12115.
30. Wen, Z.; Liu, Z.; Zhang, S.; Pan, Q. Rotation awareness based self-supervised learning for SAR target recognition with limited training samples. *IEEE Trans. Image Process.* **2021**, *30*, 7266–7279. [CrossRef] [PubMed]
31. Dangovski, R.; Jing, L.; Loh, C.; Han, S.; Srivastava, A.; Cheung, B.; Agrawal, P.; Soljacic, M. Equivariant self-supervised learning: Encouraging equivariance in representations. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021.
32. Xie, Y.; Wen, J.; Lau, K.W.; Rehman, Y.A.U.; Shen, J. What should be equivariant in self-supervised learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4111–4120.
33. Ji, H.; Gao, Z.; Zhang, Y.; Wan, Y.; Li, C.; Mei, T. Few-shot scene classification of optical remote sensing images leveraging calibrated pretext tasks. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
34. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.S.; Bai, X. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 1452–1459. [CrossRef] [PubMed]
35. Cheng, G.; Zhou, P.; Han, J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]
36. Han, J.; Ding, J.; Xue, N.; Xia, G.S. Redet: A rotation-equivariant detector for aerial object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2786–2795.
37. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*.
38. Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; et al. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 116–130. [CrossRef]
39. Cheng, G.; Wang, J.; Li, K.; Xie, X.; Lang, C.; Yao, Y.; Han, J. Anchor-free oriented proposal generator for object detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
40. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

42. Zhou, Y.; Yang, X.; Zhang, G.; Wang, J.; Liu, Y.; Hou, L.; Jiang, X.; Liu, X.; Yan, J.; Lyu, C.; et al. Mmrotate: A rotated object detection benchmark using pytorch. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 7331–7334.
43. Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 12310–12320.
44. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
45. Yang, X.; Zhang, G.; Li, W.; Wang, X.; Zhou, Y.; Yan, J. H2RBox: Horizonal Box Annotation is All You Need for Oriented Object Detection. *arXiv* **2022**, arXiv:2210.06742.