Inception Inspired Attention Hybrid Network for efficient Long-Range Dependency

*Haruna Yunusa[1] {yunusa2k2@buaa.edu.cn}, Qin Shiyin[1], Abdulrahman Hamman Adama Chukkol[2], Isah Bello[3], Adamu Lawan[1], Abdulganiyu Abdu Yusuf [2]

[1] Beihang University, [2] Beijing Institute of Technology, [3] Tianjin University, China

Abstract. The recent emergence of hybrid models has introduced another transformative approach to solving computer vision tasks, slowly shifting away from conventional CNN (Convolutional Neural Network) and ViT (Vision Transformer). However, not enough effort has been made to efficiently combine these two approaches to improve capturing long-range dependencies prevalent in complex images. In this paper, we introduce iiANET (Inception Inspired Attention Network), an efficient hybrid model designed to capture long-range dependencies in complex images. The fundamental building block, iiABlock, integrates *global* 2D-MHSA (Multi-Head Self-Attention) with Registers and convolutional layers in parallel, enabling the model to effectively leverage self-attention for capturing long-range dependencies while utilizing convolutional operations for local-detail extraction and expanded contextual understanding. Lastly, we serially integrate an ECANET (Efficient Channel Attention Network) at the end of each iiABlock to calibrate channel-wise attention for enhanced model performance. Extensive qualitative and quantitative comparative evaluation on various benchmarks demonstrates improved performance over some state-of-the-art models.

Keywords: Attention mechanism, Convolutional Neural Network, Long-range dependency, Registers

1 Introduction

The last decade has seen the rise of deep Convolutional Neural Network (CNN) architectures as the de facto standard for solving the majority of computer vision (CV) tasks, which include image classification [1, 2], object detection [3, 4] and segmentation [5] with compelling results. The prevalence of CNN architectures is not coincidental as they excel at capturing spatial features and patterns in images. However, the dominance of CNN architectures is being challenged by the emergence of ViT (Vision in Transformer) [6], presenting a transformative approach to solving CV tasks. Interestingly, this groundbreaking model outperforms many current state-of-the-art CNN-based models on the ImageNet benchmark [6] and emerges as a competitive alternative [7]. Practically, ViT works exactly like the text-based Natural Language Processing (NLP) transformers but with patch embedding. It divides the input image into patches, projects them into a high-dimensional feature space through a linear projection layer, adds positional embedding, passes them through a transformer encoder, and finally maps the output to a fixed-length vector for classification tasks.

Significantly, the key component of ViT is the self-attention mechanism [6] within the encoder, which enables the model to capture long-range dependencies by allowing each element in the input sequence to attend to all other elements, considering their relative importance [6].

While this capability allows the model to selectively focus on distantly related pixels, facilitating the efficient capture of contextual information across the entire input sequence, it encounters limitations such as increased computational complexity, reduced interpretability, data hungry and challenges in handling spatial information effectively compared to CNNs [8]. In contrast, CNN-based models, while effective at capturing local features through parameter sharing and local receptive fields, struggle with capturing long-range dependencies, limiting their ability to integrate distant pixel relationships [8]. These limitations have led to the development of hybrid models, which combine their strengths to improve performance [9].

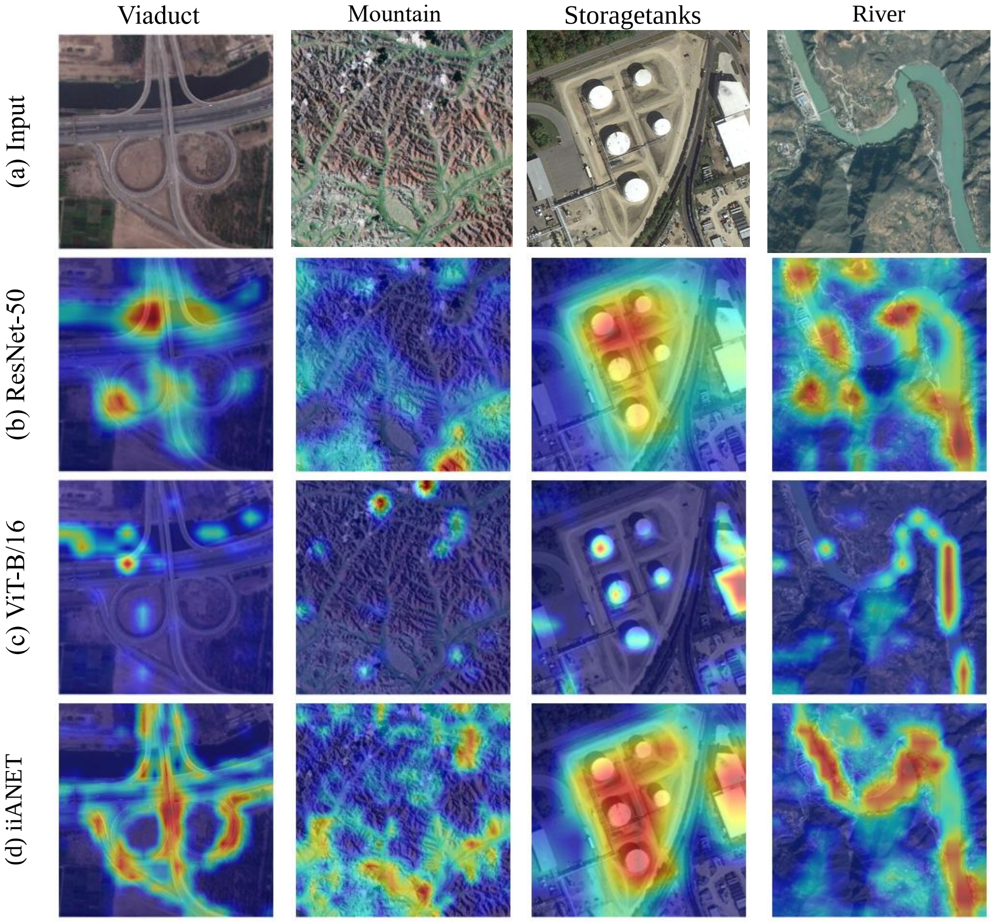


Fig. 1. Grad-CAM comparison between iiANET and other state-of-the-art models, e.g. (a) shows an aerial image of viaduct, mountain, storage tanks and river featuring complex infrastructure consisting of multiple spans, roads, and surrounding landscapes. The primary objective is to accurately detect and classify various elements to facilitate efficient maintenance, safety management, and infrastructure planning. Consequently, capturing long-range dependencies in this scenario is crucial for comprehending the spatial layout of different viaduct, mountain, storage tanks and river their interactions, and potential structural issues. (b) illustrates ResNet-50 inability to capture long-range dependencies but local features (c) demonstrate the limitations of ViT-B/16 interpretabilities, as it primarily highlights tiny spots on the image. (d) is a hybrid model, depicting notable improvements in capturing long-range dependencies, global context and improved interpretability.

