Attentional Feature Fusion

Yimian Dai[1] Fabian Gieseke[2,3] Stefan Oehmcke[3] Yiquan Wu[1] Kobus Barnard[4]

[1]College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics
[2]Department of Information Systems, University of Münster
[3]Department of Computer Science, University of Copenhagen
[4]Department of Computer Science, University of Arizona

Abstract

Feature fusion, the combination of features from different layers or branches, is an omnipresent part of modern network architectures. It is often implemented via simple operations, such as summation or concatenation, but this might not be the best choice. In this work, we propose a uniform and general scheme, namely attentional feature fusion, which is applicable for most common scenarios, including feature fusion induced by short and long skip connections as well as within Inception layers. To better fuse features of inconsistent semantics and scales, we propose a multi-scale channel attention module, which addresses issues that arise when fusing features given at different scales. We also demonstrate that the initial integration of feature maps can become a bottleneck and that this issue can be alleviated by adding another level of attention, which we refer to as iterative attentional feature fusion. With fewer layers or parameters, our models outperform state-of-the-art networks on both CIFAR-100 and ImageNet datasets, which suggests that more sophisticated attention mechanisms for feature fusion hold great potential to consistently yield better results compared to their direct counterparts. Our codes and trained models are available online[1].

1. Introduction

Convolutional neural networks (CNNs) have seen a significant improvement of the representation power by going deeper [11], going wider [36, 47], increasing cardinality [45], and refining features dynamically [14], corresponding to advances in many computer vision tasks.

Apart from these strategies, in this paper, we investigate a different component of the network, *feature fusion*, to further boost the representation power of CNNs. Whether explicit or implicit, intentional or unintentional, feature fusion is omnipresent for modern network architectures and has been studied extensively in the previous literature [36, 34, 11, 28, 21]. For instance, in the Inception-Net family [36, 37, 35], the outputs of filters with multiple sizes on the same level are fused to handle the large variation of object size. In Residual Networks (ResNet) [11, 12] and its follow-ups [47, 45], the identity mapping features and residual learning features are fused as the output via short skip connections, enabling the training of very deep networks. In Feature Pyramid Networks (FPN) [21] and U-Net [28], low-level features and high-level features are fused via long skip connections to obtain high-resolution and semantically strong features, which are vital for semantic segmentation and object detection. However, despite its prevalence in modern networks, most works on feature fusion focus on constructing sophisticated *pathways* to combine features in different kernels, groups, or layers. The feature fusion *method* has rarely been addressed and is usually implemented via simple operations such as addition or concatenation, which merely offer a fixed linear aggregation of feature maps and are entirely unaware of whether this combination is suitable for specific objects.

Recently, Selective Kernel Networks (SKNet) [19] and ResNeSt [48] have been proposed to render dynamic weighted averaging of features from multiple kernels or groups *in the same layer* based on the *global* channel attention mechanism [14]. Although such attention-based methods present nonlinear approaches for feature fusion, they still suffer from the following shortcomings:

1. *Limited scenarios:* SKNet and ResNeSt only focus on the soft feature selection in the same layer, whereas the *cross-layer* fusion in skip connections has not been addressed, leaving their schemes quite heuristic. Despite having different scenarios, all kinds of feature fusion implementations face the same challenge, in essence, that is, how to integrate features of different scales for better performance. A module that can overcome the semantic inconsistency and effectively integrate features of different scales should be able to consistently

[1]https://github.com/YimianDai/open-aff

improve the quality of fused features in various network scenarios. However, so far, there is still a lack of a generalized approach that can unify different feature fusion scenarios in a consistent manner.

2. *Unsophisticated initial integration:* To feed the received features into the attention module, SKNet introduces another phase of feature fusion in an involuntary but inevitable way, which we call *initial integration* and is implemented by addition. Therefore, besides the design of the attention module, as its input, the initial integration approach also has a large impact on the quality of fusion weights. Considering the features may have a large inconsistency on the scale and semantic level, an unsophisticated initial integration strategy ignoring this issue can be a bottleneck.
3. *Biased context aggregation scale:* The fusion weights in SKNet and ResNeSt are generated via the global channel attention mechanism [14], which is preferred for information that distributes more globally. However, objects in the image can have an extremely large variation in size. Numerous studies have emphasized this issue that arises when designing CNNs, i.e., that the receptive fields of predictors should match the object scale range [49, 31, 32, 20]. Therefore, merely aggregating contextual information on a global scale is too biased and weakens the features of small objects. This gives rise to the question if a network can dynamically and adaptively fuse the received features in a contextual scale-aware way.

Motivated by the above observations, we present the *attentional feature fusion* (AFF) module, trying to answer the question of how a unified approach for all kinds of feature fusion scenarios should be and address the problems of contextual aggregation and initial integration. The AFF framework generalizes the attention-based feature fusion from the same-layer scenario to cross-layer scenarios including short and long skip connections, and even the initial integration inside AFF itself. It provides a universal and consistent way to improve the performance of various networks, e.g., InceptionNet, ResNet, ResNeXt [45], and FPN, by simply replacing existing feature fusion operators with the proposed AFF module. Moreover, the AFF framework supports to gradually refine the initial integration, namely the input of the fusion weight generator, by iteratively integrating the received features with another AFF module, which we refer to as *iterative attentional feature fusion* (iAFF).

To alleviate the problems arising from scale variation and small objects, we advocate the idea that attention modules should also aggregate contextual information from different receptive fields for objects of different scales. More specifically, we propose the *Multi-Scale Channel Attention Module* (MS-CAM), a simple yet effective scheme to remedy the feature inconsistency across different scales for attentional feature fusion. Our key observation is that scale is not an issue exclusive to the spatial attention, and the channel attention can also have scales other than the global by varying the spatial pooling size. By aggregating the multi-scale context information along the channel dimension, MS-CAM can simultaneously emphasize large objects that distribute more globally and highlight small objects that distribute more locally, facilitating the network to recognize and detect objects under extreme scale variation.

