IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 June 2022, accepted 1 July 2022, date of publication 1 August 2022, date of current version 10 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3195907

RESEARCH ARTICLE

Adopting Attention and Cross-Layer Features for Fine-Grained Representation

SUN FAYOU, HEA CHOON NGO, AND YONG WEE SEK

Centre for Advanced Computing Technology, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Durian Tunggal, Malacca 76100, Malaysia

Corresponding author: Sun Fayou (314565679@qq.com)

ABSTRACT Fine-grained visual classification (FGVC) is challenging task due to discriminative feature representations. The attention-based methods show great potential for FGVC, which neglect that the deeply digging inter-layer feature relations have an impact on refining feature learning. Similarly, the associating cross-layer features methods achieve significant feature enhancement, which lost the long-distance dependencies between elements. However, most of the previous researches neglect that these two methods are mutually correlated to reinforce feature learning, which are independent of each other in related models. Thus, we adopt the respective advantages of the two methods to promote fine-gained feature representations. In this paper, we propose a novel CLNET network, which effectively applies attention mechanism and cross-layer features to obtain feature representations. Specifically, CL-NET consists of 1) adopting self-attention to capture long-rang dependencies for each element, 2) associating cross-layer features to reinforce feature learning,and 3) to cover more feature regions,we integrate attention-based operations between output and input. Experiments verify that CLNET yields new state-of-the-art performance on three widely used fine-grained benchmarks, including CUB-200-2011, Stanford Cars and FGVC-Aircraft. The url of our code is https://github.com/dlearing/CLNET.git.

INDEX TERMS Associating cross-layer features, attention-based operations, self-attention, CLNET.

I. INTRODUCTION

It is one giant leap of image classification with computer vision [37]–[39]. FGVC distinguishes objects of subcategories, e.g.,aircraft models [1], flower species [2], etc. It is gaint value to force on the subtle and discriminative features due to similarity in object appearances [3]–[6], [36]. However, it is a challenging task because of the difficulty of obtaining hidden features. Currently, weakly supervised learning approaches with image-level labels are the typical ways to achieve FGVC, i.e, region proposal methods, attention-based methods and transformer methods.Each method has pros and cons.

Currently, region proposal methods [7]–[11], [19], [33]–[35] rely on the local region proposing to identify the discriminative regions. Fu *et al.* [20] proposed RA-CNN, which can gradually seek out discriminative regions and merge multiple classification results to achieve image classification. Zheng *et al.* [21] proposed MA-CNN, which obtains rich image features by the feature representation of multiple local feature regions. Liu *et al.* [9] proposed filtration and distillation learning (FDL), which describes object-based features learning and region-based features learning as ''teacher'' and ''student'', respectively. FDL [9] provides better supervision for region-based features learning. However, these methods rely on complex algorithms to select discriminative parts,which makes the network difficult to train.

On the other hand, attention-based methods are gaint leap for FGVC,which utilizes attention operations to get better classification results. One strategy is the model with fixed network structure, e.g., StackedLSTM [8], TASN [12], which hinders the availability in practical use. Another strategy is to use attention mechanism to design generic block [13], [14], which can be integrate CNN conveniently. However, those generic blocks can only improve deep layers feature extraction,but perform poorly in shallow layers. On the contrary, the

The associate editor coordinating the review of this manuscript and approving it for publication was Khoa Luu.

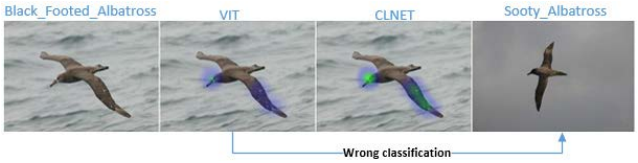

FIGURE 1. A comparison of CLNET50 and VIT on CUB-200-2011 [29].

shallow layers contain rich spatial features. Thus, different layers features are complementary. Unfortunately, existing methods neglect that the interaction of cross-layer features or only simply concatenate features [15].

Recently, transformer was used in FGVC [17], [22], [23], which relies on raw multiply self-attention weights to learn the discriminative features and performs very well. These show that the long-range dependencies among elements are great value for FGVC. However, we find that methods are still in learning region proposing. Specially, methods demand to initialize the size of the image patch with sliding window, which may split the discriminative regions to harm local region proposing. Moreover, methods only utilize K tokens with the maximum value and give up other tokens, which causes the loss of complementary parts.

