A Noise-robust Feature Fusion Model Combining Non-local Attention for Material Recognition

Chuanbo Zhou
bobozhou@stu.xjtu.edu.cn
Xi'an Jiaotong University, School of Automation Science and Engineering
Xi'an, Shaanxi 710049, China

Guoan Yang*
gayang@mail.xjtu.edu.cn
Xi'an Jiaotong University, School of Automation Science and Engineering
Xi'an, Shaanxi 710049, China

Zhengzhi Lu
lu947867114@stu.xjtu.edu.cn
Xi'an Jiaotong University, School of Automation Science and Engineering
Xi'an, Shaanxi 710049, China

Deyang Liu
yohn08@stu.xjtu.edu.cn
Xi'an Jiaotong University, School of Automation Science and Engineering
Xi'an, Shaanxi 710049, China

Yong Yang
294575885@qq.com
Xi'an Jiaotong University, School of Automation Science and Engineering
Xi'an, Shaanxi 710049, China

ABSTRACT

Material recognition, as an important task of computer vision, is hugely challenging, due to large intra-class variances and small inter-class variances between material images. To address those recognition problems, multi-scale feature fusion methods based on deep convolutional neural networks are presented, which has been widely studied in recent years. However, the past research works paid too much attention to the local features of the image, while ignoring the non-local features that are also crucial for fine image recognition tasks such as material recognition. In this paper, Non-local Attentional Feature Fusion Network (NLA-FFNet) is proposed that combines local and non-local feature of images to improve the feature representation capability. Firstly, we utilize the pre-trained deep convolutional neural network to extract the image feature. Secondly, a Multilayer Non-local Attention (MNLA) block is designed to generate a non-local attention map which represents the long-range dependencies between features of different positions. Therefore, it can achieve stronger noise-robustness of model and better ability to represent fine features. Finally, combined our Multilayer Non-local Attention block with bilinear pooling which has been proved to be effective for feature fusion, we propose a deep neural network framework, NLA-FFNet, with noise-robust multi-layer feature fusion. Experiment prove that our model can achieve a competitive classification accuracy in material image recognition, and has stronger noise-robustness at the same time.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Neural networks**.

*Corresponding author: Guoan Yang

ICIGP 2022, January 7–9, 2022, Beijing, China

ACM ISBN 978-1-4503-9546-5/22/01…$15.00
https://doi.org/10.1145/3512388.3512450

KEYWORDS

Deep convolutional neural network, Non-local attention, Material recognition, Bilinear pooling, Feature fusing

ACM Reference Format:
Chuanbo Zhou, Guoan Yang, Zhengzhi Lu, Deyang Liu, and Yong Yang. 2022. A Noise-robust Feature Fusion Model Combining Non-local Attention for Material Recognition. In *2022 the 5th International Conference on Image and Graphics Processing (ICIGP 2022), January 7–9, 2022, Beijing, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3512388.3512450

1 INTRODUCTION

Material characteristics are a very important visual clue, which is widely present on the surface of natural objects. It is visually diverse and complex, and difficult for the human eye to distinguish their differences. Due to the influence of lighting conditions, shooting angle of view, and shooting distance, the texture appearance feature of material images usually change greatly, directly leading to the low accuracy of material image recognition. Material image recognition occupies an important position in practical applications such as scene recognition, industrial inspection, and medical image recognition, which has been a hot research issue in the field of computer vision in recent years. It is generally believed that how to mine robust and detailed image features is the key to improving the recognition accuracy.

Initially, researchers proposed a variety of handcraft classic feature descriptors, such as SIFT [23], SURF [2] and so on, in order to reduce or eliminate the influence of factors such as illumination, rotation, and viewing angle. For decades, these local feature descriptors have dominated the field of computer vision. With the springing up of deep learning, Deep Convolutional Neural Networks (DCNN), driven by big data, have achieved better feature extraction and description [13][14][31]. Cimpoi [8] et al. combined Fisher Vector with CNN and proposed the FV-CNN network model. Subsequently, based on FV-CNN [26], Song et al. proposed a Locally-transferred Fisher Vector (LFV) model, which combined Fisher Vector coding and neural network in a simple and effective way and obtained lower dimensional feature descriptors than FV-CNN [26]. Lin et al. [22] proposed the Bilinear CNN (BCNN) model for texture and material recognition. The BCNN designed a bilinear structure to aggregate the pairwise feature of two independent CNNs, which adopted

outer product of feature vectors to produce a high-dimensional feature for quadratic expansion. Inspired by BCNN [22], Yu et al. [28] proposed the Hierarchical Bilinear Pooling (HBP) model in which each convolutional layer of CNN is regarded as a feature extractor of different object parts, and features of multiple convolutional layers is fused in a simple way. However, it just expands on the channel and does not pay attention to the possible gain effects of the non-local features, ignoring the influence of noise and background information when HBP extracts the feature maps of different convolutional layers.