2. Related Work

2.1. Multi-scale Attention Mechanism

The scale variation of objects is one of the key challenges in computer vision. To remedy this issue, an intuitive way is to leverage multi-scale image pyramids [27, 2], in which objects are recognized at multiple scales and the predictions are combined using non-maximum suppression. The other line of effort aims to exploit the inherent multi-scale, hierarchical feature pyramid of CNNs to approximate image pyramids, in which features from multiple layers are fused to obtain semantic features with high resolutions [10, 28, 21].

The attention mechanism in deep learning, which mimics the human visual attention mechanism [4, 7], is originally developed on a global scale. For example, the matrix multiplication in self-attention draws global dependencies of each word in a sentence [39] or each pixel in an image [6, 42, 1]. The Squeeze-and-Excitation Networks (SENet) squeeze global spatial information into a channel descriptor to capture channel-wise dependencies [14]. Recently, researchers start to take into account the scale issue of attention mechanisms. Similar to the above-mentioned approaches handling scale variation in CNNs, multi-scale attention mechanisms are achieved by either feeding multi-scale features into an attention module or combining feature contexts of multiple scales inside an attention module. In the first type, the features at multiple scales or their concatenated result are fed into the attention module to generate multi-scale attention maps, while the scale of feature context aggregation inside the attention module remains single [2, 3, 43, 5, 33, 38]. The second type, which is also referred to as multi-scale spatial attention, aggregates feature contexts by convolutional kernels of different sizes [18] or from a pyramid [18, 41] inside the attention module .

The proposed MS-CAM follows the idea of ParseNet [23] with combining local and global features in CNNs and the idea of spatial attention with aggregating multi-scale feature contexts inside the attention module, but differ in at least two important aspects: 1) MS-CAM puts forward the scale issue in channel attention and is achieved by point-wise convolution rather than kernels of different sizes. 2) instead of in the backbone network, MS-CAM aggregates local and global feature contexts inside the channel atten-

tion module. To the best of our knowledge, the multi-scale channel attention has never been discussed before.

2.2. Skip Connections in Deep Learning

Skip connection has been an essential component in modern convolutional networks. Short skip connections, namely the identity mapping shortcuts added inside Residual blocks, provide an alternative path for the gradient to flow without interruption during backpropagation [11, 45, 47]. Long skip connections help the network to obtain semantic features with high resolutions by bridging features of finer details from lower layers and high-level semantic features of coarse resolutions [15, 21, 28, 24]. Despite being used to combine features in various pathways [8], the fusion of connected features is usually implemented via addition or concatenation, which allocate the features with fixed weights regardless of the variance of contents. Recently, a few attention-based methods, e.g., Global Attention Upsample (GAU) [18] and Skip Attention (SA) [46], have been proposed to use high-level features as guidance to modulate the low-level features in long skip connections. However, the fusion weights for the modulated features are still fixed.

To the best of our knowledge, it is the Highway Networks that first introduced a selection mechanism in short skip connections [34]. To some extent, the *attentional skip connections* proposed in this paper can be viewed as its follow-up, but differs in the three points: 1) Highway Networks employ a simple fully connected layer that can only generate a scalar fusion weight, while our proposed MS-CAM generates fusion weights as the same size of feature maps, enabling dynamic soft selections in an element-wise way. 2) Highway Networks only use one input feature to generate weight, while our AFF module is aware of both features. 3) We point out the importance of initial feature integration and the iAFF module is proposed as a solution.

3. Multi-scale Channel Attention

3.1. Revisiting Channel Attention in SENet

Given an intermediate feature $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ with C channels and feature maps of size $H \times W$, the channel attention weights $\mathbf{w} \in \mathbb{R}^C$ in SENet can be computed as

$$\mathbf{w} = \sigma\left(\mathbf{g}(\mathbf{X})\right) = \sigma\left(\mathcal{B}\left(\mathbf{W}_2\delta\left(\mathcal{B}\left(\mathbf{W}_1(g(\mathbf{X}))\right)\right)\right)\right), \quad (1)$$

where $\mathbf{g}(\mathbf{X}) \in \mathbb{R}^C$ denotes the global feature context and $g(\mathbf{X}) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{X}_{[:,i,j]}$ is the global average pooling (GAP). δ denotes the Rectified Linear Unit (ReLU) [25], and $\mathcal{B}$ denotes the Batch Normalization (BN) [16]. σ is the Sigmoid function. This is achieved by a bottleneck with two fully connected (FC) layers, where $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ is a dimension reduction layer, and $\mathbf{W}_2 \in \mathbb{R}^{C\times\frac{C}{r}}$ is a dimension increasing layer. r is the channel reduction ratio.

We can see that the channel attention squeezes each feature map of size $H \times W$ into a scalar. This extreme coarse descriptor prefers to emphasize large objects that distribute globally and can potentially wipe out most of the image signal present in a small object. However, detecting very small objects stands out as the key performance bottleneck of state-of-the-art networks [32]. For example, the difficulty of COCO is largely due to the fact that most object instances are smaller than 1% of the image area [22, 31]. Therefore, global channel attention might not be the best choice. Multi-scale feature contexts should be aggregated inside the attention module to alleviate the problems arising from scale variation and small object instances.

