Learning rotation equivalent scene representation from instance-level semantics: A novel top-down perspective

Bi, Q.; You, S.; Ji, W.; Gevers, T.

DOI
10.1016/j.cviu.2023.103635

Publication date
2023

Document Version
Final published version

Published in
Computer Vision and Image Understanding

License
CC BY

Computer Vision and Image Understanding 229 (2023) 103635

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Learning rotation equivalent scene representation from instance-level semantics: A novel top-down perspective

Qi Bi [a], Shaodi You [a,*], Wei Ji [b], Theo Gevers [a]

[a] Computer Vision Research Group, University of Amsterdam, Amsterdam, Netherlands
[b] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada

ARTICLE INFO

Communicated by Nikos Paragios

Keywords:
Scene recognition
Rotation equivalent representation
Multiple instance learning
Key instance selection
Top-down Perspective

ABSTRACT

This paper focuses on rotation variant scene recognition. Different from existing rotation invariant recognition approaches which learn from either rotated images or rotated convolutional filters in a bottom-up manner, a new top-down perspective by learning is explored from instance-level semantic representation. The goal is to eliminate the convolutional feature differences in bottom-up feature propagation caused by the rotation sensitive nature of convolution operation. Our rotation equivalent convolutional neural network (RE-CNN) scheme consists of three components. Firstly, our key instance selection module highlights the instances strongly related to the scene scheme regardless of their orientation. Secondly, our key instance aggregation module builds a scene representation invariant to the position change of each instance caused by rotation. Finally, our semantic fusion module allows the framework to be organized as a whole and implements rotation regularization. Notably, our RE-CNN scheme can be adapted to existing CNNs in a plug-in-and-play manner. Extensive experiments on rotation variant scene recognition benchmarks from four domains demonstrate the state-of-the-art performance and generalization capability of the proposed RE-CNN.

1. Introduction

1.1. Problem statement

Due to changes in imaging conditions, the appearance and shape of an object in a scene may vary drastically in terms of orientation, which is usually termed as *rotation variant scenes*. Typical examples are aerial images (Ding et al., 2019; Zheng et al., 2020; Bi et al., 2020b, 2021a; Xia et al., 2018; Cheng et al., 2018) and industrial scenes (Fernandes and Cardoso, 2017; Zhang et al., 2020; Iacovacci and Lacasa, 2020), in which the texture may have a different orientation (Kylberg, 2011; Li et al., 2015). Also, the pathological regions in medical scenes can appear in a variety of orientations (Li et al., 2019; Ilse et al., 2018; Wu et al., 2020). The orientation information in such rotation variant scenarios is far more abundant than in natural scenes (Quattoni and Torralba, 2009; Almakady et al., 2020; Zhang et al., 2013; Hanbay et al., 2015) (see Fig. 1 for an intuitive example). It can cause confusion for computer vision algorithms to understand such scenes (Worrall et al., 2017; Cohen and Welling, 2016; Xia et al., 2017; Li et al., 2019; Zhang et al., 2020).

One may argue that the long-existing challenge of rotation variant scene recognition becomes trivial in the deep learning era, as the rotation based data augmentation (Simonyan and Zisserman, 2015; He et al., 2016; Szegedy et al., 2015; Ding et al., 2019; Xia et al., 2018; Cheng et al., 2018; Chen et al., 2021; Wheeler and Karimi, 2021; Bozorgtabar et al., 2019) has been widely utilized to relieve this problem. However, simply using such practical yet naive rotation based data augmentation fails to learn a discriminative feature representation from multiple angles and limits the model's generalization capability. This may lead to severe performance degradation in many rotation variant cases (Bi et al., 2020b; Xia et al., 2017; Li et al., 2019; Iacovacci and Lacasa, 2020). This phenomenon is due to the convolution operation which is sensitive to rotation, *i.e.*, rotation in-equivalent (Cohen and Welling, 2016; Barnard and Casasent, 1991; Worrall et al., 2017). As the deviation between different rotation angles accumulates in the entire feature propagation, the RoIs are often not activated properly (Bi et al., 2020b, 2021a), leading to a dramatic change in the final semantic prediction.

1.2. Motivation and objective

Many efforts have been made to learn rotation invariant deep features for rotation variant scene recognition. Existing solutions can be summarized into two categories: (1) selecting the representative feature responses from a group of CNN features extracted from rotated samples (Dmitry et al., 2016; Cheng et al., 2019; Zhang et al., 2017; Liao et al., 2018), and (2) learning from features extracted by rotated

* Corresponding author.
E-mail addresses: q.bi@uva.nl (Q. Bi), s.you@uva.nl (S. You).

https://doi.org/10.1016/j.cviu.2023.103635
Received 28 March 2022; Received in revised form 20 December 2022; Accepted 30 January 2023
Available online 3 February 2023

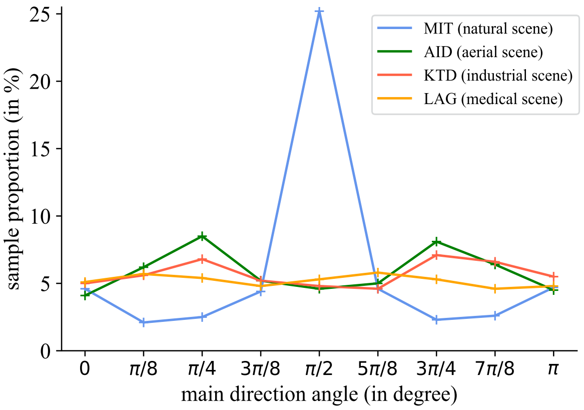


Fig. 1. Different orientation information distribution between natural image scenes and rotation variant scenes, reported in sample proportion. The main direction angle has a range of $[0, \pi)$. The statistics from generic image scenes are from MIT dataset (Quattoni and Torralba, 2009) (blue). The statistics from rotation variant image scenes are from AID (Xia et al., 2017) (green), LAG (Li et al., 2019) (orange) and KTD (Kylberg, 2011) (pink) datasets respectively. It can be clearly seen that the orientation information from rotation variant scenes are more abundant and more randomly distributed, while for natural scenes the orientation information is more gathered horizontally or vertically. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

convolution filters (Cohen and Welling, 2016; Worrall et al., 2017; Zhou et al., 2017).

As is shown in Fig. 2, since these approaches are in a bottom-up manner, *i.e.*, extracting rotation information from shallow to deep, the rotation sensitive nature of the convolution poses a bottleneck to learn a discriminative representation from different rotation angles. This hinders the understanding capability of rotation variant scenes and limits the generalization capability (Wu et al., 2020; Xia et al., 2017; Iacovacci and Lacasa, 2020).

To this end, in contrast to existing methods, we present a novel rotation equivalent scene representation learning scheme from a top-down perspective. In this scheme, it is not required to extract convolutional features from multiple rotated samples or from rotated convolution filters, and eliminates the drawback in existing bottom-up pipelines. Notably, no modification is required in the convolutional feature extraction process. We only start to build a rotation equivalent representation from high-level, which backward guides the learning process of the entire framework.

Specifically, our RE-CNN scheme introduces the classic multiple instance learning (MIL) formulation (Maron and Ratan, 1998). By describing each scene as a bag and each image patch in the scene as an instance, the relation between high-level feature maps and the scene scheme is built. The key instances in determining the scene scheme are highlighted regardless of their orientation. Also, the permutation-invariant nature of the MIL aggregation function (Zaheer et al., 2017) allows the scene scheme prediction to be invariant to the position change of image patches caused by rotation.

1.3. Contribution

Our contributions can be summarized as follows:

- We propose a rotation equivalent CNN (RE-CNN) scheme. To the best of our knowledge, it is the first work to learn rotation invariant deep features from a top-down perspective reducing the negative influence from rotation-sensitive convolution operations in existing bottom-up pipelines. More importantly, it can be easily adapted to existing CNN backbones in a plug-and-play manner.
- We propose a rotation equivalent scene scheme learning strategy by adapting the classic MIL formulation. It allows the scene scheme to be invariant to the change of instance positions. It is realized by our key instance selection (KIS), key instance aggregation (KIA) and semantic fusion (SF) modules.
- Our proposed RE-CNN substantively improves the recognition performance of rotation variant scenes, *i.e.*, up to 8.95% with only a 0.47% parameter number increase and a 1.78% prediction time increase. Extensive experiments demonstrate that our approach outperforms 24 state-of-the-art approaches on four recognition domains.

The remainder of this paper is organized as follows. Section 2 provides a detailed summary of the related work. Section 3 offers more background on multiple instance learning for a better understanding of our technical insight. Then, in Section 4, our proposed RE-CNN is introduced in detail. Section 5 reports and discusses the extensive experiments and ablation studies. Finally, the conclusion is drawn in Section 6.

2. Related work

2.1. Rotation variant scene recognition

Rotation variance scenes are common due to either the restriction of view point (*e.g.*, aerial imaging, arbitrary-oriented hand-writing digit recognition and etc.) or the unique orientation distribution (*e.g.*, medical imaging, texture recognition and etc.). Among these tasks, arbitrary-oriented hand-writing digit recognition has been investigated for a relatively long time (Dmitry et al., 2016; Zhang et al., 2017; Worrall et al., 2017; Cohen and Welling, 2016; Zhou et al., 2017). Unfortunately, for other large-scale or real-world scenarios such as aerial, industrial and medical imaging, this challenge has not been well tackled and the recognition capability remains to be boosted.

To be specific, for aerial scenario, as the imaging sensor carried by airplane or satellite is bird-view, the objects are posed in arbitrary orientations in an scene. Recent works of aerial image understanding tend to highlight these key local regions regardless of the orientation (Xia et al., 2017, 2018; Bi et al., 2020b, 2021a; Cheng et al., 2018; Bi et al., 2020a,c; Wang et al., 2021). Although such solutions usually lead to an obvious performance gain compared with the baselines and former works, the rotation invariant scene representation has not been widely discussed in aerial imaging community (Han et al., 2021).

Different from traditional medical imaging dealing with X-ray and ultrasound data where the body and organ is presented in a fixed order, in fundus image the eyeball is circle-shaped, and the pathological regions can be posed in an arbitrary orientation (Ilse et al., 2018; Li et al., 2019; Ghamdi et al., 2019; Diaz-Pinto et al., 2019; Wu et al., 2020). However, as fundus disease recognition is only drawing attention in the past few years, the varied orientation of these fundus pathological regions is still not considered so far in the medical imaging community.

In industrial imaging, texture recognition is a typical task that demands rotation invariant scene representation, as the texture can be posed in arbitrary orientation. Before the deep learning era, texture recognition with rotation-invariant hand-crafted features has been thoroughly investigated (Hanbay et al., 2016; Zhao et al., 2012; Sifre and Mallat, 2013; Schmidt and Roth, 2012; Takacs et al., 2010). However, the generalization capability of these hand-crafted features is still significantly inferior to features learnt by deep learning models (Zhang et al., 2020; Iacovacci and Lacasa, 2020).

To summarize, although the challenge of rotation variance has existed and been investigated for a long time, till now few works in the computer vision community have attempted to tackle the rotation variant challenge in such more complicated real-world large-scale scenarios.

