ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414323

DIFFERENTIAL CONVOLUTION FEATURE GUIDED DEEP MULTI-SCALE MULTIPLE INSTANCE LEARNING FOR AERIAL SCENE CLASSIFICATION

*Beichen Zhou, Jingjun Yi, Qi Bi**

School of Remote Sensing and Information Engineering, Wuhan University, China
corresponding author: q_bi@whu.edu.cn

ABSTRACT

Aerial image classification is challenging for current deep learning models due to the varied geo-spatial object scales and the complicated scene spatial arrangement. Thus, it is necessary to stress the key local feature response from a variety of scales so as to represent discriminative convolutional features. In this paper, we propose a deep multi-scale multiple instance learning (DMSMIL) framework to tackle the above challenges. Firstly, we develop a differential multi-scale dilated convolution feature extractor to exploit the different patterns from different scales. Then, the deep features of each scale are fed into a multiple instance learning module to generate a bag-level probability prediction. Lastly, probability predictions from all the MIL branches are fused to generate the final semantic prediction. Extensive experiments on three widely-utilized aerial scene classification benchmarks demonstrate that our proposed DMSMIL outperforms the state-of-the-art approaches by a large margin.

***Index Terms*—** Deep multi-scale multiple instance learning, differential dilated convolution features, semantic prediction fusion, scene classification, aerial image

1. INTRODUCTION

With more and more available sensors deployed on platforms such as the unmanned aerial vehicles and satellites, aerial images have drawn increasing attention [1] for the applications such as aerial scene classification [2, 3], geo-spatial object detection [4, 5] and land cover mapping [6]. Among these tasks, aerial scene classification deals with the scientific problem on how to build a discriminative feature representation for aerial images, which guides the feature extraction process of other aerial tasks such as object detection and segmentation.

Although deep learning approaches especially convolutional neural networks (CNNs) have boosted the performance of scene classification significantly [4], aerial scene classification remains challenging mainly due to some special characteristics of aerial images when compared with ground images.

One of the major differences between them is the largely varied object sizes in aerial images (See Fig. 1 (a) as an example). Sensors from aerial platforms are often quite different in terms of the spatial resolution and the photography height [1]. However, the wide range of object sizes usually causes more difficulties to interpret the aerial scene [7, 8].

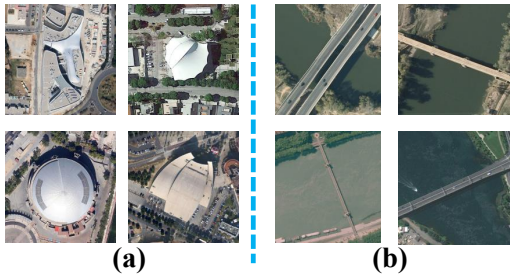

Fig. 1. Some characteristics of aerial scenes different from ground scenes. (a) Buildings are of a variety of sizes; (b) Bridges are posed at arbitrary orientations.

On the other hand, the bird view of aerial images indicates that the objects can be located anywhere in an aerial scene at an arbitrary orientation [9, 1] (See Fig. 1 (b) for an example). In contrast, objects in ground images are up-ward and are often located at the photo center. Thus, how to highlight the key local regions in an aerial scene is the key to improve the recognition performance and still remains challenging[1, 10].

To tackle the above challenges, in this paper, we adopt the concept of multiple instance learning (MIL) under the current deep learning paradigm [1, 11] while taking the above unique characteristics of aerial scenes into account. To be specific, each aerial scene is regarded as a bag and each scene is assumed to consist of a series of instances, i.e., image patch. In this way, whether each image patch belongs to a certain scene category or not can be identified, so that the key local regions of an aerial scene can be perceived more precisely in a fine-grained manner compared with current CNNs.