Specifically, previous hybrid designs have aimed to enhance capturing long-range dependencies for various CV tasks [10, 11, 12]. However, the design of hybrid models introduces additional design complexities [8], and computational costs compared to monolithic models [9], while also potentially leading to information loss due to feature fusion of distinct models [8]. Lastly, not much effort has been given to designing efficient hybrid models for capturing long-range dependencies in complex images, which remains largely unaddressed.

This study introduces a novel hybrid model, termed iiANET (Inception Inspired Attention Network), designed to address the limitations of existing monolithic models and enhance the capture of long-range dependencies in complex images. Drawing inspiration from the Inception [18] model, iiANET integrates parallel CNN modules and *global* MHSA (Multi-Head Self-Attention) mechanism with Registers to efficiently capture long-range dependencies. By leveraging the complementary strengths of CNNs and transformers, iiANET aims to provide a simple yet effective solution for enhancing the understanding of complex visual scenes with long-range dependencies e.g., classification on the AID (Aerial Image Dataset) [24], iiANET demonstrates superior performance with an accuracy of 80.57%, outperforming ResNet-50, ViT-B/224, and DiNAT-B with accuracies of 71.93%, 69.93% and 79.12% respectively. This highlight iiANET effectiveness in capturing long-range dependency in challenging datasets.

The contributions of this paper are summarized as follows:

- A novel, efficient hybrid model for capturing long-range dependencies in complex images while maintaining less computational cost, and first hybrid model to integrate Registers.
- iiANET design is adaptable, can serve as backbone for detection and segmentation tasks.
- Extensive experimental results on commonly used benchmarks demonstrate that iiANET outperforms some existing state-of-the-art methods.

2 Related Work

CNN Methods. There has been previous attempt to enhance CNNs ability to capture long-range dependencies in images for various tasks. Donahue et al. (2015) [13] introduces Long-term Recurrent Convolutional Network (LRCN) by fusing CNNs and LSTM to improve capturing long-range dependencies in images, in contrast, Yu et al. (2017) [14] propose Dilated Residual Network (DRN) using several dilation rates to expand the receptive fields for capturing further intricate details in images, and Yu & Koltun, (2016) [15] designs a model based on the Dilated Convolution (DC) for improving the capturing of global context in semantic segmentation. These unique approaches have significantly contributed to advancing the capacity of CNNs in capturing long-range dependencies. However, LRCN shows increased computational complexity due to recurrent connections. DRN struggles with losing fine-grained spatial details due to variation in dilation rates and DC introduces gridding artifacts due to dilations.

ViT Methods. The emergence of ViT models [6, 31] has provided another groundbreaking approach to solving CV tasks, achieving state-of-the-art results and significantly advancing the field. Due to the attention mechanism within the ViT encoder, it excels at capturing long-range dependencies. However, the attention mechanisms within ViT exhibit quadratic computational complexity demanding substantial computational resources and data, due to their inherent weak inductive bias compared to CNN models.

Hybrid Methods. Inspired by the limitation of both CNN and ViT models, this approach aims to improve the performance of CV tasks by combining the strengths of CNN feature extraction and ViT ability to capture long-range dependencies [8]. Zhang et al. (2022) [16] design an efficient long-range attention network (ELAN) that utilizes a group-wise multi-scale self-attention mechanism to capture long-range dependencies in super-resolution tasks. Guo et al. (2021) [10], on the other hand, introduced CMT (CNN Meet Vision Transformers) to enhance the capture of long-range dependencies by integrating ViT into the CMT blocks. Srinivas et al. (2021) [11] proposed BoTNet (Bottleneck Transformers) to achieve similar goals by replacing the convolutional block with MHSA in Resnet last block. However, both ELAN and BoTNet introduce the self-attention mechanism at later stages of the network, where the spatial dimension is smaller, resulting in a trade-off between efficiency and effectiveness compared to earlier stages with wider spatial features. Additionally, CMT-L faces memory usage limitations. Furthermore, while various other designs have been proposed with significant improvements in accuracies [27, 28, 29, 30], these methods are structurally complex, leading to information loss due to fusing distinct methods and increased computational costs.

CNN vs RNN vs Self-Attention. CNNs are well-suited for capturing local short details in images, but they struggle with long-range dependencies [8, 9]. RNNs are effective for capturing long-range dependencies, but their sequential design makes them less parallelizable, leading to slower training times [17]. ViTs, with their self-attention mechanism, excel in capturing long-range dependencies and are highly parallelizable. However, they demand more memory due to the self-attention mechanism [6]. To the best of our knowledge, there has not been much effort in exploring the effective combination of CNN and ViT to enhance the capture of long-range dependencies in complex images. This lack of exploration may be attributed to several factors, including the increased design complexity of combining distinct models, higher computational costs compared to monolithic models, potential information loss due to fusing distinct features, and interpretability issues. In this paper, we aim to explore the potential of designing a hybrid model that effectively captures long-range dependencies in images.

3 Method

3.1 iiANET Architectural Overview

iiANET, a novel hybrid network architecture, aims to improve the challenge of capturing long-range dependencies prevalent in complex images by leveraging the combination of dilated convolution [15], MBConv2 [19], *global* 2D-MHSA [11] with Register [35], and ECANET [20]. The synergy enables iiANET to effectively capture intricate pattern, local spatial feature, long-range dependencies, and channel-wise dependencies, leading to improved performance.

