Attentional Feature Fusion

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

[1] Nanjing University of Aeronautics and Astronautics
[2] University of Münster
[3] University of Copenhagen
[4] University of Arizona

November 26, 2020

Our Work in A Nutshell

Nested Optimization on the Fusion of Received Features

Generalizing the Channel Attention to **Multi-Scale** Cases

Feature Fusion Is Omnipresent in Modern Networks

Linear Aggregation Approaches

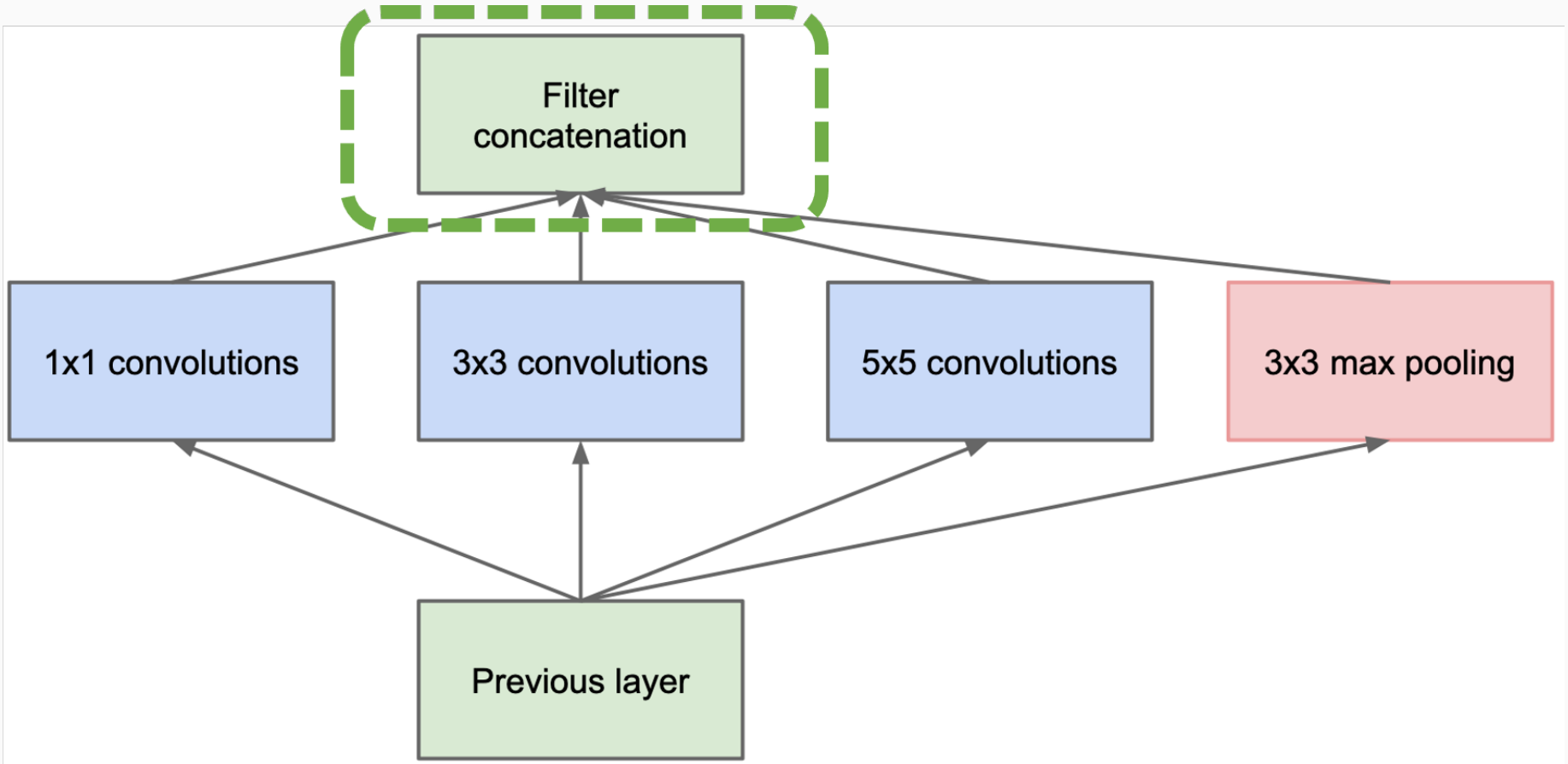


Figure 1: Concatenation in Inception Module [1] – In the Same Layer

Linear Aggregation Approaches

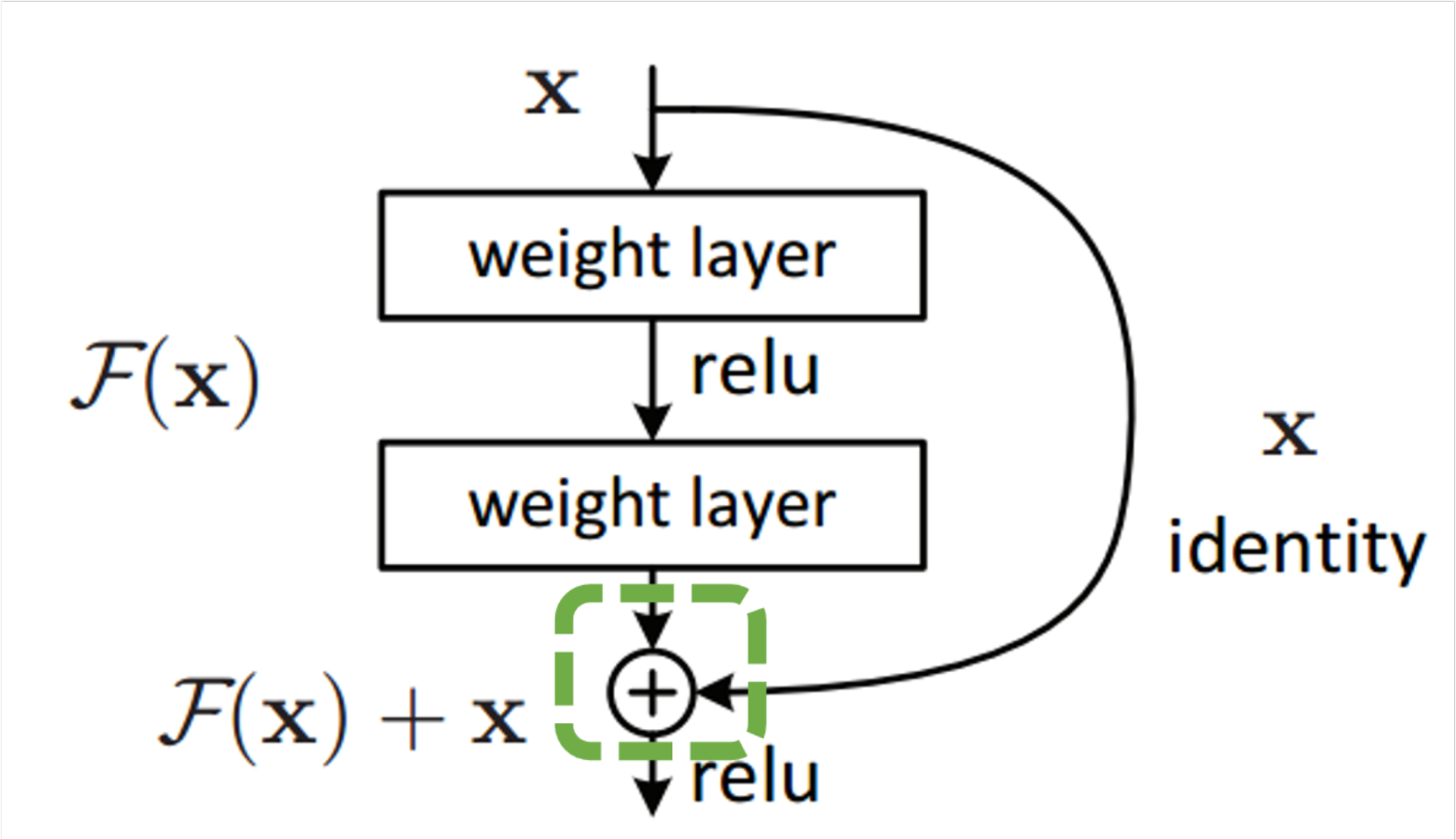


Figure 2: Addition in Residual Block [2] – Short Skip Connection

Linear Aggregation Approaches

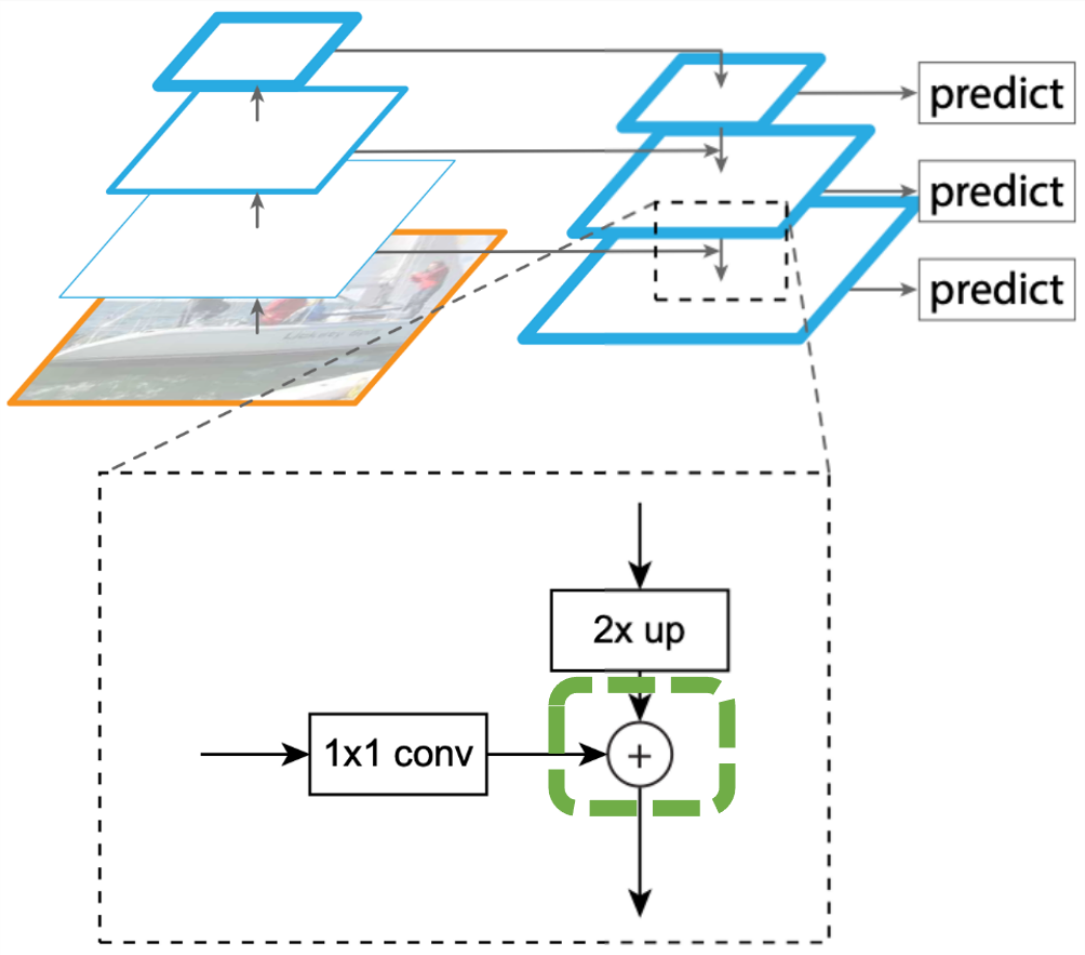


Figure 3: Addition in Feature Pyramid [3] – Long Skip Connection

