DMPCONV: DECOUPLING MULTI-BRANCH POINTWISE CONVOLUTIONS FOR LIGHT-WEIGHT REMOTE SENSING SCENE CLASSIFICATION

Jianlong Hou[1,2,3,4], Zhi Guo[1,2,*], Youming Wu[1,2], Wenhui Diao[1,2], Yingchao Feng[1,2,3,4],

[1]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China
[2]Key Laboratory of Network Information System Technology, Aerospace Information Research Institute,
Chinese Academy of Sciences, Beijing, China - (guozhi, whdiao)@mail.ie.ac.cn, youming_wu@yeah.net
[3]University of Chinese Academy of Sciences, Beijing, China - (houjianlong18, fengyingchao17)@mails.ucas.edu.cn
[4]School of Electronic, Electrical and Communication Engineering,University of Chinese Academy of Sciences, Beijing, China

KEY WORDS: Remote Sensing Scene Classification, Light-weight Model, Inferencing-free, Multi-branch Architectures,Neural Network.

ABSTRACT:

The use of multi-branch architectures in off-the-shelf light-weight residual series neural networks can significantly improve their performance in remote sensing scene classification tasks. However, such architectures come at the expense of an increased number of parameters and calculations. In this paper, we propose the Decoupling Multi-branch Pointwise Convolutions (DMPConv), which works without a corresponding increase in parameters and calculations during inferencing, and at the same time, can maintain the same performance improvement ability as the multi-branch architectures. DMPConv can be decoupled into two states, the training-time DMPConv and the inferencing-time DMPConv. The training-time DMPConv enhances the expressivity of the network by using weighted multi-branch 1×1 convolutions. After training, we use structural reconstruction to convert the training-time DMPConv to the inferencing-time DMPConv, which has the same form as vanilla 1×1 convolution, so as to realize the inferencing-free. Extensive experiments were conducted on multiple remote sensing scene classification benchmarks, including Aerial Image data set and NWPU-RESISC45 data set to demonstrate the superiority of DMPConv.

1. INTRODUCTION

Remote sensing scene classification is the automatic assignment of semantic labels to each item in remote sensing images (Xia et al., 2017). The process has recently received massive attention because it plays a vital role in many applications, such as geographic target detection, land use classification, and urban planning. However, remote sensing images have a wide visual range and a complex spatial structure with high intra-class and low inter-class variability, and due to the continuous increase in the volume of remote sensing image data as well, accurate and efficient scene classification in remote sensing has become a very challenging issue.

Recently, convolutional neural networks (CNNs) have achieved remarkable progress in remote sensing scene classification tasks. This development has enabled emerging technology such as drones, satellite platforms, vehicle platforms, Internet-of-Things (IoT) devices, and robots. However, these devices are often limited in terms of memory storage and computing power, which hinders the actual deployment of top-performing CNNs (Szegedy et al., 2017, He et al., 2016) with highly complex trainable parameters on these resource-constrained platforms.

Regarding the above-mentioned challenge, considerable effort has been put into designing light-weight residual networks for emerging applications. In recent years, a popular basic component named depthwise separable convolution is welcomed to design light-weight residual series models, such as MobileNet (Howard et al., 2017, Sandler et al., 2018) and ShuffleNet (Zhang et al., 2018b). In addition, ShiftResNet (Wu et al., 2018) utilizes shift operations to remove the constraint imposed by depthwise separable convolutions, thus attaining competitive accuracy with fewer parameters. However, when these off-the-shelf light-weight residual series models remain unable to meet our specific needs with regards to accuracy and efficiency, new models need to be redesigned at the cost of numerous man-hours or GPU hours. In recent years, the superior performance of multi-branch architectures have been verified, such as in Multilevel ResNets (Zhang et al., 2018a), which can be incorporated into residual series networks to further improve their accuracy on remote sensing images. However, multi-branch architectures are very expensive in both parameters and calculations, each of which grows multiplies with respect to branch numbers and may exceed the resource limitations of the embedded platforms, thus significantly affecting the efficiency of these models.

In order to alleviate this problem, a novel multi-branch architecture was proposed, where he performance of the off-the-shelf light-weight residual series networks on remote sensing images can be improved without a corresponding increase in parameters and calculations during inferencing. More specifically, a novel multi-branch architecture called Decoupling Multi-branch Pointwise Convolutions (DMPConv) was designed, which can be decoupled into two states of training-time DMPConv and inferencing-time DMPConv. The training-time DMPConv utilizes weighted multi-branch 1×1 convolutions to enhance the performance of the off-the-shelf light-weight residual series networks. The multi-branch architectures focus on different feature subspaces to aggregate more information and diverse features, which is helpful to extract remote sensing image features with rich and diverse backgrounds, the high similarity of different scenes, high resolution, color and texture, and complex and varied imaging methods. After training,