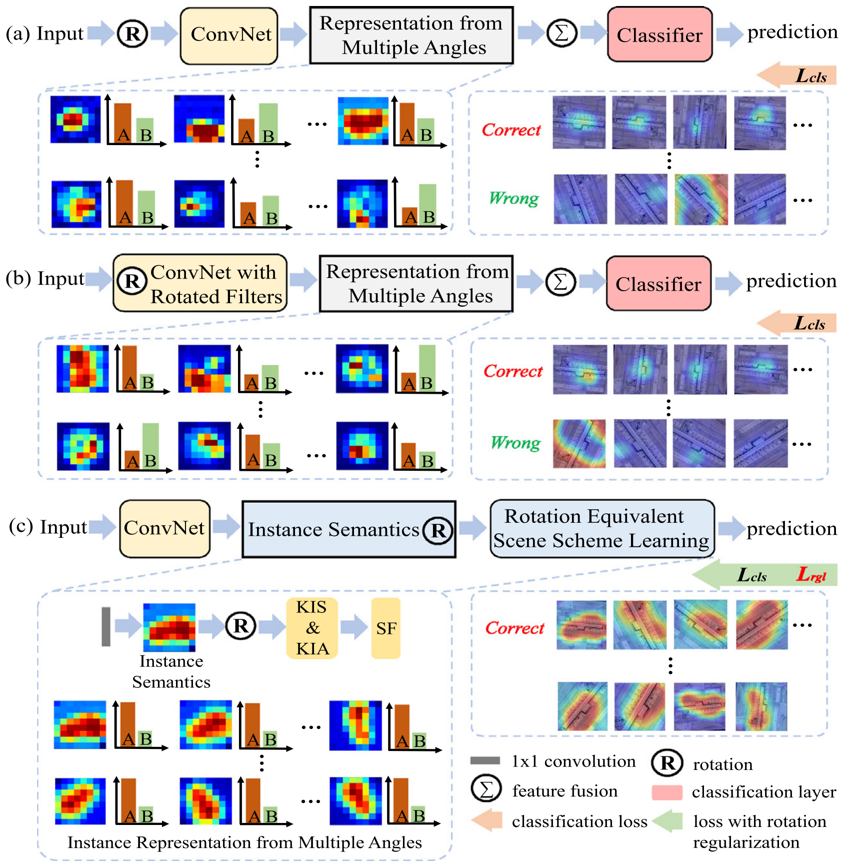


Fig. 2. (a) & (b): Existing rotation invariant scene recognition pipelines in a bottom-up manner by learning from either rotated samples (Zhang et al., 2017; Cheng et al., 2019; Dmitry et al., 2016) or from rotated convolution filters (Cohen and Welling, 2016; Worrall et al., 2017; Zhou et al., 2017; Marcos et al., 2017); (c) Our RE-CNN pipeline in a novel top-down pipeline. KIS: key instance selection module; KIA: Key instance aggregation module; SF: semantic fusion module; A and B denotes the correct and wrong category for intuitive illustration.

2.2. Rotation invariant features & Down-stream tasks

Existing CNN based methods exploited rotation invariant feature representation for a relatively long time. Generally speaking, these approaches can be generally divided into two categories, that is, selecting representative feature responses from rotated samples (Dmitry et al., 2016; Zhang et al., 2017) and using rotated convolution for feature extraction (Zhou et al., 2017; Worrall et al., 2017; Cohen and Welling, 2016; Marcos et al., 2017).

To be specific, Dmitry et al. extracted convolutional features from eight different rotation angles and selected the max point-wise response from these eight representations as the rotation invariant representation (Dmitry et al., 2016). Zhang et al. designed binary filters to generate the convolutional features from different angles, and then conducted a linear combination to generate the final rotation invariant representation (Zhang et al., 2017). On the other hand, to design rotated filters, circular harmonics transformation (Worrall et al., 2017), group theories (Cohen and Welling, 2016), active rotating learning (Zhou et al., 2017) and rotation equivariant vector field (Marcos et al., 2017) have been investigated.

In summary, both strategies are in the bottom-up manner. Their flaw lies in that the rotation insensitive nature of convolution effects negatively on the entire feature extraction process. Thus, it is hard for such methods to activate the RoIs properly when posed in arbitrary orientations, especially in more complicated scenarios.

Therefore attention has been shifted towards down-stream applications such as multi-orientation object detection and segmentation. Instead of learning rotation equivalent representations, the performance of such down-stream detection and segmentation tasks mainly relies on the orientation sensitive region proposal strategies (Ding et al., 2019; Liao et al., 2018; Xu et al., 2020; Jiang et al., 2017; Yang et al., 2021; Han et al., 2021) and rotation-angle-aware loss functions (Cheng et al., 2019; Mou et al., 2019; Yang and Yan, 2020; Qian et al., 2021; Zheng et al., 2020).

2.3. Multiple instance learning

Multiple instance learning is initially designed to deal with weakly-annotated data, as it formulates an object as a bag, and the bag consists of a set of instances which do not have specific labels (Maron and Ratan, 1998; Saad and Mubarak, 2010; Wang et al., 2015, 2013a). Each instance is only labeled as either positive or negative. These weak annotations are utilized to compute the final bag category.

Classic MIL describes the relation between a bag and its instances. It allows us to compute more robust representations and is applied in visual tasks such as image classification (Tang et al., 2017b), object detection (Wang et al., 2012; Tang et al., 2017a), tracking (Babenko et al., 2009) and saliency detection (Zhang et al., 2016; Wang et al., 2013b).

In the past few years, deep multiple instance learning (deep MIL) is drawing increasingly attention. Wang et al. uses the mean and max pooling operation as an instance aggregation function (Wang et al., 2016). Then, gated attention based (Ilse et al., 2018) and channel-spatial attention based (Bi et al., 2020b) deep MIL is studied. The insight of the attention based deep MIL lies in that the aggregation of instance representation is averaged by the attention weights. In this way, the generated bag probability distribution becomes more robust (Ilse et al., 2018; Bi et al., 2020b; Yu et al., 2021) than the mean or max pooling based deep MIL (Wang et al., 2016). Recently, a multi-scale form of deep MIL is proposed (Zhou et al., 2021; Bi et al., 2022).

Using deep MIL into the rotation invariant representation learning is not trivial or straightforward. The rotation sensitive nature of convolutions leads to feature variances caused by different rotation angles. Thus, how to learn a robust instance representation from these varied convolutional features is an important question to be addressed.

3. Preliminary

3.1. Classic MIL formulation

In classic multiple instance learning, an object is formulated as a bag consisting of a set of instances. Assume a bag has label Y, and each instance $\{x_s\}$ of the bag has weakly-annotated labels y_s ($y_s = 0, 1$). The bag label Y is given by

$$Y = \begin{cases} 0 & \text{if} \sum y_s = 0 \\ 1 & \text{else.} \end{cases} \tag{1}$$

3.2. Probability distribution assumption

In classic MIL, the bag probability distribution is binary, *i.e.*, either 0 (false) or 1 (true). In contrast, in deep MIL, the bag probability distribution Y_p is assumed to be continuous in $[0, 1]$ to circumvent the gradient vanishing problem (Ilse et al., 2018).

3.3. Deep MIL for multi-class recognition

We assume that bag label Y belongs to the lth bag category if and only if the bag probability of the lth bag category Y_{p_l} is the maximum among $Y_{p_1}, Y_{p_2}, \dots, Y_{p_l}, \dots, Y_{p_N}$, where N denotes the number of bag categories. This is defined as follows:

$$Y = \begin{cases} 1 & \text{if } Y_{p_l} = \max\{Y_{p_1}, \dots, Y_{p_l}, \dots, Y_{p_N}\} \\ 0 & \text{else.} \end{cases} \tag{2}$$

3.4. MIL aggregation function

In MIL, an aggregation function is needed to bridge the gap between the instance representation and the bag representation. We adopt the instance space paradigm of MIL so that the instance representation can be directly aggregated to the bag probability distribution. The construction of bag-level probability distribution Y_p is a two-step process with transformations f and g given by:

$$Y_p = g(f(\{x_s\})), \tag{3}$$

where f refers to the transformation to instance representation, and g denotes the MIL aggregation function which directly obtains the bag probability Y_p.

3.5. T-equivalent transformation

For a group of transformations T, a function F is T-*equivalent* (Maron et al., 2020; Han et al., 2021) if

$$F(t(x)) = F(x), \tag{4}$$

for all $t \in T$.

3.6. Rotation equivalent bag scheme prediction

For our task, the bag (scene) scheme needs to be invariant to changes caused by the rotation transformation T. From the formulation in Eq. (4), t is a rotation operation with a certain rotation angle in the transformation set T. Our objective is to design such a transformation F to predict the scene scheme invariant to the rotation angle.

3.7. Objective

Eq. (2) shows that the MIL aggregation function needs to meet the aforementioned T-*equivalent* requirement for rotation transformation T. Hence, function F needs to be permutation-invariant (Worrall et al., 2017; Maron et al., 2020).

3.8. Permutation-invariant MIL aggregator

The MIL aggregation function itself tolerates the possible order changes of the instances so that the bag scheme remains unchanged. It has shown that the MIL aggregation function is permutation -invariant (Wang et al., 2016; Ilse et al., 2018; Zaheer et al., 2017), which is beneficial to generate a rotation equivalent bag scheme prediction.

4. Methodology

4.1. Framework overview

Fig. 3 demonstrates the framework of our proposed rotation equivalent convolutional neural network (RE-CNN). Firstly, the convolutional feature maps from the backbone are flattened by a transitional layer and the multi-angle class confident maps (MACCMs) are computed (in Section 4.2). Then, for each CCM from a rotation angle, the key local regions relevant to the scene scheme are selected by our key instance selection (KIS) module (in Section 4.3). Later on, the key instance aggregation (KIA) module fuses these instance representations in a rotation insensitive manner. This ensures that the scene scheme is invariant to a change of instance positions (in Section 4.4). Lastly, our semantic fusion module (in Section 4.5) and the corresponding loss function (in Section 4.6) minimizes the semantic variance from different rotation angles and allows the entire framework to be optimized as a whole.

4.2. Multi-angle class confidence representation

CCMs from multiple rotation angles contain abundant orientation information. CCMs rotated by a certain angle correspond to the sample rotated by the same angle due to the same receptive field of a CNN. Learning from rotated CCMs eliminates the weaknesses of existing bottom-up rotation invariant scene recognition pipelines, which are negatively influenced by the rotation sensitive nature of the convolution operation.

As shown in Fig. 3, a 1 × 1 convolutional layer with weight matrix W_1 and bias matrix b_1, also termed as *transitional layer* in our framework, is used to generate the CCM X_1 from the extracted convolutional feature X. Assume there are N scene categories and $\otimes$ denotes the convolution operator, then X_1 also has N channels, each corresponding to the feature response of a category, and is given by:

$$X_1 = W_1^{(1\times1,N)} \otimes X + b_1^{(1\times1,N)}. \tag{5}$$

Then, the CCM X_1 is rotated by multiple rotation angles θ_i with an interval of $\pi/4$. The set of multi-angle class confidence maps (MACCMs) $\{X_1^{\theta_i}\}$ is defined by:

$$\{X_1^{\theta_i}\} = \{X_1^0, X_1^{\pi/4}, X_1^{\pi/2}, \dots, X_1^{\pi\cdot i/4}, \dots, X_1^{2\pi}\}, \tag{6}$$

where $i = 0, 1, 2, \dots, 8$.