3.2. Aggregating Local and Global Contexts

In this part, we depict the proposed multi-scale channel attention module (MS-CAM) in detail. The key idea is that the channel attention can be implemented in multiple scales by varying the spatial pooling size. To maintain it as lightweight as possible, we merely add the local context to the global context inside the attention module. We choose the point-wise convolution (PWConv) as the local channel context aggregator, which only exploits point-wise channel interactions for each spatial position. To save parameters, the local channel context $\mathbf{L}(\mathbf{X}) \in \mathbb{R}^{C\times H\times W}$ is computed via a bottleneck structure as follows:

$$\mathbf{L}(\mathbf{X}) = \mathcal{B}\left(\mathrm{PWConv}_2\left(\delta\left(\mathcal{B}\left(\mathrm{PWConv}_1(\mathbf{X})\right)\right)\right)\right) \quad (2)$$

The kernel sizes of PWConv_1 and PWConv_2 are $\frac{C}{r}\times C\times 1\times 1$ and PWConv_2 is $C\times\frac{C}{r}\times 1\times 1$, respectively. It is noteworthy that $\mathbf{L}(\mathbf{X})$ has the same shape as the input feature, which can preserve and highlight the subtle details in the low-level features. Given the global channel context $\mathbf{g}(\mathbf{X})$ and local channel context $\mathbf{L}(\mathbf{X})$, the refined feature $\mathbf{X}' \in \mathbb{R}^{C\times H\times W}$ by MS-CAM can be obtained as follows:

$$\mathbf{X}' = \mathbf{X}\otimes\mathbf{M}(\mathbf{X}) = \mathbf{X}\otimes\sigma\left(\mathbf{L}(\mathbf{X})\oplus\mathbf{g}(\mathbf{X})\right), \quad (3)$$

where $\mathbf{M}(\mathbf{X}) \in \mathbb{R}^{C\times H\times W}$ denotes the attentional weights generated by MS-CAM. $\oplus$ denotes the broadcasting addition and $\otimes$ denotes the element-wise multiplication.

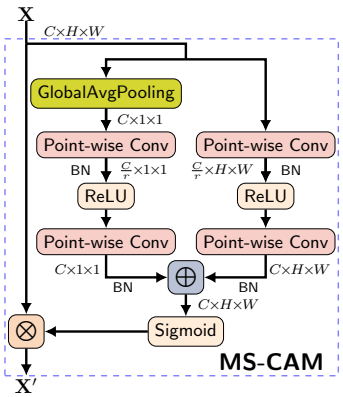


Figure 1: Illustration of the proposed MS-CAM

4. Attentional Feature Fusion

4.1. Unification of Feature Fusion Scenarios

Given two feature maps $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{C\times H\times W}$, by default, we assume $\mathbf{Y}$ is the feature map with a larger receptive field. More specifically,

1. *same-layer scenario*: $\mathbf{X}$ is the output of a 3×3 kernel and $\mathbf{Y}$ is the output of a 5×5 kernel in InceptionNet;
2. *short skip connection scenario*: $\mathbf{X}$ is the identity mapping, and $\mathbf{Y}$ is the learned residual in a ResNet block;
3. *long skip connection scenario*: $\mathbf{X}$ is the low-level feature map, and $\mathbf{Y}$ is the high-level semantic feature map in a feature pyramid.

Based on the multi-scale channel attention module $\mathbf{M}$, *Attentional Feature Fusion* (AFF) can be expressed as

$$\mathbf{Z} = \mathbf{M}(\mathbf{X} \uplus \mathbf{Y}) \otimes \mathbf{X} + (1 - \mathbf{M}(\mathbf{X} \uplus \mathbf{Y})) \otimes \mathbf{Y}, \quad (4)$$

where $\mathbf{Z} \in \mathbb{R}^{C\times H\times W}$ is the fused feature, and $\uplus$ denotes the initial feature integration. In this subsection, for the sake of simplicity, we choose the element-wise summation as initial integration. The AFF is illustrated in Fig. 2(a), where the dashed line denotes $\mathbf{1} - \mathbf{M}(\mathbf{X} \uplus \mathbf{Y})$. It should be noted that the fusion weights $\mathbf{M}(\mathbf{X} \uplus \mathbf{Y})$ consists of real numbers between 0 and 1, so are the $1 - \mathbf{M}(\mathbf{X} \uplus \mathbf{Y})$, which enable the network to conduct a soft selection or weighted averaging between $\mathbf{X}$ and $\mathbf{Y}$.

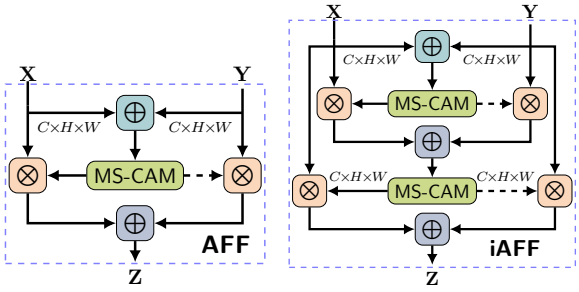


Figure 2: Illustration of the proposed AFF and iAFF

We summarize different formulations of feature fusion in deep networks in Table 1. $\mathbf{G}$ denotes the global attention mechanism. Although there are many implementation differences among multiple approaches for various feature fusion scenarios, once being abstracted into mathematical forms, these differences in details disappear. Therefore, it is possible to *unify these feature fusion scenarios with a carefully designed approach*, thereby improving the performance of all networks by replacing original fusion operations with this unified approach.

From Table 1, it can be further seen that apart from the implementation of the weight generation module $\mathbf{G}$, the state-of-the-art fusion schemes mainly differ in two crucial points: (a) the context-awareness level. Linear approaches like addition and concatenation are entirely contextual unaware. Feature refinement and modulation are non-linear, but only partially aware of the input feature maps. In most cases, they only exploit the high-level feature map. Fully context-aware approaches utilize both input feature maps for guidance at the cost of raising the initial integration issue. (b) Refinement vs modulation vs selection. The sum of weights applied to two feature maps in soft selection approaches are bound to 1, while this is not the case for refinement and modulation.