Moreover, taking the wide range of object sizes in aerial scenes into account, we extend the deep MIL framework into multiple scales. Under each scale, a bag-level (scene-level) semantic prediction is generated directly from the instances, which is a major difference from the currently-existing deep MIL framework [12, 11]. Also, for each scale, the inputted features come from the aforementioned differential dilated

convolutional features, in the hope of utilizing the discriminative features from different scales.

Our contribution can be summarized as below.

(1) We propose a deep multi-scale multiple instance learning (DMSMIL) framework. It can be embedded in current ConvNets in a trainable manner. *To the best of our knowledge*, it is the first deep multi-scale MIL that directly generates bag-level predictions from instances on each scale and fuses them to get the final semantic prediction.

(2) We develop a differential dilated convolutional feature extraction strategy. After extracting a multi-scale dilated convolutional feature representation, the l-1 norm of features from two adjacent scales is calculated, in the hope of exploiting more discriminative feature representation.

(3) The above techniques are combined as a whole to tackle the aforementioned challenges for aerial scene classification. It achieves the state-of-the-art performance on three widely-utilized aerial scene classification benchmarks.

2. RELATED WORK

2.1. Multiple Instance Learning

In multiple instance learning (MIL), each object for classification is regarded as a bag, and each bag consists of a series of instances. If a bag contains at least one positive instance, then the bag is judged as a positive bag. Otherwise, the bag is negative [13]. Since each instance is only labeled as true or false, MIL is qualified to deal with the weakly-annotated data.

Before the development of deep learning approaches, MIL was usually regarded as a kind of classifier after the feature extraction process [14, 15]. After the rapid utilization of deep learning approaches, MIL now has the trend to be combined with CNNs in a trainable manner [12, 16].

To make MIL trainable in deep learning frameworks, recently Ilse et al. assume the bag-level probability distributes as a Bernoulli distribution with the probability $\theta(p) \in [0, 1]$ [11]. However, the challenge remains as in [11] the deep MIL is conducted in the embedding space, where the instance representation fails to generate a bag representation directly. In this case, it needs several fully-connected layers to generate the bag-level (scene-level) probability distribution and the semantic representation capability remains to be enhanced.

2.2. Aerial Scene Classification

Due to the strong feature representation capability, deep learning approaches usually outperform the traditional hand-crafted feature based approaches by a large margin [4]. However, as is discussed in Section 1, the complexity of aerial images indicate that a more discriminative feature representation is needed for deep learning approaches. Thus, recent trends for aerial scene classification usually exploit multi-scale [7] or multi-level [10, 8] convolutional features and enhance the feature response of key local regions [17, 1].

Compared with former works of multi-scale convolutional feature representation, we further exploit the differences of the two convolutional features from adjacent scales and highlight the key local regions strongly relevant to the semantic label to build a more discriminative feature representation.

3. METHODOLOGY

3.1. Framework Overview

Fig. 2 gives a brief illustration of our proposed DMSMIL framework. It firstly extracts the differential dilated convolutional features (Section 3.2) to build a more discriminative multi-scale feature representation. Then, the feature representation is fed into a trainable multi-scale multi-instance learning module (Section 3.3), each scale of which converts the instance representation directly into the bag-level probability distribution. Finally, a classification module (Section 3.4) fuses the bag/semantic probability distribution from different scales for classification.

3.2. Differential Dilated Convolutional Features

Our backbone is the widely-utilized VGGNet-16 in the aerial image community. We derived the convolutional features from the last convolutional layer (denoted as X) to extract the differential convolutional features.

Firstly, we utilize a series of dilated convolution operators $\{d_i(\cdot)\}$ with the same 3×3 window size but different dilated rate r to extract multi-scale dilated convolutional features. Specifically, we empirically adopt four scales with the dilated rate $r = 1, 3, 5, 7$ respectively. Note that, there are 256 channels for each dilated convolution operator.