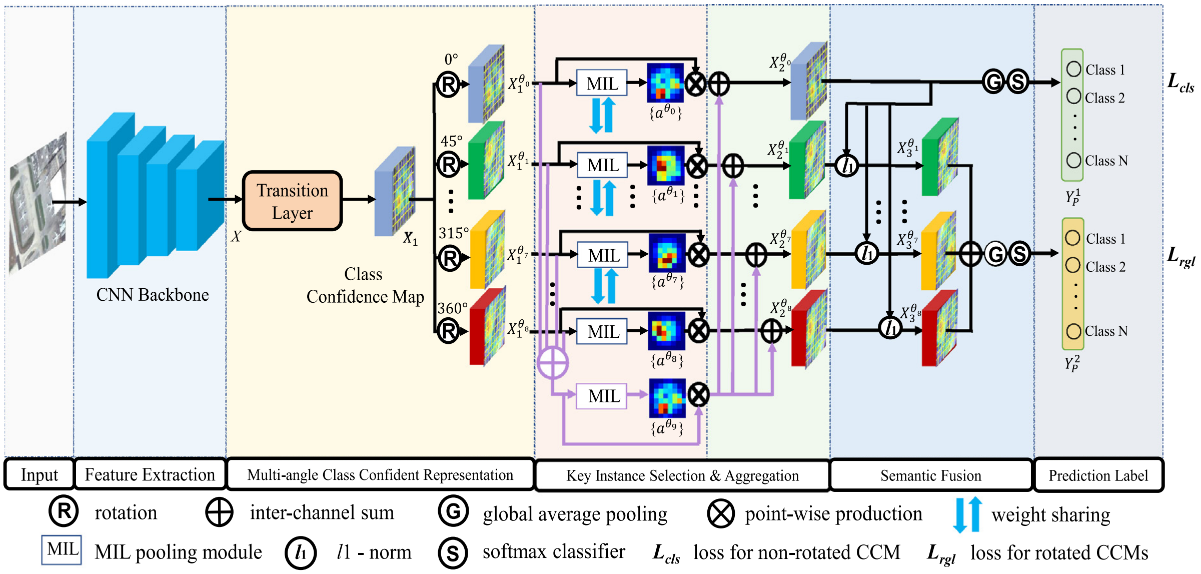


Fig. 3. Framework of our proposed RE-CNN.

4.3. Key instance selection module

To properly activate the key regions regardless of their orientation, as shown in Fig. 3, the key instance selection (KIS) module consists of two branches, *i.e.*, one learns the key instance representation insensitive to rotation, and the other learns representation sensitive to rotation.

The rotation insensitive representation is expected to be robust to a shift of orientation. Following (Maron et al., 2020; Dmitry et al., 2016), the tolerance of rotation is implemented by a weight-sharing feature extraction module with the input from a different rotation. The first branch only has a weight-sharing spatial attention based deep MIL module, aiming to compute an instance spatial weight matrix $\{a_{w,h}^{\theta_i}\}$ for each element in our MACCMs $\{X_1^{\theta_i}\}$.

Overall, it helps to fully exploit the rotation insensitive representation for the scene scheme. This instance weight distribution $\{a_{w,h}^{\theta_i}\}$ provides a description on how each instance contributes to the scene scheme. Higher weights are assigned to instances which are relevant to the scene scheme and vise versa and is given by:

$$\{a_{w,h}^{\theta_i}\} = \mathrm{softmax}(W_2 \otimes X_1^{\theta_i} + b_2), \tag{7}$$

where W_2 and b_2 denote the weight and bias matrix of the 1×1 convolutional layer in this weight-sharing deep MIL module, softmax denotes the softmax function and (w, h) marks the position of a certain instance in the N channel $W \times H$-sized instance representation.

The aim of the second branch is to learn the instance representation sensitive to the rotation. Thus, the rotation features from all orientations need to be included. This is obtained by the sum of the instance representations from each rotation. Specifically, this objective is obtained by a single spatial attention based deep MIL module, which extracts another instance spatial weight matrix $\{\beta_{w,h}\}$. The input of this branch X_1^{sum} is the sum of $X_1^{\theta_i}$, which is calculated as

$$X_1^{sum} = \sum_{i=0}^{8} X_1^{\theta_i}, \tag{8}$$

where $i = 0, 1, \ldots, 8$. The distribution of key regions on X_1^{sum} is more scattered, as the position of many key regions changes due to rotation. Thus, this branch is capable to perceive the rotation sensitive representation while maintaining the scene scheme.

Then, the instance spatial weight matrix $\{\beta_{w,h}\}$, derived from the deep MIL module, is computed by

$$\{\beta_{w,h}\} = \mathrm{softmax}(W_3 \otimes X_1^{sum} + b_3), \tag{9}$$

where W_3 and b_3 are the weight and bias matrix of the 1×1 convolutional layer in this deep MIL module.

4.4. Key instance aggregation module

Before generating the scene probability distribution, it is required to aggregate the above rotation insensitive and sensitive instance representation. The key instance aggregation (KIA) module allows the scene scheme from these aggregated representations invariant to the change of instance positions caused by rotation.

First, the instance weight distribution $\{a_{w,h}^{\theta_i}\}$ from the above weight-sharing deep MIL module has a point-wise product on the instance representations $\{X_1^{\theta_i}\}$ emphasizing the contribution of key instances in determining the scene scheme.

Specifically, assume $1 \le w \le W$, $1 \le h \le H$ and l denotes the lth channel corresponding to the lth scene category for N categories. Also assume $\cdot$ denotes the element-wise production. Then, the feature response of the lth dimension of instance $X'^{\theta_i}_{1,(w,h,l)}$ is accentuated by

$$X'^{\theta_i}_{1,(w,h,l)} = a_{w,h}^{\theta_i} \cdot X^{\theta_i}_{1,(w,h,l)}. \tag{10}$$

Similarly, for the rotation sensitive instance representation $X^{sum}_{1,(w,h,l)}$, the feature response of the lth dimension of instance $X'^{sum}_{1,(w,h,l)}$ is accentuated by

$$X'^{sum}_{1,(w,h,l)} = \beta_{w,h} \cdot X^{sum}_{1,(w,h,l)}. \tag{11}$$

Then, as demonstrated in Fig. 3, the instance representation $X_2^{\theta_i}$ is aggregated over the sum of $X'^{\theta_i}_1$ and the rotation insensitive representation X'^{sum}_1, given by

$$X_2^{\theta_i} = X'^{\theta_i}_1 + X'^{sum}_1. \tag{12}$$

In this way, both the summed rotation sensitive representation X'^{sum}_1 and the rotation insensitive representation $X'^{\theta_i}_1$ from each rotation angle are incorporated by $X_2^{\theta_i}$.

4.5. Semantic fusion module

The aim of this module is two-fold: (1) convert the instance representation to the scene probability distribution in a rotation equivalent manner, and (2) guide the convolution parameter learning process to tolerate rotation variance. In this way, two scene probability distributions, *i.e.*, $\{Y_{p_l}^1\}$ and $\{Y_{p_l}^2\}$, are generated.

Table 1
Summary of four rotation variant scenarios involved in our experiments, including brief descriptions, corresponding benchmarks, evaluation protocols, baselines, sample numbers, scene category numbers and inputted image sizes.

Scenes	Description	Benchmark	Evaluation protocol	Baseline	#number	#category	Input size
Aerial	Bird view	AID (Xia et al., 2017)	50% training, 10 independent runs	ResNet50	10,000	30	256 × 256
Fundus	Arbitrary-oriented pathological regions on cycle-shaped background	LAG (Li et al., 2019)	Five-cross test accuracy	ResNet50	4,850	2	256 × 256
Texture	Arbitrary-oriented	KTD (Kylberg, 2011)	Five-cross test accuracy	ResNet50	4,480	28	256 × 256
Hand-writing digits	Arbitrary-oriented	MNIST-rot (LeCun et al., 1998) MNIST-rot-12k (LeCun et al., 1998)	Five-cross test error	Four-layer CNN	70,000 60,000	10	32×32

For $\{Y^1_{p_l}\}$, only the non-rotated instance representation $X^{\theta_0}_2$ is considered as follows

$$Y^1_{p_l} = \mathrm{softmax}(\sum_{w=1}^{W}\sum_{h=1}^{H} X^{\theta_0}_{2,(w,h,l)}). \tag{13}$$

Note that that this process is the MIL aggregator g in Eq. (3). The MIL aggregator is permutation-invariant and hence the change of instance position caused by rotation does not effect the scene scheme prediction.

On the other hand, $\{Y^2_{p_l}\}$ is the difference between $X^{\theta_i}_2$ $(i = 1, \ldots, 8)$ and $X^{\theta_0}_2$ given by:

$$Y^2_{p_l} = \mathrm{softmax}(\sum_{w=1}^{W}\sum_{h=1}^{H}(\sum_{i=1}^{8} \| X^{\theta_i}_{2,(w,h,l)} - X^{\theta_0}_{2,(w,h,l)} \|_1)), \tag{14}$$

where $\| \cdot \|_1$ denotes the l-1 norm function.

4.6. Semantic fusion loss

The two-branch semantic fusion module matches with a specific loss function L consisting of a classification term $\mathscr{L}_{cls}$ and a rotation regularization term $\mathscr{L}_{rgl}$. The scene probability distribution $\{Y^1_{p_l}\}$ is directly used by the classification loss $\mathscr{L}_{cls}$, which is calculated as

$$\mathscr{L}_{cls} = -\frac{1}{N}\sum_{l=1}^{N}[Y_l \log Y^1_{p_l} + (1 - Y_l)\log(1 - Y^1_{p_l})], \tag{15}$$

where Y_l is the true label of a scene.

$Y^2_{p_l}$ describes the potential difference among $X^{\theta_i}_2$ $(i = 1, \ldots, 8)$. It regularizes the convolutional feature learning process despite the impact of different orientations. This regularization term $\mathscr{L}_{rgl}$ is given by:

$$\mathscr{L}_{rgl} = -\frac{1}{N}\sum_{l=1}^{N}[Y_l \log Y^2_{p_l} + (1 - Y_l)\log(1 - Y^2_{p_l})]. \tag{16}$$

Finally, our semantic fusion loss function $\mathscr{L}$ is the combination of the $\mathscr{L}_{cls}$ and $\mathscr{L}_{rgl}$ terms and calculated as follows:

$$\mathscr{L} = \mathscr{L}_{cls} + \alpha\mathscr{L}_{rgl}, \tag{17}$$

where α is a hyper-parameter to balance the impact of two terms. Empirically, we set $\alpha = 5 \times 10^{-4}$.

4.7. Discussion on rotation invariance

Assume that $X'^{sum}_{1_{\sigma(1)}}, \ldots, X'^{sum}_{1_{\sigma(W\times H)}}$ and $X'^{\theta_i}_{1_{\sigma(1)}}, \ldots, X'^{\theta_i}_{1_{\sigma(W\times H)}}$ is an arrangement of $\{X'^{sum}_{1,(w,h)}\}$ and $\{X'^{\theta_i}_{1,(w,h)}\}$. Then, from instance-level, the rotation $t(\cdot)$ (defined in Eq. (4)) generates an arrangement of $\{X'^{sum}_{1,(w,h)}\}$ and $\{X'^{\theta_i}_{1,(w,h)}\}$, presented as:

$$t : \{X'^{sum}_{1,(w,h)}\} \to \{X'^{sum}_{1_{\sigma(1)}}, \ldots, X'^{sum}_{1_{\sigma(W\times H)}}\}, \tag{18}$$

$$t : \{X'^{\theta_i}_{1,(w,h)}\} \to \{X'^{\theta_i}_{1_{\sigma(1)}}, \ldots, X'^{\theta_i}_{1_{\sigma(W\times H)}}\}. \tag{19}$$

The attention based deep MIL module (Eqs. (7) and (9)) intends to assign a weight for each instance regardless of its spatial position. In other words, we have:

$$\{a^{\theta_i}_{\sigma(w,h)}\} = \mathrm{softmax}(W_2 \otimes t(X^{\theta_i}_1) + b_2), \tag{20}$$

$$\{\beta_{\sigma(w,h)}\} = \mathrm{softmax}(W_3 \otimes t(X^{sum}_1) + b_3). \tag{21}$$

In this case, the sum-wise instance aggregation is invariant to the spatial position change, where we have:

$$\sum_{w=1}^{W}\sum_{h=1}^{H} a^{\theta_i}_{w,h} \cdot X^{\theta_i}_{1,(w,h,l)} = \sum_{w=1}^{W}\sum_{h=1}^{H} a^{\theta_i}_{\sigma(w,h)} X^{\theta_i}_{1_{\sigma(w,h),l}}, \tag{22}$$

$$\sum_{w=1}^{W}\sum_{h=1}^{H} \beta_{w,h} \cdot X^{sum}_{1,(w,h,l)} = \sum_{w=1}^{W}\sum_{h=1}^{H} \beta_{\sigma(w,h)} X^{sum}_{1_{\sigma(w,h),l}}. \tag{23}$$

Eqs. (22) and (23) indicate that the image-level prediction from the instance-level fusion (in Eqs. (12) and (13)) is unchanged despite the order change of instances, which guarantees the rotation invariance.