Therefore, in order to obtain a more noise-robust feature representation, building on [22][28], we propose a Non-local Attentional Feature Fusion Network (NLA-FFNet) which combines the local and non-local features of multiple convolutional layers. The NLA-FFNet can obtain a more noise-robust feature descriptor by considering the non-local similarity of Multi-layers of CNN. The main contributions of this paper are as follows:

1. Based on the non-local block [27], we proposed two different multilayer attention modules combined with the non-local module, named MNLA-1 and MNLA-2 respectively, which can be easily applied to classic CNN architectures;

2. We proposed a deep neural network model, NLA-FFNet, for material recognition by applying MNLA-1 and MNLA-2, which has competitive classification accuracy and strong noise-robustness on the DTD [7] and MINC [4] datasets.

2 RELATE WORK

2.1 Deep feature fusion

In DCNN, it is an important way to improve the performance of the network model by fusing the multi-scale feature which is extracted from multi-layers. CNN's shallow features, with high resolution, contain more contour and texture information, but are poor in expressing semantic information of images; while deep features, with low resolution, have stronger semantic information, but are poor in perception of texture details. Therefore, it's key to efficiently fusing shallow and deep features. With the aim of synthesizing more discriminative fusion features, the feature fusion strategies can be mainly divided into two categories, Concatenation and Add: The feature fusion methods used in [16][1][6] are directly concatenating features in dimensionality, that is, if the dimensions of the two input features x and y are p and q, the dimension of the fusing feature of concatenation is $p + q$; while in [21][11], the two feature vectors, x and y, are directly added and combined into a complex vector, which is element-by-element addition of input features. However, Concatenation and Add only perform first-order linear fusion for the feature vector, ignoring the second-order information of feature that was shown to be a highly effective for image classification and semantic segmentation [8]. In [22], Lin et al. proposed BCNN, in which the bilinear pooling can express local features more efficiently by synthesizing the second-order fusion information. Subsequently, Kim et al. [17] proposed a low-rank bilinear model using Hadamard product in order to simplify the computational complexity of BCNN [22].

Based on the low-rank bilinear model [17], we propose the non-local attention feature fusion module including MNLA block, as shown in Figure 1. The innovations of our feature fusion method are shown in: 1) We fused the features of different convolutional layers, while BCNN [22] only performed feature fusion for the last single convolutional layer; 2) We enhance model's ability to express local and global information by use of the non-local attention block, while BCNN only express single-layer features through the backbone network.

2.2 Non-local attention

When facing complex visual scenes in real world, the human always focus on certain specific areas that are most prominent through rapid scanning of eye movement [9]. This selective visual attention mechanism has been widely used in various fields of computer vision such as image recognition, object detection, image segmentation, etc. [5][18], [16]. Hu et al. [13] proposed a squeeze-and-excitation (SE) block, which calibrates the weights of different channels of the feature map and adaptively adjust the channel importance for image classification. Following Hu et al., Woo et al. proposed a convolutional block attention module (CBAM) [3]. CBAM divides the attention process into two parts, channel attention and spatial attention, to focus on the most important spatial area of the image. However, these methods, as a kind of local attention, only select the most important part of the entity in the global scope, which may discard some important material entities, resulting in poor model performance when there are multiple entities in one image.

Therefore, Wang [27] proposed a non-local attention model, which calculates the weighted average of all pixels to ensure that distant pixels can also contribute to the final prediction, thereby the important entities are contained. Non-local attention has been successfully used to improve the performance of natural image recognition [30] and semantic segmentation [15], by enhancing the long-range dependencies of visual features.

3 METHODOLOGY

3.1 Non-local attentional feature fusion network architecture

In image classification and image segmentation tasks, DCNN shows great advantage that DCNN has stronger feature representation capability and better performance, compared with traditional neural networks. When faced with large-scale and more complex data, the convolutional layer in DCNN can automatically learn and extract hierarchical features. At the same time, DCNN can also build a deeper convolutional layer to allow the network model to better extract the more discriminative and more noise-robust deep features in the image. Based on the above analysis, our model, NLA-FFNet, also uses the CNN model as the image feature extractor.

The overall architecture of NLA-FFNet, as shown in Figure 1, is divided into three parts: the first part uses VGG-D [25] as the image feature extractor, named backbone network whose the specific parameters will be given in section 4.1. The main task of backbone is to extract the multi-scale features of the image through VGG-D which is well pre-trained; the second part is the multi-layer projection based on MNLA block, whose purpose is to use the non-local attention block to obtain more non-local (NL) detailed information when fusing the extracted multi-layer features, and significantly eliminate the influence of noise on network performance; the third