Linear Aggregation Approaches

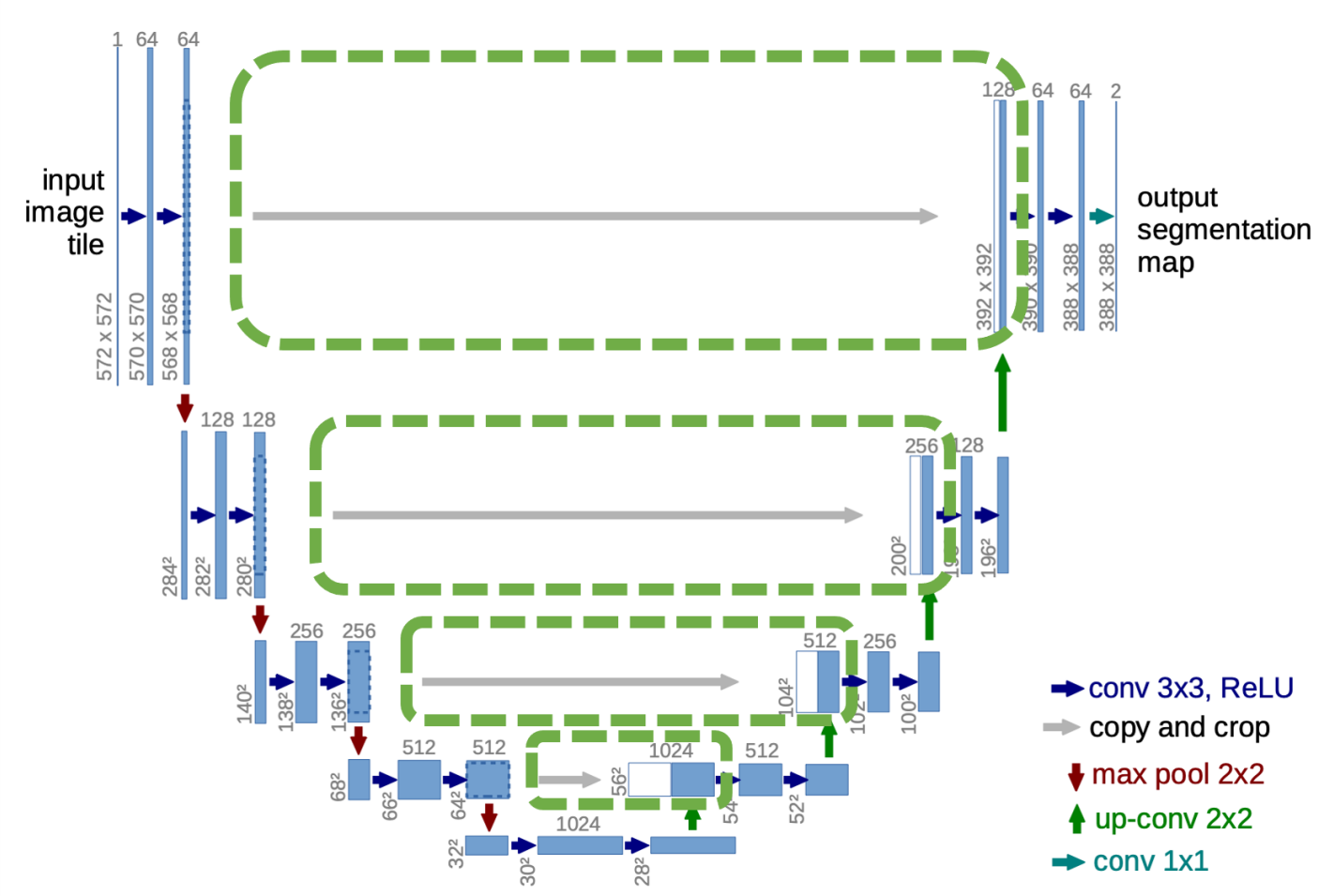


Figure 4: Concatenation in U-Net [4] – Long Skip Connection

Prior Work – Non-linear Approach

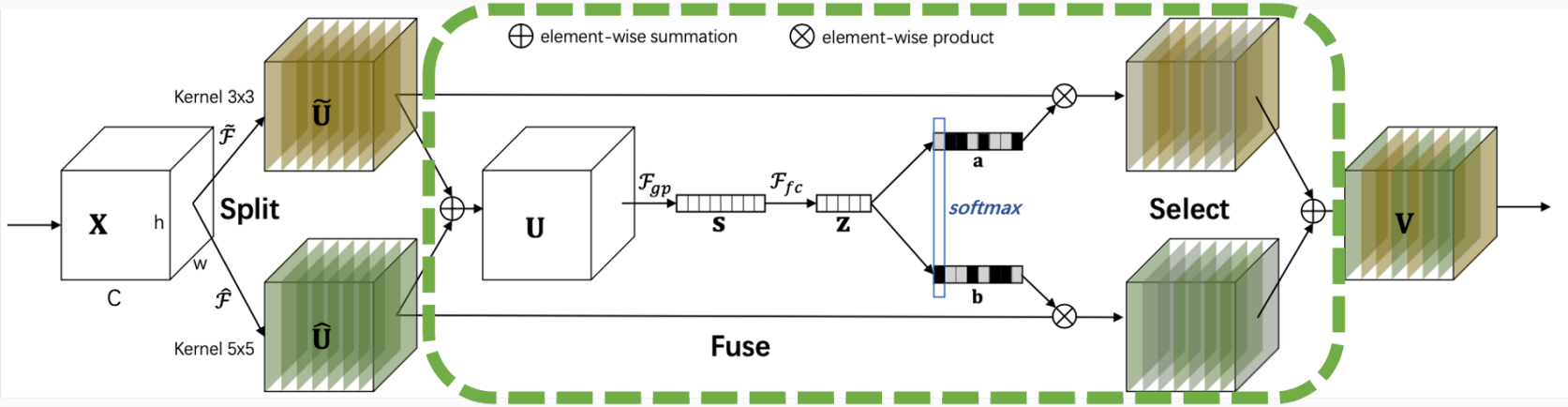


Figure 5: Selective Kernel Networks [5] – **In the Same Layer**

Fusing Same-Layer Multiple Branches Using Softmax Attention (Squeeze-and-Excitation Networks Style [6]).