5. Experiments and analysis

5.1. Dataset

Five rotation variant recognition datasets from four different image domains are used to validate the effectiveness of our RE-CNN framework and summarized in Table 1.

5.1.1. Aerial Image Dataset (AID)

The *bird view* of aerial sensors results in aerial scenes posed *in arbitrary orientations*. AID dataset is a large-scale aerial scene classification benchmark with 30 categories and 10,000 samples in total (Xia et al., 2017).

5.1.2. Large Age Gap (LAG)

Images of glaucoma pathological parts are *at any position and arbitrary orientation* along with the circle-shaped optic disc. LAG, is a newly-released glaucoma recognition benchmark containing 1710 glaucoma and 3140 non-glaucoma samples (Li et al., 2019).

5.1.3. Kylberg Texture Dataset (KTD)

Texture is an important recognition cue in industrial applications and a major challenge for texture recognition is its *arbitrary orientation*. *KTD* is a 28-class texture recognition dataset with 160 samples per class (Kylberg, 2011).

5.1.4. MNIST-rot & MNIST-rot-12k

These are two standard benchmarks to validate rotation robustness (LeCun et al., 1998; Dmitry et al., 2016). MNIST-rot contains 60k training samples and 10k test samples. MNIST-rot-12k is more challenging to validate the generalization ability with only 10k training samples and 50k test samples.

The last two benchmarks (MNIST-rot & MNIST-rot-12k) are traditional small-sized standard benchmarks to validate the rotation invariant representation, while the first three benchmarks (AID, LAG, KTD) are from more challenging real-world large-scale rotation variant scenes.

5.2. Evaluation protocols & Implementation details

5.2.1. Evaluation metrics

Following the former works (Dmitry et al., 2016; Zhang et al., 2017; Zhou et al., 2017; Cohen and Welling, 2016; Worrall et al., 2017), *test error* of five-fold experiments (denoted as err) is reported on *MNIST-rot* and *MNIST-rot-12k*.

On *AID*, the evaluation protocol (Xia et al., 2017) randomly selects 50% samples as the training set and the remaining samples as the test set. The *mean* and *standard deviation* of the *overall accuracy* (denoted as OA) from ten independent runs are reported (Xia et al., 2017). Following the common evaluation protocol, both mean and variance are presented in two-decimal format (Xia et al., 2017).

On *LAG* and *KTD*, the evaluation protocols report on test accuracy (denoted as Acc) from five-fold cross-validation experiments (Li et al., 2019; Kylberg, 2011).

5.2.2. Baseline

On *MNIST-rot* and *MNIST-rot-12k*, our baseline is the same as Dmitry et al. (2016), Zhang et al. (2017), Zhou et al. (2017), Cohen and Welling (2016), Worrall et al. (2017), which is a naive four-layer CNN with multiple 3 × 3 filters. On *AID*, *LAG*, and *KTD*, ResNet-50 (He et al., 2016) serves as the baseline for its wide utilization in the computer vision community.

For the comparison with state-of-the-art methods on each community, following the existing protocols the test samples are not rotated (in Section 5.3). In contrast, to fully evaluate the performance of the former rotation invariant methods, the test samples are rotated under a variety of settings (in Section 5.4).

5.2.3. Hyper-parameter settings

For fair evaluation, the baseline ResNet-50 on AID, LAG and KTD datasets is implemented by ourselves under the same hyper-parameter settings as the proposed RE-CNN. The batch size of all our experiments is set to 64. The Adam optimizer is used. The initial learning rate is 5×10^{-5} and is divided by 10 every 20 epochs. The training process terminates after 60 epochs. To overcome potential over-fitting, L_2 normalization with a relative importance weight of 5×10^{-4} is used. Moreover, the dropout rate is set to 0.2 for all the experiments.

The backbone on *MNIST-rot* and *MNIST-rot-12k* is a naive four-layer CNN. Its performance is directly cited from the corresponding reference. All the hyper-parameter settings of our RE-CNN are the same as Dmitry et al. (2016), Zhang et al. (2017), Zhou et al. (2017), Cohen and Welling (2016), Worrall et al. (2017).

5.2.4. Parameter initialization

For all experiments, except for *MNIST-rot* and *MNIST-rot-12k*, the pre-trained model on ImageNet is used as the initial parameters for the backbone. For the rest parts of our RE-CNN, random initialization is utilized for the weight parameters with a standard deviation of 0.001. All bias parameters are set to zero for initialization. For the experiments on *MNIST-rot* and *MNIST-rot-12k*, the parameter initialization is the same as used in Dmitry et al. (2016), Zhang et al. (2017), Zhou et al. (2017), Cohen and Welling (2016), Worrall et al. (2017).

Table 2
Classification accuracy of our proposed RE-CNN and other approaches on the AID dataset. Results presented in the form of 'average±deviation' from ten independent runs (Xia et al., 2017); Metrics presented in %. The ResNet-50 result is implemented under the same hyper-parameter settings as the RE-CNN. The performance of the state-of-the-art methods is directly cited from the corresponding references.

Method	Publication & Year	Accuracy
ResNet-50 (baseline)	CVPR_2016	91.72 ± 0.17
SPPNet (Han et al., 2017)	RS_2018	91.45 ± 0.38
MSCP (He et al., 2018)	TGRS_2018	94.42 ± 0.17
ARCNet (Wang et al., 2018)	TGRS_2018	93.10 ± 0.55
DCNN (Cheng et al., 2018)	TGRS_2018	96.89 ± 0.10
APNet (Bi et al., 2020c)	GRSL_2020	92.15 ± 0.29
MIDCNet (Bi et al., 2020b)	TIP_2020	92.53 ± 0.18
RANet (Bi et al., 2020a)	NC_2020	92.35 ± 0.19
DSENet (Wang et al., 2021)	TGRS_2021	94.50 ± 0.30
MS2AP (Bi et al., 2021b)	NC_2021	94.82 ± 0.20
DMSMIL (Zhou et al., 2021)	ICASSP_2021	95.65 ± 0.22
LSENet (Bi et al., 2021a)	TIP_2021	94.41 ± 0.16
RE-CNN (ours)	2022	**96.95** ± **0.14**

5.2.5. Development environment

All the experiments are implemented on a workstation with an Intel® Core™ i7-10700K CPU and 64 GB memory. Two GeForce RTX 2080 SUPER GPUs are utilized for acceleration.

5.3. Comparison with rotation variant recognition methods

This subsection reports the performance of our RE-CNN on the three large-scale rotation variant recognition benchmarks (AID, LAG and KTD) and compares it with current rotation variant scene recognition methods.

5.3.1. On AID

The performance of our RE-CNN and other state-of-the-art methods on the AID benchmark is listed in Table 2. It is shown that the proposed RE-CNN outperforms all these methods by a large margin. The close performance of DCNN (Cheng et al., 2018) may be caused by the additional pair-wise supervision used by DCNN, which is stronger than conventional deep learning pipelines and RE-CNN.

As recent work in aerial imaging tends to highlight key regions in an aerial scene regardless of their orientation, these methods are still incapable of providing rotation invariant scene representation. In contrast, our RE-CNN learns the rotation invariant representation from a top-down manner, and thus enhances the model's generalization capability.

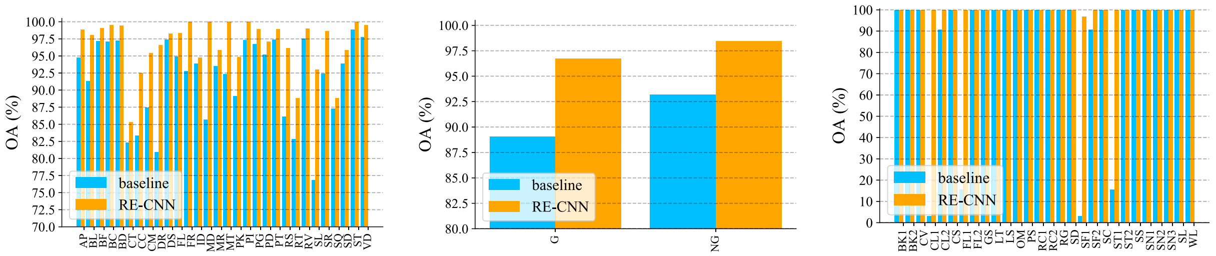
5.3.2. On LAG

Table 3 shows the performance of our RE-CNN and current fundus disease recognition methods on the LAG dataset. It can be derived that our RE-CNN significantly outperforms existing state-of-the-art methods for fundus image disease recognition.

As research in high-resolution fundus image disease recognition only intensified over the past few years, only a few methods consider the rotation variance problem in this domain. Our RE-CNN not only highlights small and tiny pathological regions, but also learns the rotation invariant scene representation.

5.3.3. On KTD

Table 4 lists the performance of our RE-CNN and other latest deep learning based methods on the KTD benchmark. Texture recognition requires rotation invariant features and extensive effort is made before the deep learning era. However, the recognition capability of these hand-crafted features (Hanbay et al., 2016; Zhao et al., 2012; Sifre and Mallat, 2013; Schmidt and Roth, 2012; Takacs et al., 2010) is inferior to deep learning based methods. Therefore, they are not listed in Table 4. It is shown that our RE-CNN outperforms the latest texture recognition methods, as it learns the rotation invariant features from a novel top-down manner.



(a) on AID

(b) on LAG

(c) on KTD

Fig. 4. Per-category classification accuracy of the baseline and RE-CNN on AID (a), LAG (b) and KTD (c) benchmarks respectively. Metrics presented in %. In (a): AP—airport; BL—bare land; BF—baseball field; BC—beach; BD—bridge; CT—center; CC—church; CM—commercial; DR—dense residential; DS—desert; FL—farmland; FR—forest; ID—industrial; MD—meadow; MR—medium residential; MT—mountain; PK—park; PI—parking; PG—playground; PD—pond; PT—port; RS—railway station; RT—resort; RV—river; SL—school; SR—sparse residential; SQ—square; SD—stadium; ST—storage tanks; VD—viaduct. In (b): NG: no glaucoma; G: glaucoma. In (c): BK1—blanket1; BK2—blanket2; CV—canvas; CL1—ceilings1; CL2—ceilings2; CS—cushion; FL1—floor1; FL2—floor2; GS—grass; LT—lentils; LS—linseeds; OM—oatmeals; PS—pearlsugar; RC1—rice1; RC2—rice2; RG—rug; SD—sand; SF1—scarf1; SF2—scarf2; SC—screen; ST1—seat1; ST2—seat2; SS—sesameseeds; SN1—stone1; SN2—stone2; SN3—stone3; SL—stoneslab; WL—wall.

Table 3
Classification accuracy of our proposed RE-CNN and other approaches on the LAG dataset. Results presented are five-cross test accuracy (Li et al., 2019); Metrics presented in %. The ResNet-50 result is implemented under the same hyper-parameter settings as the RE-CNN. The performance of the state-of-the-art methods is directly cited from the corresponding references.