[*] Corresponding author

https://doi.org/10.5194/isprs-annals-V-3-2022-233-2022

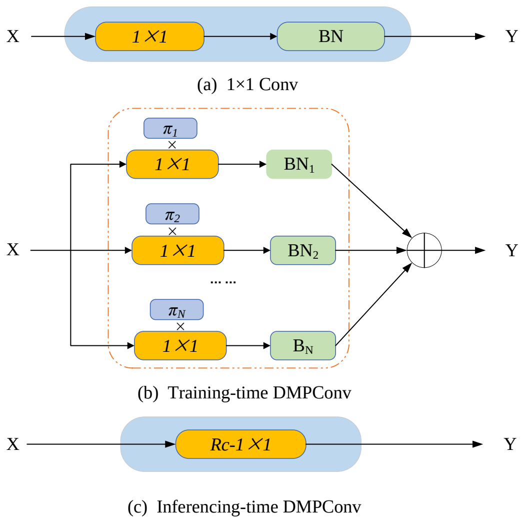


Figure 1. Overview of DMPConv.

the training-time DMPConv is equivalently converted into the corresponding inferencing-time DMPConv via structural reconstruction of the weighted multi-branch 1×1 convolution kernels with batch normalization to 1×1 convolution kernels. Due to the additivity principle of 1×1 convolution kernel fusion and batch normalization (BN) branch fusion, the reconstructed inferencing-time DMPConv does not incur additional computational and parameter burdens during inferencing as opposes to the 1×1 convolution. The general framework of DMPConv is shown in Fig.1. Furthermore, DMPConv can be applied as a drop-in replacement to the residual series blocks of the vanilla 1×1 convolutions (He et al., 2016, Howard et al., 2017, Wu et al., 2018) and 1×1 group convolutions (Zhang et al., 2018b, Ma et al., 2018). Next, decoupling multi-branch 1×1 convolutions neural networks (denoted as DMP-CNNs) can construct two states of the training-time DMP-CNNs and inferencing-time DMP-CNNs via the decoupled DMPConvs.

The experiments show that substituting the 1×1 convolution with DMPConv improves the performance of popular 2D light-weight residual series CNN backbones, such as MobileNet (Howard et al., 2017, Sandler et al., 2018), ShuffleNetV1 (Zhang et al., 2018b), and ShiftResNet (Wu et al., 2018), when applied to two widely used public remote sensing image data sets: the Aerial Image (Xia et al., 2017) and the NWPU-RESISC45 (Cheng et al., 2017a). The following are the key contributions of this article.

1. A novel weighted multi-branch 1×1 convolutions architecture, denoted as DMPConv, used to enhance the performance of off-the-shelf light-weight residual series networks on remote sensing images, is presented.
2. DMPConv can be decoupled into training-time DMPConv and inferencing-time DMPConv by structural reconstruction, which is a key factor in achieving inferencing-free.
3. Comprehensive experiments on three benchmark data sets prove the effectiveness of DMPConv.

2. RELATED WORK

In this section, we give a brief literature review of this paper, including prior works on light-weight CNNs design and multi-branch convolutional networks.

2.1 Light-weight CNNs.

In recent years, high-quality deep neural networks are increasingly run on resource-constrained platforms, creating an urgent need for further research on light-weight CNNs (He and Sun, 2015), which can create an optimal balance between accuracy and efficiency. SqueezeNet (Hu et al., 2018) reduces the required parameters and computations significantly, by extensively utilizing the Fire module, in which the output of a 3×3 convolution and a 1×1 convolution are concatenated in the channel dimension. Xception (Chollet, 2017), MobileNetV1 (Howard et al., 2017) and MobileNetV2 (Sandler et al., 2018) adopt depthwise separable convolutions as alternatives to standard convolutions, which greatly improves computational efficiency. However, when these off-the-shelf light-weight networks cannot meet our specific needs, many resources are required to redesign the networks. In order to alleviate the above problems, multi-branch architectures have been proposed in recent years, which makes it more convenient to improve the accuracy of a series of off-the-shelf light-weight neural networks.

2.2 Multi-branch convolutional networks.

Multi-branch representation has shown to be successful in Inception models (Szegedy et al., 2017), where each branch is carefully customized, to aggregate more information and myriad features. Multi-branch representation, in which each branch is carefully customized, has been particularly successful in Inception models (Szegedy et al., 2017) in the capacity to aggregate a variety of data and features. ResNets (He et al., 2016) add shortcut connections to layers, in a process called pure identity mapping. Inspired by all of these methods, multi-branch weighted 1×1 convolutions are utilized to optimize our structure. Notably, our structures can be converted into a vanilla 1×1 convolution during inferencing, distinguishing them from other Multi-branch designs.

3. APPROACH

This paper proposes the DMPConv, which can improve the performance of off-the-shelf light-weight residual series networks without an increase in the parameters and calculations required during inferencing. DMPConv has two different states, the training-time DMPConv and inferencing-time DMPConv. The training-time DMPConv uses weighted multi-branch 1×1 convolutions to improve performance of the networks on remote sensing images with complex scenes. Through structural reconstruction technology, the training-time DMPConv can be equivalently transformed into the inferencing-time DMPConv, thereby achieving no additional computing burden during inferencing.

In following sections, the training-time DMPConv structure is first introduced, before describing the structural reconstruction for inferencing-free.