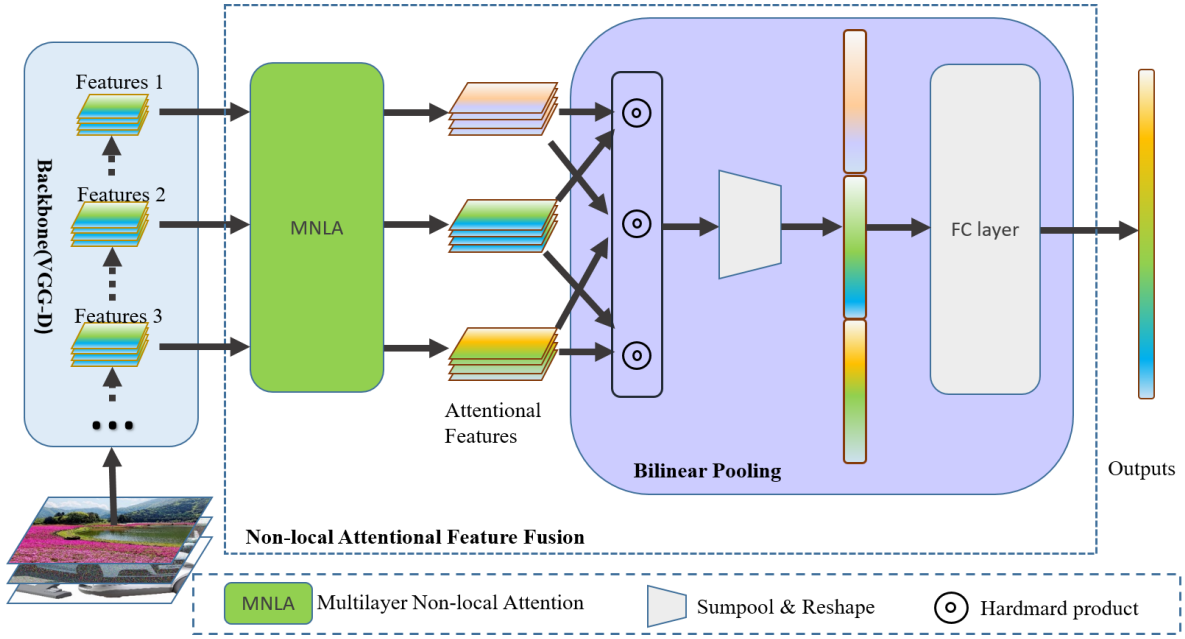


Figure 1: the architecture of NLA-FFNet

part is the bilinear pooling model [17]: it combines the attention feature maps of the MNLA block in pairs, and obtains the vector with second-order information after the two features are fused through bilinear pooling. Finally, the output, a class vector, is synthesized.

3.2 Non-local attentional feature fusion

3.2.1 Multilayer Non-local Attention (MNLA) Block. In order to obtain the long-range dependencies of the feature map, the non-local block [27] enhances the information of the location correlation by aggregating the information of other locations in the long distance. We assume that the feature map of the single convolutional layer is $\mathbf{X} \in R^{hw \times c}$, where h, w, c denote the height, width and channel number of the feature map respectively. Let $\mathbf{x} = (x_1, x_2, x_3, \ldots x_{hw})^T \in R^{hw \times 1}$ as the column vector of the feature map $\mathbf{X}$. Then the output vector of the non-local block can be expressed as

$$z_i = x_i + W_z \sum_{j=1}^{hw} \frac{f\left(x_i, x_j\right)}{\mathbb{N}(\mathbf{x})} \left(W_\theta x_j\right), i = 1, 2, \ldots, hw \qquad (1)$$

where $f\left(x_i, x_j\right)$ denotes the correlation similarity of the vector $\mathbf{x}$ at the position i, j; W_θ denotes the linear mapping matrix at the position j of $\mathbf{x}$; $\mathbb{N}(\mathbf{x})$ represents the correlation similarity regularization factor of $\mathbf{x}$; W_z denotes the linear mapping matrix which can be implemented by convolutional operation (e.g., 1x1 convolution layer).

Let the correlation feature be $\omega_{ij} = \frac{f(x_i, x_j)}{\mathbb{N}(\mathbf{x})}$, which represents the correlation of the feature vector $\mathbf{x}$ at position i, j. Following the selection in [27], ω_{ij} can be denoted as $\omega_{ij} = \frac{\langle W_\varphi x_i, W_\phi x_j \rangle}{N}$, where $N = hw$ is the dimensionality of the feature vector $\mathbf{x}$; $W_\varphi x_i, W_\phi x_j$ are the linear mapping matrixes and $\langle , \rangle$ denotes the inner product.

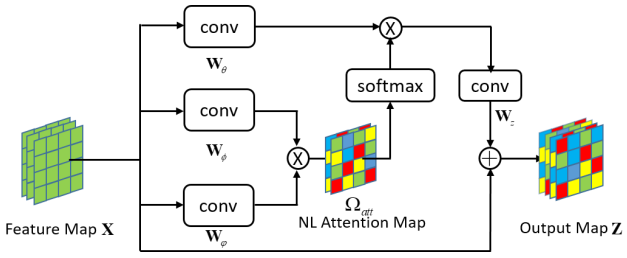


Figure 2: illustration of non-local attention.