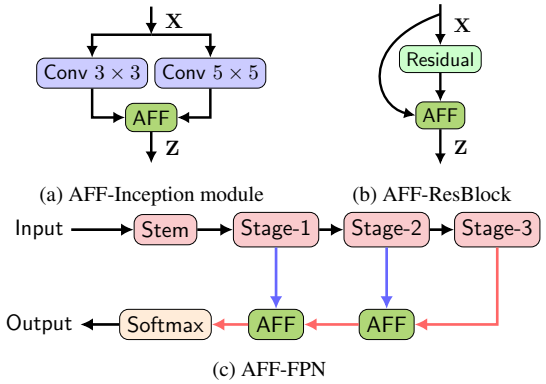


Figure 3: The schema of the proposed AFF-Inception module, AFF-ResBlock, and AFF-FPN. The blue and red lines denote channel expansion and upsampling, respectively.

4.2. Iterative Attentional Feature Fusion

Unlike partially context-aware approaches [18], fully context-aware methods have an inevitable issue, namely how to initially integrate input features. As the input of the attention module, the initial integration quality may profoundly affect final fusion weights. Since it is still a feature fusion problem, an intuitive way is to have another attention module to fuse input features. We call this two-stage approach *iterative Attentional Feature Fusion* (iAFF), which is illustrated in Fig. 2(b). Then, the initial integration $\mathbf{X} \uplus \mathbf{Y}$ in Eq. (4) can be reformulated as

$$\mathbf{X} \uplus \mathbf{Y} = \mathbf{M}(\mathbf{X} + \mathbf{Y}) \otimes \mathbf{X} + (1 - \mathbf{M}(\mathbf{X} + \mathbf{Y})) \otimes \mathbf{Y} \quad (5)$$

4.3. Examples: InceptionNet, ResNet, and FPN

To validate the proposed AFF/iAFF as a uniform and general scheme, we choose ResNet, FPN, and InceptionNet as examples for the most common scenarios: short and long skip connections as well as the same layer fusion. It is straightforward to apply AFF/iAFF to existing networks by replacing the original addition or concatenation. Specifically, we replace the concatenation in the InceptionNet module as well as the addition in ResNet block (ResBlock) and FPN to obtain the attentional networks, which we call AFF-Inception module, AFF-ResBlock, and AFF-FPN, respectively. This replacement and the schemes of our proposed architectures are shown in Fig. 3. The iAFF is a particular case of AFF, so it does not need another illustration.

Table 1: A brief overview of different feature fusion strategies in deep networks.

Context-aware	Type	Formulation	Scenario & Reference	Example
None	Addition	$\mathbf{X}+\mathbf{Y}$	Short Skip [11, 12], Long Skip [24, 21]	ResNet, FPN
	Concatenation	$\mathbf{W_A}\mathbf{X}_{:,i,j}+\mathbf{W_B}\mathbf{Y}_{:,i,j}$	Same Layer [36], Long Skip [28, 15]	InceptionNet, U-Net
Partially	Refinement	$\mathbf{X}+\mathbf{G}(\mathbf{Y})\otimes\mathbf{Y}$	Short Skip [14, 13, 44, 26]	SENet
	Modulation	$\mathbf{G}(\mathbf{Y})\otimes\mathbf{X}+\mathbf{Y}$	Long Skip [18]	GAU
	Soft Selection	$\mathbf{G}(\mathbf{X})\otimes\mathbf{X}+(\mathbf{1}-\mathbf{G}(\mathbf{X}))\otimes\mathbf{Y}$	Short Skip [34]	Highway Networks
Fully	Modulation	$\mathbf{G}(\mathbf{X},\mathbf{Y})\otimes\mathbf{X}+\mathbf{Y}$	Long Skip [46]	SA
	Soft Selection	$\mathbf{G}(\mathbf{X}+\mathbf{Y})\otimes\mathbf{X}+(1-\mathbf{G}(\mathbf{X}+\mathbf{Y}))\otimes\mathbf{Y}$	Same Layer [19, 48]	SKNet
		$\mathbf{M}(\mathbf{X}\uplus\mathbf{Y})\otimes\mathbf{X}+(1-\mathbf{M}(\mathbf{X}\uplus\mathbf{Y}))\otimes\mathbf{Y}$	Same Layer, Short Skip, Long Skip	***ours***

5. Experiments

For experimental evaluation, we resort to the following benchmark datasets: CIFAR-100 [17] and ImageNet [29] for image classification in the same-layer InceptionNet and short-skip connection ResNet scenarios as well as StopSign (a subset of COCO dataset [22]) for semantic segmentation in the long-skip connection FPN scenario. The detailed settings are listed in Table 2. b is the ResBlock number in each stage used to scale the network by depth. Note that our CIFAR-100 experiments classify images into 20 superclasses, not 100 classes. It is a default setting of the CIFAR100 class in MXNet/Gluon. We didn't notice it until a bug issue in our github repo at the camera ready day. However, since all the CIFAR-100 experiments are conducted on the same class number, our conclusion drawn from the experiment results still hold. For more implementation details, please see the supplementary material and our code.

5.1. Ablation Study

5.1.1 Impact of Multi-Scale Context Aggregation

To study the impact of multi-scale context aggregation, in Fig. 4, we construct two ablation modules "Global + Global" and "Local + Local", in which the scales of the two contextual aggregation branches are set as the same, either global or local. The proposed AFF is dubbed as "Global + Local" here. All of them have the same parameter number. The only difference is their *context aggregation scale*.

Table 3 presents their comparison on CIFAR-100, ImageNet, and StopSign on various host networks. It can be seen that the multi-scale contextual aggregation (Global + Local) outperforms single-scale ones in all settings. The results suggest that the multi-scale feature context is vital for the attentional feature fusion.