Then, the convolutional feature from each scale is implemented on a l-1 norm (denoted as $||\cdot||_1$) with the convolutional feature from its adjacent scale. Also, the original convolutional features X is regarded as the 0^{th} scale. To be specific, the four differential dilated convolutional features X_1, X_2, X_3, X_4 are calculated as

$$X_1 = \|d_1(X) - X\|_1, \tag{1}$$

$$X_2 = \|d_2(X) - d_1(X)\|_1, \tag{2}$$

$$X_3 = \|d_3(X) - d_2(X)\|_1, \tag{3}$$

$$X_4 = \|d_4(X) - d_3(X)\|_1. \tag{4}$$

3.3. Multi-scale Multiple Instance Learning Module

Each differential dilated convolutional feature X_i ($i = 1, 2, 3, 4$) is fed into a deep MIL module f_i. In general, f_i converts X_i into a set of bag-level probability distribution Y_i. This process can be presented as

$$Y_i = f_i(X_i). \tag{5}$$

To be specific, this process firstly utilizes a 1×1 convolutional layer whose channel number is the same as the number of bag categories N to convert the convolutional features X into an instance-level feature representation X'. Assume

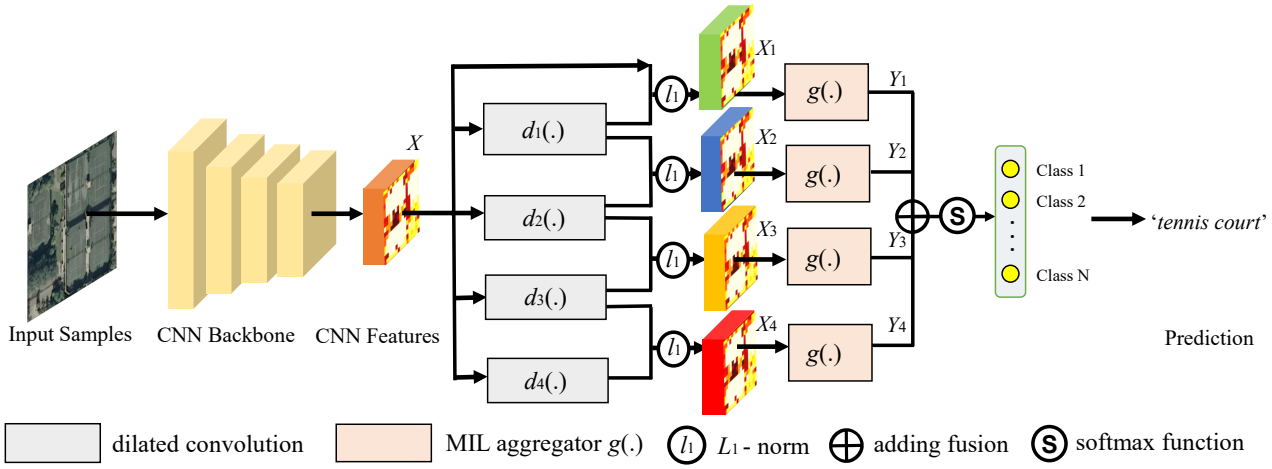

Fig. 2. Demonstration of our proposed deep multi-scale multiple instance learning (DMSMIL) framework.

the width and height of the instance representation is W and H respectively, then X' has a shape of $W \times H \times N$. This indicates that when calculating the bag probability, in each channel $W \times H$ instances are involved. To convert this instance representation into bag representation, an aggregation function $g(\cdot)$ is needed so that the bag-level probability distribution has a shape of $1 \times 1 \times N$, presented as

$$Y_i = g(X'). \tag{6}$$

Since all patches of an aerial scene have the potential to contain key information in determining the scene label (bag label) regardless of their positions, we average all the instance feature responses in each channel to get the corresponding bag-level feature response, which can be regarded as the MIL aggregator $g(\cdot)$ in Eq.(6). From the view of each channel, this process can be presented as

$$Y_{i,k} = \frac{\sum_{w=1}^{W}\sum_{h=1}^{H} X'_{w,h,k}}{WH}, \tag{7}$$

where k denotes the k^{th} channel in Y_i and $k = 1, 2, ..., N$. Note that, for VGG backbone, we have $W = 14$ and $H = 14$.