To address above challenges, we propose a novel FGVC model CLNET to reinforce the discriminative features extraction. We observe that self-attenton method and cross-layer method can mutually reinforce fine-grained feature learning. As a result of this, the CLNET consists of self-attention module, high-level features module and associating cross-layer features module. First, the self-attention module is non-local block [26], which can obtain long-distance dependencies for each element. Second, the high-level feature module is DBT block [13], which can effectively learn deep semantic representations. Finally, associating cross-layer features module is R^3Net [27], which integrates different layers features to achieve feature representation saliency enhancement.

To the best of our knowledge, the discriminative features of the object will be lost in the propagation CNNs. To minimize the loss of valuable information,we adopt a combination of self-attention and cross-layer to achieve enhanced feature representations. In order to exploit the dependencies between all pixels, we use non-local block [26], [13] to reinforce deep feature representations. To utilize the complementarity of different convolution layer, we employ R^3Net [27] to achieve discriminative feature learning. Meanwhile, we are the first to adopt the self-attention mechanism and cross-layer features to achieve FGVC. Our CLNET outperforms existing vision transformer VIT [16], TransFG [17] models on the benchmark datasets, as shown in **FIGURE 1**. Our contributions are summarized as follows:

1) We propose a novel model CLNET which demonstrates the effectiveness of associating self-attention mechanism with cross-layer features.

2) Currently, CLNET still outperforms vision transformers in FGVC tasks.

II. RELATED WORK

In this section, we review the existing FGVC works, which are relevant to our research. To overcome the challenging for FGVC, the research methods are composed of attention models and feature fusion models.

A. ATTENTION METHODS

Feature learning is an important role for FGVC. Due to the subtle differences among subcategories, we only utilize CNN to extract deep semantic features, which hinders further the representation learning. To address above problems, Lin *et al.* [3] proposed bilinear pooling model, which adopts two CNN to get the pairwise feature, and then uses outer product to generate high dimensional vector. Hu *et al.* [18] proposed SENET, which caculates the weight of each channel to enhance significant features for realizing feature recalibration. Zheng *et al.* [12] adopted trilinear attention module to extract fine attention map and designed an attention-based sampler to highlight the discriminative regions. Woo *et al.* [14] proposed the Convolutional Block Attention Module (CBAM) model, which is general module that combines spatial attention and channel attention. Zheng *et al.* [13] proposed the deep bilinear transformer (DBT), which learns fine-grained feature representation by semantic grouping and intra-group interaction,and CNN performs well with DBT blocks. Note that Dosovitskiy *et al.* [16] proposed VIT, which is first to apply transformer to image classification.Meanwhile, He *et al.* [17] proposed the first transformer model of FGVC (i.e.TransFG), which uses the raw attention weights to select the discriminative regions of the image. However, attention-based generic blocks, e.g.,CBAM,DBT,etl. merely utilize deep semantic information,and VIT-based models hava to face super large-scale training dataset. Instead, our model with non-local [26] operator achieves global attention,w hich is a flexible network framework.

B. FEATURE FUSION METHODS

Due to the success of resnet [28], resnet-based models are widely used in visual tasks. while one convolution layer contains limited discriminative features for FGVC, thus researchers try to utilize multi-layer features for feature extraction. These approaches rely on the interaction of cross-layer features to increase attention to the region of interest. In general, low-level contains rich spatial features and the object location is accurate. On the contrary, high-level is only rich in semantic features. Long *et al.* [24] used feature representations of different convolutional layers to achieve better image segmentation. Yu *et al.* [15] proposed HBP, which refines the feature representation capabilities by cross-layer bilinear pooling. Qi *et al.* [25] proposed high resolution remote sensing image road extraction algorithm based on multi-feature fusion, which fuses spectral features with spatial features to improve the recognition performance of road meshes. However, inter-layer merely simply linear

TABLE 1. An illustration of integrating non-Local and DBT into resnet50.

Stage	Output	ResNet-50	DBTNET-50	CLNET50
I	112×112	7 × 7, 64, stride 2		
II	56 × 56	3 × 3 max pool, stride 2		
		$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\\\text{Non-}localblock\end{bmatrix}$
III	28 × 28	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\\\text{Non-}localblock\end{bmatrix}$
IV	14 × 14	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}\text{DBT block}\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}\text{DBT block}\\3\times3,256\\1\times1,1024\\\text{Non-}localblock\end{bmatrix}$
V	7 × 7	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}\text{DBT block}\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}\text{DBT block}\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$

pooling calculations are unable to get sufficient feature relationship. Therefore, our model refine cross-layer feature for saliency detection with R3NET [27].