Therefore, we can get the matrix form of non-local operation

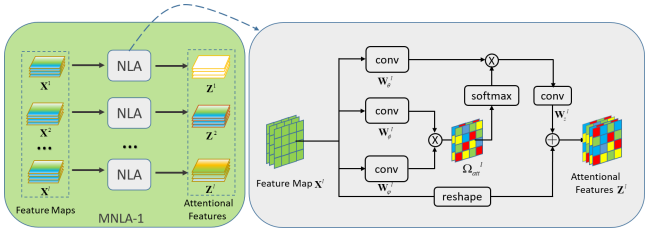
$$\begin{cases} \mathbf{Z} = \mathbf{X} + \mathbf{W}_z \Omega_{\text{att}} \mathbf{W}_\theta \mathbf{X} \\ \Omega_{\text{att}} = \frac{\langle \mathbf{W}_\phi \mathbf{X}, \mathbf{W}_\varphi \mathbf{X} \rangle}{N} \end{cases} \qquad (2)$$

where $\mathbf{X}, \mathbf{Z} \in R^{N \times c}$ denote the feature maps of input and output respectively; $\mathbf{W}_z, \mathbf{W}_\phi, \mathbf{W}_\varphi, \mathbf{W}_\theta \in R^{N \times N}$ are all learnable mapping matrixes, which can be realized by convolution operation, and $\Omega_{att} \in R^{N \times N}$ is a non-local attention map, as shown in Figure 2.

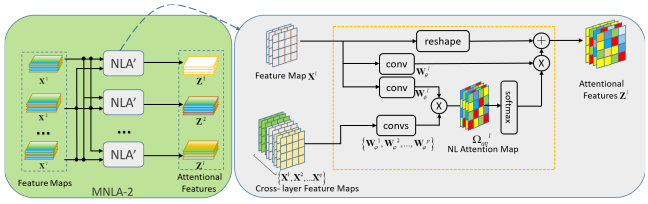
A. Parallel-Projection-Based multi-layer non-local attention

Following Eq. (2), we propose a non-local attention block based on parallel projection for multi-layer feature maps. For a feature map in the l-th convolutional layer, h^l, w^l and c^l denote its height, width and the number of channels respectively. And we denote the feature map as $\mathbf{X}^l \in R^{h^l w^l \times c^l}$. Then, for each map $\mathbf{X}^l$ as shown in Figure 3(a), the output $\mathbf{Z}^l \in R^{h^l w^l \times c^l}$ of non-local operation can be obtained respectively

$$\begin{cases} \mathbf{Z}^l = \mathbf{X}^l + \mathbf{W}_z^l \Omega_{att}^l \mathbf{W}_\theta^l \mathbf{X}^l \\ \Omega_{att}^l = \frac{\left\langle \mathbf{W}_\phi^l \mathbf{X}^l, \mathbf{W}_\varphi^l \mathbf{X}^l \right\rangle}{N^l} \end{cases}, l = 1, 2, \ldots \qquad (3)$$



(a) Multilayer Parallel Projection



(b) Multilayer Cross Projection

Figure 3: illustration of multi-layer non-local attention.

where $N^l = h^l w^l$ is the number of positions in $\mathbf{X}^l$ and $\mathbf{W}_z^l, \mathbf{W}_\phi^l, \mathbf{W}_\varphi^l, \mathbf{W}_\theta^l \in R^{N^l \times N^l}$ are all learnable mapping matrixes.

B. Cross-Projection-Based multi-layer non-local attention

Deep features are sensitive to large objects in the image because they have a larger receptive field. On the contrary, shallow features are more sensitive to smaller targets which contain more detailed information, such as image contours and texture information. So as to get trade-off between the two kinds of features, we design the non-local attention based cross projection, as shown in Figure 3(b), which can weight different feature maps of multiple layers and improve the representation ability of details information. According to Eq. (2), we define the Cross-Projection-Based multi-layer non-local attention as

$$\mathbf{Z}^l = \mathbf{X}^l + \sum_{p \neq l} \mathbf{W}_z{}^p \Omega_{att}{}^p \mathbf{W}_\theta{}^p \mathbf{X}^p, l = 1, 2, \ldots, p = 1, 2, \ldots \quad (4)$$

Different from the Parallel-projection-based multi-layer non-local attention above, the cross-projection- based non-local block aggregate features with other different convolutional layers, and therefore can learn more discriminative and detailed multi-scale features which contribute to reducing the impact of meaningless background noise.

3.2.2 Bilinear Pooling. The bilinear pooling can express the local feature information fusion more precisely through the synthesis of two feature vectors, as shown in Figure 1. Let $\mathbf{z}_1, \mathbf{z}_2 \in R^{N \times 1}$ as input column vectors, f_i as output value and b_i as bias. Thereby the bilinear pooling model can be define as

$$f_i = \mathbf{z}_1^T \mathbf{W}_i \mathbf{z}_2 + b_i, i = 1, 2, \ldots, d \quad (5)$$

where $\mathbf{W}_i$ denotes the low-rank matrix that can be factorized as $\mathbf{W}_i = \mathbf{U}_i \mathbf{V}_i^T$ [17]. T denotes a transpose of matrix.