Method	Publication & Year	Accuracy
ResNet-50 (baseline)	CVPR_2016	91.75
DAENet (Fu et al., 2018)	TMI_2018	93.88
MSCNN (Li et al., 2019)	CVPR_2019	92.20
semiCNN (Ghamdi et al., 2019)	ICASSP_2019	95.01
AG-CNN (Li et al., 2019)	CVPR_2019	95.30
CGAN (Diaz-Pinto et al., 2019)	TMI_2019	93.77
L2T-KT (Wu et al., 2020)	MICCAI_2020	96.04
RE-CNN (ours)	2022	**97.98**

Table 4
Classification accuracy of our proposed RE-CNN and other approaches on the KTD dataset. Results presented are five-cross test accuracy (Kylberg, 2011); Metrics presented in %. The ResNet-50 result is implemented under the same hyper-parameter settings as the RE-CNN. The performance of the state-of-the-art methods is directly cited from the corresponding references.

Method	Publication & Year	Accuracy
ResNet-50 (baseline)	CVPR_2016	91.75
TCNN (Andrearczyk and Whelan, 2016)	PRL_2017	96.70
CoHOG (Hanbay et al., 2016)	NC_2017	98.02
Deep LBP (Fernandes and Cardoso, 2017)	arXiv_2018	96.81
IVG-LD (Iacovacci and Lacasa, 2020)	TPAMI_2020	95.80
KGW (Zhang et al., 2020)	TPAMI_2020	93.10
M2-CNN (Aggarwal and Kumar, 2021)	MTA_2021	96.36
RE-CNN (ours)	2022	**99.95**

5.3.4. Per-category classification accuracy

Fig. 4 lists the per-category classification accuracy of the baseline and our RE-CNN on the AID, LAG and KTD benchmarks respectively. It can be derived that by learning a rotation invariant feature representation from these scenes, the per-category recognition performance is significantly increased when compared to the CNN baseline.

5.3.5. Visualization

Fig. 5 shows a number of samples from the three large-scale benchmarks. The key instances and the key regions related to the scene scheme have higher feature responses after they are processed by our RE-CNN, regardless of their orientation. This may be one of the reasons for its superior performance. Also, the interpretable feature maps indicate that the representation learnt by our pipeline has the potential to be transferred into the down-stream detection and segmentation tasks for more rotation robust feature representation.

To understand the impact of RE-CNN on the low-level convolution features, Fig. 6 provides visualized low-level features from the first block of the ResNet-50 backbone. The cases with and without the proposed top-down rotation invariant learning scheme on the AID benchmark are provided. The low-level convolutional feature maps are resized and overlaid to the samples for clarity. Without the proposed scheme, the generic convolutional features tend to be randomly scattered over the entire image. In contrast, with the proposed scheme, the low-level convolutional features tend to highlight the corner or edge of the key objects in the scene, which contain more abundant rotation information. This observation may explain the performance gain from 91.72% (baseline) to 96.95% (RE-CNN), as the rotation information is important to understand the rotation variant scenes.

5.4. Comparison with rotation invariant methods

This subsections compares and discusses the performance of our top-down RE-CNN and existing bottom-up rotation invariant scene representation learning methods, namely, TI-pooling (Dmitry et al., 2016), RILBCNN (Zhang et al., 2017), ORN (Zhou et al., 2017), H-Net (Worrall et al., 2017), RotEqNet (Marcos et al., 2017) and P4CNN (Cohen and Welling, 2016). Moreover, the performance of baselines (for details please refer to Table 1) and two commonly-used data augmentation approaches (rotating samples to 45, 90 and 135 degrees, denoted as *four-angle augmt.*; rotating each sample to a random angle, denoted as *random augmt.*) is also reported for reference.

Note that: (1) Existing recognition methods that theoretically generate a rotation invariant representation are only validated on small-sized standard benchmarks *MNIST-rot* and *MNIST-rot-12k*; (2) The performance on these two benchmarks is saturated. Hence, for fair comparison, (1) On three large-scale recognition benchmarks (AID, LAG and KTD), the results of the above methods are re-implemented with default settings and are under the same ResNet-50 backbone; (2) On *MNIST-rot* and *MNIST-rot-12k*, the same baseline as former works (Dmitry et al., 2016; Zhang et al., 2017; Zhou et al., 2017; Worrall et al., 2017; Marcos et al., 2017; Cohen and Welling, 2016) is adopted.

5.4.1. On overall accuracy

Table 5 lists all the experimental results of these rotation invariant scene representation methods. Some interesting observations can be found.

- Our RE-CNN leads to a performance gain on all these five benchmarks, indicating its effectiveness when applied to multiple image domains especially large-scale scenes and compared to existing rotation invariant recognition methods in a bottom-up learning manner.

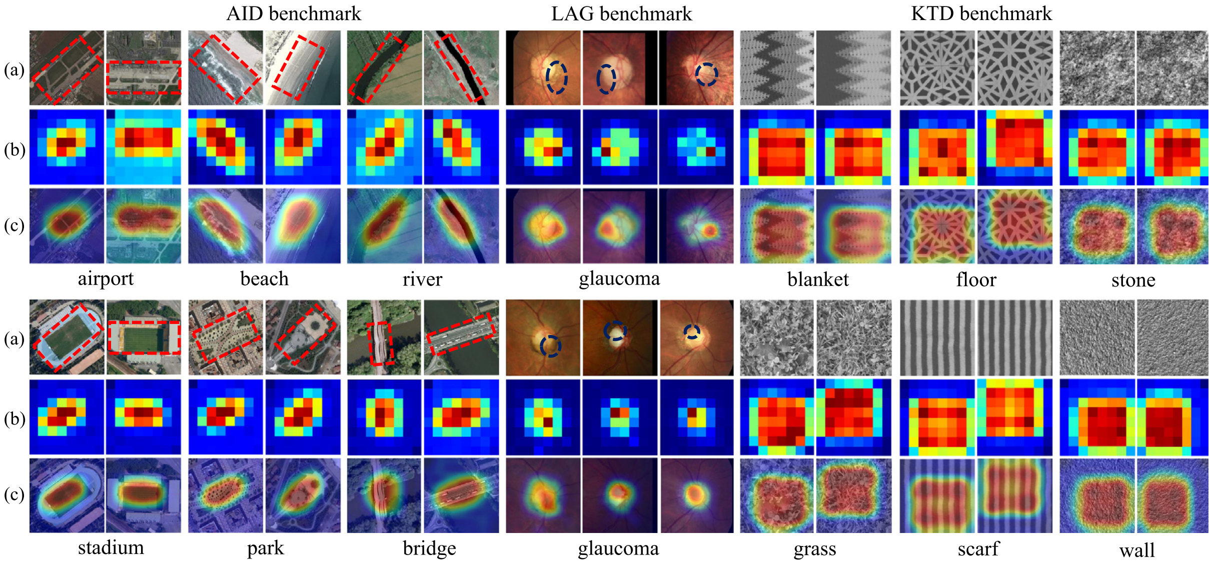


Fig. 5. Visualized samples from our RE-CNN. (a) samples from aerial, medical and industrial scenes; (b) instance-level semantic response; (c) heatmap based on instance response.

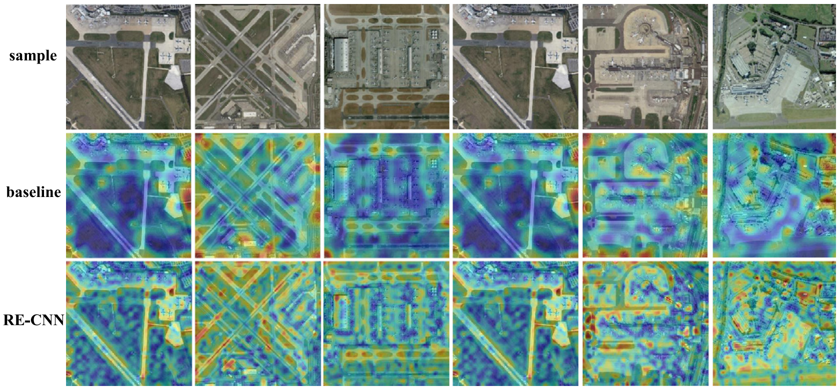
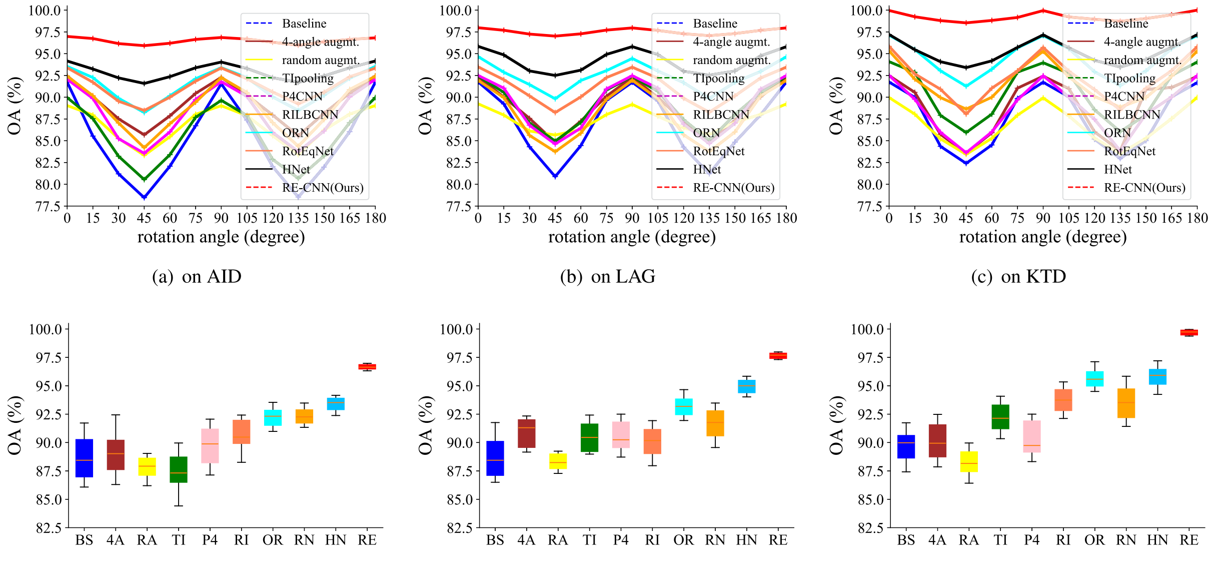


Fig. 6. Comparison of low-level features from the backbone without (denoted as baseline) and with (denoted as RE-CNN) the proposed top-down rotation invariant learning scheme. The RE-CNN enforces the low-level convolution features to focus more on the edges and corners of the key objects, containing more abundant rotation information.

Table 5
Performance comparison of our RE-CNN and other rotation invariant approaches on two standard small-scale benchmarks and three large-scale visual recognition tasks (test err: test error required by LeCun et al., 1998; OA: Overall accuracy of ten independent runs required by *AID* Xia et al., 2017; Acc: classification accuracy of five-fold cross validation required by *LAG* Li et al., 2019 and *KTD* Kylberg, 2011; Metrics presented in %). Baseline: For *MNIST-rot* and *MNIST-rot-12k*, the baseline is a four-layer CNN following (Dmitry et al., 2016; Zhou et al., 2017; Zhang et al., 2017; Cohen and Welling, 2016; Worrall et al., 2017) and the results are directly cited from the corresponding references; For *AID*, *LAG* and *KTD*, the baseline is ResNet50 implemented under the same hyper-parameter settings as the RE-CNN; '–' denotes not reported.