5.1.2 Impact of Feature Integration Type

Further, we investigate which feature fusion strategy is the best in Table 1. For fairness, we re-implement these approaches based on the proposed MS-CAM for attention weights. Since MS-CAM are different from their original attention modules, we add a prefix of "MS-" to these newly

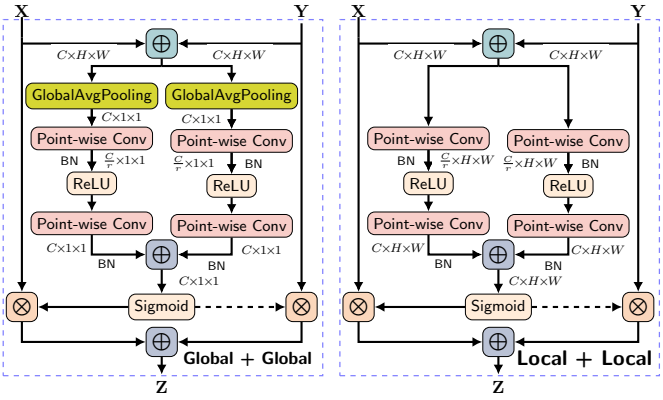
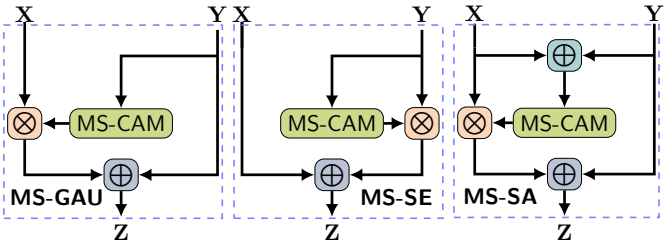


Figure 4: Architectures for the ablation study on the impact of **contextual aggregation scale**.

implemented schemes. To keep the parameter budget the same, here the channel reduction ratio r in MS-GAU, MS-SE, MS-SA, and AFF is 2, while r in iAFF is 4.



(a) MS-GAU (b) MS-SE (c) MS-SA

Figure 5: Architectures for ablation study on the impact of feature integration strategies

Table 4 provides the comparison results in three scenarios, from which it can be seen that: 1) compared to the linear approach, namely addition and concatenation, the nonlinear fusion strategy with attention mechanism always offers better performance; 2) our fully context-aware and selective strategy is slightly but consistently better than the others, suggesting that it should be preferred for multiple feature integration; 3) the proposed iAFF approach is significantly better than the rest in most cases. The results strongly demonstrate our hypothesis that the early integration quality has a large impact on the attentional feature fu-

Table 2: Experimental settings for the networks integrated with the proposed AFF/iAFF.

Task	Dataset	Host Network	Fusing Scenario	r	Epochs	Batch Size	Optimizer	Learning Rate	Learning Rate Mode	Initialization
Image Classification	CIFAR-100	Inception-ResNet-20-b	Same Layer	4	400	128	Nesterov	0.2	Step, $\gamma = 0.1$	Kaiming
		ResNet-20-b	Short Skip	4	400	128	Nesterov	0.2	Step, $\gamma = 0.1$	Kaiming
		ResNeXt-38-32x4d	Short Skip	16	400	128	Nesterov	0.2	Step, $\gamma = 0.1$	Xavier
	ImageNet	ResNet-50	Short Skip	16	160	128	Nesterov	0.075	Cosine	Kaiming
Semantic Segmentation	StopSign	ResNet-20-b + FPN	Long Skip	4	300	32	AdaGrad	0.01	Poly	Kaiming

Table 3: Comparison of **contextual aggregation scales** in attentional feature fusion given the same parameter budget. The results suggest that a mix of scales should always be preferred inside the channel attention module.

Aggregation Scale	InceptionNet on CIFAR-100				ResNet on CIFAR-100				ResNet + FPN on StopSign				ResNet on ImageNet
	$b=1$	$b=2$	$b=3$	$b=4$	$b=1$	$b=2$	$b=3$	$b=4$	$b=1$	$b=2$	$b=3$	$b=4$	
Global + Global	0.735	0.766	0.775	0.789	0.754	0.796	0.811	0.821	0.911	0.923	0.936	0.939	0.777
Local + Local	0.746	0.771	0.785	0.787	0.754	0.794	0.808	0.814	0.895	0.919	0.921	0.924	0.780
Global + Local	**0.756**	**0.784**	**0.794**	**0.801**	**0.763**	**0.804**	**0.816**	**0.826**	**0.924**	**0.935**	**0.939**	**0.944**	**0.784**

Table 4: Comparison of **context-aware level** and **feature integration strategy** in feature fusion *given the same parameter budget*. The results suggest that a fully context-aware and selective strategy should always be preferred for feature fusion. If no problem in optimization, we should adopt the iterative attentional feature fusion without hesitation for better performance.

Fusion Type	Context	Strategy	InceptionNet (Same Layer)				ResNet (Short Skip)				ResNet + FPN (Long Skip)			
			$b=1$	$b=2$	$b=3$	$b=4$	$b=1$	$b=2$	$b=3$	$b=4$	$b=1$	$b=2$	$b=3$	$b=4$
Add	None	\	0.720	0.753	0.771	0.782	0.740	0.786	0.797	0.808	0.895	0.920	0.925	0.928
Concatenation	None	\	0.725	0.749	0.772	0.779	0.742	0.782	0.793	0.798	0.897	0.909	0.925	0.939
MS-GAU	Partially	Modulation	0.751	0.774	0.788	0.795	0.766	0.803	0.815	0.819	0.917	0.926	0.937	0.941
MS-SENet	Partially	Refinement	0.752	0.780	0.790	0.798	0.765	0.799	0.814	0.820	0.915	0.929	0.940	0.940
MS-SA	Fully	Modulation	0.756	0.779	0.790	0.798	0.761	0.801	0.814	0.822	0.920	0.932	0.938	0.941
AFF (***ours***)	Fully	Selection	*0.756*	*0.784*	*0.794*	*0.801*	*0.763*	*0.804*	*0.816*	**0.826**	*0.924*	*0.935*	*0.939*	*0.944*
iAFF (***ours***)	Fully	Selection	**0.774**	**0.801**	**0.808**	**0.814**	**0.772**	**0.807**	**0.822**	/	**0.927**	**0.938**	**0.945**	**0.953**

sion, and another level of attentional feature fusion can further improve the performance. However, this improvement may be obtained at the cost of increasing the difficulty in optimization. We notice that when the network depth increases as b changes from 3 to 4, the performance of iAFF-ResNet did not improve but degraded.