3.4. Semantic Prediction Fusion

The final bag probability distribution Y is the adding fusion of the bag probability distribution Y_i from each scale. This process can be presented as

$$Y = softmax(\sum_{i=1}^{4} Y_i), \tag{8}$$

where *softmax* denotes the softmax classifier. Finally, the cross-entropy loss function is chosen for optimization.

4. EXPERIMENT AND ANALYSIS

4.1. Dataset, Evaluation Protocols and Settings

Dataset. We validate our approach on three widely-utilized aerial scene classification datasets, that is, UC Merced [18], AID [9] and NWPU [19] respectively.

Evaluation protocols. For all the three datasets, the evaluation protocol is to have 10 independent runs [18, 9, 19] with the below training ratios. For *UCM*, the training ratios are 50% and 80% [18]. For *AID*, the training ratios are 20% and 50% [9]. For *NWPU*, the ratios are 10% and 20% [19].

Parameter settings. The batch size of our DMSMIL is 32, and we choose the Adam optimizer to optimize the entire framework. The learning rate is firstly 5×10^{-5} and is divided by 10 every 20 epochs. The training process does not terminate until 60 epochs are finished. Moreover, to fasten the convergence, we utilize the pre-trained model on ImageNet as our initial parameter values. To overcome the over-fitting problem, we use L_2 normalization with 5×10^{-4} and the dropout rate is set 0.2 in all solutions. For the instance-level classifier, the channel number of 1×1 convolutional layer equals to the scene category number of each dataset. To be specific, it is 21, 30 and 45 for UCM, AID and NWPU dataset.

Development environment. All our experiments were implemented on a 32GB-memory workstation with Intel(R) Xeon(R) E5-2630 v3 CPU and Titan970Ti GPU.

4.2. Comparison with the State-of-the-art Approaches

Table. 1 lists the performance of our DMSMIL and other SOTA approaches on the above three benchmarks under six different experiments.

It can be seen that our approach outperforms all the SOTA approaches and the corresponding baseline models in five out of six experiments with an obvious improvement. In AID 50% experiment, our approach only performs a little bit worse than the D-CNN[22].

Since some current SOTA approaches also utilize the VGG-16 model as the backbone [21, 22, 8], the major reasons for our approach's effectiveness can be attributed to our

Table 1. Comparison of our DMSMIL and other SOTA approaches (Metrics are presented in % and are in the form of *'mean accuracy ± standard deviation'* following the evaluation protocols [18, 9, 19].).

Method	UCM		AID		NWPU	
	50%	80%	20%	50%	10%	20%
AlexNet [9]	93.98±0.67	95.02±0.81	86.86±0.47	89.53±0.31	76.69±0.21	79.85±0.13
VGGNet-16 [9]	94.14±0.69	95.21±1.20	86.59±0.29	89.64±0.36	76.47±0.18	79.79±0.15
GoogLeNet [9]	92.70±0.60	94.31±0.89	83.44±0.40	86.39±0.55	76.19±0.38	78.48±0.26
SPP-Net [7]	94.77±0.46	96.67±0.94	87.44±0.45	91.45±0.38	82.13±0.30	84.64±0.23
MIDC-Net [1]	95.41±0.40	97.40±0.48	88.51±0.41	92.95±0.17	86.12±0.29	87.99±0.18
RA-Net [20]	94.79±0.42	97.05±0.48	88.12±0.43	92.35±0.19	85.72±0.25	87.63±0.28
TEX-Net [21]	94.22±0.50	95.31±0.69	87.32±0.37	90.00±0.33	—-	—-
D-CNN [22]	—-	98.93±0.10	90.82±0.16	**96.89±0.10**	89.22±0.50	91.89±0.22
MSCP [8]	—-	98.36±0.58	91.52±0.21	94.42±0.17	85.33±0.17	88.93±0.14
FV [10]	—-	98.57±0.34	—-	—-	—-	—-
ARCNet [17]	96.81±0.14	99.12±0.40	88.75±0.40	93.10±0.55	—-	—-
DMSMIL (ours)	**99.09±0.36**	**99.45±0.32**	**93.98±0.17**	95.65±0.22	**91.93±0.16**	**93.05±0.14**

differential dilated convolutional features and deep multi-scale multiple instance learning.