3.1 Feature Expression Enhancement with Training-time DMPConv

Modern convolutional neural networks (Deng et al., 2009, Szegedy et al., 2015a) are commonly composed of successive building blocks. The evolutionary ResNet (He et al., 2016) introduces residual structures into the building blocks to target the

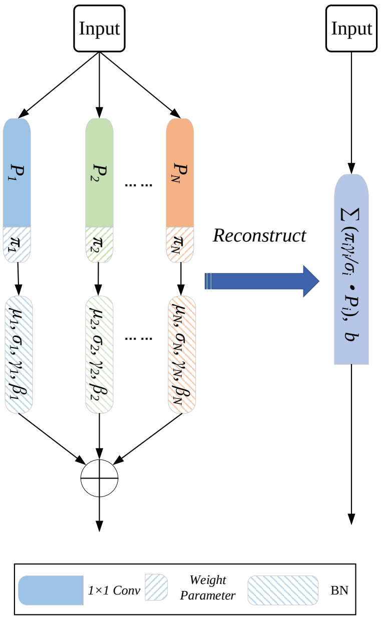

Input
Input
P_1
P_2
P_N
π_1
π_2
π_N
$\mu_1, \sigma_1, \gamma_1, \beta_1$
$\mu_2, \sigma_2, \gamma_2, \beta_2$
$\mu_N, \sigma_N, \gamma_N, \beta_N$
Reconstruct
$\sum(\pi_i\gamma_i/\sigma_i \bullet P_i)$, b
1×1 Conv
Weight Parameter
BN

Figure 2. Structural reconstruction of a DMPConv.

problem of vanishing gradients in deep neural networks. Light-weight residual series CNNs (Zhang et al., 2015) perform excellently, but in practice, deeper and wider light-weight residual series networks tend to perform better. However, ResNeXt (Xie et al., 2017), GoogLeNet (Szegedy et al., 2015b), and ACNet (Ding et al., 2019) demonstrate that cardinality is also a critical factor. ResNeXt (Xie et al., 2017) demonstrates that a higher cardinality can effectively induce accuracy gain than deeper or wider network structure, especially when increasing depth and width begin to yield diminishing returns on off-the-shelf networks. This is due to increasing the cardinality allows the network to focus on information from different representation subspaces.

Light-weight residual series networks mainly apply 1×1 convolutions to achieve a reduction in computational costs. However, due to the limited capacity of 1×1 convolutions, weighted multi-branch 1×1 convolutions are used to improve networks expression capacity are considered instead, as it increases cardinality.

To this end, we introduce the training-time DMPConv, which utilizes weighted multi-branch 1×1 convolutions, transforming every 1×1 convolution in the residual series blocks into a parallel weighted multi-branch structure.

In the formula, $X, Y \in R^{c \times h \times w}$ denote the input and output feature tensor of the weighted multi-branch 1×1 convolutions, respectively. While h and w denote the spatial dimensions, and c represents the channel numbers. π_i is the learnable weight for each branch, P_i is the i-th branch of 1 × 1 convolution kernel, and Y_i is the corresponding result. The input tensor for each P_i is exactly the same. The result of each weighted multi-branch 1×1 convolutions can be computed according to Eq.1, where $*$ denotes the 2D convolution operator.

$$Y = \sum_{i=1}^{N} Y_i = \sum_{i=1}^{N} \pi_i (P_i * X) \tag{1}$$

Due to the weighted multi-branch 1×1 convolutions architecture, our proposed modules tend to obtain richer feature information than the 1×1 convolution. Multi-branch architectures can be said to be extremely similar to the idea of integrated learning, which increase the expressiveness and robustness of the trained networks.

3.2 Structural Reconstruction for Inferencing-time DMPConv

The weighted multi-branch 1×1 convolutions structure is adopted to improve accuracy. However, the increase in parameters and calculations decreases the efficiency of the models. When training is completed, each training-time DMPConv is transformed into a 1×1 convolutional layer, which is the form of the inferencing-time DMPConv, and produces the same output as training-time DMPConv. In this way, a more powerful network, which requires no additional calculations during inferencing can be obtained. Thus, inferencing-free is mainly achieved through two steps, namely, weighted multi-branch 1×1 convolutions reconstruction and BN branches reconstruction. The details can be seen in Algorithm 1.

3.2.1 Weighted Multi-branch 1×1 Convolutions Reconstruction: The homogeneous property of convolutions is very useful. Several 1×1 convolution kernels, each with a consistent size, operate on the same input, which produces an output with the same resolution. As illustrated in Fig.2, the weighted parameters can be directly added to the corresponding positions to acquire an equivalent kernel with the same output. This computation can be written as in Eq.2. Additionally, $\sum$ represents the element-wise addition of the kernel parameters.

$$Y = \sum_{i=1}^{N} \pi_i (P_i * X) = X * (\sum_{i=1}^{N} \pi_i P_i) \tag{2}$$

3.2.2 BN Branches Reconstruction: In modern CNNs, BN (Ioffe and Szegedy, 2015) is commonly employed to prevent overfitting and accelerate the training. A linear scaling transformation follows a BN layer to improve the representation capability as a common practice. Let X, $B_{[i]:,:,k}$ be the input and k-th output layer filter in the i-th branch, respectively. For the k-th output layer filter in the i-th branch, the corresponding output feature map channel can be formulated as

$$B_{[i]:,:,k} = (\pi_i (X * P_i^{(k)}) - \mu_{[i]k}) \frac{\gamma_{[i]k}}{\sigma_{[i]k}} + \beta_{[i]k} \tag{3}$$

where $\mu_{[i]k}$ and $\sigma_{[i]k}$ represent the channel-wise mean and standard deviation of BN. $\gamma_{[i]k}$ and $\beta_{[i]k}$ respectively denote the trainable scaling and bias parameters. Eq.3 can then be converted to Eq.4, which produces identical outputs, that can be easily verified.