- A Non-linear Dynamic Selection Mechanism

Shortcomings and Motivation

1. Limited Scenarios
    - Cross-Layer Feature Fusion Has Not Been Discussed
2. Unsophisticated Initial Integration
    - SKNet Introduces Another Phase of Feature Fusion
    - This Initial Integration Has a Large Impact on Fusion Quality
3. Biased Context Aggregation Scale
    - Channel Attenion in SENet Biased to Large Objects
    - **Can Channel Attenion Be Multi-Scale?**

Contributions

1. To Tackle The Issue of Limited scenarios
    - Extending to Short and Long Skip Connections
    - Unifying All Kinds of Feature Fusion Scenarios
2. To Tackle the Issue of Unsophisticated Initial Integration
    - Iteratively Integrating Features with Another Attention Module
3. To Tackle the Issue of Biased Context Aggregation Scale
    - Introducing Scale into Channel Attention by Varying the Pooling Size
    - Multi-Scale Channel Attention Module (MS-CAM)

Method – Multi-Scale Channel Attention Module

Observation: Spatial Pooling Size Reflects the Scale of Channel Scale

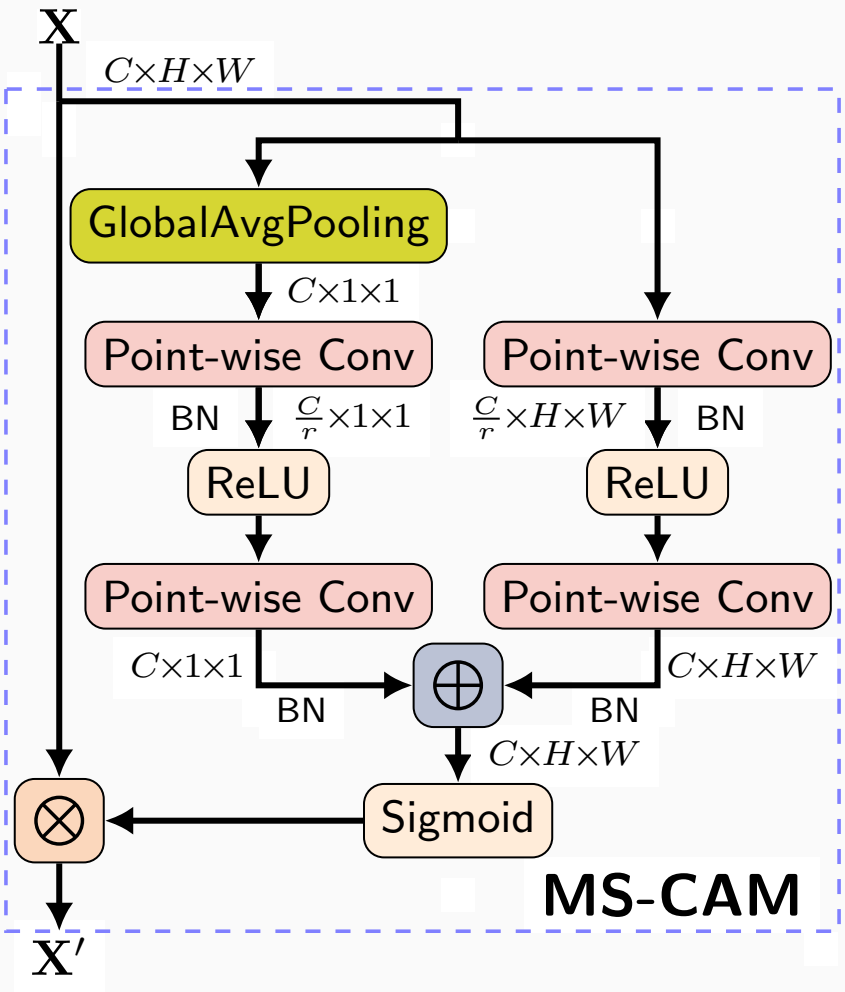


Figure 6: Multi-Scale Channel Attention Module (MS-CAM)