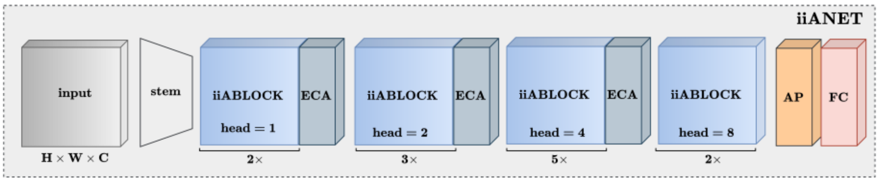


Fig. 2. iiANET architectural overview

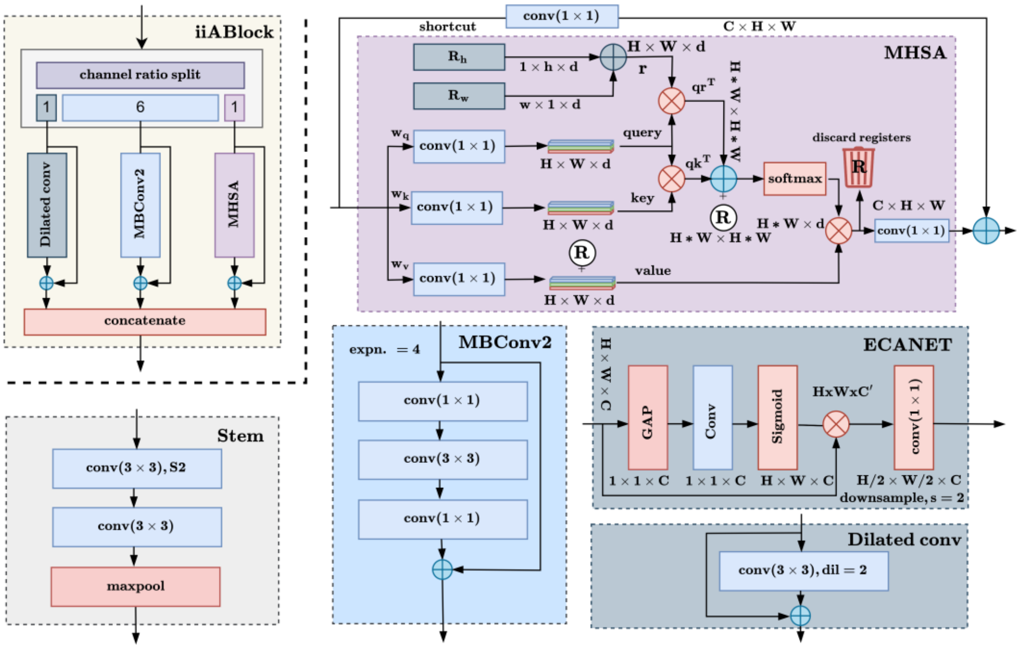


Fig. 3. iiANET architectural components

3.2 iiANET Architectural Components

Stem (*Initial stage*) given the higher resolutions of complex images, this component serves to compress the computational costs of iiANET by shrinking the spatial dimensions of the input image to half, trading off spatial details for improved model efficiency and basic feature extraction [21]. Let the input image be $x \in \mathbb{R}^{H \times W \times 3}$ we apply two sequential 3 × 3 convolutional layers, each followed by a batch-norm and ReLU activation function, with the initial layer using a stride of 2.

MBConv2 (*Efficient convolutional block*) iiABlock utilizes MBConv2 in parallel to improve computational efficiency and enhance local feature extraction, consist of 1×1 convolutions for dimensionality reduction, 3×3 depth-wise separable convolution for spatial information extraction, and 1×1 convolution for projection. Notably, this block is limited to capturing local context with fixed kernel, making it less effective in understanding prevalent global context in complex images [12], e.g. road, viaduct, bridge. Given the depth-wise operation in eqn. (1).

$$y_i = \sum_{j \in L(i)} w_{i-j} * x_j \qquad (1)$$

y_i calculates the output at position i by taking a weighted sum of x_i as input, where $x_i, y_i \in R^D$. The weights w_{i-j} determine the contribution of each input to the output, and $L(i)$ represents a local neighborhood, typically a 3×3 grid centered around i. The small size of $L(i)$ limits the receptive field's ability to capture intricate details, particularly in complex images with prevalent long-range dependencies. To overcome this limitation of MBConv2 in capturing global context, a *global* MHSA has been integrated into the iiABlock to enhance the model's capability to capture long-range dependencies in complex images.

***Global* 2D-MHSA** (*Capturing global context and long-range dependencies*) to improve iiABlock comprehensive contextual understanding across the whole image, we utilize *global* MHSA in 2D, which captures long-range dependencies and contextual relationships. Unlike the standard MHSA, the input is processed as a whole, enabling the MHSA mechanism to capture global dependencies. Given a 2D input feature denoted as $x \in \mathbb{R}^{H\times W\times C}$, we reshape it into $x \in \mathbb{R}^{HW\times d}$, d denote input feature dimensions, then apply a linear transformation to queries $Q = xW_Q$, keys $K = xW_K$ and values $V = xW_V$. To explain how it captures long-range dependencies. Let's consider a single token i attending to all other tokens in the sequence, attention Z with h head is denoted as in equation (2).

$$Z(Q_i, K, V)_h = softmax\left(\frac{Q_i K^T}{\sqrt{d_k^h}}\right)V \tag{2}$$

Here, Q_i interacts with all keys in K across the entire sequence, unlike the standard MHSA, which attends within a limited context window. This enables the model to consider the relationships between all tokens regardless of their positional distance, capturing the global context and long-range dependencies prevalent in complex images. The softmax operation normalizes these scores and produces attention weights, which, when applied to the value matrix $V,$ compute the final attended values. However, this interaction is ordering agnostic and doesn't capture positional relationships in the input sequence. Therefore, in image data where spatial information is essential, integrating positional encodings is necessary to complement the power of MHSA mechanism.

Relative Position Encoding [11] MHSA is permutation equivariant with no positional encoding. This characteristic limit its representational power, particularly for computer vision tasks involving highly structured data like images. Notably, it is added to the input image representation before the MHSA is applied, and it is used to guide the attention weights to focus on relevant pixels based on their relative positions in the input image.

$$Z(Q_i, K, V)_h = softmax\left(\frac{Q_i K^T + Q_i R}{\sqrt{d_k^h}}\right)V \tag{3}$$

Where R is a trainable matrix, lastly, reshape $Z(x)_h$ back to its original spatial shape of $x \in \mathbb{R}^{H\times W\times C}$. This addresses MHSA's order agnostic nature, enhancing its representation power.

Registers (*Improving interpretability*) while self-attention mechanism significantly improves network ability to capture long-range dependencies, it struggles with poor interpretability [35]. We instead add additional learnable tokens to mitigate the prevalent artifacts in the attention mechanism caused by high norms in image areas with low information during inference or training, similar to the implementation by Darcet et al. [35]. In this case, it is a global MHSA where the attention mechanism has a single input image, in contrast to having several patches. We initialize the register tokens for query and keys as $\mathcal{R}_{qk} \in \mathbb{R}^{N\times HW\times \mathrm{HW}}$, where N is the number of register tokens and HW is the spatial dimension, then value register tokens as $\mathcal{R}_v \in \mathbb{R}^{N\times(\frac{D}{\mathrm{head}})\times HW}$, where D is the dimension. The operations expand both $\mathcal{R}_{qk}$, $\mathcal{R}_v$ to $\mathbb{R}^{B\times N\times \mathrm{H}\times W}$ and $\mathbb{R}^{N\times\frac{D}{head}\times HW}$, respectively, eqn. 4, 5, effectively creating B copies of the register tokens for each batch, where B is the batch size. This step ensures that each batch has its own set of register tokens, facilitating batch-wise parallel processing in the attention mechanism.

$$\mathcal{R}_{qk} = \mathrm{repeat}(\mathcal{R}'_{qk}, nhw \rightarrow bnhw', b = B) \tag{4}$$

$$\mathcal{R}_{v} = \mathrm{repeat}(\mathcal{R}'_{v}, nhw \rightarrow bnhw', b = B) \tag{5}$$