$$B_{[i]:,:,k} = X * \frac{\gamma_{[i]k}}{\sigma_{[i]k}} \pi_i P_i^{(k)} + (\beta_{[i]k} - \mu_{[i]k} \frac{\gamma_{[i]k}}{\sigma_{[i]k}}) \tag{4}$$

Every BN and its preceding conv layer are then reconstructed into a conv with a bias. b_k and $P^{(k)}$ are the bias term and convolution kernel after reconstitution, respectively. They can be formulated as

$$b_k = \sum_{i}^{N} \beta_{[i]k} - \frac{\alpha_{[i]k}}{\gamma_{[i]k}} \mu_{[i]k} \tag{5}$$

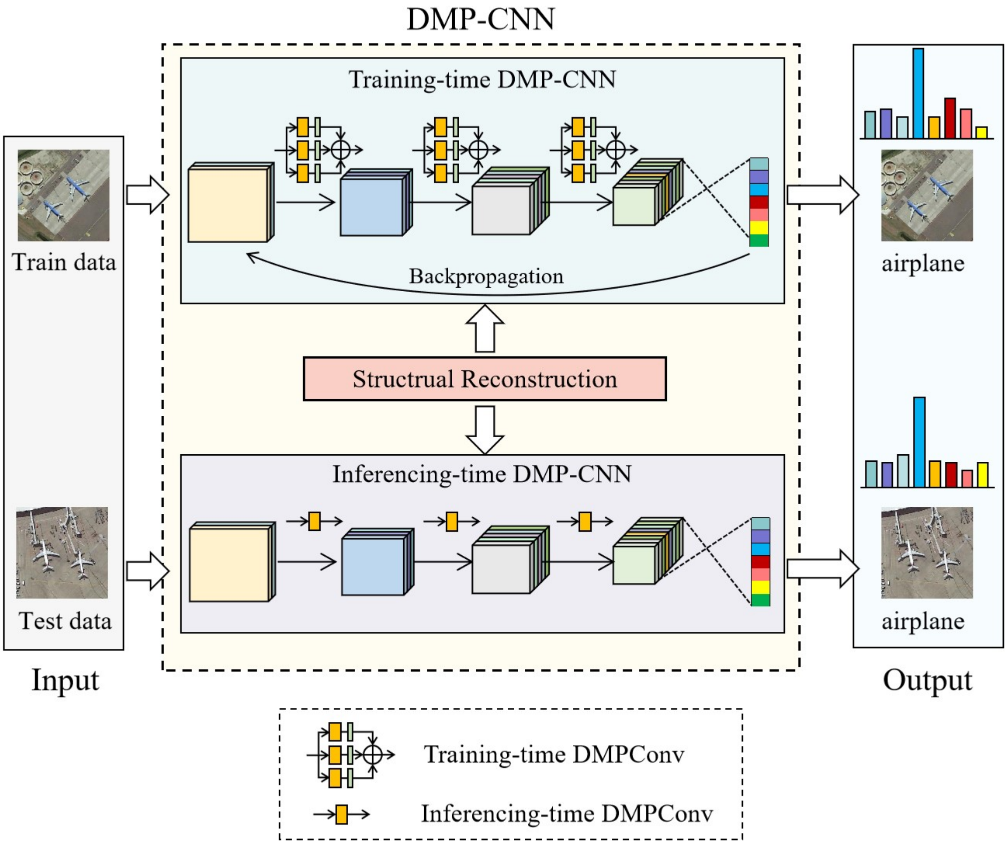


Figure 3. Framework of DMP-CNN.

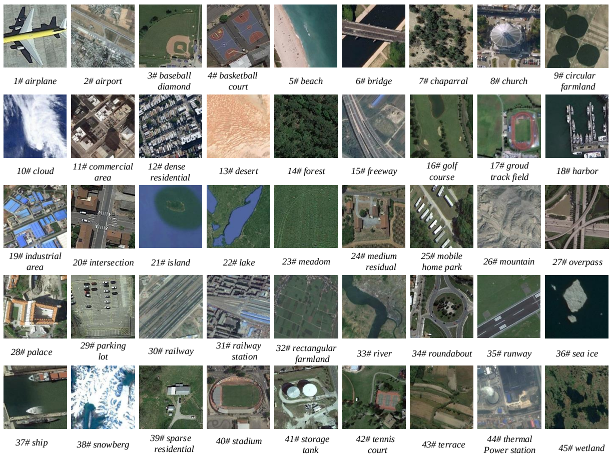


Figure 4. some examples from the NWPU-RESISC45 data set.

$$P^{(k)} = \sum_{i}^{N}(\pi_i \frac{\gamma_{[i]k}}{\sigma_{[i]k}} P_i^{(k)}) \tag{6}$$

Thus the output for the weighted multi-branch 1×1 convolutions can also be calculated as:

$$B_{:,:,k} = X * P^{(k)} + b_k \tag{7}$$

3.3 From CNNs to DMP-CNNs

Decoupling Multi-branch Pointwise Convolutions can easily be employed as a drop-in replacement for 1×1 convolutions in architectures of light-weight residual series. In this paper, we use the prefix DMP- for the networks that employ Decoupling Multi-branch Pointwise Convolutions. Fig.3 illustrates the framework of DMP-CNN.