We denote the output, class vector, as $\mathbf{f} = (f_1, f_2, \ldots, f_d) \in R^{d \times 1}$ where d denotes the number of category. Then, as for the image classification task, the low-rank bilinear pooling can be defined as

$$\mathbf{f} = \mathbf{z}_1^T \mathbf{U} \mathbf{V}^T \mathbf{z}_2 + \mathbf{b} = \mathbf{P}^T \left(\mathbf{U}^T \mathbf{z}_1 \circ \mathbf{V}^T \mathbf{z}_2 \right) + \mathbf{b} \quad (6)$$

where $\mathbf{P} \in R^{N \times d}$ denotes the all-in-one matrix (all elements are 1) and $\mathbf{U}, \mathbf{V} \in R^{N \times N}$ are the learnable parameters of feature projection; $\mathbf{b} \in R^{d \times 1}$ is learnable bias vector; $\circ$ denotes Hadamard product which is element-wise multiplication

Table 1: Details of Our Basic Backbone Network. In the table, the input image size is 224x224.

Modules	Blocks	Basic Layers		Output size
		Conv Size	Conv Number	
VGG-D	Block0	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}$	2	$112 \times 112 \times 64$
	Block1	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}$	2	$56 \times 56 \times 128$
	Block2	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}$	3	$28 \times 28 \times 256$
	Block3	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \\ 3\times3, 256 \end{bmatrix}$	3	$14 \times 14 \times 512$
	Block4	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \\ 3\times3, 256 \end{bmatrix}$	3	$7 \times 7 \times 512$
classifier	FC	$fc\ layer \times 3 => n$		n classes

4 EXPERIMENT

4.1 Implementation and training details

4.1.1 Implementation. We choose VGG-D [25], shown in Table 1, as backbone network of NLA-FFNet. The convolutional layer of VGG-D is divided into 5 blocks, among which Block0 and Block1 contain two convolutional layers with 3x3 convolution kernels, and Block2, Block3, and Block4 all contain three 3x3 convolutional layers. The classifier part of VGG-D is composed of three full connection layers. In NLA-FFNet, we contain all the convolutional layers and discard classifier part of backbone. In addition, batch normalization and wavelet pooling are also applied to reinforce the convergence performance of the network, similar to the implementation details of [20]. For a fair comparison, we constructed the VGG-D network, named baseline, for CIFAR, DTD and MINC data sets with the same trick. Besides, both the baseline and the NLA-FFNet use network parameters that have been fully pre-trained in the ImageNet dataset [24], in order to speed up the learning process.

4.1.2 Training. The two MNLA blocks, MNLA-1 and MNLA-2, are added into NLA-FFNet respectively for training and testing which follow the same rules of data augmentation. For the CIFAR dataset, the input image is directly resized to a fixed size (32x32, 64x64 and128x128), while for the DTD and MNC dataset, images are resize to 256×256 and randomly crop patches to 224×224. The training images of all datasets are further augmented via horizontal flip (p = 0.5) and normalization.

The training procedure is divided into two stages: in the first stage, the parameters of backbone of NLA-FFNet are frozen to adjust the parameters of the non-local attention feature fusion module in which the initial learning rate is 0.1; in the second stage, we cancel

Table 2: Results of the Ablation Experiments on CIFAR10. MNLA-1, MNLA-2 are represent of the parallel and cross projection of NL attention structure (section 3.2).

Method	Feature Fusion	Input Image Size		
		32 × 32	64 × 64	128 × 128
Baseline	N/A	91.94%	94.78%	95.37%
NLA-FFNet-v1	MNLA-1	93.35%	95.70%	95.85%
NLA-FFNet-v2	MNLA-2	93.64%	95.84%	95.91%

the freezing of parameters of NLA-FFNet and fine-tune the entire network of NLA-FFNet whose learning rate is set to 0.01. Our model is trained using the stochastic gradient descent (SGD) optimizer whose optimization momentum, weight-decay and batch size are set to 0.9, 1e-5 and 12 respectively. We do not use the validation dataset during training, and directly perform on the test dataset. We train NLA-FFNet for 120 epochs on a PC (Nvidia GeForce GTX2080Ti, RAM: 64GB), and finally save the best trained model.

4.2 Effectiveness of NLA-FFNet

The main contribution of our work is to propose a network based on the non-local attention feature fusion module. In order to verify the impact of different non-local attentional feature fusion structures on image recognition, we conduct ablation experiments on our network architecture, in which the general image recognition datasets, CIFAR10 and CIFAR100 [19], are chosen. The CIFAR10 consists of 60,000 color images with a size of 32×32 pixels in 10 categories and each category includes 5000 training images and 1000 testing images respectively. While the CIFAR100 is more complex which includes 100 classes, each with 500 training images and 100 testing images.

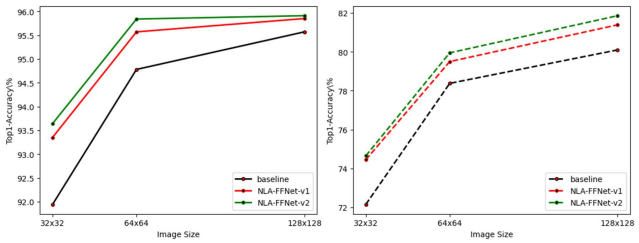


Figure 4: the Top1-accruacy of different image size. The left and right figure represent the experiment on CIFAR10 and CIFAR100, respectively.