III. METHOD

In this section, we introduce the proposed CLNET, which contains three modules, i.e., long-range dependencies module, deep semantic information extraction module, and associating cross-layer feature module.

An overview of the proposed CLNET is show in **FIGURE 2.** Note that we show the framework of the bakbone in Table 1. From **FIGURE 2,** it can be observed that we integrate non-local blocks [26] and DBT blocks [13] in the resnet [28] as backbone. Subsequently, the integrated feature consists of 1) refine each layer feature, 2) integrate high-level features(H) and integrate low-level features(L).Next,we take H and L as R^3Net [27] inputs to reinforce the salient features. Meanwhile, R^3Net uses H as the saliency feature, which shows that applying supervision signals to H can obtain better saliency features. Finally, the last saliency map (S_n) is used as input to the fully convolutional network (FCN [24]) to get the classification results via softmax.

A. LONG-RANGE DEPENDENCIES MODULE

If the global features can be effectively used in the FGVC, the image classification performance can be further improved. To address this problem, we utilize non-local block [26], which captures the long-rang dependencies between any two positions. Furthermore, experiments show that non-local block [26] can be integrated into resnet [28] and perform well, so it will be good to enhance feature learning. Specifically, the non-local block is as follow:

$$z_i = w_z y_i + x_i \tag{1}$$

where y_i is a non-local operation, x_i is an input feature, w_z is a convolution operation and the output channels are equal to x_i,z_i is a residual connection. The detailed calculation process is shown in **FIGURE 3.**

B. DEEP SEMANTIC INFORMATION EXTRACTION MODULE

Currently, attention mechanism plays important role for deep feature representations. Typically, DBT [13] adopts semantic grouping and intra-group bilinear interaction to promote feature learning. To verify the performance of DBT, ablation studies on integrated stages show that DBT block is insufficient to get low-level features, whereas it is effective for the extraction of high-level features. Concurrently, dbtnet [13] is a model built by integrating DBT on resnet. Thus, we merely add non-local blocks on dbtnet [13] as backbone. Meanwhile, the loss function of DBTnet is as follow:

$$L_{DBTNET} = L_c + \lambda \sum_{b}^{B} L_g^{(b)} \tag{2}$$

C. ASSOCIATING CROSS-LAYER FEATURE MODULE

1) OPTIMIZED INTEGRATED FEATURES

As far as we know, we are the first to apply R^3Net [28] to the FGVC. Specifically, R^3Net only uses upsampling feature maps and concatenation feature maps. To alleviate the issue, we use attention mechanism, which can better restrain the features with weak correlation and enhance the features with strong correlation. To be specific, we utilize convolutional block attention module [14] (CBAM) to achieve better fine-grained feature representation, as shown below

$$\begin{aligned} O^1 &= M_c(I) \otimes I \\ O^2 &= M_s(O^1) \otimes O^1 \end{aligned} \tag{3}$$

where $\otimes$ represents element-wise multiplication, $I \in Rc^*h^*w$ is input feature map, $Mc \in Rc^*1^*1$ is channel feature map, $Ms \in R1^*H^*W$ is spatial feature map, O2 is final output. In short, the calculation formulas of Mc and Ms are:

$$M_c(I) = \sigma((MLP(AvgPool(I))) + MLP(MaxPool(I))) \tag{4}$$

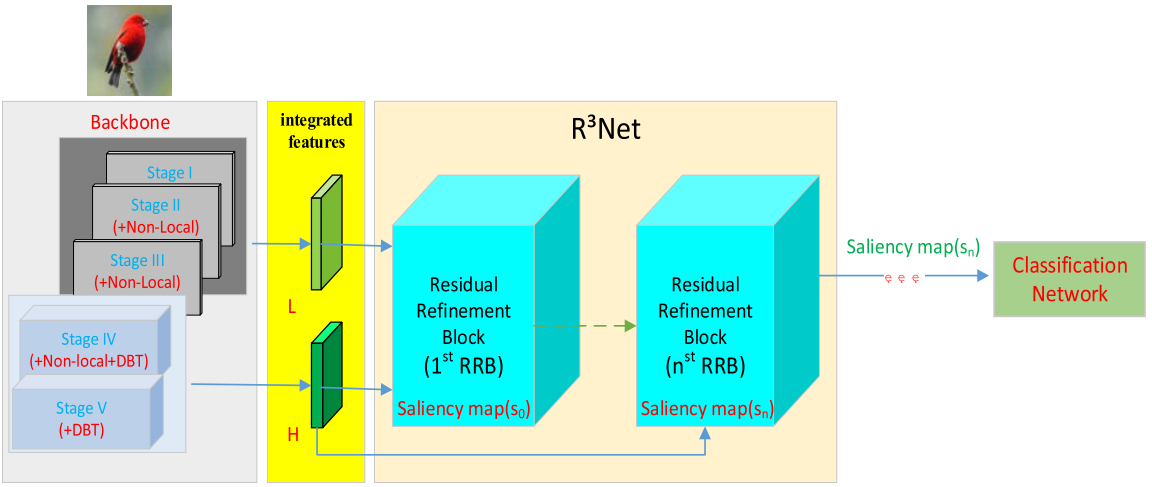

FIGURE 2. Overview of the proposed CLNET. Resnet is the backbone. The low-level features consist of stage I, stage II and stage III, and the high-level features consist of stage IV and stage V.L is low-level Integrated Features.H is high-level Integrated Features. Non-blocks are integrated into stage II, stage III and stage IV, and DBT blocks are integrated into stage IV and stage V. RESNET integrated with non-blocks and DBT blocks is the backbone network.

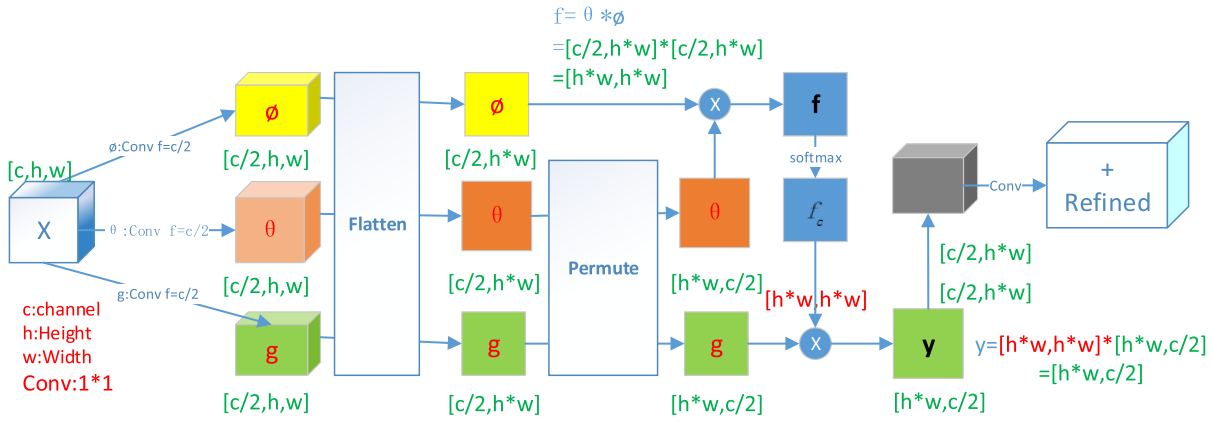

FIGURE 3. The computational process of non-local block.

where σ denotes the sigmoid function,MLP is multi-layer perceptron.

$$M_s(O^1) = \sigma(f^{7\times7}([AvgPool(O^1)]; [MaxPool(O^1)])) \quad (5)$$

where f 7*7 is a convolution operation with the filter size of 7*7.

2) SALIENT FEATURE NETWORK

Saliency detection aims to search salient region in the image, i.e.,the region of interest RoI). In this paper, we use R^3Net.

The optimized network structure is shown in **FIGURE 4.**

D. NETWORK ARCHITECTURE

We propose that non-local blocks and DBT blocks can be integrated into resnet, as shown in **Table 1**.

IV. EXPERIMENTS

In this section, we evaluate and analyze the performance of CLNET on three fine-grained benchmarks.

TABLE 2. Detailed statistics of the three datasets used in this paper.