4. EXPERIMENTS

In order to validate the performance of DMPConv, numerous experiments are performed on two commonly used remote sensing image classification benchmarks: Aerial Image (Xia et al., 2017) and NWPU-RESISC45 (Cheng et al., 2017a). Each experiment was repeated five times to reduce the effect of random initialization. All experiments were conducted with the PyTorch (Paszke et al., 2019) library.

4.1 Data sets Description

4.1.1 Aerial Image data set: The data set contains 10,000 scenes whose resolution is 600×600. The number of classes is 30, and the sensing spatial resolution varies from 0.5 to 8m. Each class has 220 to 420 images.

4.1.2 NWPU-RESISC45 data set: The data set is comprised of 31,500 images with 45 classes, and each class consists of 700 images with a spatial resolution ranging from about 30 to 0.2 m/pixel and fixed at 256×256 pixels. The NWPU-RESISC45 is the biggest remote sensing scene classification benchmark in both the types of classes and the number of images, making this data set the most challenging. Some examples of images are shown in Fig.4.

Model	Dual-DMP	*1×1*	Training ratio	
			10%	20%
ShuffleNet 2x		✓	87.25	90.18
ShuffleNet 2x	✓		**88.85**	**91.48**
ShiftResNet110-3		✓	87.43	90.45
ShiftResNet110-3	✓		**89.62**	**92.17**
MobileNetV2		✓	88.15	90.44
MobileNetV2	✓		**89.89**	**92.82**

Table 1. Ablation study on NWPU-RESISC45 data set.

4.2 Data Set Setting and Parameter Setting

As in previous studies, datasets are split into training images and testing images. For the Aerial Image data set (Xia et al., 2017), the ratio of the training are set to 20% and 50%, respectively, while in the NWPU-RESISC45 data set (Cheng et al., 2017a), the ratio of the training are set to 10% and 20%. These settings are identical to previous studies in the literature (Cheng et al., 2018). The data augmentation techniques scheme for training is adhered to as follows: random center cropping, random rotating 15 degrees and left-right cropping before feeding it into the network for training.

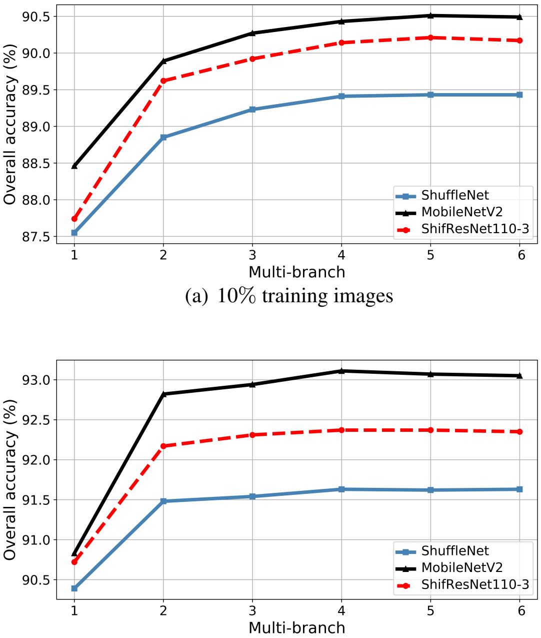
The method is performed on various light-weight residual series CNN architectures, which are pre-trained on ImageNet. Several representative benchmark models, including ShuffleNet 2x, ShiftResNet110-3, and MobileNetV2, are trained with Stochastic Gradient Descent (SGD). The hyperparameter settings in paper (Liu et al., 2019) is followed, where init lr = 0.001 before being set to 0.0001 for the last ten epochs, and training proceeds for 120 epochs with a batch size of 36, weight decay of $5e^{-4}$, and Nesterov momentum of 0.9. The correlative DMP-CNNs use identical training settings. Overall accuracy (OA) is adopted to assess the scene classification performance of these methods in the subsequent experiments.

4.3 Ablation Study

4.3.1 DMPConv vs 1×1 Convolution: In our networks, the Decoupling Multi-Branch Pointwise Convolutions (DMPConv) structure is employed to improve network performance. To validate the effectiveness of the DMPConv, ablation experiments are performed (see Table.1). For a clear comparison, the performance of the proposed Decoupling Dual-Branch Pointwise Convolutions (Dual-DMP) structure is first evaluated.

The ablation experiments are conducted on the largest and most challenging scene data set (i.e., NWPU-RESISC45 data set(Cheng et al., 2017a)). As shown in the Table.1, for the NWPU-RESISC45 data set (10% training images), the ShuffleNet 2x equipped Dual-DMP shows an improved OA from 87.25% to 88.85% (1.60% ↑), while the NWPU-RESISC45 data set (20% training images) increased from 90.18% to 91.48% (1.30% ↑). The OA for MobileNetV2 increased by 1.74% on NWPU-RESISC45 (10% training images) and increased by 2.38% on NWPU-RESISC45 (20% training images). Notably, compared to using 1×1 convolutions, the use of Dual-DMP achieves a greater improvement, which can also be demonstrated by ShiftResNet110-3 in Table.1.