Method – Attentional Feature Fusion

It Is Still a Feature Fusion Problem

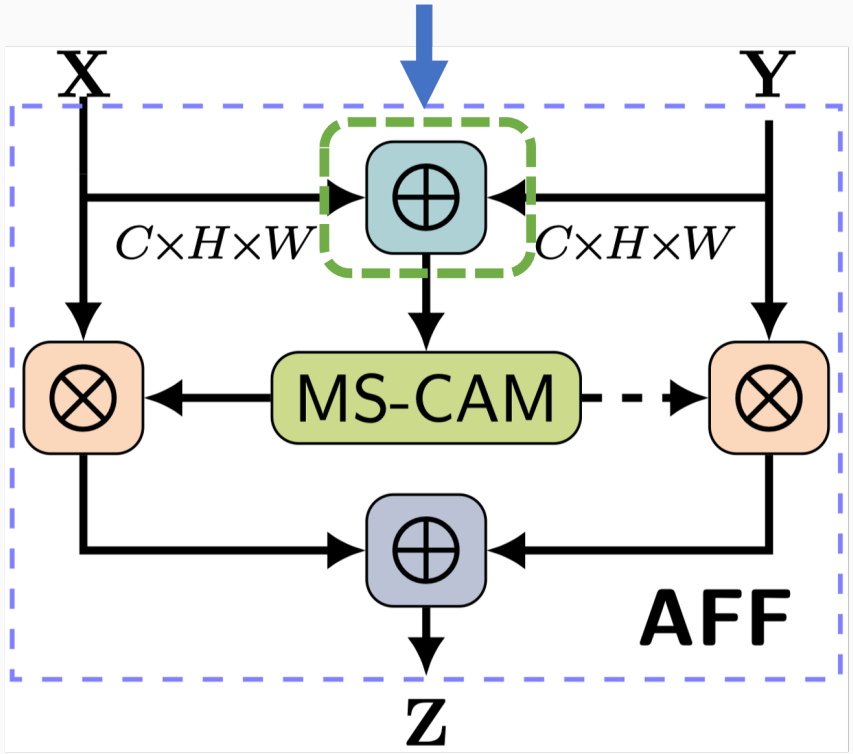


Figure 7: Attentional Feature Fusion (AFF)

Method – Iterative Attentional Feature Fusion

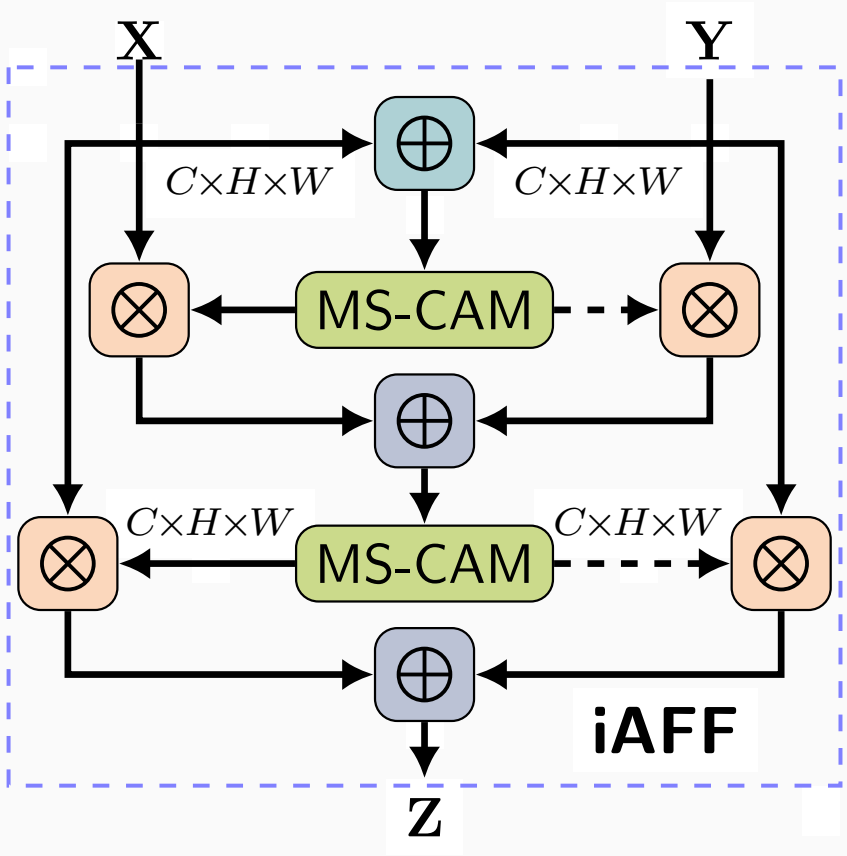
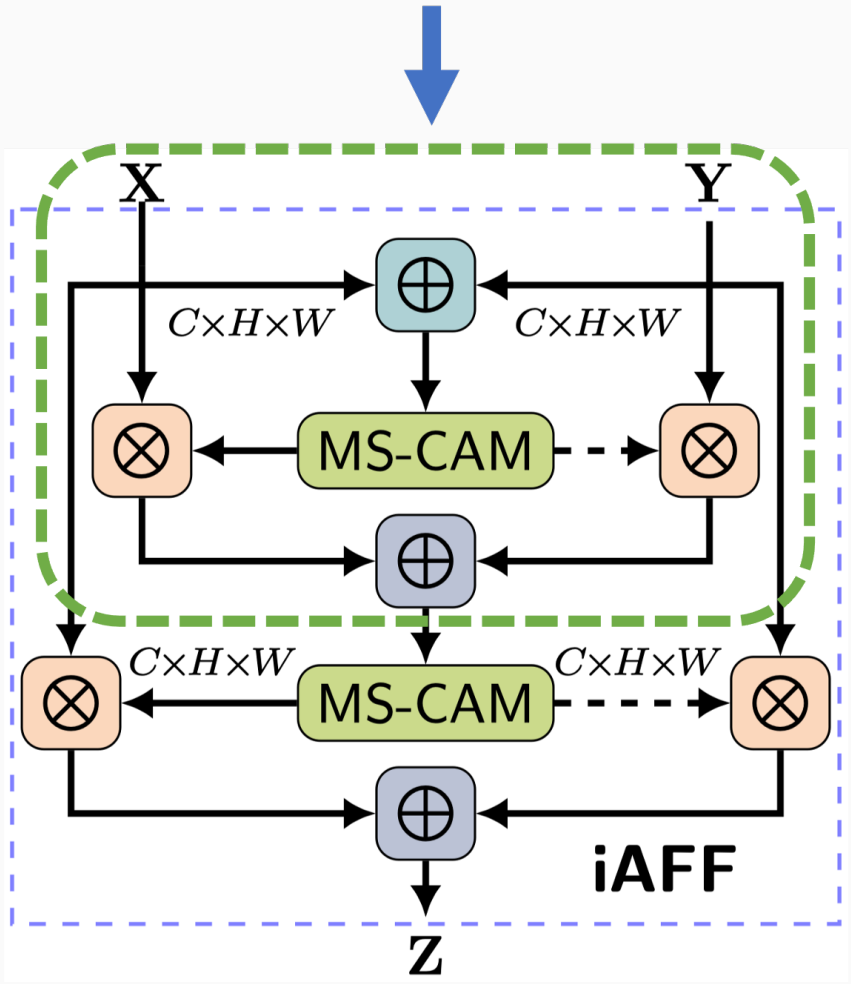