5.1.3 Impact on Localization and Small Objects

To study the impact of the proposed MS-CAM on object localization and small object recognition, we apply Grad-CAM [30] to ResNet-50, SENet-50, and AFF-ResNet-50 for the visualization results of images from the ImageNet dataset, which are illustrated in Fig. 6. Given a specific class, Grad-CAM results show the network's attended regions clearly. Here, we show the heatmaps of the predicted class, and the wrongly predicted image is denoted with the symbol ✖. The predicted class names and their softmax scores are also shown at the bottom of heatmaps.

From the upper part of Fig. 6, it can be seen clearly that the attended regions of the AFF-ResNet-50 highly overlap with the labeled objects, which shows that it learns well to localize objects and exploit the features in object regions. On the contrary, the localization capacity of the baseline ResNet-50 is relatively poor, misplacing the center of attended regions in many cases. Although SENet-50 are able to locate the true objects, the attended regions are overlarge including many background components. It is because SENet-50 only utilizes the global channel attention, which is biased to the context of a global scale, whereas the proposed MS-CAM also aggregates the local channel context, which helps the network to attend the objects with fewer background clutters and is also beneficial to the small object recognition. In the bottom half of Fig. 6, we can clearly see that AFF-ResNet-50 can predict correctly on the small-scale objects, while ResNet-50 fails in most cases.

5.2. Comparison with State-of-the-Art Networks

To show that the network performance can be improved by replacing original fusion operations with the proposed

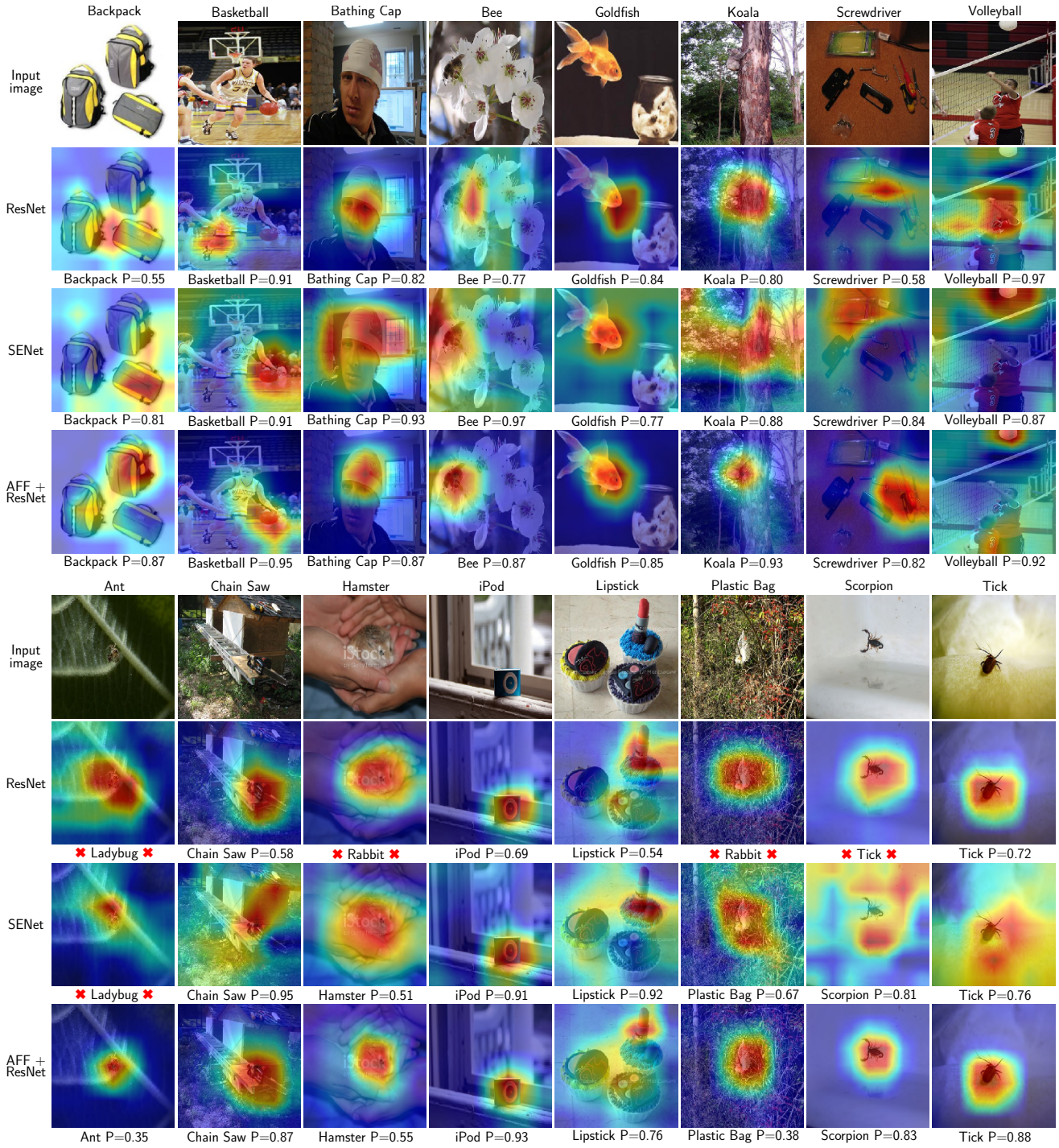
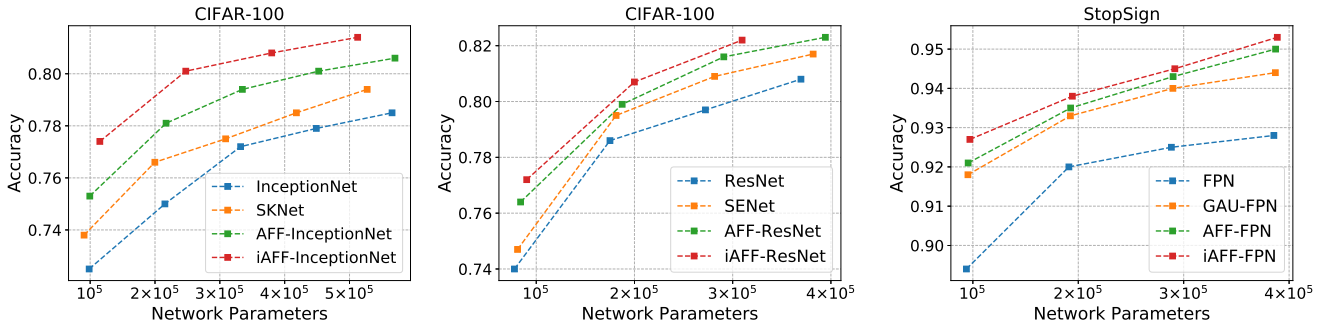