4.3.2 Effects of Different Number of Branches: For further exploration on the benefit of the DMPConv structure, experiments on a spectrum of branch numbers from a candidate set of $\{1, 2, 3, 4, 5, 6\}$ are conducted. As shown in Fig.5, when branch=1, DMP-CNNs degrade to the original light-weight networks, which obtain the lowest OA on data sets including NWPU-RESISC45 (10% training images) and NWPU-RESISC45 (20% training images). In particular, Fig.5 demonstrates that the performance in DMP-CNNs with branches less than 4 improves with an increased number of branches, while the rate of improvement slows down. As branches more than 4, performance stops growing and may deteriorate slightly as the number of branches increases due to parameter redundancy. Besides, all DMP-CNNs models are able to achieve the best balance between accuracy and computational burden on the NWPU-RESISC45 data set (10% training images) and the NWPU-RESISC45 data set (20% training images), when branch = 4. Thus, the branch is set to 4 in the following experiments.



(a) 10% training images

(b) 20% training images

Figure 5. Performance of DMP-CNNs with different branch numbers on the NWPU-RESISC45 data set.

Model	OA	
	20%	50%
fine-tuned ShuffleNet 2x	92.07±0.16	95.03±0.10
DMP-ShuffleNet 2x	**92.36±0.14**	**95.84±0.10**
fine-tuned ShiftResNet110-3	90.92±0.17	94.77±0.15
DMP-ShiftResNet110-3	**91.43±0.13**	**95.93±0.13**
fine-tuned MobileNetV2	92.55±0.12	95.21±0.10
DMP-MobileNetV2	**92.77±0.11**	**96.22±0.10**

Table 2. Performance improvement of ShuffleNet, ShiftResNet110-3 and MobileNetV2 on Aerial Image data set.

4.4 Experimental Results and Comparisons

In order to clearly demonstrate the effectiveness of our method, DMPConv was embedded in several representative models including ShuffleNet 2x, ShiftResNet110-3 and MobileNetV2. The AID data set(Xia et al., 2017) and the NWPU-RESISC45 data set(Cheng et al., 2017a) are used in the following experiments.

Model	Inferencing Parameters (M)	Inferencing Flops (M)	OA 10%	OA 20%
fine-tuned ShuffleNet 2x	5.63	524	87.25±0.22	90.18±0.14
DMP-ShuffleNet 2x	5.61	524	**89.23±0.18**	**91.63±0.12**
fine-tuned ShiftResNet110-3	0.59	96	87.43±0.20	90.45±0.16
DMP-ShiftResNet110-3	0.53	96	**89.92±0.16**	**92.37±0.18**
fine-tuned MobileNetV2	3.50	315	88.15±0.17	90.44 ±0.16
DMP-MobileNetV2	3.47	315	**90.27±0.13**	**93.11±0.10**

Table 3. Performance improvement of ShuffleNet, ShiftResNet110-3 and MobileNetV2 on NWPU-RESISC45 data set.

Backbone	Method	Inferencing Parameters	OA 10%	OA 20%
Alexnet	BoVW(Cheng et al., 2017b)	-	55.22±0.39	59.22±0.39
VGG16	BoVW(Cheng et al., 2017b)	-	82.65±0.31	84.32±0.17
Alexnet	MSCP(He et al., 2018)	-	81.70±0.23	85.58±0.16
VGG16	MSCP(He et al., 2018)	-	85.33±0.21	88.93±0.14
Alexnet	Fine-tuning	60M	80.66±0.29	84.74±0.31
VGG16	Fine-tuning	130M	87.76±0.10	91.67±0.12
Alexnet	DCNN(Cheng et al., 2018)	60M	85.56±0.20	87.24±0.12
VGG16	DCNN(Cheng et al., 2018)	130M	89.22±0.50	91.89±0.22
Alexnet	SCCov(He et al., 2020)	6M	84.33±0.26	87.30±0.23
VGG16	SCCov(He et al., 2020)	13M	89.30±0.35	92.10±0.25
MobileNetV2	Fine-tuned	3.50M	88.15±0.17	90.44 ±0.16
MobileNetV2	DMPConv	**3.47M**	**90.27±0.13**	**93.11±0.10**

Table 4. Comparison of OAs (%) obtained on the NWPU-RESISC45 data set.