Figure 8: iterative Attentional Feature Fusion (iAFF)

Method – Iterative Attentional Feature Fusion

Replacing $\oplus$ With Another Level of Attention



Examples: InceptionNet, ResNet, and FPN

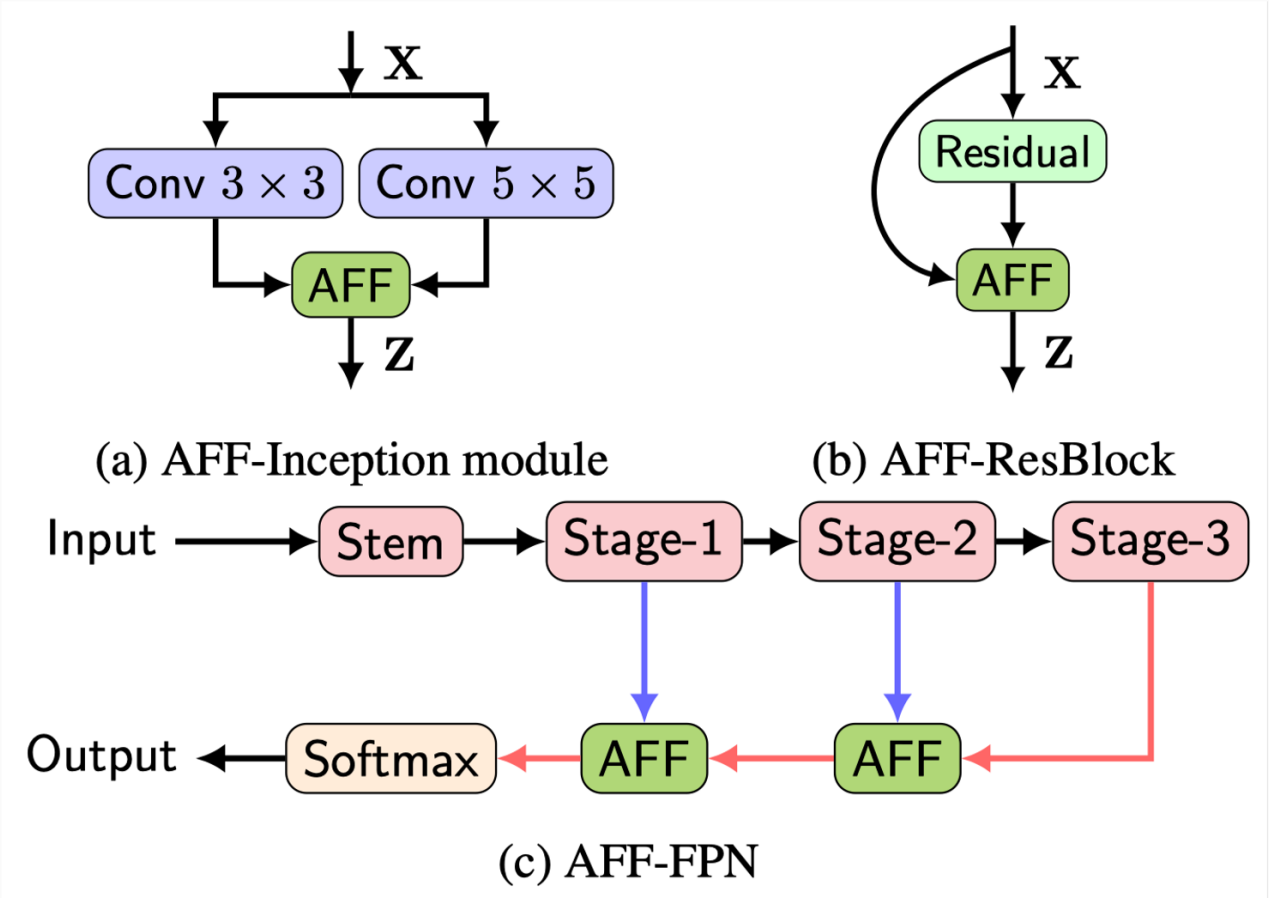


Figure 9: AFF-Inception module, AFF-ResBlock, and AFF-FPN

Ablation Study on Impact of Multi-Scale Context Aggregation

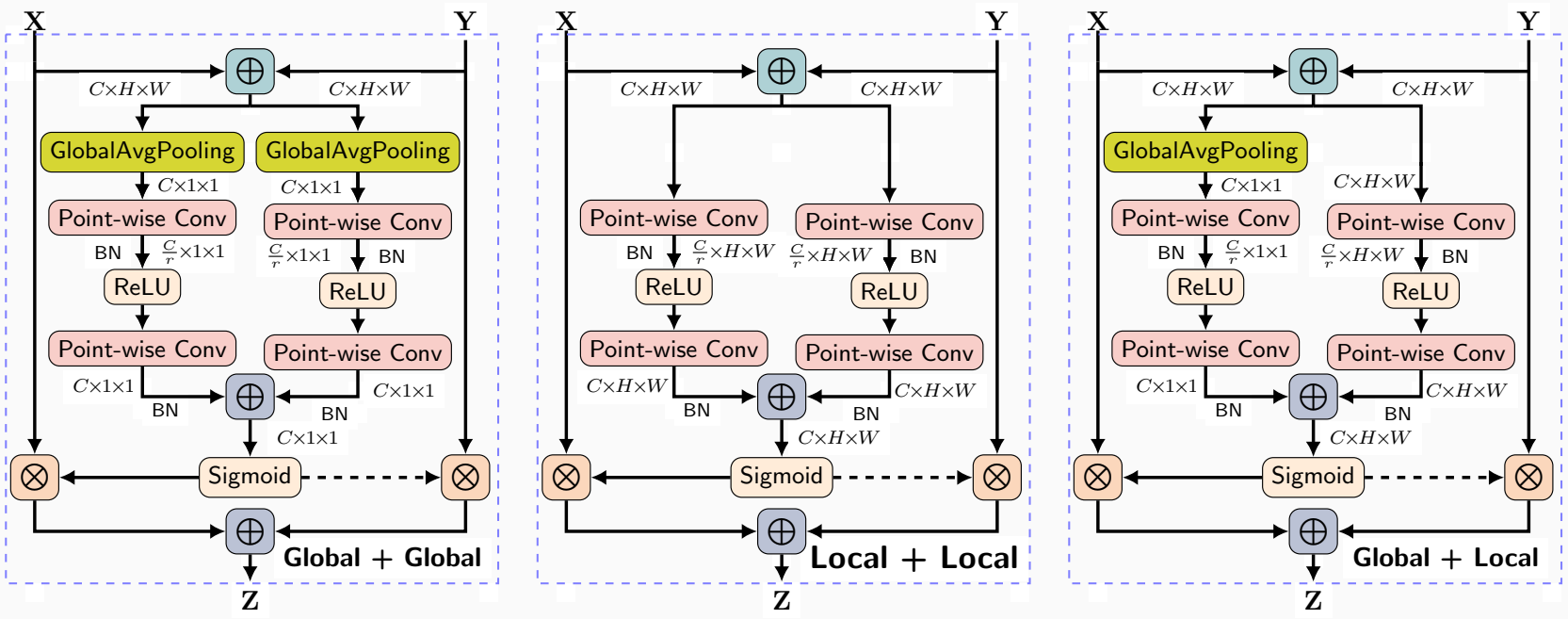


Figure 10: Ablation study on the impact of contextual aggregation scale.

Ablation Study on Impact of Multi-Scale Context Aggregation

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**

Results: Comparison to Other Networks

Table 1: Comparison on **ImageNet**

Architecture	top-1 err.	Params
ResNet-101 [2]	23.2	42.5 M
Efficient-Channel-Attention-Net-101 [7]	21.4	42.5 M
Attention-Augmented-ResNet-101 [8]	21.3	45.4 M
SENet-101 [6]	20.9	49.4 M
Gather-Excite-θ^+-ResNet-101 [9]	20.7	58.4 M
Local-Importance-Pooling-ResNet-101 [10]	20.7	42.9 M
iAFF-ResNet-50 (ours)	***20.2***	***35.1 M***
iAFF-ResNeXt-50-32x4d (ours)	***19.8***	***34.7 M***