Figure 6: Network visualization with Grad-CAM. The comparison results suggest that the proposed MS-CAM is beneficial to the object localization and small object recognition.

attentional feature fusion, we compare the AFF and iAFF modules with other attention modules based on the same host networks in different feature fusion scenarios. Fig. 7 illustrates the comparison results with a gradual increase in network depth for all networks. It can be seen that: 1) Comparing SKNet / SENet / GAU-FPN with AFF-InceptionNet / AFF-ResNet / AFF-FPN, we can see that our AFF or iAFF integrated networks are better in all scenarios, which shows that our (iterative) attentional feature fusion approach not only has superior performance, but a good generality. We



(a) InceptionNet (Same layer) (b) ResNet (Short skip connection) (c) FPN (Long skip connection)

Figure 7: Compassion with baseline and other state-of-the-art networks with a gradual increase of network depth.

believe the improved performance comes from the proposed multi-scale channel contextual aggregation inside the attention module. 2) Comparing the performance of iAFF-based networks with AFF-based networks, it should be noted that the proposed iterative attentional feature fusion scheme can further improve the performance. 3) By replacing the simple addition or concatenation with the proposed AFF or iAFF module, we can get a more efficient network. For example, in Fig. 7(b), iAFF-ResNet ($b = 2$) achieves similar performance with the baseline ResNet ($b = 4$), while only 54% of the parameters were required.

Last, we validate the performance of AFF/iAFF based networks with state-of-the-art networks on ImageNet. The results are listed in Table 5. The results show that the proposed AFF/iAFF based networks can improve performance over the state-of-the-art networks under much smaller parameter budgets. Remarkably, on ImageNet, the proposed iAFF-ResNet-50 outperforms Gather-Excite-θ^+-ResNet-101 [13] by 0.3% with only 60% parameters. These results indicate that the feature fusion in short skip connections matters a lot for ResNet and ResNeXt. Instead of blindly increasing the depth of the network, we should pay more attention to the quality of feature fusion.

Table 5: Comparison on **ImageNet**

Architecture	top-1 err.	Params
ResNet-101 [11]	23.2	42.5 M
Efficient-Channel-Attention-Net-101 [40]	21.4	42.5 M
Attention-Augmented-ResNet-101 [1]	21.3	45.4 M
SENet-101 [14]	20.9	49.4 M
Gather-Excite-θ^+-ResNet-101 [13]	20.7	58.4 M
Local-Importance-Pooling-ResNet-101 [9]	20.7	42.9 M
AFF-ResNet-50 (ours)	***20.9***	***30.3 M***
AFF-ResNeXt-50-32x4d (ours)	***20.8***	***29.9 M***
iAFF-ResNet-50 (ours)	***20.4***	***35.1 M***
iAFF-ResNeXt-50-32x4d (ours)	***20.2***	***34.7 M***

6. Conclusion

We generalize the concept of attention mechanisms as a selective and dynamic type of feature fusion to most scenarios, namely the same layer, short skip, and long skip connections as well as information integration inside the attention mechanism. To overcome the semantic and scale inconsistency issue among input features, we propose the multi-scale channel attention module, which adds local channel contexts to the global channel-wise statistics. Further, we point out that the initial integration of received features is a bottleneck in attention-based feature fusion, and it can be alleviated by adding another level of attention that we call iterative attentional feature fusion. We conducted detailed ablation studies to empirically verify the individual impact of the context-aware level, the feature integration type, and the contextual aggregation scales of our proposed attention mechanism. Experimental results on both the CIFAR-100 and the ImageNet dataset show that our models outperform state-of-the-art networks with fewer layers or parameters per network, which suggests that one should pay attention to the feature fusion in deep neural networks and that more sophisticated attention mechanisms for feature fusion hold the potential to consistently yield better results.

Acknowledgement

The authors would like to thank the editor and anonymous reviewers for their helpful comments and suggestions, and also thank @takedarts on Github for pointing out the bug in our CIFAR-100 code. This work was supported in part by the National Natural Science Foundation of China under Grant No. 61573183, the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) under Grant No. 201900029, the Nanjing University of Aeronautics and Astronautics PhD short-term visiting scholar project under Grant No. 180104DF03, the Excellent Chinese and Foreign Youth Exchange Program, China Association for Science and Technology, China Scholarship Council under Grant No. 201806830039.