Then integrate the register tokens into $Q_i K_{\mathcal{R}}^T = Q_i K^T + \mathcal{R}_{qk}$ and $V_{\mathcal{R}} = V + \mathcal{R}_v$ matrices before computing the attention, where $Z_{\mathcal{R}}$ is the final 2D-MHSA with Register, then the register token is discarded, eqn. 6.

$$Z_{\mathcal{R}}(Q_i, K, V)_h = \mathrm{softmax}\left(\frac{Q_i K_{\mathcal{R}}^T + Q_i R}{\sqrt{d_k^h}}\right) V_{\mathcal{R}} \tag{6}$$

ECANET (*Channel-wise recalibrations and down-sampling*) Long-range dependencies in CV tasks can span the entire feature space along spatial and channel dimensions. However, traditional CNNs and attention mechanisms primarily focus on spatial adaptivity and may overlook channel adaptability. To address this limitation, we integrate ECANET into the iiANET architecture after each stage. ECANET directly calculates channel attention weights based on the inter-channel relationships within the feature map. This enables the model to adaptively recalibrate channel-wise information flow, enhancing its ability to capture complex dependencies. Compared to SENET, a widely-used channel-wise attention, ECANET offers improved efficiency, scalability and accuracy. Given an input of $x \in \mathbb{R}^{H \times W \times C}$, an adaptive average pooling operation p is applied across the spatial dimensions H, W, resulting in a pooled tensor $p \in \mathbb{R}^{1 \times 1 \times C}$. Subsequently, a 1×1 convolutions is performed on p, followed by the application of a sigmoid activation function to derive attention weights Z ranging between $\{0, 1\}$. These weights represent the relative significance of each channel. Finally, an element-wise multiplication operation to compute attended features, $ECA = x \otimes Z$. Lastly, we down-sample the feature maps by applying a 1×1 convolutional block with stride 2 after each stage.

iiABlock (*Merging dilated convolution, MBConv2 and global 2D-MHSA*) serves as the fundamental building block of iiANET, integrating three key components, a dilated convolution, MBConv2, and 2D-MHSA with Registers in parallel to enhance computational efficiency while effectively capturing local details, global context and long-range spatial dependencies prevalent in complex images. Then, concatenates the outputs of these components, ensuring that the network retains the information captured by each component and allowing it to selectively focus on the most relevant features. By default, we employed a channel ratio of $r = (1{:}\,6{:}\,1)$ for each component, respectively. Let the input be $x \in \mathbb{R}^{H \times W \times \mathrm{C}}$, then C is divided into r denoted by $i, j,$ and k, such that $C = i + 6j + k$, we can express x as $\mathrm{x} \in \mathrm{R}^{\mathrm{H \times W \times (i+6j+k)}}$. By defining $i = \frac{C}{8}$, $j = 6 \times \frac{C}{8}$, and $k = \frac{C}{8}$, we have $x \in \mathbb{R}^{H \times W \times (\frac{C}{8} + 6\frac{C}{8} + \frac{C}{8})}$. Substituting the values of $i, j,$ and k, we obtain $x \in \mathbb{R}^{H \times W \times (\frac{C}{8} + \frac{3C}{4} + \frac{C}{8})}$. We then split x to $x_1 \in \mathbb{R}^{H \times W \times \frac{C}{8}}$, $x_2 \in \mathbb{R}^{H \times W \times \frac{3C}{4}}$, and $x_3 \in \mathbb{R}^{H \times W \times \frac{C}{8}}$. where, $\mathrm{f}_1, \mathrm{f}_2,$ and f_3 represent the corresponding components. The iiABlock is defined as the concatenations of $f_1(x_1),\ f_2(x_2)$, and $f_3(x_3)$ (Fig. 3).

4 Experimental Results and Comparisons

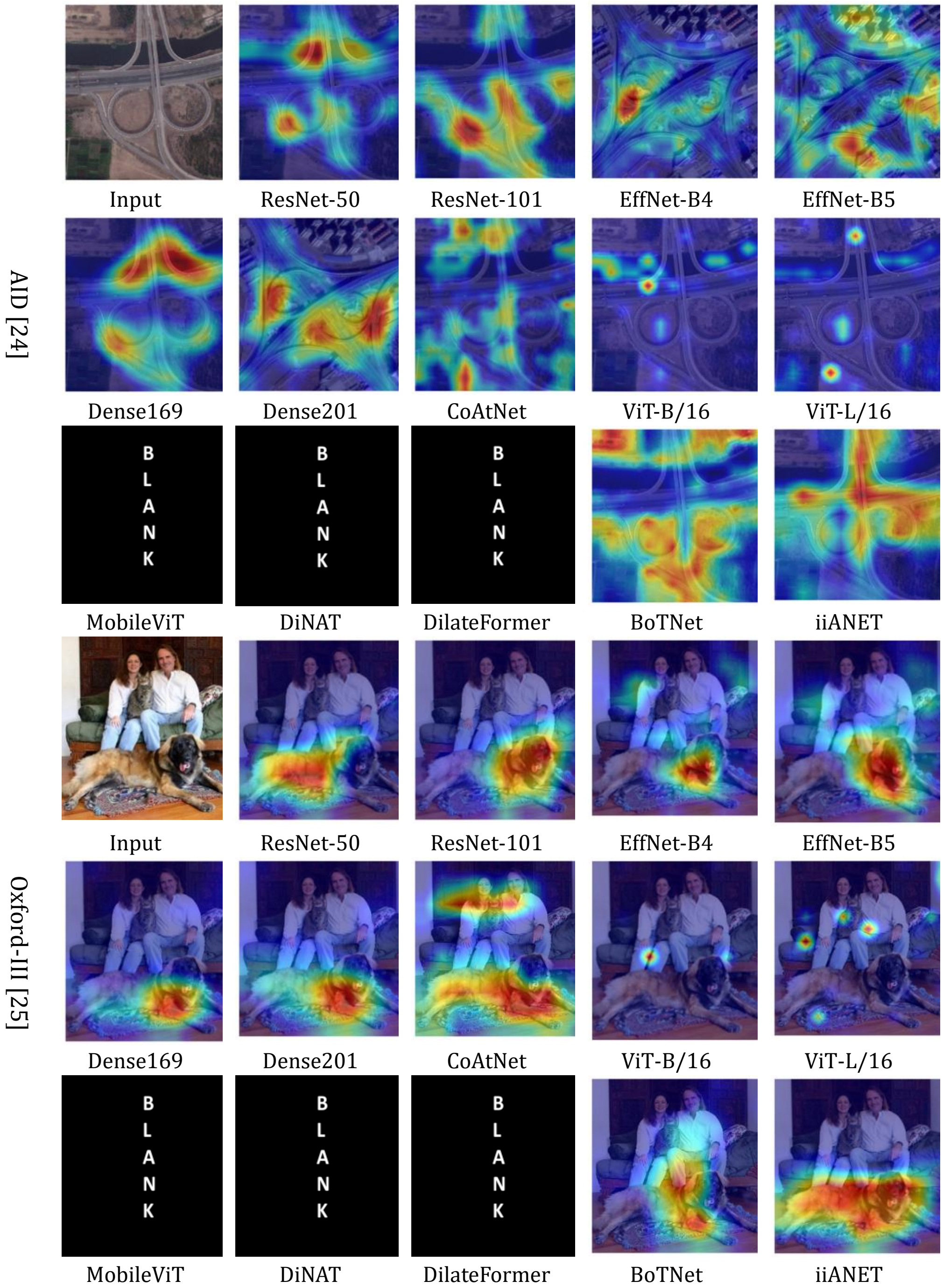
In this section, we evaluate the performance of iiANET both qualitatively and quantitatively on various widely recognized benchmark datasets. We compare it with state-of-the-art classification models and asses its effectiveness as a backbone in object detection and segmentation tasks. These models are based on CNN, ViT and hybrid architectures. Our aim is to validate the effectiveness of iiANET in capturing long-range dependencies in complex images compared to established approaches.