4.3. Ablation Study

Our ablation study consists of the following five cases, that is, VGG backbone (denoted as VGG), differential dilated convolutional feature representation from VGG (denoted as VGG+DDC), VGG backbone with current single-scale deep MIL (denoted as VGG+SMIL), differential dilated convolutional feature representation from VGG with deep single-scale MIL (denoted as VGG+DDC+SMIL), and differential dilated features with our deep multi-scale MIL (ours, denoted as VGG+DDC+MSMIL). From Table 2, it can be seen that:

Table 2. Ablation study of our DMSMIL approach on NWPU dataset (Metrics are presented in % and are in the form of *'mean accuracy ± standard deviation'* following the evaluation protocols in [18, 9, 19].).

	50%	80%
VGG	76.69±0.21	79.85±0.13
VGG+DDC	85.36±0.18	88.73±0.17
VGG+SMIL	87.23±0.16	89.42±0.19
VGG+DDC+SMIL	89.02±0.15	91.35±0.18
VGG+DDC+MSMIL (ours)	**91.93±0.16**	**93.05±0.14**

(1) Our approach (VGG+DDC+MSMIL) achieves the best performance, while the benefits of using solely differential dilated convolutional features (VGG+DDC) and solely current single-scale deep MIL (VGG+SMIL) is both obvious when compared with only using the VGG backbone (VGG).

(2) Using deep multi-scale MIL (VGG+DDC+MSMIL) leads to an obvious performance boost compared with using single-scale deep MIL (VGG+DDC+SMIL), and so it is when using differential dilated convolutional features (VGG+DDC and VGG+DDC+SMIL) compared with the traditional convolutional features (VGG and VGG+SMIL).

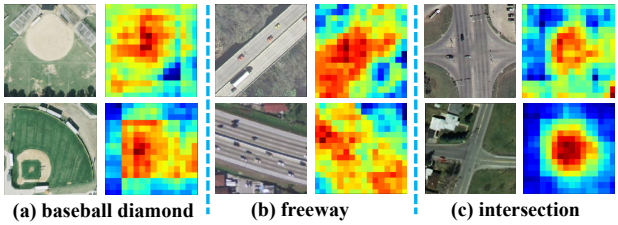

Fig. 3. Visualized feature maps of some samples processed by our deep multi-scale multiple instance learning (DMSMIL) framework.

4.4. Visualized Samples

Fig. 3 offers some visualized feature response maps when processed by our DMSMIL framework, which comes from the adding fusion of X_1, X_2, X_3 and X_4. It can be seen that most of the key local regions in an aerial scene can be properly activated and have much higher feature responses than other regions irrelevant to the scene label. It indicates that our DMSMIL framework could be transferred to the tasks such as aerial image object detection and segmentation.

5. CONCLUSION

In this paper, we propose a deep multi-scale multiple instance learning (DMSMIL) framework for aerial scene recognition, taking the wide range of object sizes and the complicated object distribution into account. It firstly builds a discriminative feature representation from our differential dilated convolutional feature extractor. Then, the features from each scale is fed into a deep MIL module, which directly converts the instance representation into bag-level probability distribution. Finally, the bag probability distributions from all the scales are fused together to get the final prediction. Experiments on three datasets validate the effectiveness of our DMSMIL.