Results: Visualization

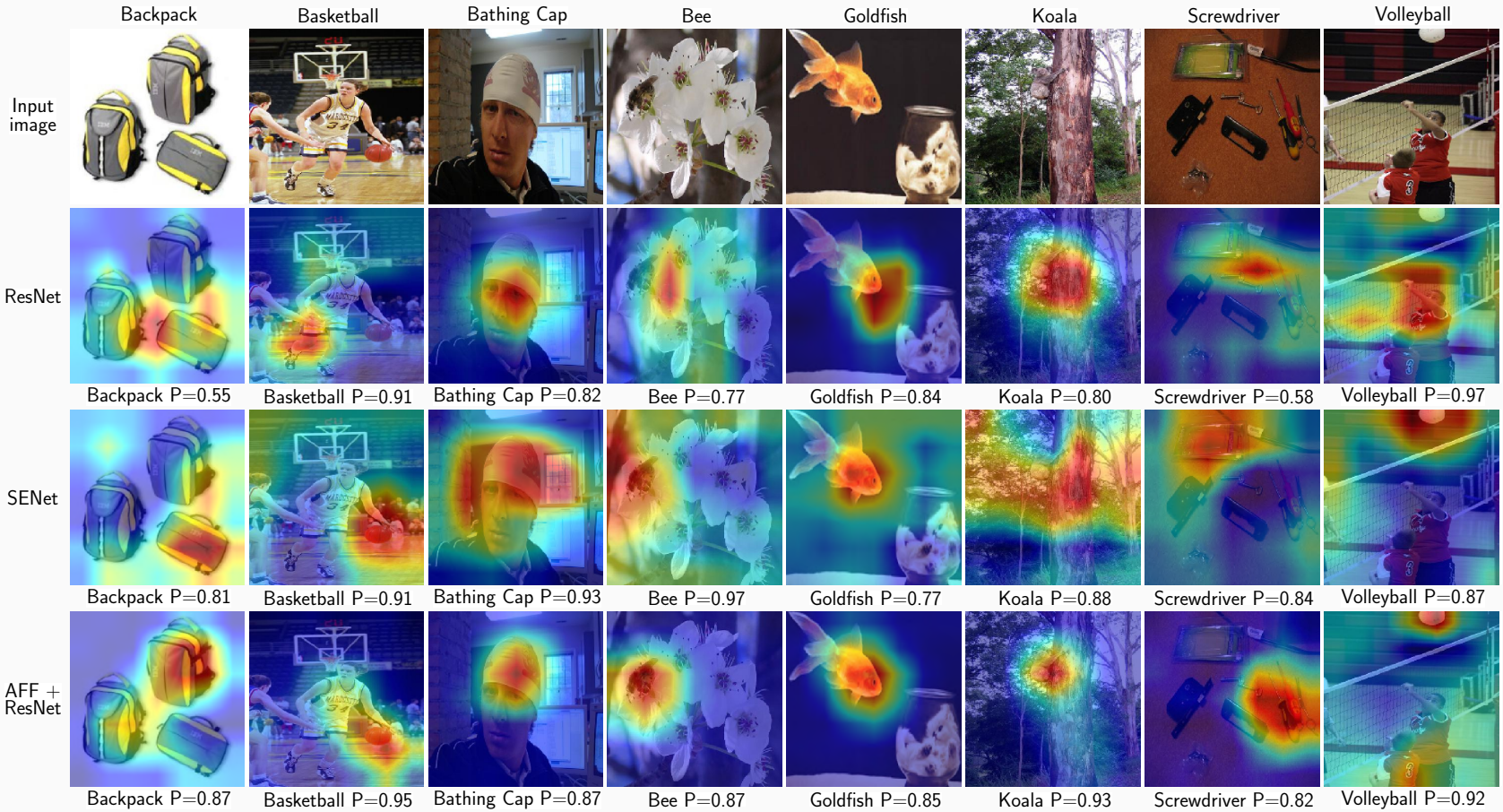


Figure 11: Better Localization

Results: Visualization

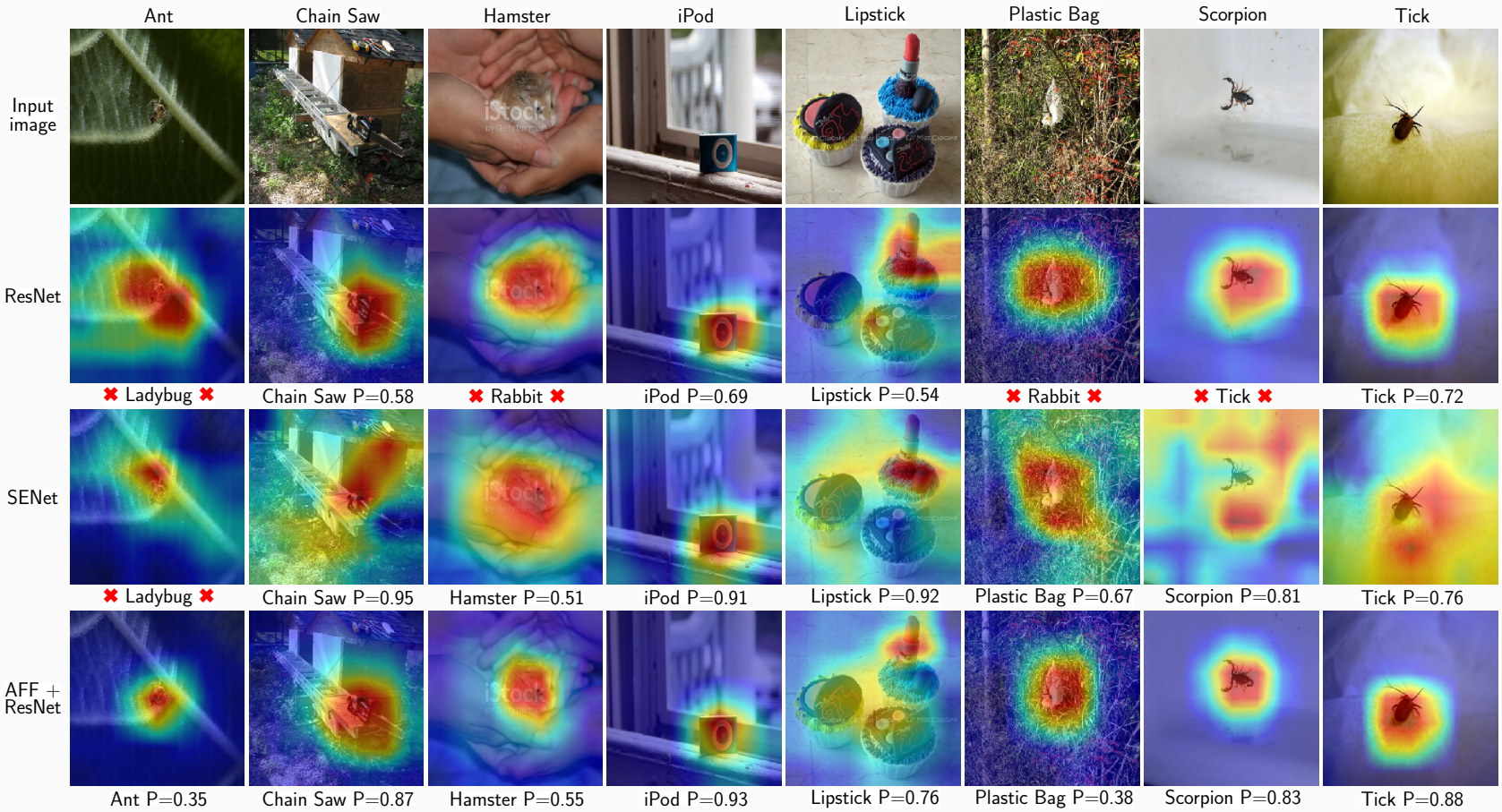


Figure 12: Especially Better Localization on Small Objects

Codes and Trained Models

Our codes and trained models is available at

`https://github.com/YimianDai/open-aff`

Thank you!

References i

[1] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[3] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 936–944.

References ii

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 2015, pp. 234–241.

[5] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 510–519.

[6] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141.

[7] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11 534–11 542.

References iii

[8] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in 2019 IEEE International Conference on Computer Vision (ICCV), October 2019, pp. 3286–3295.

[9] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," in Annual Conference on Neural Information Processing Systems (NeurIPS), 2018, pp. 9423–9433.

[10] Z. Gao, L. Wang, and G. Wu, "LIP: local importance-based pooling," in 2019 IEEE International Conference on Computer Vision (ICCV). IEEE, 2019, pp. 3354–3363.