Datasets and Metrics. We conduct extensive experiments on diverse, complex and widely used image datasets. These include AID [24], which comprises 10,000 images of 30 scene classes. The Oxford-III datasets [25] consist of 4,978 images featuring 37 breeds of cats and dogs. Then, RLD datasets (Rice Leaf Disease) [26] with 5,932 images of four disease categories. In addition to these complex images, we also considered the COCO-2017 datasets [32] and IMAGENET1K [39] to assess iiANET generalization and robustness. Note, metrics used for quantitative evaluation include top-1/top-5 accuracy, AP (Average Precision), FLOPs and throughputs providing comprehensive insights into iiANET performance across various tasks and datasets.

Experimental Setup. The experiments were conducted on a Linux-based system with hardware specifications including an Intel Core i7 8700k processor, 2 NVIDIA Titan XP 12GB GPUs, and 32GB of RAM. The training process involved 90/150 epochs with a batch size of 16, utilizing the AdamW optimizer with an initial learning rate set at 0.0001 and a decay rate of 0.05. To ensure a fair comparison, given that our comparison models were not originally trained on the AID, Oxford-III, and RLD dataset, we have re-trained the models using the default settings as provided by the authors.

4.1 Qualitative Evaluation and Comparison: iiANET visual inspection

We applied Grad-CAM [23] on the final layer of iiANET and other models to conduct qualitative analysis, with visualization results presented in Fig. 4. Comparative assessments were performed against several state-of-the-art models, including ResNet-50/100 [1], EfficientNet-B4/B5 [2], DenseNet-169/201 [33], ViT-B/L-16 [6], Coatnet-3 [12], and BoTNet [11]. Our observations show that CNN-based models exhibit a strong local neighborhood heat-map around the object of interest due to their local small receptive field. Conversely, ViT-based models display tiny spots of strong heatmap, indicating interpretability issues. Hybrid models demonstrate improved capture of long-range dependencies in complex images and offer better interpretability. However, iiANET stands out for its superior precision in outlining complex objects while minimizing background interference, this can potentially enhance accuracy and reliability in applications where long-range dependencies is prevalent like medical imaging, autonomous driving, remote sensing, and security surveillance.


AID [24]
Input
ResNet-50
ResNet-101
EffNet-B4
EffNet-B5
Dense169
Dense201
CoAtNet
ViT-B/16
ViT-L/16
BLANK
BLANK
BLANK
MobileViT
DiNAT
DilateFormer
BoTNet
iiANET
Oxford-III [25]
Input
ResNet-50
ResNet-101
EffNet-B4
EffNet-B5
Dense169
Dense201
CoAtNet
ViT-B/16
ViT-L/16
BLANK
BLANK
BLANK
MobileViT
DiNAT
DilateFormer
BoTNet
iiANET

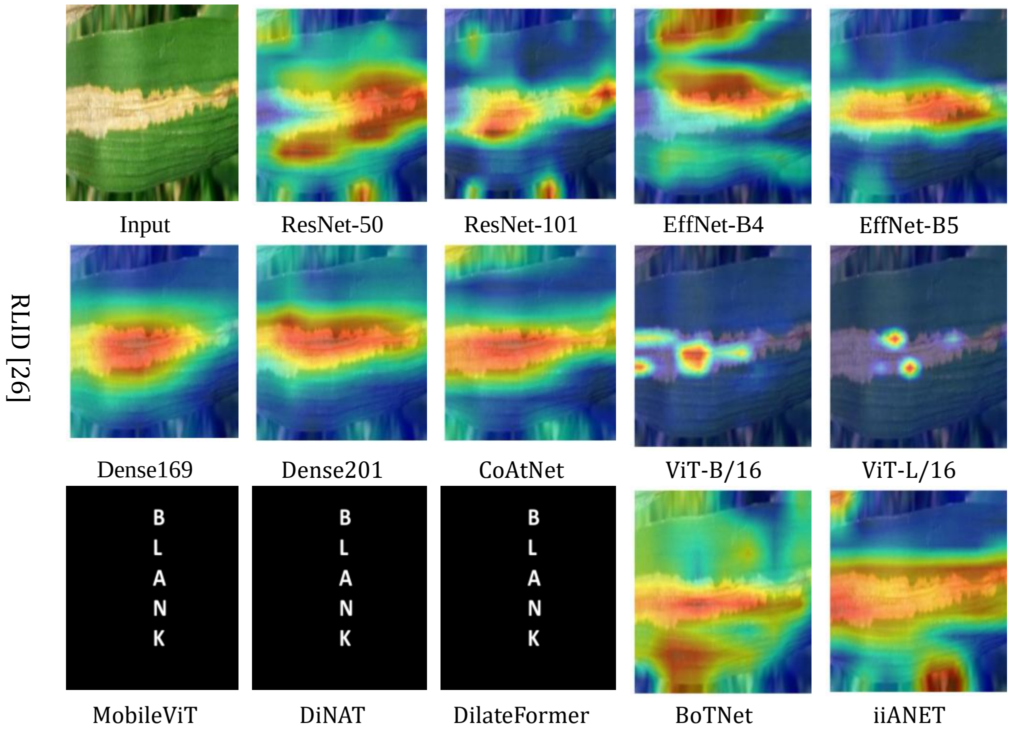


Fig. 4. Visual inspection of iiANET and comparison other state-of-the-art models using Grad-CAM [23], the images contain heatmaps of complex objects.