Dataset	Total	Class	Train	Test
CUB-200-2011[29]	11788	200	5994	5794
Stanford Cars[30]	16185	196	8144	8041
FGVC-Aircraft[1]	10200	102	6667	3533

A. EXPERIMENTS SETUP

1) DATASETS

To evaluate the effectiveness of CLNET, we conducted experiments on tree widely used datasets, including CUB-200-2011 [29], Stanford Cars [30] and FGVC-Aircraft [1]. The detailed description of quantity,category numbers and the standard training/testing splits can be found in **Table 2**.

2) IMPLEMENTATION

The pytorch was used as deep learning freamwork.Our CLNET is trained on 3 GPU (i.e, GeForce RTX 2070 8GB). We adopt the common setting to pre-train CLNET on

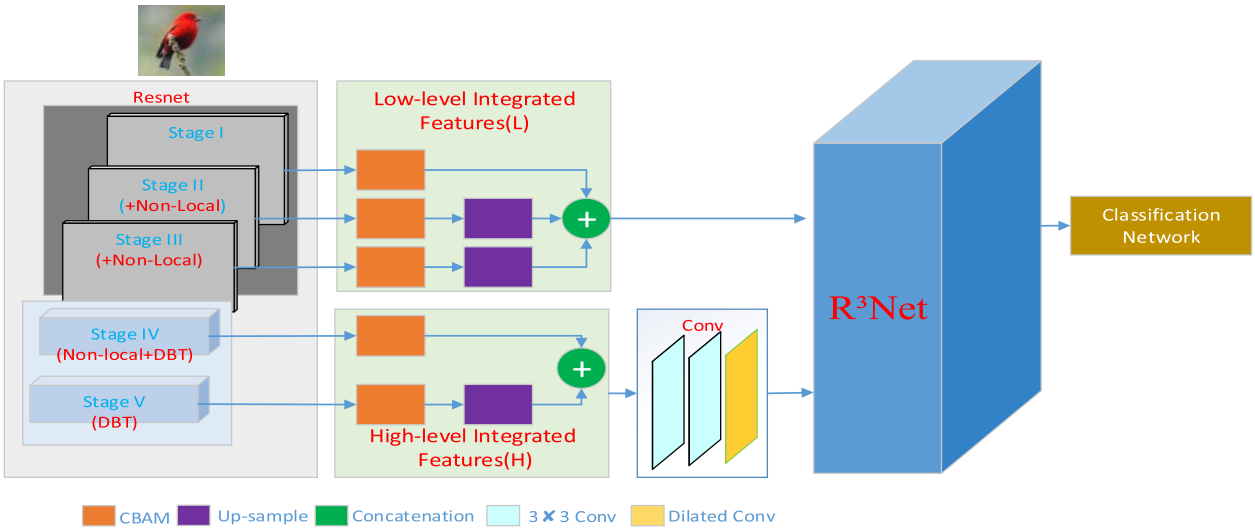

FIGURE 4. An illustration of optimized integrated features.

ImageNet [31]. To speed up the training process and reduce the over-fitting issue,we utilize the well-trained R^3Net [27] on MSRA10K. It can be observed that the loss function of CLNET is as follows:

$$L = L_{DBTNET} \tag{6}$$

Concurrently, referring to methods of most FGVC models,we used SGD optimizer without momentum and weight decay,and the batch size was set to 48. Note that for a fair comparison, loss function(L) has the same parameters with DBTNET [13].

B. PERFORMANCE COMPARISON

To verify the advantage of CLNET,we compared it with other state-of-the-art (SOAT) networks on three benchmark datasets. From **Table 3**, it can be observed that CLNET achieves SOAT competitive performance on CUB-200-2011 [29], Stanford Cars [30] and FGVC-Aircraft [1].

Specifically, the 3th column of Table 3 shows the comparison results on CUB [29]. For resnet-101 based method, we compare CLNET101 to the SOAT StackedL STM,and achieve 2.7% improvements. Moreover, CLNET101 gets 1.4% performace gain compared to TransFG [17] and reaches 93.1% accuracy. HBP [15] uses hierarchical bilinear pooling to extract discriminant features, which ignores low-level features. FDL [9] utilizes the filtration learning with discrimination matching method to locate discriminative regions, which ignores the associations between global features of the image. StackedLSTM [8] uses LSTM for image classification, and its fixed network structure hinders application in practice. TransFG uses raw attention weights to select discriminant regions of the image, whereas the initial image segmentation ignores the association between pixels. However, our CLNET maintains simplicity and robustness.