As for CIFAR10, compared to the baseline model, the NLA-FFNet can improve the Top-1 accuracy by 1.41% with MNLA-1, and 1.7% with MNLA-2 when input image size is 32x32, as shown in Table 2. It can be easily seen that our approach greatly improve classification performance. Similar results can be observed on CIFAR100 (shown in Table 3), which demonstrates the effective and validity of the proposed methods. Besides, our model has a higher degree of improvement than baseline in CIAFR100 which is shown in Figure 4.

Table 3: Results of the ablation experiments on CIFAR100

Method	Feature Fusion	Input Image Size		
		32 × 32	64 × 64	128 × 128
Baseline	N/A	72.15%	78.38%	80.10%
NLA-FFNet-v1	MNLA-1	74.45%	79.50%	81.39%
NLA-FFNet-v2	MNLA-2	74.65%	79.96%	81.86%

Table 4: Test accuracy (Top1-accuracy) compared with other methods.

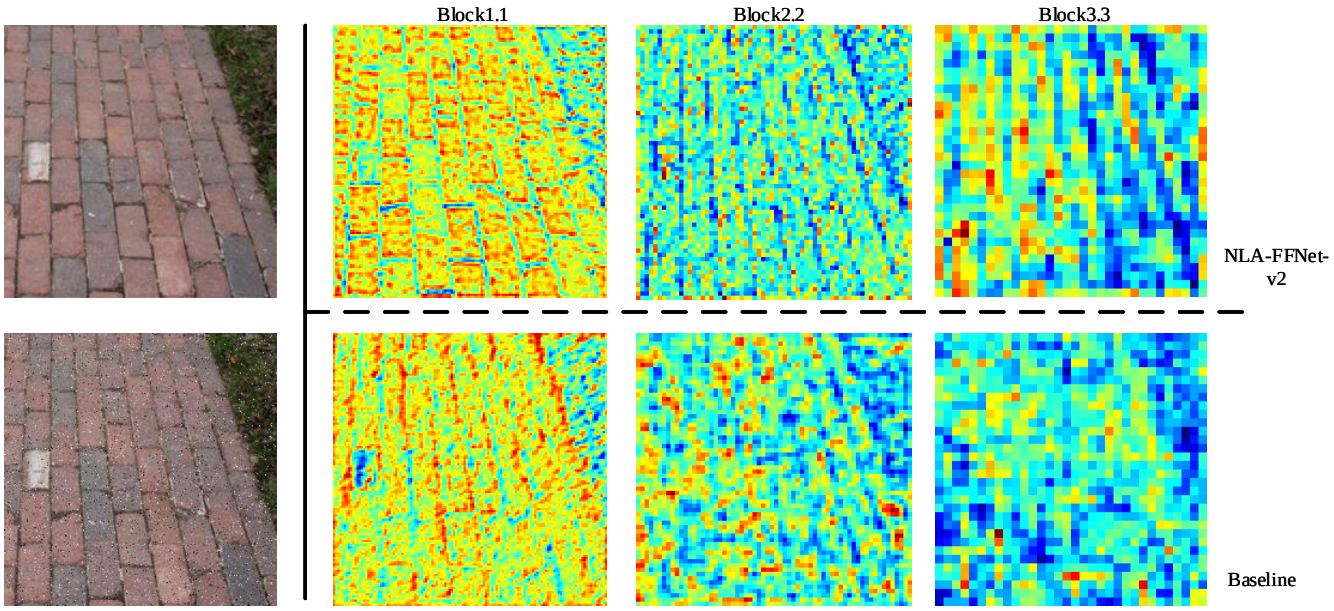
Method	DTD	MINC-2500
baseline[25]	65.17%	76.87%
FV-CNN[8]	72.30%	63.10%
Deep-TEN[29]	69.60%	80.40%
BCNN[22]	72.90%	N/A
Compact BCNN[12]	67.70%	N/A
FASON[10]	72.90%	N/A
NLA-FFNet-v2(ours)	73.04%	79.94%