4.2 Quantitative Evaluation and Comparison

Table 1 shows the performance metrics of iiANET model, showcasing its superior accuracy while maintaining efficiency with fewer parameters and FLOPs compared to ResNet-50/100 [1], EfficientNet-B4/B5 [2], DenseNet-169/201 [33], ViT-B/L-16 [6], Coatnet-3 [12], and BoT-Net [11]. Notably, iiANET achieves an outstanding accuracy of 80.57% on the AID dataset with just 34 GFLOPs, and slightly surpasses others with top-1 accuracies of 81.8% on the Oxford-III dataset and 82.4% on the RLD dataset. These results highlight the effectiveness of the iiABlock in capturing long-range dependencies in complex images. Importantly, iiANET consistent outperformance of other models by a significant margin highlights its potential for enhancing various applications reliant on accurate and efficient image analysis, including medical imaging, autonomous driving, and remote sensing.

Table 2 Classification result comparison on AID, OXFORD-III and RLD datasets.

Dataset	Backbone	Size	Train	#Params	FLOPs	Throughput	top-1 acc.	top-5 acc.
IMAGENET1K [39]	ResNet-50 [1]	224^2	90	25.6 M	7.71 G	-	76.0 %	93.0 %
	ResNet-101 [1]	224^2	90	44.5 M	14.58G	-	78.0%	94.0 %
	EffNet-B4 [2]	224^2	90	19.3 M	2.86G	-	82.9%	96.4 %
	EffNet-B5 [2]	224^2	90	30.4 M	4.49G	-	83.6%	96.7 %
	Dense169 [36]	224^2	90	14.1 M	5.81G	-	76.2%	93.2 %
	Dense 201 [36]	224^2	90	20.0 M	7.35G	-	77.42%	93.6 %
	ViT-B/16 [6]	224^2	150	86.6 M	16.86G	-	77.91%	93.6 %
	ViT-L/16 [6]	224^2	150	304.3 M	59.69G	-	76.53%	93.2 %
	MobileViT-S [37]	256^2	90	6M	2G	-	77.0%	94.6 %
	DiNAT-B [29]	224^2	90	90M	13.7G	764	84.40%	-
	Dilate-B [38]	224^2	120	48M	9.96G	-	84.9%	-
	BoT50[11]	256^2	90	25.6M	3.18G	-	84.4%	-
	CoAtNet-3[28]	224^2	90	168M	32.53G	-	84.5%	-
	iiANET[ours]	299^2	90	25.2M	8.22G	-	79.34%	94.71 %
AID [24]	ResNet-50	224^2	90	25.6 M	7.71 G	93.32	71.93 %	93.27 %
	ResNet-101	224^2	90	44.5 M	14.58G	-	68.93 %	93.37 %
	EffNet-B4	224^2	90	19.3 M	2.86G	-	63.57 %	90.70 %
	EffNet-B5	224^2	90	30.4 M	4.49G	81.85	65.73 %	92.07 %
	Dense169	224^2	90	14.1 M	5.81G	-	70.70 %	93.83 %
	Dense 201	224^2	90	20.0 M	7.35G	76.38	71.60 %	94.37 %
	ViT-B/16	224^2	150	86.6 M	16.86G	48.01	69.93 %	93.27 %
	ViT-L/16	224^2	150	304.3 M	59.69G	59.11	67.27 %	92.77 %
	MobileViT-S	256^2	90	6M	2G	1986	66.76 %	91.23 %
	DiNAT-B	224^2	90	90M	13.7G	764	79.12 %	93.27 %
	Dilate-B	224^2	120	48M	9.96G	-	78.39 %	94.69 %
	BoT50	256^2	90	25.6M	3.18G	95.20	72.50 %	94.27 %
	CoAtNet-3	224^2	90	168M	32.53G	27.92	80.17%	94.93 %
	iiANET [ours]	299^2	90	25.2M	8.22G	37.21	80.57%	95.67 %
OXFORD-III [25]	ResNet-50	224^2	90	25.6M	7.71G	119.75	59.07%	86.61 %
	ResNet-101	224^2	90	44.5M	14.58G	105.80	59.25%	87.38 %
	EffNet-B4	224^2	90	19.3M	2.86G	106.95	45.86%	75.85 %
	EffNet-B5	224^2	90	30.4M	4.49G	96.844	47.54%	78.92 %
	Dense169	224^2	90	14.1M	5.81G	104.88	58.98%	87.61 %
	Dense 201	224^2	90	20.0M	7.35G	98.82	62.55%	86.84 %
	ViT-B/16	224^2	150	86.6M	16.86G	-	51.23%	82.31 %
	ViT-L/16	224^2	150	304.3M	59.69G	-	53.19%	83.28 %
	MobileViT-S	256^2	90	6M	2G	1986	51.89 %	84.52 %
	DiNAT-B	224^2	90	90M	13.7G	764	69.37 %	92.09 %
	Dilate-B	224^2	120	48M	9.96G	-	64.85 %	88.18 %
	BoT50	256^2	90	25.6M	3.18G	118.97	64.13%	90.00%

	CoAtNet-3	224^2	90	168M	32.53G	43.80	67.57%	91.36%
	iiANET [ours]	299^2	90	25.2M	8.22G	51.44	74.04%	93.98%
RLD [26]	ResNet-50	224^2	90	25.6M	7.71G	188.72	93.63%	-
	ResNet-101	224^2	90	44.5M	14.58G	157.76	94.89%	-
	EffNet-B4	224^2	90	19.3M	2.86G	159.55	92.71 %	-
	EffNet-B5	224^2	90	30.4M	4.49G	134.40	92.92 %	-
	Dense169	224^2	90	14.1M	5.81G	158.78	89.41 %	-
	Dense 201	224^2	90	20.0M	7.35G	141.14	91.34 %	-
	ViT-B/16	224^2	150	86.6M	16.86G	-	83.79 %	-
	ViT-L/16	224^2	150	304.3M	59.69G	-	82.21 %	-
	MobileViT-S	224^2	90	6M	2G	1986	87.65 %	-
	DiNAT-B	224^2	90	90M	13.7G	764	94.28 %	-
	Dilate-B	224^2	120	48M	9.96G	-	91.66 %	-
	BoT50	256^2	90	25.6M	3.18G	191.57	93.95%	-
	CoAtNet-3	224^2	90	168M	32.53G	50.18	94.45%	-
	iiANET [ours]	299^2	90	25.2M	8.22G	59.176	94.79%	-

Object detection on COCO val2017 in Table 2 we experiment iiANET as a backbone on Faster-RCNN on the COCO dataset, our results shows that iiANET demonstrates outstanding performance in terms of mAP across all evaluation metrics. Specifically, iiANET achieves mAP of 62.6% and 63.0% for AP val2017 and AP test2017 indicating its proficiency in detecting objects with high precision and recall.