Then, we analyze the results of the cars [30] experiment, which shows that CLNET is in absolute advantage. Meanwhile, we observe that the models that use CNN as the backbone perform better than transform in this dataset. Directly, our model improves by 1.9% compared to TransFG [17] in terms of accuracy metric even if the backbone network is resnet-50. Our analysis is related to the fact that the image noise in this dataset is less, and the discriminative features of the image are easy to extract.

Similarly, the performance of CLNET in aircraft dataset is also excellent. Due to the subtle differences between the objects in this dataset, image classification is difficult. Currently, a few FGVC models use this dataset for experiments. MACNN [21] obtains feature regions by generating multiple significant feature regions. FDL [9] conducts region proposing via filtration learning. Experiments show that the accuracy of CLNET50 is improved by 1.69% compared with FDL and reaches 95.06% accuracy.

Since the non-local blocks and cross-layer features fusion are adopted,the efficiency of the CLNET should be analyzed. We use memory size and the efficiency of images processing as metrics to compute the complexity of the proposed method.

From **Table 4**, it can be seen that the performance of CLNET are lower than that of renet50. We believe that this is because of the relatively complex network structure of CLNET. Concurrently, we can see that CLNET has completely surpassed the vision transformer which also proves the superiority of CLNET.

C. ABLATION STUDIES

We conducted ablation studies on CLNET50 to illustrate the impact of different model structures on accuracy metric. Ablation studies have the same effects on the three datasets, so these experiments were done only on the CUB-200-2011 [29] dataset.

TABLE 3. Comparison results on CUB-200-2011, stanford cars, FGVC-aircraft.

Method	Backbone	Accuracy(%)		
		CUB-200-2011	Stanford Cars	FGVC-Aircraft
RACNN[20]	VGG19	85.2	92.5	-
MACNN[21]	VGG19	86.5	92.8	89.9
TASN[12]	VGG19	86.1	92.4	-
FDL[9]	VGG19	86.84	91.52	-
MAMC[32]	Resnet-50	86.2	93.0	-
NTS-Net[10]	Resnet-50	87.5	93.3	91.4
DBTNet[13]	Resnet-50	87.5	94.1	91.2
TASN[12]	Resnet-50	87.9	93.8	-
DBTNet[13]	Resnet-101	88.1	94.5	91.6
FDL [9]	DenseNet161	89.09	94.02	91.27
FDL [9]	Resnet-50	-	-	93.37
HBP[15]	VGG-16	87.1	93.7	90.3
StackedLSTM[8]	GoogleNet	90.4	-	-
ViT[16]	ViT-B_16	90.3	93.7	-
TransFG[17]	ViT-B_16	91.7	94.8	
CLNET50	Resnet-50	92.4	96.7	95.06
CLNET101	Resnet-101	93.1	97.4	-

TABLE 4. Performance comparison results on CUB-200-2011.

	images/s	model size	Acc.(%)
Resnet50[28]	2301	23.92M	85.49
FDL [9]	1791	28.51M	88.35
HBP[15]	1826	17.5M	87.15
VIT[16]	799	86.6M	90.3
CINET	1672	46.3M	92.4

TABLE 5. DBTNET50 adds non-local block at different stages.

Stage	Accuracy (%)
baseline	85.1
I	85.4
II	86.1
II +III	87.5
II +III +IV	89.9
II +III +IV+V	90.1

TABLE 6. CLNET50 adds non-local block at different stages.

Stage	Accuracy (%)
baseline	89.6
I	89.7
II	90.3
II +III	91.5
II +III +IV	92.4
II +III +IV+V	92.6

TABLE 7. Comparison of CBAM [14] in classification accuracy.

Block	Accuracy (%)
no cbam	91.5
+cbam	92.4

1) NON-LOCAL BLOCKS

It is important to verify the effectiveness of non-local blocks in CLNET.

First, in order to clarify the power of non-local blocks, we add non-local block into DBTNET50 [13] to demonstrate the influence on deep feature learning.

In **Table 5**, baseline is resnet50 without non-local blocks. We can observe that deeply integrating non-local block into stage II and stage III brings 0.7% accuracy gains compared to baseline. Meanwhile, integrating non-local block into stage V can not significantly improve the performance. Thus, we integrate non-local block into Stage II, Stage III and Stage IV in DBTNet50 [13].