6. REFERENCES

[1] Q. Bi, K. Qin, Z. Li, H. Zhang, K. Xu, and G.-S. Xia, "A multiple-instance densely-connected convnet for aerial scene classification," in *IEEE Trans Image Process.*, 2020, vol. 29, pp. 4911–4926.

[2] Q. Bi, K. Qin, H. Zhang, J. Xie, Z. Li, and K. Xu, "Apdcnet: Attention pooling-based convolutional neural network for aerial scene classification," in *IEEE Geosci. Remote Sens. Lett.*, 2020, vol. 17, pp. 1603–1607.

[3] Q. Bi, H. Zhang, and K. Qin, "Multi-scale stacking attention pooling for remote sensing scene classification," in *Neurocomputing*, 2021, vol. 436, pp. 147–161.

[4] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *IEEE Comput. Vis. Pattern Recognit.*, 2017.

[5] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for detecting oriented objects in aerial images," in *IEEE Comput. Vis. Pattern Recognit.*, 2019.

[6] X. Tong, G. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution rs images using transferable deep models," in *Remote Sens. Environ.*, 2020, vol. 237, p. 111322.

[7] X. Han, Y. Zhong, L. Cao, and L. Zhang, "Pre-trained alexnet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification," in *Remote Sens.*, 2017, vol. 9, p. 848.

[8] N. He, L. Fang, S. Li, A. Plaza, and J. Plaza, "Remote sensing scene classification using multilayer stacked covariance pooling," in *IEEE Trans. Geosci. Remote Sens.*, 2018, vol. 99, pp. 1–12.

[9] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, and L. Zhang, "Aid: A benchmark dataset for performance evaluation of aerial scene classification," in *IEEE Trans. Geosci. Remote Sens.*, 2017, vol. 55, pp. 4911–4926.

[10] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multi-layer features of convolutional neural networks for remote sensing scene classification," in *IEEE Trans. Geosci. Remote Sens.*, 2017, vol. 55, pp. 5653 – 5665.

[11] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Int. Conf. Mach. Learn.*, 2018.

[12] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," in *Pattern Recognit.*, 2016, vol. 74, pp. 15–24.

[13] M. Zhang and Z. Zhou, "Improve multi-instance neural networks through feature selection," in *Neural Process. Lett.*, 2004, vol. 19, pp. 1–10.

[14] P. Tang, X. Wang, B. Feng, and W. Liu, "Learning multi-instance deep discriminative patterns for image classification," in *IEEE Trans Image Process.*, 2017, vol. 26, pp. 3385–3396.

[15] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," in *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016, vol. 5, pp. 865–878.

[16] H. Yang, T. Zhou, J. Cai, and Y. Ong, "Miml-fcn+: Multi-instance multi-label learning via fully convolutional networks with privileged information," in *IEEE Computer. Vis. Pattern Recognit.*, 2017.

[17] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of vhr remote sensing images," in *IEEE Trans. Geosci. Remote Sens.*, 2018, vol. 57, pp. 1155–1167.

[18] Y. Yang and N. Shawn, "Geographic image retrieval using local invariant features," in *IEEE Trans. Geosci. Remote Sens.*, 2013, vol. 51, pp. 818–832.

[19] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," in *Proc. IEEE*, 2017, vol. 10, pp. 1–19.

[20] Q. Bi, K. Qin, H. Zhang, Z. Li, and K. Xu, "Radcnet: A residual attention based convolution network for aerial scene classification," in *Neurocomputing*, 2020, vol. 377, pp. 345–359.

[21] M. HRao, F. Khan, J. Weijer, M. Molinier, and J. Laaksonen, "Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification," in *ISPRS J. Photogramm. Remote Sens.*, 2017, vol. 138, pp. 74–85.

[22] G. Cheng, C. Yang, X. Yao, and J. Guo, L.and Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns," in *IEEE Trans. Geosci. Remote Sens.*, 2018, vol. 56, pp. 2811 – 2821.