4.3 NLA-FFNet for material recognition

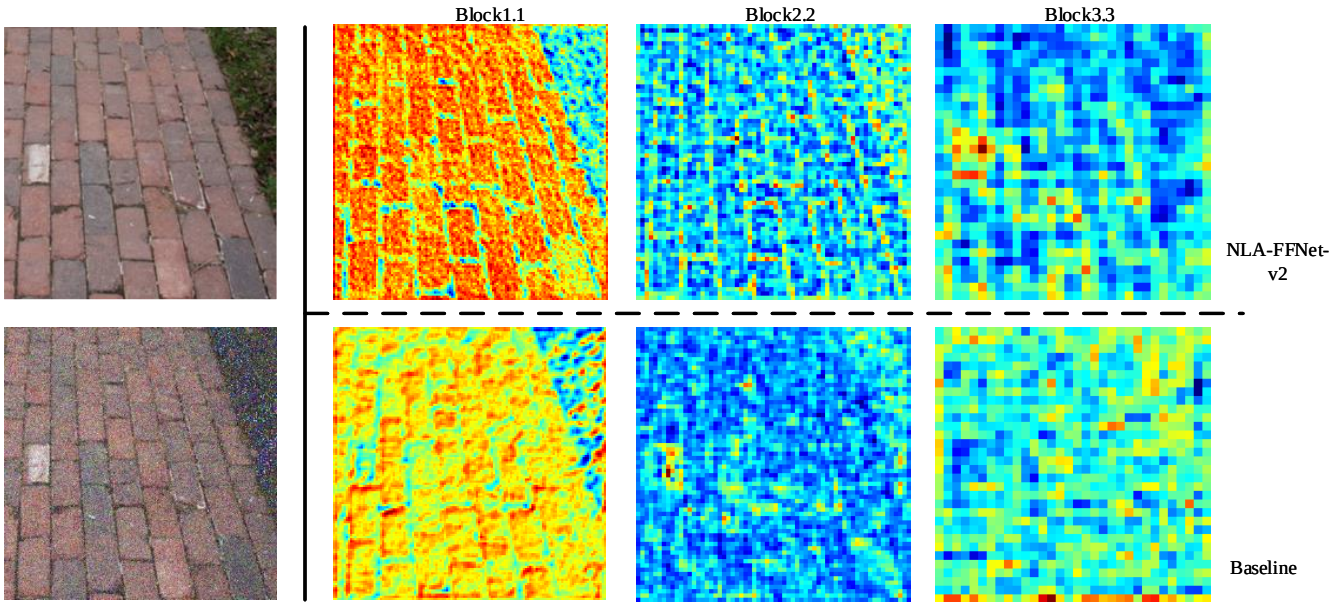
To show the generality of NLA-FFNet for material recognition, we experiment on two material/texture recognition datasets: Describable Textures Database (DTD) [7] and Materials in Context Database (MINC-2500) [4]. The proposed method was compared to other state-of-the-art material classification methods as well as the baseline model [20] with batch normalization and wavelet pool. Overall, NLA-FFNet performed better or comparably than FV-CNN [8], Deep-TEN [29], BCNN [22], Compact BCNN [12], FASON [10], as shown in Table 4. For DTD, proposed method, NLA-FFNet, achieves slightly better Top1-accuracy which improves by 0.14% compared with FASON [10] and BCNN [22]. This is because the DTD data set contains more images with homogeneous textures. That is to say, non-local attention module can get more non-local similar patches to further enhance the query feature. As for the MINC-2500 dataset, NLA-FFNet still outperforms baseline [25] and FV-CNN [8] with 3.07% and 16.84% increment of top1-accuracy respectively, while NLA-FFNet almost achieves the same Top1-accuracy of Deep-TEN [29], it is only 0.46% lower in accuracy. It should be noted that most MINC images only have textures of interest at local in which the global information, such as non-local feature, is difficult to for the model to enhance representation capability because the non-local attention block can only obtain a minority of non-local features. Nevertheless, our model still performed comparably an accuracy and becomes more noise-robust.

4.4 Noise-robustness and visualization

We visualize the feature maps to verify the noisy-robustness of NLA-FFNet. Firstly, clean input image with size of 224x224 is corrupted by Gaussian noise and Salt and pepper noise respectively. Then, after these noisy images are feed into our network, we extract and visualize feature maps of convolutional layer of baseline and NLA-FFNet, respectively. The result is shown in Figure 5, in which Gaussian noise's variance (mean equals 0 in default) is 0.02 and



(a) Gaussian noise, variance $\sigma^2 = 0.02$



(b) Salt and pepper noise, noisy ponit rate $r = 0.02$

Figure 5: illustration of non-local attention. The visualization of feature maps of noisy image. The Block2.2 is represent of 2nd convolutional layer of Block2 of backbone (Table 1).

the rate of salt and pepper is 0.02. As shown in Figure 5(a), we can find that our method could suppress the noise and maintain the object structure better in different convolutional layer, compared with baseline model. And it's following the same discovery when input image is corrupted by salt and pepper noise in Figure5(b).

5 CONCLUSION

In this paper, we proposed NLA-FFNet, a CNN architecture combined with non-local attention block, fusing long-range dependencies information to build up more noise-robust and detailed feature representation. We also designed two multilayer non-local attention block with parallel projection (MNLA-1) and cross projection

(MNLA-2), which can easily embedded into feature fusing model. We have demonstrated the validity and noisy-robustness of the proposed method through ablation and quantitative study on three datasets. In the future, we will further explore the application of MNLA to other image tasks such as segmentation.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 61673314 and 61573272, in part by the National Key Research and Develop program of China under Grant 2018YFB1700104.