Secondly, to further verify the influence of non-local block on the model classification results, we conduct more experiments with non-local blocks adding to different stages.

In **Table 6**, we define the baseline as CLNET50 without non-local blocks. We can observe the impact of adding non-local blocks at different stages on the classification results. Specifically, if we add non-local blocks at every stage, the accuracy can improve by 3.0%. Even if we only add non-local blocks in Stage II, there is 0.7% improvement over baseline. However, it can be observed that integrating non-local blocks into stage I and stage V cannot significantly improve the performance. Our analysis is that stage I has less semantic information and stage V has fewer low-level features. In addition, adding blocks to clnet50 will increase the amount of computation.Thus we abandon the blocks in stage I and stage V.

2) OPTIMIZED INTEGRATED FEATURES

In order to obtain more discriminative features, we use channel attention and spatial attention.Thus,our model integrates CBAM [14] blocks. From **Table 7**, it can be observed that integrating CBAM [14] blocks in the model can improve by 0.9% accuracy gains,which means that CBAM blocks improve feature represent ions.

As each module has different functions,the order of CBAM blocks may affect the overall performance. From **Table 8**,

FIGURE 5. Visualization experiments of attention map. The first row shows input images, the second row shows attention map of DBTnet50, the third row shows the results of DBTnet50 + Non_local blocks, and the fourth row shows the results of CLNET50.

TABLE 8. The impact of CBAM blocks location on classification metric.

Location	Accuracy (%)
before integrated features	92.4
after integrated features	92.33

TABLE 9. The impact of associating cross-layer features.

Method	Accuracy (%)
baseline	90.4
+R^{3}Net	92.4

TABLE 10. Comparison of different integrated features as saliency map.

Method	Accuracy (%)
Low-level Integrated Features(L)	91.7
High-level Integrated Features(H)	92.4

it can be observed that CBAM-first achieves a 0.07% improvement, but the impact is not significant.

3) CROSS-LAYER FEATURE FUSION

In order to show the advantages of the associating cross-layer features, R^{3}Net [27] is used to illustrate the performance gain.

In **Table 9**, the DBTnet50 [13] network integrates non-blocks as baseline. From **Table 9**, it can be observed that the associating cross-layer features through R^{3}Net [27] can improve by 2% gains. Thus different levels of features can be mutual benefit and enhance the information of the region of interest.

From **Table 10**, we know that H as saliency map brings 0.7% accuracy gains.It confirms that high-level features retain more discriminative features.

D. VISUALIZATION EXPERIMENTS

We randomly select two images from each dataset. The visualization result of CLNET50 is shown in **FIGURE 5**. To investigate the advantages of the CLNET50, we conduct experiments by gradually integrating different modules.

Specifically, the 3th row shows that the model identifies multiple discriminative parts of the object. Instead, CLNET50 is able to focus multiple attention regions and reinforce the feature representation as shown in 4th row.

V. CONCLUSION

In this work, we propose a novel FGVC model CLNET to enhance the learning of fine-grained features. Currently, we are the first to propose using global features to refine semantic features, and associate cross-layer features to reinforce saliency features.Extensive experiments demonstrate that CLNET is able to achieve state-of-the-art performance on various FGVC tasks. In addition, visualization experiments prove the interpretability and effectiveness of the model.

With the results achieved by CLNET, it shows the great potential of adopting attention and cross-layer features in FGVC tasks. Since the complex model structure of CLNET, we will study on using methods such as knowledge distillation to compress model to further improve efficiency. In addition,we will explore utilizing cross-layer feature fusion in vision transformers.

REFERENCES

[1] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," 2013, *arXiv:306.5151*.

[2] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2008, pp. 722–729.

[3] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1449–1457.

[4] T.-Y. Lin and S. Maji, "Improved bilinear pooling with CNNs," 2017, *arXiv:1707.06772*.

[5] P. Li, J. Xie, Q. Wang, and W. Zuo, "Is second-order information helpful for large-scale visual recognition?" in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2070–2078.

[6] P. Li, J. Xie, Q. Wang, and Z. Gao, "Towards faster training of global covariance pooling networks by iterative matrix square root normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 947–955.