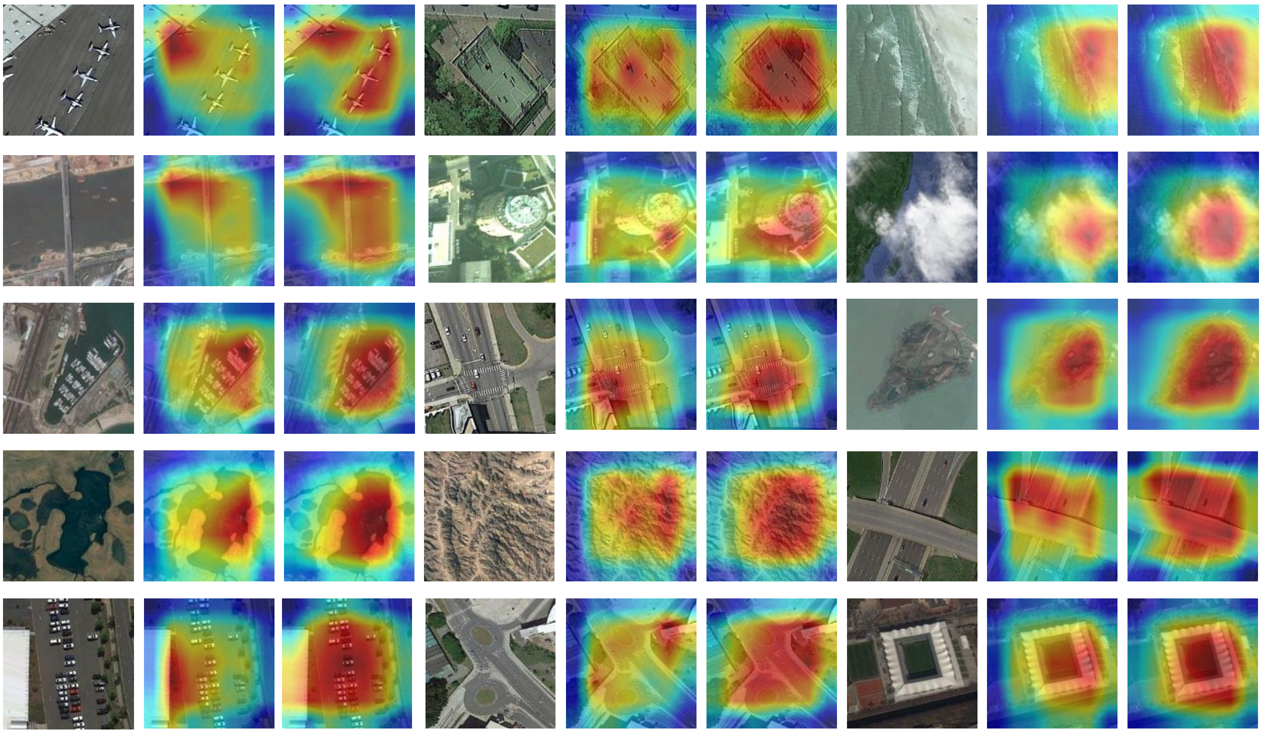


Figure 6. Examples of the CAMs generated from the NWPU-RESISC45 data set. We show the original images, CAMs generated from fine-tuned MobileNetV2 and DMP-MobileNetV2, respectively.

4.4.1 Experiment on Aerial Image Data Set: The proposed method is evaluated on the AID dataset. Table 2 shows the results. The adoption of DMPConv significantly enhances the classification accuracy in each of the above light-weight residual series networks. For instance, when training rate (Tr) = 20%, the OA of ShuffleNet 2x, ShiftResNet110-3 and MobileNetV2 is lifted by 0.29%, 0.51% and 0.22%, respectively.

4.4.2 Experiment on NWPU-RESISC45 Data Set: Lastly, the proposed approach is applied to the most extensive and difficult scene data sets (i.e., NWPU-RESISC45 data set) for contrast with related fine-tuned models. Table.3 shows a summary of experimental results: DMP-CNNs show significant improvement with 10% and 20% training ratios compared to related fine-tuned models, demonstrating their effectiveness.

For instance, when training rate (Tr) = 20%, with MobileNetV2 as the backbone, the DMP-MobileNetV2 achieves 93.11%classification accuracy, while the fine-tuned method obtains only 90.44% classification accuracy.

In conclusion, DMPConv can be used to enhance OA by a large margin under the same computational complexity and memory. The increased performance is essentially a free benefit for the end-users.

4.5 Visualization Experiment

CAMs (Zhou et al., 2016) can reveal discriminative object segments the CNN has detected. To better understand our method, CAM is used to determine if the network can identify the appropriate image segments belonging to the true class. Fig. 6 shows the CAMs produced by the fine-tuned MobileNetV2 and DMP-MobileNetV2. The original images are from the NWPU-RESISC45 data set. We observed that the semantic object corresponding to the true class could be highlighted by both the fine-tuned MobileNetV2 and DMP-MobileNetV2, indicating that CNN can localize and identify objects. However, the CAMs created by DMP-MobileNetV2 have a wider and more accurate range of highlights, allowing them to better cover semantic objects. This is due to the weighted multi-branch 1×1 convolution used in our network, which enables the network to discriminate more complex information.

4.6 Comparisons With Other Methods

To further demonstrate the effectiveness of DMPConv, we compare DMP-MobileNetV2 with several state-of-the-art models on the NWPU-RESISC45 dataset. The comparison models include the fine-tuned Alexnet and VGG16, the BoVW (Cheng et al., 2017b) with Alexnet and VGG16, the MSCP (He et al., 2018) with Alexnet and VGG16, the DCNN (Cheng et al., 2018) with Alexnet and VGG16 and the SCCov (He et al., 2020) with Alexnet and VGG16.

Table 4 reports the performance of different models. As shown in this table, when training rate (Tr) = 20%, our DMP-MobileNetV2 achieves 93.11%, which surpasses the classfication accuracy obtained by second-best model—SCCov(He et al., 2020) with VGG16 by 1.01%. Similarly, when training rate (Tr) = 10%, our DMP-MobileNetV2 also obtained the best performance with an OA of 90.27% that is 0.97% higher than the OA of the second-best model-SCCov(He et al., 2020) with VGG16. In addition, our model has the fewest parameters. Furthermore, by contrast to the fine-tuned MobileNetV2, the gain of our DMP-MobileNetV2 on OA can be achieved without extra parameters, computational and storage space requirements during inferencing, such that end-users can enjoy additional performance improvements at no extra cost.