References

[1] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. Attention augmented convolutional networks. In *2019 IEEE International Conference on Computer Vision (ICCV), Seoul, Korea (South)*, pages 3286–3295, October 2019.

[2] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA*, pages 3640–3649, 2016.

[3] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L. Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA*, pages 5669–5678. IEEE Computer Society, 2017.

[4] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA*, pages 8554–8564. Computer Vision Foundation / IEEE, 2019.

[5] Yang Feng, Deqian Kong, Ping Wei, Hongbin Sun, and Nanning Zheng. A benchmark dataset and multi-scale attention network for semantic traffic light detection. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand*, pages 1–8. IEEE, 2019.

[6] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA*, pages 3146–3154, 2019.

[7] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. JL-DCF: joint learning and densely-cooperative fusion framework for RGB-D salient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA*, pages 3049–3059. IEEE, 2020.

[8] Keren Fu, Qijun Zhao, Irene Yu-Hua Gu, and Jie Yang. Deepside: A general deep framework for salient object detection. *Neurocomputing*, 356:69–82, Sep 2019.

[9] Ziteng Gao, Limin Wang, and Gangshan Wu. LIP: local importance-based pooling. In *2019 IEEE International Conference on Computer Vision (ICCV), Seoul, Korea (South)*, pages 3354–3363. IEEE, 2019.

[10] Bharath Hariharan, Pablo Andrés Arbeláez, Ross B. Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA*, pages 447–456. IEEE Computer Society, 2015.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA*, pages 770–778, 2016.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands*, pages 630–645, 2016.

[13] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Andrea Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. In *Annual Conference on Neural Information Processing Systems (NeurIPS) 2018, Montréal, Canada*, pages 9423–9433, 2018.

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA*, pages 7132–7141, 2018.

[15] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA*, pages 2261–2269, 2017.

[16] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *the 32nd International Conference on Machine Learning (ICML), Lille, France*, pages 448–456, 2015.

[17] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[18] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. In *British Machine Vision Conference (BMVC) 2018, Newcastle, UK*, pages 1–13, 2018.

[19] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA*, pages 510–519, 2019.

[20] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhao-Xiang Zhang. Scale-aware trident networks for object detection. In *2019 IEEE International Conference on Computer Vision (ICCV), Seoul, Korea (South)*, pages 6053–6062, 2019.

[21] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA*, pages 936–944, 2017.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *13th European Conference on Computer Vision (ECCV), Zurich, Switzerland*, pages 740–755, Cham, 2014.

[23] Wei Liu, Andrew Rabinovich, and Alexander C. Berg. Parsenet: Looking wider to see better. *CoRR*, abs/1506.04579, 2015.

[24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA*, pages 3431–3440, 2015.

[25] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *the 27th International Conference on Machine Learning (ICML), Haifa, Israel*, ICML'10, pages 807–814, USA, 2010.

[26] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. BAM: bottleneck attention module. In *British*

Machine Vision Conference (BMVC) 2018, Newcastle, UK, pages 1–14, 2018.

[27] Shaoqing Ren, Kaiming He, Ross B. Girshick, Xiangyu Zhang, and Jian Sun. Object detection networks on convolutional feature maps. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(7):1476–1481, 2017.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany*, pages 234–241, 2015.

[29] 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. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[30] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2020.

[31] Bharat Singh and Larry S. Davis. An analysis of scale invariance in object detection - SNIP. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA*, pages 3578–3587, June 2018.

[32] Bharat Singh, Mahyar Najibi, and Larry S. Davis. SNIPER: efficient multi-scale training. In *Annual Conference on Neural Information Processing Systems (NeurIPS) 2018, Montréal, Canada*, pages 9333–9343, 2018.

[33] Ashish Sinha and Jose Dolz. Multi-scale self-guided attention for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, pages 1–14, Apr 2020.

[34] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. In *Annual Conference on Neural Information Processing Systems (NeurIPS) 2015, Montreal, Quebec, Canada*, pages 2377–2385, 2015.

[35] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, California, USA*, pages 4278–4284, 2017.

[36] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA*, pages 1–9, 2015.

[37] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA*, pages 2818–2826, 2016.

[38] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Annual Conference on Neural Information Processing Systems (NeurIPS) 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[40] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA*, pages 11534–11542, 2020.

[41] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven C. H. Hoi, and Ali Borji. Salient object detection with pyramid attention and salient edges. In *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA*, pages 1448–1457. Computer Vision Foundation / IEEE, 2019.

[42] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA*, pages 7794–7803, 2018.

[43] Yi Wang, Haoran Dou, Xiaowei Hu, Lei Zhu, Xin Yang, Ming Xu, Jing Qin, Pheng-Ann Heng, Tianfu Wang, and Dong Ni. Deep Attentive Features for Prostate Segmentation in 3D Transrectal Ultrasound. *IEEE Transactions on Medical Imaging*, 38(12):2768–2778, Apr 2019.

[44] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. In *15th European Conference on Computer Vision (ECCV), Munich, Germany*, pages 3–19, 2018.

[45] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA*, pages 5987–5995, 2017.

[46] Weitao Yuan, Shengbei Wang, Xiangrui Li, Masashi Unoki, and Wenwu Wang. A skip attention mechanism for monaural singing voice separation. *IEEE Signal Processing Letters*, 26(10):1481–1485, 2019.

[47] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Richard C. Wilson, Edwin R. Hancock, and William A. P. Smith, editors, *British Machine Vision Conference (BMVC) 2016, York, UK*. BMVA Press, 2016.

[48] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. ResNeSt: Split-Attention Networks. *arXiv e-prints*, page arXiv:2004.08955, Apr. 2020.

[49] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. S3FD: Single shot scale-invariant face detector. In *2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy*, Oct 2017.