	MNIST-rot	MNIST-rot-12k	AID	LAG	KTD
	Test err	Test err	OA	Acc	Acc
Baseline	2.82	4.34	91.72 ± 0.17	91.75	91.74
Four-angle augmnt.	–	–	92.56 ± 0.15	92.44	92.48
Random augmnt.	–	–	88.75 ± 0.23	89.04	89.96
TI-pooling (Dmitry et al., 2016)	–	2.20	89.95 ± 0.16	92.47	94.08
P4CNN (Cohen and Welling, 2016)	2.28	–	92.06 ± 0.18	92.50	92.50
RILBCNN (Zhang et al., 2017)	1.85	3.05	92.42 ± 0.11	91.93	95.33
ORN (Zhou et al., 2017)	1.42	2.25	93.55 ± 0.17	94.65	97.12
H-Net (Worrall et al., 2017)	1.69	–	94.16 ± 0.19	95.74	97.21
RotEqNet (Marcos et al., 2017)	1.58	2.31	93.48 ± 0.21	93.48	95.82
RE-CNN (ours)	**1.31**	**1.92**	**96.95 ± 0.14**	**97.98**	**99.95**



(a) on AID (b) on LAG (c) on KTD

(d) on AID (e) on LAG (f) on KTD

Fig. 7. Classification performance comparison between two commonly-used rotation based data augmentation strategies and other rotation invariant scene recognition methods when all the test samples are rotated by a specific angle ((a), (b) and (c)) and by a random angle ((d), (e) and (f)). In (a), (b) and (c), *four-angle augmnt.* and *random augmnt.* denotes the rotation based data augmentation when samples are rotated by 0, 45, 90 or 135 degrees and rotated by random angles respectively. In (d), (e) and (f), BS: baseline, 4A: four-angle rotation based data augmentation, RA: random rotation based data augmentation, TI: TI-pooling (Dmitry et al., 2016), P4: P4CNN (Cohen and Welling, 2016), RI: RILBCNN (Zhang et al., 2017), OR: ORN (Zhou et al., 2017), RN: RotEqNet (Marcos et al., 2017); HN: HNet (Worrall et al., 2017); RE: RE-CNN.

- Both our top-down pipeline and existing bottom-up pipelines outperform the commonly-used rotation based data augmentation strategies in learning rotation invariant representation. Moreover, four-angle augmentation can slightly improve the overall recognition capability, but random rotation augmentation decreases the overall recognition capability.

The reason is that the convolution operation is sensitive to rotation. Hence, the feature representation from different angles can vary. Sometimes it is difficult for existing bottom-up methods to learn a more robust rotation invariant representation. The case of random rotation based data augmentation is also similar, as the features from a variety of rotation angles vary too much and the overall recognition capability declines.

To show how the proposed top-down rotation invariant learning scheme outperforms the existing bottom-up schemes, a visualization is given in Fig. 8. The proposed RE-CNN learns the rotation invariant representation from the instance-level representation. Thus, the last-layer feature maps from the other six bottom-up schemes are averaged and normalized for comparison. It is shown that although all six bottom-up methods provide different responses to different rotation angles, five out of six methods have a scattered activation over the image. They do not properly activate the key local regions in the image. In contrast, the proposed RE-CNN not only has different response to different rotation angles, but also activates the key local regions properly despite the rotation.

However, directly investigating the overall classification performance is still not sufficient to fully evaluate a model's capability of learning rotation invariant representations. Hence, the following two subsections consider the model's performance when all the test samples are rotated by a specific angle and by random angles.

5.4.2. On specific rotation angle

Fig. 7(a), (b) and (c) demonstrate the recognition performance variation when all test samples from AID, LAG and KTD are rotated by a specific angle. For full testing, the rotation range is $[0, \pi]$ with an interval of $\pi/12$.

It is shown that our RE-CNN not only outperforms existing rotation invariant recognition approaches on every specific rotation angle, but also demonstrates a more stable performance on all rotation angles. As the backbone is kept the same, the effectiveness of our RE-CNN may is provided by the novel top-down rotation invariant representation learning scheme. It bypasses the feature variance caused by the convolution operation in the feature extraction stage.

5.4.3. On random rotation angle

Fig. 7(d), (e) and (f) demonstrate the recognition performance fluctuation when each sample from AID, LAG and KTD is rotated by a random angle. In this way, every sample contains a different orientation. To provide results that are statistically significant and representative, such observations are based on 20 independent runs.

It is shown in Fig. 7 that our RE-CNN has the least fluctuation among these rotation invariant recognition approaches. This observation is also not difficult to explain, as the flaw of existing bottom-up rotation invariant recognition approaches is obvious. The difference of feature representation from multiple rotation angles exists and accumulates in the feature extraction process, and it lows down the generalization capability. In contrast, our top-down scheme bypasses this problem, and thus has stronger generalization capability.

5.5. Ablation studies

Our RE-CNN consists of a backbone, a transitional layer (TL), a key instance selection (KIS) module, a key instance aggregation (KIA) module and a semantic fusion (SF) module. The ablation studies investigate the performance gain of each component and all results are listed in Table 6. Note that, for fair comparison, in all the cases without SF, the scene probability distribution is directly generated from a global average pooling followed by a softmax function, and the cross-entropy loss function is utilized.

5.5.1. Effect of TL

The experiments on both AID and LAG benchmarks indicate that simply using KL to generate MACCMs only slightly improves the classification performance. This observation also demonstrates that more advanced solutions are needed to solve the rotation invariant problem instead of simply rotating samples as augmentation.

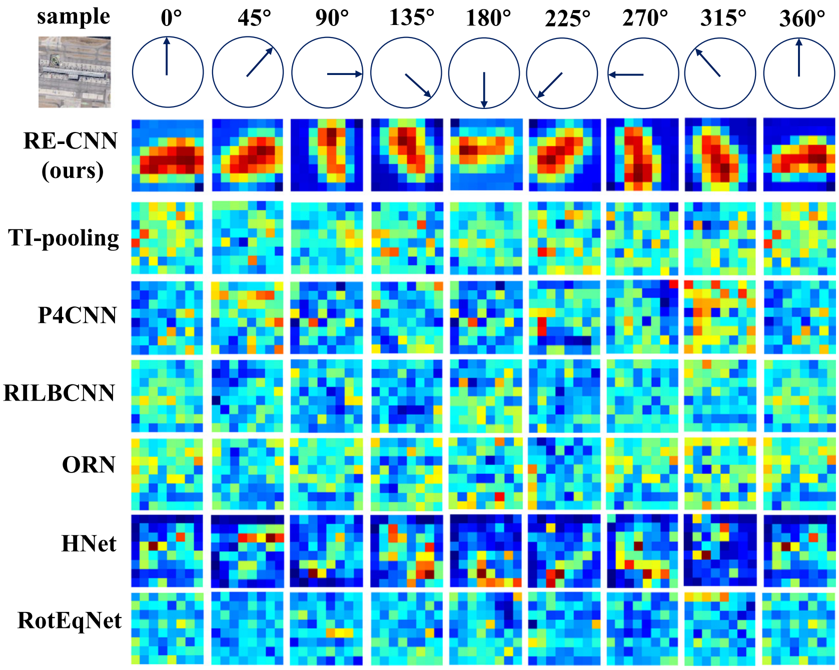


Fig. 8. Feature maps from different rotation angles learnt by the proposed top-down RE-CNN and the other six bottom-up rotation invariant learning methods. The instance-level top-down scheme is more effective to highlight the key local regions regardless of the rotation angle.

Table 6
Ablation study of our RE-CNN on the AID and LAG dataset (OA: Overall Accuracy required by Xia et al., 2017; Acc: Five-fold test accuracy required by Li et al., 2019; Metrics presented in %; ResNet: Backbone ResNet-50; TL: Transitional layer for MACCMs; KIS: Key instance selection module; KIA: Key instance aggregation module; SF: semantic fusion module).

Module					AID	LAG
ResNet	TL	KIS	KIA	SF	OA	Acc
✓					91.72 ± 0.17	91.75
✓	✓				92.05 ± 0.22	92.56
✓	✓	✓			93.53 ± 0.13	94.45
✓	✓	✓	✓		95.39 ± 0.10	96.34
✓	✓			✓	93.43 ± 0.15	94.21
✓	✓	✓		✓	95.26 ± 0.17	96.19
✓	✓	✓	✓	w.o L_{rgl}	96.30 ± 0.19	97.23
✓	✓	✓	✓	✓	**96.95 ± 0.14**	**97.98**

5.5.2. Effect of KIS

Two comparison pairs on AID indicate that the utilization of KIS leads to a performance gain of 1.48% and 1.83% respectively. Similarly, the improvement on LAG is 1.89% and 1.98% respectively. The effectiveness of KIS may be explained that our deep MIL module stresses the region of interest (RoI) in an scene regardless of the rotation angle. Hence, the representation can be more insensitive to the changes caused by rotation.

5.5.3. Effect of KIA

Two comparison pairs on AID and LAG demonstrate that the performance gain of KIA are 1.86%, 1.69% and 1.89%, 1.79% respectively. The function of KIA is important as it aggregates both the rotation sensitive and insensitive representation from KIS in a rotation equivalent manner. Bear in mind that the permutation invariant nature of MIL aggregation function allows the scene representation invariant to the position change of instances caused by rotation.

For an intuitive understanding, some instance representations from different rotation angles are visualized in Fig. 9 (I), where the key instances are activated properly regardless of the orientation. Also, some heatmaps processed by either only KIS or by both KIS and KIA are displayed in Fig. 9 (II), reflecting our KIA helps activate the RoIs more accurately.

5.5.4. Effect of SF

Among three comparison pairs on either using or not using our SF module, the performance gain on AID dataset varies from 1.38% to 1.73%. Similarly, the performance gain on LAG dataset varies from 1.64% to 1.74%. Generally speaking, our SF can not only stress the contribution of key instances but also regularize the entire learning process to be rotation tolerable.

5.5.5. Effect of regularization loss

The loss function of the proposed RE-CNN is a combination of the conventional classification loss term and the regularization loss (Eq. (15)). The impact of the regularization loss, which is based on the difference between feature representations from different rotation angles, is also investigated. When the RE-CNN framework only has the classification loss, the performance on AID and LAG declines 0.65% and 0.75% respectively. The regularization loss helps the representation from different rotation angles to align to the same semantic label of the scene, and thus can benefit the model's robustness to some extent.

5.6. Generalization capability test

To validate the generalization capability of our RE-CNN scheme, we report its performance when embedded into three conventional

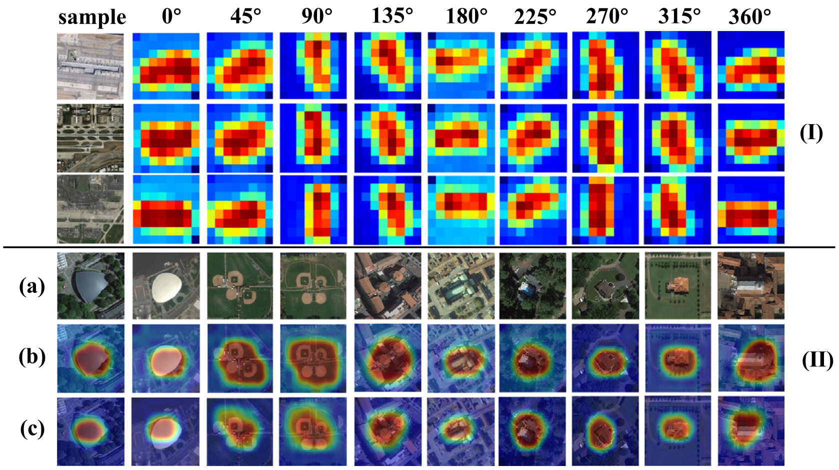


Fig. 9. (I) Instance-level semantics from multiple rotation angles are activated properly. (II) Samples (a) and the corresponding heat maps when processed only by the KIS module (b) or by both the KIS and KIA modules (c).