Table 2 Object detection on COCO dataset Faster R-CNN

Backbone	Object Detector	AP^b val2017	AP^b test2017
ResNet-50 [45]	Faster R-CNN	36.7	37.9
FD-SwinV2-G [46]	HTC++	-	64.2
Florence-CoSwin-H [47]	DyHead	62.0	62.4
Swin-L [48]	DINO	63.2	63.3
BEiT-3 [49]	ViTDet	-	63.7
Swin V2-G [50]	HTC++	62.5	63.1
iiANET [ours]	YOLOv8	62.6	63.0

Instance segmentation on COCO val2017, we also experiment the effectiveness of iiANET in capturing long-range dependencies on instance segmentation tasks, Table 3 demonstrates outstanding performance in instance segmentation tasks. It achieves an impressive AP box of 45.3% and AP mask of 39.5, indicating its capability to accurately segment instances in images with varying complexities and occlusions.

Table 3. Instance Segmentation on COCO dataset with Mask R-CNN

Model	Backbone	$\mathrm{AP}^{\mathrm{box}}$	$\mathrm{AP}^{\mathrm{box}}_{50}$	$\mathrm{AP}^{\mathrm{box}}_{75}$	$\mathrm{AP}^{\mathrm{mask}}$	$\mathrm{AP}^{\mathrm{mask}}_{50}$	$\mathrm{AP}^{\mathrm{mask}}_{75}$
			Mask R-CNN – 1x schedule				
Mask R-CNN [3]	ResNet-50 [1]	38.0	58.6	41.4	34.4	55.1	36.7
	PVT-M [40]	42.0	64.4	45.6	39.0	61.6	42.1
	TRT-ViT-C [41]	44.7	66.9	48.8	40.8	63.9	44.0
	Focal-T [42]	44.8	67.7	49.2	41.0	64.7	44.2
	UniFormer-S/h14 [43]	45.6	68.1	49.7	41.6	64.8	45.0
	Swin-T [44]	42.2	64.6	46.2	39.1	61.6	42.0
	Dilate-S [38]	45.8	68.2	50.1	41.7	65.3	44.7
	BoT50 [11]	43.7	-	-	37.9	-	-
	iiANET [ours]	45.3	65.1	49.8	39.5	58.9	58.0

4.3 Ablation

We performed series of ablation studies to investigate iiANET capacity from different aspects, we use image classification tasks.

Table 4. Ablation studies on various variant of iiABlock

Settings		Model (components)	Size	Top-1 accuracy
AID [24]	(a)	MBConv2	299	69.23%
	(b)	MBConv + Dilated Conv	299	74.85%
	(c)	MBConv + Dilated Conv + MHSA	299	80.57%
	(d)	Dilated + MHSA	299	67.01%
	(e)	MBConv + MHSA	299	78.72%

Table 4 we showcase the superiority of our model against several other variants and variant (c) shows the highest performance.

Table 5. Ablation on ratio and MHSA head size

Settings	Model (components)	Size	Top-1 accuracy
AID [24]	iiANET		
	Ratio: 1.6.1 → 2.4.2	299	74.21%
	Head size: 8 → 16	299	76.85%

Table 5 shows increasing MHSA head size from 8 to 16 significantly affects the model's performance, although we observed increased in GPU speed by 3% to 5%. And modifying the iiABlock ratio size from 1.6.1 to 2.4.2 also affects models' performance and increases computational cost.

Registers. the table 6 below shows the effect of adding registers to the 2D-MHSA mechanism on the model accuracy.

Table 6. Ablation studies on Registers

Dataset	Number of Registers	Top-1 accuracy
AID [24]	0	80.57%
	1	80.61%
	2	80.62%
	4	80.77%

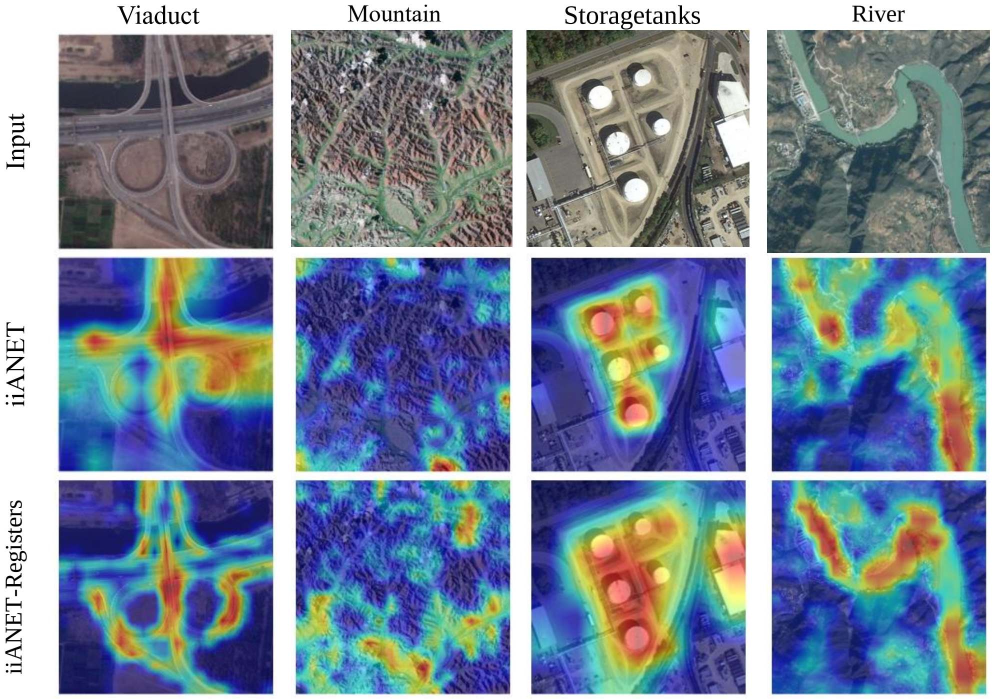


Fig. 6. Visual effect of registers on iiANET on the AID [24]

5 Conclusion

In this work, we introduce iiANET, a novel hybrid model designed to efficiently improve long-range dependencies in complex images by integrating convolutional layers and the MHSA mechanism with registers in parallel. Comprehensive qualitative and quantitative results show significant improvements in capturing long-range dependencies compared to some previous state-of-the-art models. Additionally, we validate the performance of our model across diverse datasets and highlight its potential as a backbone in object detection and segmentation models.