5. CONCLUSIONS

This paper proposes an efficient DMPConv for remote sensing image classification task. DMPConv can be decoupled into training-time DMPConv and inferencing-time DMPConv. With our structural reconstruction method, DMPConv is able to improve the performance of the off-the-shelf light-weight residual series CNNs without additional computational and parameter burdens during inferencing. Experimental results show that DMPConv is able to improve the performance of various light-weight residual series CNNs on UC Merced Land-Use data set, Aerial Image data set, and NWPU-RESISC45 data set.

REFERENCES

Cheng, G., Han, J., Lu, X., 2017a. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE*, 105(10), 1865–1883. https://doi.org/10.1109/JPROC.2017.2675998.

Cheng, G., Li, Z., Yao, X., Guo, L., Wei, Z., 2017b. Remote Sensing Image Scene Classification Using Bag of Convolutional Features. *IEEE Geosci. Remote. Sens. Lett.*, 14(10), 1735–1739. https://doi.org/10.1109/LGRS.2017.2731997.

Cheng, G., Yang, C., Yao, X., Guo, L., 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. https://doi.org/10.1109/TGRS.2017.2783902.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, IEEE Computer Society, 1800–1807.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F., 2009. Imagenet: A large-scale hierarchical image database. *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, IEEE Computer Society, 248–255.

Ding, X., Guo, Y., Ding, G., Han, J., 2019. Acnet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks. *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, IEEE, 1911–1920.

He, K., Sun, J., 2015. Convolutional neural networks at constrained time cost. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, IEEE Computer Society, 5353–5360.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, IEEE Computer Society, 770–778.

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. https://doi.org/10.1109/TGRS.2018.2845668.

He, N., Fang, L., Li, S., Plaza, J., Plaza, A., 2020. Skip-Connected Covariance Network for Remote Sensing Scene Classification. *IEEE Trans. Neural Networks Learn. Syst.*, 31(5), 1461–1474. https://doi.org/10.1109/TNNLS.2019.2920374.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *CoRR*, abs/1704.04861. http://arxiv.org/abs/1704.04861.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, IEEE Computer Society, 7132–7141.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. F. R. Bach, D. M. Blei (eds), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, JMLR Workshop and Conference Proceedings, 37, JMLR.org, 448–456.

Liu, X., Zhou, Y., Zhao, J., Yao, R., Liu, B., Zheng, Y., 2019. Siamese Convolutional Neural Networks for Remote Sensing Scene Classification. *IEEE Geosci. Remote. Sens. Lett.*, 16(8), 1200–1204. https://doi.org/10.1109/LGRS.2019.2894399.

Ma, N., Zhang, X., Zheng, H., Sun, J., 2018. Shufflenet V2: practical guidelines for efficient CNN architecture design. V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (eds), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, Lecture Notes in Computer Science, 11218, Springer, 122–138.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: An imperative style, high-performance deep learning library. H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, R. Garnett (eds), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 8024–8035.

Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A., Chen, L., 2018. Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation. *CoRR*, abs/1801.04381. http://arxiv.org/abs/1801.04381.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A. A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. S. P. Singh, S. Markovitch (eds), *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, AAAI Press, 4278–4284.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015a. Going deeper with convolutions. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, IEEE Computer Society, 1–9.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015b. Going deeper with convolutions. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, IEEE Computer Society, 1–9.

Wu, B., Wan, A., Yue, X., Jin, P. H., Zhao, S., Golmant, N., Gholaminejad, A., Gonzalez, J., Keutzer, K., 2018. Shift: A zero flop, zero parameter alternative to spatial convolutions. *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, IEEE Computer Society, 9127–9135.

Xia, G., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X., 2017. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Trans. Geosci. Remote. Sens.*, 55(7), 3965–3981. https://doi.org/10.1109/TGRS.2017.2685945.

Xie, S., Girshick, R. B., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, IEEE Computer Society, 5987–5995.

Yang, Y., Newsam, S. D., 2010. Bag-of-visual-words and spatial extensions for land-use classification. D. Agrawal, P. Zhang, A. E. Abbadi, M. F. Mokbel (eds), *18th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2010, November 3-5, 2010, San Jose, CA, USA, Proceedings*, ACM, 270–279.

Zhang, K., Sun, M., Han, T. X., Yuan, X., Guo, L., Liu, T., 2018a. Residual Networks of Residual Networks: Multilevel Residual Networks. *IEEE Trans. Circuits Syst. Video Technol.*, 28(6), 1303–1314. https://doi.org/10.1109/TCSVT.2017.2654543.

Zhang, X., Zhou, X., Lin, M., Sun, J., 2018b. Shufflenet: An extremely efficient convolutional neural network for mobile devices. *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, IEEE Computer Society, 6848–6856.

Zhang, X., Zou, J., Ming, X., He, K., Sun, J., 2015. Efficient and accurate approximations of nonlinear convolutional networks. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, IEEE Computer Society, 1984–1992.

Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, IEEE Computer Society, 2921–2929.