Table 7
Performance of our RE-CNN on different backbones on the AID dataset (OA: Overall Accuracy required by Xia et al., 2017; Metric presented in %; Para. num.: Parameter numbers; presented in million; FPS: Frame Per Second.).

	OA	Para. num.	FPS
VGG	90.64 ± 0.14	15.43	245.82
Ours with VGG	**95.02** ± **0.13**	15.52	232.41
	↑ 4.83%	↓ 0.58%	↓ 5.77%
ResNet	91.72 ± 0.17	23.46	422.30
Ours with ResNet	**96.95** ± **0.14**	23.57	414.94
	↑ 5.70%	↓ 0.47%	↓ 1.74%
Inception	91.40 ± 0.19	6.64	704.22
Ours with Inception	**95.67** ± **0.17**	6.65	700.28
	↑ 4.67%	↓ 0.15%	↓ 0.56%
Swin-T	98.30 ± 0.04	27.63	213.47
Ours with Swin-T	**99.23** ± **0.18**	27.75	198.72
	↑ 0.95%	↓ 0.43%	↓ 6.91%
ViTAEv2	98.22 ± 0.09	18.82	336.12
Ours with ViTAEv2	**99.21** ± **0.23**	18.91	314.66
	↑ 1.01%	↓ 0.48%	↓ 6.38%

CNN backbones, namely, VGGNet-16 (Simonyan and Zisserman, 2015), ResNet-50 (He et al., 2016), Inception-V2 (Szegedy et al., 2015), and two latest backbones, namely, Swin-T (Liu et al., 2021) and ViTAEv2 (Wang et al., 2022) (denoted as VGG, ResNet, Inception, Swin-T and ViTAEv2 in Table 7) on the AID benchmark. Apart from the required overall accuracy metric (Xia et al., 2017), the parameter number and the frame per second are also reported to evaluate its impact on parameter number and inference time.

From all the outcomes on Table 7, it is clearly seen that our RE-CNN framework significantly boosts the recognition performance on not only the three classic CNN backbones but also the two latest state-of-the-art backbones, only slightly increasing the parameter number and prediction time. Hence, our RE-CNN scheme can be easily adapted to existing CNN backbones with a generic performance boost.

5.7. Sensitivity analysis

5.7.1. Influence of the sampling interval for θ_i

The sampling interval for our θ_i is 45 degrees in our framework by default. It is also interesting to observe the influence of sampling interval on the overall recognition performance. A larger interval leads to a less number of MACCMs while a smaller interval leads to a larger number of MACCMs. To investigate this impact, we test the cases when the sampling interval is 15, 30, 45, 60 and 75 degrees on AID benchmark, while all the other default settings keep the same.

Table 8
Influence of sampling interval (presented in degree) for θ_i on the performance of our RE-CNN on the AID benchmark (OA: Overall Accuracy required by Xia et al., 2017; Metrics presented in %).

Sampling interval (in degree)	OA
15	96.87 ± 0.18
30	96.89 ± 0.16
45	**96.95** ± **0.14**
60	96.90 ± 0.15
75	96.75 ± 0.17

Table 8 lists all the results. It can be seen that when the interval is 15, 30, 45 and 60 leads to an outcome of 96.89%, 96.95% and 96.90%, indicating that there is no significant difference regarding the interval. When the sampling interval is too large, feature responses may be not sufficient for the entire model in the learning phase, and thus the performance shows a slight decline.

Our RE-CNN only utilizes the instance representation rotated by 45, 90, 135, 180, 225, 270, 315 and 360 degrees respectively. But it still demonstrates a stable performance on a variety of rotation angles. The effectiveness lies in two-folds: (1) The permutation-invariant nature of MIL aggregation function allows the scene scheme invariant to the position change of instances caused by rotation. Thus, the specific rotation angle does not influence much on the recognition performance. (2) As the high-level feature maps are often small in sizes (e.g., 8 × 8 in ResNet), the rotated high-level feature maps are not that sensitive to specific rotation angles.

5.7.2. Influence of the hyper-parameter α

Our loss function has two terms L_{cls} and L_{rgl}, which are balanced by a hyper-parameter α. Table 9 shows the impact of α on classification results.

It can be seen that when α is set 5×10^{-3}, 10^{-4}, 10^{-5}, the performance of our RE-CNN is relatively stable. However, when α is either too large or too small, 5×10^{-2} or 5×10^{-6}, the performance of our model degrades. A too-small α indicates that the model does not fully learn the rotation robust features from MACCM, while a too-large α may overwhelm the impact of the original scene representation.

5.7.3. Influence of network initialization

Two network initialization settings, namely, optimization and weight initialization, may have an impact on the classification performance of RE-CNN. Multiple variations of these settings are studied on the KTD dataset (Kylberg, 2011).

Table 9
Influence of hyper-parameter α on the performance of our RE-CNN on the AID benchmark (OA: Overall Accuracy required by Xia et al., 2017; Metrics presented in %).

α value	OA
5×10^{-2}	95.99 ± 0.18
5×10^{-3}	96.62 ± 0.16
5×10^{-4}	**96.95 ± 0.14**
5×10^{-5}	96.74 ± 0.15
5×10^{-6}	96.58 ± 0.17

Table 10
Influence of a different optimization on the performance of our RE-CNN on the KTD benchmark (Acc: Five-fold test accuracy required by Kylberg, 2011; Metrics presented in %).

Optimization	SGD	SGDM	Adam
Acc	99.79	99.84	99.95

Table 11
Influence of different standard deviations of weight initialization on the performance of our RE-CNN on the KTD benchmark (Acc: Five-fold test accuracy required by Kylberg, 2011; Metrics presented in %). The optimizer is fixed as the Adam in all experiments.

Standard deviation	1×10^{-5}	1×10^{-4}	1×10^{-3}	1×10^{-2}
Acc	99.84	99.84	99.95	99.95

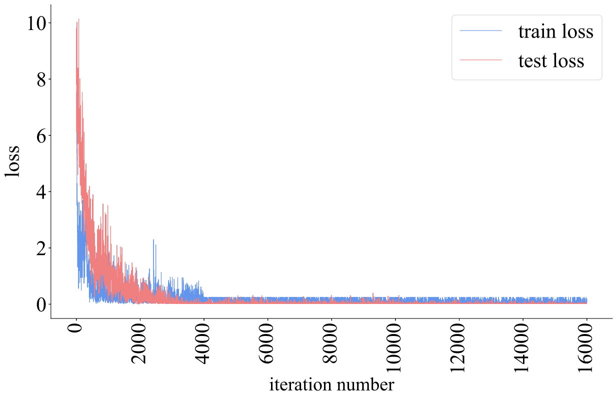


Fig. 10. The curves of training and test losses on KTD (Kylberg, 2011). Both curves become stable after 4000 iterations and do not shown over-fitting.

Table 10 reports the five-fold test accuracy (denoted by Acc) when using stochastic gradient descent (SGD), stochastic gradient descent of momentum (SGDM) and Adam optimizer. Generally, the performance of RE-CNN is not influenced by a different setting of the optimizer.

Table 11 reports the five-fold test accuracy (denoted by Acc) when the standard deviation of weight initialization varies from 1×10^{-5} to 1×10^{-2}. It is shown that the weight initialization has very little influence on the performance of the proposed RE-CNN.

The curves of training and test losses are shown in Fig. 10. After about 4000 iterations, both the training and test losses on KTD dataset become stable. It shows that there is no indication of over-fitting in the learning process.

6. Conclusion

In this paper, we proposed a RE-CNN framework for rotation variant scene recognition. Compared with existing rotation invariant scene recognition methods learning in a bottom-up manner, the RE-CNN scheme uses the classic MIL formulation and learns in a novel top-down manner. It not only eliminates the problem caused by the rotation sensitive nature of convolution operations in the existing bottom-up pipelines, but also accentuates the key regions in a scene regardless of their orientation. Furthermore, by exploiting the permutation-invariant characteristic of the MIL aggregation function, it allows the scene scheme prediction invariant to a position change of instances caused by rotation. Extensive experiments demonstrate that our RE-CNN outperforms 24 representative SOTA approaches on five rotation variant scene benchmarks.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

The authors would like to thank the editors and the anonymous reviewers for their insightful suggestions, which significantly improve the quality of our manuscript. This paper is supported by University of Amsterdam with Contract No. 000000177430.

References

Aggarwal, A., Kumar, M., 2021. Image surface texture analysis and classification using deep learning. Multimedia Tools Appl. 80, 1289–1309.
Almakady, Y., Mahmoodi, S., Conway, J., Bennett, M., 2020. Rotation invariant features based on three dimensional Gaussian Markov random fields for volumetric texture classification. Comput. Vis. Image Underst. 194, 102931.
Andrearczyk, V., Whelan, P., 2016. Using filter banks in Convolutional Neural Networks for texture classification. Pattern Recognit. Lett. 84, 63–69.
Babenko, B., Yang, M., Belongie, S., 2009. Visual tracking with online multiple instance learning. In: IEEE Conf. Comput. Vis. Pattern Recog. CVPR, pp. 983–990.
Barnard, E., Casasent, D., 1991. Invariance and neural nets. IEEE Trans. Neural Netw. 2 (5), 498–508.
Bi, Q., Qin, K., Zhang, H., Li, Z., Xu, K., 2020a. RADC-Net: A residual attention based convolution network for aerial scene classification. Neurocomputing 377, 345–359.
Bi, Q., Qin, K., Zhang, H., Li, Z., Xu, K., Xia, G.-S., 2020b. A multiple-instance densely-connected ConvNet for aerial scene classification. IEEE Trans. Image Process. 29, 4911–4926.
Bi, Q., Qin, K., Zhang, H., Xia, G.-S., 2021a. Local semantic enhanced ConvNet for aerial scene classification. IEEE Trans. Image Process. 30, 6498–6511.
Bi, Q., Qin, K., Zhang, H., Xie, J., Li, Z., Xu, K., 2020c. APDCNet: Attention pooling-based convolutional neural network for aerial scene classification. IEEE Geosci. Remote Sens. Lett. 17 (9), 1603–1607.
Bi, Q., Zhang, H., Qin, K., 2021b. Multi-scale stacking attention pooling for remote sensing scene classification. Neurocomputing 436, 147–161.
Bi, Q., Zhou, B., Qin, K., Ye, Q., Xia, G.-S., 2022. All grains, one scheme (AGOS): Learning multigrain instance representation for aerial scene classification. IEEE Trans. Geosci. Remote Sens. 60, 1–17.
Bozorgtabar, B., Mahapatra, D., von Tengg-Kobligk, H., Poellinger, A., Ebner, L., Thiran, J., Reyes, M., 2019. Informative sample generation using class aware generative adversarial networks for classification of chest Xrays. Comput. Vis. Image Underst. 184, 57–65.
Chen, J., Hu, J., Li, S., 2021. Learning to locate for fine-grained image recognition. Comput. Vis. Image Underst. 206, 103184.
Cheng, G., Han, J., Zhou, P., Xu, D., 2019. Learning rotation-invariant and Fisher discriminative convolutional neural networks for object detection. IEEE Trans. Image Process. 28 (1), 265–278.
Cheng, G., Yang, C., Yao, X., Lei, G., Han, J., 2018. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs. IEEE Trans. Geosci. Remote Sens. 56 (5), 2811–2821.
Cohen, T., Welling, M., 2016. Group equivariant convolutional networks. In: Int. Conf. Mach. Learn. ICML, pp. 2990–2999.
Diaz-Pinto, A., Colomer, A., Naranjo, V., Morales, S., Xu, Y., Frangi, A., 2019. Retinal image synthesis and semi-supervised learning for glaucoma assessment. IEEE Trans. Med. Imaging 38 (9), 2211–2218.
Ding, J., Xue, N., Long, Y., Xia, G.-S., Lu, Q., 2019. Learning RoI transformer for detecting oriented objects in aerial images. In: IEEE Conf. Comput. Vis. Pattern Recog. CVPR, pp. 2844–2853.
Dmitry, L., Nikolay, S., Joachim, M., P., M., 2016. TI-POOLING: transformation-invariant pooling for feature learning in Convolutional Neural Networks. In: IEEE Conf. Comput. Vis. Pattern Recog. CVPR, pp. 289–297.