Limitations. While iiANET excels in capturing long-range dependencies, it may not perform as effectively on datasets like ImageNet-1k, where objects are more localized and do not exhibit global dependencies. Therefore, it is less suited for tasks involving images with limited long-range contextual information.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.
2. Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." In International conference on machine learning, pp. 6105-6114. PMLR, 2019.
3. Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems 28 (2015).
4. Redmon, Joseph, and Ali Farhadi. "YOLO9000: better, faster, stronger." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263-7271. 2017.
5. Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431-3440. 2015.
6. Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
7. Han, Kai, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang et al. "A survey on vision transformer." IEEE transactions on pattern analysis and machine intelligence 45, no. 1 (2022): 87-110.Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016).
8. Yunusa, Haruna, Shiyin Qin, Abdulrahman Hamman Adama Chukkol, Abdulganiyu Abdu Yusuf, Isah Bello, and Adamu Lawan. "Exploring the Synergies of Hybrid CNNs and ViTs Architectures for Computer Vision: A survey." arXiv preprint arXiv:2402.02941 (2024).
9. Khan, Asifullah, Zunaira Rauf, Anabia Sohail, Abdul Rehman Khan, Hifsa Asif, Aqsa Asif, and Umair Farooq. "A survey of the vision transformers and their CNN-transformer based variants." Artificial Intelligence Review 56, no. Suppl 3 (2023): 2917-2970.
10. Guo, Jianyuan, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. "Cmt: Convolutional neural networks meet vision transformers." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12175-12185. 2022.
11. Srinivas, Aravind, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. "Bottleneck transformers for visual recognition." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16519-16529. 2021.
12. Dai, Zihang, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. "Coatnet: Marrying convolution and attention for all data sizes." Advances in neural information processing systems 34 (2021): 3965-3977.
13. Donahue, Jeffrey, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. "Long-term recurrent convolutional networks for visual recognition and description." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625-2634. 2015.
14. Yu, Fisher, Vladlen Koltun, and Thomas Funkhouser. "Dilated residual networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 472-480. 2017.
15. Yu, Fisher, and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions." arXiv preprint arXiv:1511.07122 (2015).
16. Zhang, Xindong, Hui Zeng, Shi Guo, and Lei Zhang. "Efficient long-range attention network for image super-resolution." In European Conference on Computer Vision, pp. 649-667. Cham: Springer Nature Switzerland, 2022.
17. Banerjee, Imon, Yuan Ling, Matthew C. Chen, Sadid A. Hasan, Curtis P. Langlotz, Nathaniel Moradzadeh, Brian Chapman et al. "Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification." Artificial intelligence in medicine 97 (2019): 79-88.

18. Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9. 2015.
19. Sandler, Mark, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. "Mobilenetv2: Inverted residuals and linear bottlenecks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510-4520. 2018.
20. Wang, Qilong, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. "ECA-Net: Efficient channel attention for deep convolutional neural networks." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11534-11542. 2020.
21. Bello, Irwan, William Fedus, Xianzhi Du, Ekin Dogus Cubuk, Aravind Srinivas, Tsung-Yi Lin, Jonathon Shlens, and Barret Zoph. "Revisiting resnets: Improved training and scaling strategies." Advances in Neural Information Processing Systems 34 (2021): 22614-22627.
22. Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132-7141. 2018.
23. Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization." In Proceedings of the IEEE international conference on computer vision, pp. 618-626. 2017.
24. Xia, Gui-Song, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. "AID: A benchmark data set for performance evaluation of aerial scene classification." IEEE Transactions on Geoscience and Remote Sensing 55, no. 7 (2017): 3965-3981.
25. Parkhi, Omkar M., Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. "Cats and Dogs." In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3696-3703. IEEE.
26. Sethy, Prabira Kumar, Nalini Kanta Barpanda, Amiya Kumar Rath, and Santi Kumari Behera. "Deep feature-based rice leaf disease identification using support vector machine." Computers and Electronics in Agriculture 175 (2020): 105527.
27. Peng, Zhiliang, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. "Conformer: Local features coupling global representations for visual recognition." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 367-376. 2021.
28. Dai, Zihang, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. "Coatnet: Marrying convolution and attention for all data sizes." Advances in neural information processing systems 34 (2021): 3965-3977.
29. Hassani, Ali, and Humphrey Shi. "Dilated neighborhood attention transformer." arXiv preprint arXiv:2209.15001 (2022).
30. Wu, Haiping, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. "Cvt: Introducing convolutions to vision transformers." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 22-31. 2021.
31. Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin transformer: Hierarchical vision transformer using shifted windows." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 10012-10022. 2021.
32. Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pp. 740-755. Springer International Publishing, 2014.
33. Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely connected convolutional networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708. 2017.
34. Srinivas, Aravind, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. "Bottleneck transformers for visual recognition." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16519-16529. 2021.
35. Darcet, Timothée, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. "Vision transformers need registers." arXiv preprint arXiv:2309.16588 (2023).

36. Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely connected convolutional networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708. 2017.
37. Mehta, Sachin, and Mohammad Rastegari. "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer." arXiv preprint arXiv:2110.02178 (2021).
38. Jiao, Jiayu, Yu-Ming Tang, Kun-Yu Lin, Yipeng Gao, Andy J. Ma, Yaowei Wang, and Wei-Shi Zheng. "Dilateformer: Multi-scale dilated transformer for visual recognition." IEEE Transactions on Multimedia 25 (2023): 8906-8919.
39. Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In 2009 IEEE conference on computer vision and pattern recognition, pp. 248-255. Ieee, 2009.
40. Hassani, Ali, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. "Neighborhood attention transformer." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6185-6194. 2023.
41. Wang, Wenhai, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 568-578. 2021.
42. Xia, Xin, Jiashi Li, Jie Wu, Xing Wang, Xuefeng Xiao, Min Zheng, and Rui Wang. "TRT-ViT: TensorRT-oriented vision transformer." arXiv preprint arXiv:2205.09579 (2022).
43. Yang, Jianwei, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. "Focal self-attention for local-global interactions in vision transformers." arXiv preprint arXiv:2107.00641 (2021).
44. Li, Kunchang, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. "Uniformer: Unifying convolution and self-attention for visual recognition." IEEE Transactions on Pattern Analysis and Machine Intelligence 45, no. 10 (2023): 12581-12600.
45. Carion, Nicolas, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. "End-to-end object detection with transformers." In European conference on computer vision, pp. 213-229. Cham: Springer International Publishing, 2020.
46. Wei, Yixuan, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. "Contrastive learning rivals masked image modeling in fine-tuning via feature distillation." arXiv preprint arXiv:2205.14141 (2022).
47. Yuan, Lu, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu et al. "Florence: A new foundation model for computer vision." arXiv preprint arXiv:2111.11432 (2021).
48. Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin transformer: Hierarchical vision transformer using shifted windows." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 10012-10022. 2021.
49. Wang, Wenhui, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal et al. "Image as a foreign language: Beit pretraining for all vision and vision-language tasks." arXiv preprint arXiv:2208.10442 (2022).
50. Liu, Ze, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning et al. "Swin transformer v2: Scaling up capacity and resolution." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12009-12019. 2022.