REFERENCES

[1] Jawadul H. Bappy, Cody Simons, Lakshmanan Nataraj, B. S. Manjunath, and Amit K. Roy-Chowdhury. 2019. Hybrid LSTM and Encoder–Decoder Architecture for Detection of Image Forgeries. *IEEE Transactions on Image Processing* 28, 7 (2019), 3286–3300. https://doi.org/10.1109/TIP.2019.2895466

[2] Tinne Bay, Herbert Tuytelaars and Luc Van Gool. 2006. SURF: Speeded Up Robust Features. In *Computer Vision – ECCV 2006*. Springer, Berlin, Heidelberg, 404–417. https://doi.org/10.1007/11744023_32

[3] Tinne Bay, Herbert Tuytelaars and Luc Van Gool. 2018. CBAM: Convolutional Block Attention Module. In *Computer Vision – ECCV 2018*. Springer, Cham, 3–19. https://doi.org/10.1007/978-3-030-01234-2_1

[4] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. 2015. Material recognition in the wild with the Materials in Context Database. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3479–3487. https://doi.org/10.1109/CVPR.2015.7298970

[5] Sean Bell, C. Lawrence Zitnick, Kavita Bala, and Ross Girshick. 2016. Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2874–2883. https://doi.org/10.1109/CVPR.2016.314

[6] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. 2018. MaskLab: Instance Segmentation by Refining Object Detection with Semantic and Direction Features. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4013–4022. https://doi.org/10.1109/CVPR.2018.00422

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing Textures in the Wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 3606–3613. https://doi.org/10.1109/CVPR.2014.461

[8] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. 2015. Deep filter banks for texture recognition and segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3828–3836. https://doi.org/10.1109/CVPR.2015.7299007

[9] Maurizio Corbetta and Gordon L. Shulman. 2002. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience* 3, 3 (2002), 201–215. https://doi.org/10.1038/nrn755

[10] Xiyang Dai, Joe Yue-Hei Ng, and Larry S. Davis. 2017. FASON: First and Second Order Information Fusion Network for Texture Recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6100–6108. https://doi.org/10.1109/CVPR.2017.646

[11] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. 2020. Few-Shot Object Detection With Attention-RPN and Multi-Relation Detector. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4012–4021. https://doi.org/10.1109/CVPR42600.2020.00407

[12] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. 2016. Compact Bilinear Pooling. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 317–326. https://doi.org/10.1109/CVPR.2016.41

[13] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. 2020. Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 8 (2020), 2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[15] Zilong Huang, Xinggang Wang, Yunchao Wei, Lichao Huang, Humphrey Shi, Wenyu Liu, and Thomas S. Huang. 2020. CCNet: Criss-Cross Attention for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), 1–14. https://doi.org/10.1109/TPAMI.2020.3007032

[16] Artur Jordao, Ricardo Kloss, and William Robson Schwartz. 2018. Latent HyperNet: Exploring the Layers of Convolutional Neural Networks. In *2018 International Joint Conference on Neural Networks (IJCNN)*. 1–7. https://doi.org/10.1109/IJCNN.2018.8489506

[17] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Hadamard Product for Low-rank Bilinear Pooling. In *The 5th International Conference on Learning Representations*. 1–14.

[18] Tao Kong, Anbang Yao, Yurong Chen, and Fuchun Sun. 2016. HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 845–853. https://doi.org/10.1109/CVPR.2016.98

[19] Alex Krizhevsky. 2009. *Learning Multiple Layers of Features from Tiny Images*. Master's thesis. University of Toronto, Canada.

[20] Qiufu Li, Linlin Shen, Sheng Guo, and Zhihui Lai. 2021. WaveCNet: Wavelet Integrated CNNs to Suppress Aliasing Effect for Noise-Robust Image Classification. *IEEE Transactions on Image Processing* 30 (2021), 7074–7089. https://doi.org/10.1109/TIP.2021.3101395

[21] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. 2017. RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5168–5177. https://doi.org/10.1109/CVPR.2017.549

[22] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. 2018. Bilinear Convolutional Neural Networks for Fine-Grained Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 6 (2018), 1309–1322. https://doi.org/10.1109/TPAMI.2017.2723400

[23] David G. Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60, 2 (2004), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[25] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-scale Image Recognition. In *International Conference on Learning Representations*. 1–14.

[26] Yang Song, Fan Zhang, Qing Li, Heng Huang, Lauren J. O'Donnell, and Weidong Cai. 2017. Locally-Transferred Fisher Vectors for Texture Classification. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 4922–4930. https://doi.org/10.1109/ICCV.2017.526

[27] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local Neural Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7794–7803. https://doi.org/10.1109/CVPR.2018.00813

[28] Chaojian Yu, Xinyi Zhao, Qi Zheng, Peng Zhang, and Xinge You. 2018. Hierarchical Bilinear Pooling for Fine-Grained Visual Recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 574–589.

[29] Hang Zhang, Jia Xue, and Kristin Dana. 2017. Deep TEN: Texture Encoding Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2896–2905. https://doi.org/10.1109/CVPR.2017.309

[30] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. 2020. Exploring Self-Attention for Image Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10073–10082. https://doi.org/10.1109/CVPR42600.2020.01009

[31] Wu Zifeng, Shen Chunhua, and van den Hengel Anton. 2019. Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. *Pattern Recognition* 90 (2019), 119–133. https://doi.org/10.1016/j.patcog.2019.01.006