Fernandes, K., Cardoso, J., 2017. Deep local binary patterns. arXiv:1711.06597.

Fu, H., Cheng, J., Xu, Y., Zhang, C., Wong, D., Liu, J., Cao, X., 2018. Disc-aware ensemble network for glaucoma screening from fundus image. IEEE Trans. Med. Imaging 37 (11), 2493–2501.

Ghamdi, M., Li, M., M., A.-M., M., S., 2019. Semi-supervised transfer learning for convolutional neural networks for glaucoma detection. In: ICASSP. pp. 3812–3816.

Han, J., Ding, J., Xue, N., Xia, G.-S., 2021. ReDet: A rotation-equivariant detector for aerial object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. CVPR, pp. 2786–2795.

Han, X., Zhong, Y., Cao, L., Zhang, L., 2017. Pre-trained AlexNet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification. Remote Sens. 9 (8), 848.

Hanbay, K., Alpaslan, N., Talu, M., Hanbay, D., 2016. Principal curvatures based rotation invariant algorithms for efficient texture classification. Neurocomputing 199, 77–89.

Hanbay, K., Alpaslan, N., Talu, M., Hanbay, D., Karci, A., Kocamaz, A., 2015. Continuous rotation invariant features for gradient-based texture classification. Comput. Vis. Image Underst. 132, 87–101.

He, N., Fang, L., Li, S., Plaza, A., Plaza, J., 2018. Remote sensing scene classification using multilayer stacked covariance pooling. IEEE Trans. Geosci. Remote Sens. 56 (12), 6899–6910.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. CVPR, pp. 770–778.

Iacovacci, J., Lacasa, L., 2020. Visibility graphs for image processing. IEEE Trans. Pattern Anal. Mach. Intell. 42 (4), 974–987.

Ilse, M., Tomczak, J., Welling, M., 2018. Attention-based deep multiple instance learning. In: Int. Conf. Mach. Learn. ICML, 80, pp. 2127–2136.

Jiang, Y., Zhu, X., Wang, X., Yang, S., Li, W., Wang, H., Fu, P., Luo, Z., 2017. R2CNN: Rotational region CNN for orientation robust scene text detection. arXiv:1706.09579.

Kylberg, G., 2011. The Kylberg Texture Dataset v. 1.0. Technical Report, Centre for Image Analysis, Swedish University of Agricultural Sciences and Uppsala University, URL: http://www.cb.uu.se/~gustaf/texture/.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.

Li, C., Duan, G., Zhong, F., 2015. Rotation invariant texture retrieval considering the scale dependence of Gabor wavelet. IEEE Trans. Image Process. 24 (8), 2344–2354.

Li, L., Xu, M., Wang, X., Jiang, L., Liu, H., 2019. Attention based glaucoma detection: A large-scale database and CNN model. In: IEEE Conf. Comput. Vis. Pattern Recog. CVPR, pp. 10571–10580.

Liao, M., Zhu, Z., Shi, B., Xia, G.-S., Bai, X., 2018. Rotation-sensitive regression for oriented scene text detection. In: IEEE Conf. Comput. Vis. Pattern Recog. CVPR, pp. 5909–5918.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. CVPR, pp. 10012–10022.

Marcos, D., Volpi, M., Komodakis, N., Tuia, D., 2017. Rotation equivariant vector field networks. In: Int. Conf. Comput. Vis. ICCV, pp. 5048–5057.

Maron, H., Litany, O., Chechik, G., Fetaya, E., 2020. On learning sets of symmetric elements. In: Int. Conf. Mach. Learn. ICML, pp. 6734–6744.

Maron, O., Ratan, A., 1998. Multiple-instance learning for natural scene classification. In: Int. Conf. Mach. Learn. ICML, pp. 341–349.

Mou, L., Hua, Y., Zhu, X., 2019. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In: Int. Conf. Comput. Vis. ICCV, pp. 12416–12425.

Qian, W., Yang, X., Peng, S., Yan, J., Guo, Y., 2021. Learning modulated loss for rotated object detection. AAAI 2458–2466.

Quattoni, A., Torralba, A., 2009. Recognizing indoor scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. CVPR, pp. 413–420.

Saad, A., Mubarak, S., 2010. Human action recognition in videos using kinematic features and multiple instance learning. IEEE Trans. Pattern Anal. Mach. Intell. 32 (2), 288–303.

Schmidt, U., Roth, S., 2012. Learning rotation-aware features: From invariant priors to equivariant descriptors. In: IEEE Conf. Comput. Vis. Pattern Recog. CVPR, pp. 2050–2057.

Sifre, L., Mallat, S., 2013. Rotation, scaling and deformation invariant scattering for texture discrimination. In: IEEE Conf. Comput. Vis. Pattern Recog. CVPR, pp. 1233–1240.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: ICLR.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: IEEE Conf. Comput. Vis. Pattern Recog. CVPR, pp. 1–9.

Takacs, G., Chandrasekhar, V., Tsai, S., Chen, D., Grzeszczuk, R., Girod, B., 2010. Unified real-time tracking and recognition with rotation-invariant fast features. In: IEEE Conf. Comput. Vis. Pattern Recog. CVPR, pp. 934–941.

Tang, P., Wang, X., Bai, X., Liu, W., 2017a. Multiple instance detection network with online instance classifier refinement. In: IEEE Conf. Comput. Vis. Pattern Recog. CVPR, pp. 2843–2851.

Tang, P., Wang, X., Feng, B., Liu, W., 2017b. Learning multi-instance deep discriminative patterns for image classification. IEEE Trans. Image Process. 26 (7), 3385–3396.

Wang, Q., Liu, S., Chanussot, J., Li, X., 2018. Scene classification with recurrent attention of VHR remote sensing images. IEEE Trans. Geosci. Remote Sens. 57 (2), 1155–1167.

Wang, Q., Si, L., Zhang, D., 2012. A discriminative data-dependent mixture-model approach for multiple instance learning in image classification. In: Eur. Conf. Comput. Vis. ECCV, pp. 660–673.

Wang, X., Wang, B., Bai, X., Liu, W., Tu, Z., 2013a. Max-margin multiple-instance dictionary learning. In: Int. Conf. Mach. Learn. ICML, pp. 846–854.

Wang, X., Wang, S., Ning, C., Zhou, H., 2021. Enhanced feature pyramid network with deep semantic embedding for remote sensing scene classification. IEEE Trans. Geosci. Remote Sens. 59 (9), 7918–7932.

Wang, X., Yan, Y., Tang, P., Bai, X., Liu, W., 2016. Revisiting multiple instance neural networks. Pattern Recognit. 74, 15–24.

Wang, Q., Yuan, Y., Yan, P., Li, X., 2013b. Saliency detection by multiple-instance learning. IEEE Trans. Cybern. 43 (2), 660–672.

Wang, D., Zhang, J., Du, B., Xia, G.-S., Tao, D., 2022. An empirical study of remote sensing pretraining. IEEE Trans. Geosci. Remote Sens. 1. http://dx.doi.org/10.1109/TGRS.2022.3176603.

Wang, X., Zhu, Z., Yao, C., Bai, X., 2015. Relaxed multiple-instance SVM with application to object discovery. In: Int. Conf. Comput. Vis. ICCV, pp. 1224–1232.

Wheeler, B., Karimi, H., 2021. A semantically driven self-supervised algorithm for detecting anomalies in image sets. Comput. Vis. Image Underst. 213, 103279.

Worrall, D., Garbin, S., Turmukhambetov, D., Brostow, G., 2017. Harmonic networks: Deep translation and rotation equivariance. In: IEEE Conf. Comput. Vis. Pattern Recog. CVPR, pp. 5028–5037.

Wu, J., Yu, S., Chen, W., Ma, K., Fu, R., Liu, H., Di, X., Zheng, Y., 2020. Leveraging undiagnosed data for glaucoma classification with teacher-student learning. In: MICCAI. pp. 731–740.

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Marcello, P., Zhang, L., 2018. DOTA: A large-scale dataset for object detection in aerial images. In: IEEE Conf. Comput. Vis. Pattern Recog. CVPR, pp. 3974–3983.

Xia, G.-S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., 2017. AID: A benchmark dataset for performance evaluation of aerial scene classification. IEEE Trans. Geosci. Remote Sens. 55 (7), 3965–3981.

Xu, Y., Fu, M., Wang, Q., Wang, Y., Chen, K., Xia, G.-S., Bai, X., 2020. Gliding vertex on the horizontal bounding box for multi-oriented object detection. IEEE Trans. Pattern Anal. Mach. Intell. 43 (4), 1452–1459.

Yang, X., Hou, L., Zhou, Y., Wang, W., Yan, J., 2021. Dense label encoding for boundary discontinuity free rotation detection. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 15819–15829.

Yang, X., Yan, J., 2020. Arbitrary-oriented object detection with circular smooth label. In: Eur. Conf. Comput. Vis. (ECCV). pp. 677–694.

Yu, S., Ma, K., Bi, Q., Bian, C., Ning, M., He, N., Li, Y., Liu, H., Zheng, Y., 2021. Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 45–54.

Zaheer, M., Kottur, S., Ravanbhakhsh, S., Poczos, B., Smola, A., 2017. Deep sets. In: Adv. Neural Inform. Process. Syst. (NeurIPS). pp. 3394–3404.

Zhang, X., Liu, L., Xie, Y., Chen, J., Wu, L., Pietikainen, M., 2017. Rotation invariant local binary convolution neural networks. In: Int. Conf. Comput. Vis. ICCV, pp. 1210–1219.

Zhang, D., Meng, D., Han, J., 2016. Co-saliency detection via a self-paced multiple-instance learning framework. IEEE Trans. Pattern Anal. Mach. Intell. 39 (5), 865–878.

Zhang, Z., Wang, M., Nehorai, A., 2020. Optimal transport in reproducing kernel Hilbert spaces: Theory and applications. IEEE Trans. Pattern Anal. Mach. Intell. 42 (7), 1741–1754.

Zhang, J., Zhao, H., Liang, J., 2013. Continuous rotation invariant local descriptors for texton dictionary-based texture classification. Comput. Vis. Image Underst. 117 (1), 56–75.

Zhao, G., Ahonen, T., Matas, J., Pietikainen, M., 2012. Rotation-invariant image and video description with local binary pattern features. IEEE Trans. Image Process. 21 (4), 1465–1477.

Zheng, Z., Zhong, Y., Wang, J., Ma, A., 2020. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In: IEEE Conf. Comput. Vis. Pattern Recog. CVPR, pp. 4096–4105.

Zhou, Y., Ye, Q., Qiu, Q., Jiao, J., 2017. Oriented response networks. In: IEEE Conf. Comput. Vis. Pattern Recog. CVPR, pp. 519–528.

Zhou, B., Yi, J., Bi, Q., 2021. Differential convolution feature guided deep multi-scale multiple instance learning for aerial scene classification. In: ICASSP. pp. 4595–4599.