[7] Y. Ding, Y. Zhou, Y. Zhu, Q. Ye, and J. Jiao, "Selective sparse sampling for fine-grained image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6599–6608.

[8] W. Ge, X. Lin, and Y. Yu, "Weakly supervised complementary parts models for fine-grained image classification from the bottom up," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3034–3043.

[9] C. Liu, H. Xie, Z.-J. Zha, L. Ma, L. Yu, and Y. Zhang, "Filtration and distillation: Enhancing region attention for fine-grained visual categorization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 11555–11562.

[10] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and A. L. Wang, "Learning to navigate for fine-grained classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 420–435.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[12] H. Zheng, J. Fu, Z. J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5012–5021.

[13] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Learning deep bilinear transformation for fine-grained image representation," 2019, *arXiv:1911.03621*.

[14] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[15] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, "Hierarchical bilinear pooling for fine-grained visual recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 595–610.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[17] J. He, J. N. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, and C. Wang, "TransFG: A transformer architecture for fine-grained recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 852–860.

[18] J. Hu, L. Shen, and G. Sun, "Squeeze- and-excitation networks," 2019, *arXiv:1709.01507*.

[19] S. Fayou, H. C. Ngo, and Y. W. Sek, "Combining multi-feature regions for FineGrained image recognition," *Int. J. Image, Graph. Signal Process.*, vol. 14, no. 1, pp. 15–25, Feb. 2022.

[20] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4438–4446.

[21] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5209–5217.

[22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[23] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6881–6890.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[25] Y. Yang, Y. Li, and Q. H. Zhao, "High resolution remote sensing image road extraction algorithm based on multi-feature fusion," *J. Transp. Syst. Eng. Inf. Technol.*, vol. 20, no. 1, pp. 111–116, 2020.

[26] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[27] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, "R^3Net: Recurrent residual refinement network for saliency detection," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 684–690.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[29] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200–2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.

[30] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Sydney, NSW, Australia, Dec. 2013, pp. 554–561.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, R. Karpathy, A. Khosla, M. Bernstein, C. Alexander Berg, and L. Fei-Fei, "ImageNet Large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[32] M. Sun, Y. Yuan, F. Zhou, and E. Ding, "Multi-attention multi-class constraint for fine-grained image recognition," in *Proc. ECCV*, 2018, pp. 805–821.

[33] X. He and Y. Peng, "Fine-grained image classification via combining vision and language," 2017, *arXiv:1704.02792*.

[34] X. He and Y. Peng, "Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4075–4081.

[35] X. He, Y. Peng, and J. Zhao, "Which and how many regions to gaze: Focus discriminative regions for fine-grained visual categorization," *Int. J. Comput. Vis.*, vol. 127, no. 9, pp. 1235–1255, Sep. 2019.

[36] X. He, Y. Peng, and L. Xie, "A new benchmark and approach for fine-grained cross-media retrieval," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1740–1748.

[37] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2020.

[38] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

[39] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

SUN FAYOU is currently pursuing the Ph.D. degree with the Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka. His research interests include computer vision, generative adversarial networks, and Information network security.

HEA CHOON NGO received the bachelor's degree in computer science (software development) from Universiti Teknikal Malaysia Melaka (UTeM), in 2004, the master's degree in information technology from The University of New South Wales (UNSW), Sydney, in 2007, and the Ph.D. degree in computer science from Universiti Sains Malaysia (USM), in 2016. He is currently a Senior Lecturer with the Department of Intelligent Computing and Analytics, Faculty of Information and Communication Technology, UTeM, where he is also a Faculty Member. He is also a member of the Computational Intelligence and Technologies Laboratory, under the Centre for Advanced Computing Technology, UTeM. His research interests include computational intelligence, data science and analytics, planning and scheduling, optimization, health informatics, and intelligent systems.

YONG WEE SEK received the bachelor's degree in statistics from Unversiti Kebangsaan Malaysia (UKM), in 1999, the master's degree in information technology from Universiti Putra Malaysia (UPM), in 2001, and the Ph.D. degree in business information system from RMIT University Melbourne, Australia, in 2017. He is currently a Senior Lecturer with the Department of Intelligent Computing and Analytics, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM). He is also a member of the Computational Intelligence and Technologies Laboratory, under the Centre for Advanced Computing Technology, UTeM. His research interests include operation research, information systems, web based and multimedia learning, and mathematics.