OverLoCK: An Overview-first-Look-Closely-next ConvNet with Context-Mixing Dynamic Kernels (Supplementary Material)

Meng Lou Yizhou Yu
School of Computing and Data Science, The University of Hong Kong
loumeng@connect.hku.hk, yizhouy@acm.org

A. More Ablation Studies

On the basis of the training settings outlined in Section 4.4, we additionally conduct a series of in-depth ablation experiments to meticulously examine the impact of every component in our proposed method.

Impact of Kernel Sizes. We compared the performance under various settings of kernel sizes, as outlined in Table 6 (the definition of the kernel size in our proposed method is given in Section 3.3). The results indicate that the configuration $\{[17, 15, 13], [7], [13, 7]\}$ yields the optimal performance on both image classification and semantic segmentation tasks. Further enlarging the kernels does not lead to additional improvements.

Table A. Ablation study of the kernel size setting.

Kernel Sizes	# F (G)	# P (M)	Top-1 (%)	mIoU (%)
$\{[19, 17, 15], [7], [15, 7]\}$	2.8	16.5	80.7	43.8
$\{[17, 15, 13], [7], [13, 7]\}$	2.6	16.4	**80.8**	**43.8**
$\{[13, 11, 9], [7], [9, 7]\}$	2.6	16.3	80.5	43.5
$\{[9, 9, 7], [7], [7, 7]\}$	2.6	16.1	80.6	43.3
$\{[7, 7, 7], [7], [7, 7]\}$	2.5	16.1	80.4	43.1

Impact of Stage Ratio. The *Stage Ratio* means the ratio between the number of blocks in the last stage of Base-Net and the number of blocks in the first stage of Focus-Net. In the default setting of the OverLoCK model, the stage ratio is 1:2 with the intention of allocating more network blocks to Focus-Net for extracting robust contextual information. In this section, we investigate the impact of *Stage Ratio*. Apart from the default setting of 1:2, we further set *Stage Ratio* to 1:1 and 1:3 while maintaining the total number of network blocks constant. The results presented in Table B demonstrate that a *Stage Ratio* of 1:2 yields the best outcomes. We posit that this is because a too small *Stage Ratio* results in insufficient number of blocks in Focus-Net, thereby hindering the extraction of discriminative deep features. Conversely, an excessively large *Stage Ratio* leads to a shortage of blocks in Base-Net, thereby providing insufficient contextual guidance.

Table B. Ablation study of different stage ratio settings.

Stage Ratio	# F (G)	# P (M)	Top-1 (%)	mIoU (%)
1:1	2.7	16.1	80.4	42.9
1:2	2.6	16.4	**80.8**	**43.8**
1:3	2.7	15.9	80.6	43.6

Impact of Channel Reduction Factor. In the default configuration of the OverLoCK model, we employ a 1×1 convolution to reduce the number of output channels of Overview-Net by a factor of 4 and concatenate this result with the output of Base-Net before forwarding it to Focus-Net. We term this reduction as the *Channel Reduction Factor (CRF)*. Therefore, the value of *CRF* determines the number of channels in the *context prior*, thereby influencing the guidance capability. In this regard, we investigate the effects of different *CRF* settings. It is important to note that during the adjustment of *CRF*, we also modify the number of channels of Focus-Net to maintain similar complexities across different model variants. The results in Table C demonstrate that *CRF*=4 yields the optimal performance.

Table C. Ablation study of channel reduction factor settings.

CRF	# F (G)	# P (M)	Top-1 (%)	mIoU (%)
2	2.6	16.1	80.5	42.9
4	2.6	16.4	**80.8**	**43.8**
6	2.7	16.6	80.7	43.4
8	2.7	16.7	80.6	43.0

Impact of Auxiliary Loss. To explore the effects of applying the auxiliary loss to Overview-Net, we adjust the weight of the auxiliary loss, drawing inspiration from prior research [10]. Given that the architectures of the models in this comparison are consistent, we opt not conduct further

experiments on segmentation tasks for the sake of simplicity. The results presented in Table D indicate that the utilization of an auxiliary loss improves accuracy, while varying the weight of the auxiliary loss does not lead to a notable impact on performance. This observation aligns with findings in previous study [10].

Table D. Ablation study of auxiliary loss.

Aux Loss Ratio	0	0.2	0.4	0.8	1.0
Top-1 (%)	80.4	80.7	**80.8**	80.7	80.7

Effectiveness of our DDS-based Top-down Network. To evaluate the effectiveness of the proposed DDS, we reconstruct our OverLoCK-XT model as a standard hierarchical network. To be specific, we eliminate the top-down attention mechanism by removing the Overview-Net while keeping the same types of layers in the Base-Net and Focus-Net. To maintain comparable complexity with other models, the number of channels and layers in the four stages are set to [64, 112, 256, 360] and [2, 2, 9, 4], respectively. This model is denoted as the "Hierarchical Model". Additionally, we compare it with the "Baseline" model in Table 6 which is a fully static ConvNet. As shown in Table E, the "Hierarchical Model" results in a noticeable performance drop, demonstrating the effectiveness of our DDS-based top-down context guidance. However, when compared with the "Baseline" model, it still exhibits significant advantages, clearly indicating the superiority of our proposed dynamic convolution module.

Table E. Effectiveness of the proposed DDS-based top-down network.

Method	# F (G)	# P (M)	Top-1 (%)	mIoU (%)
Baseline Model	2.6	16.3	78.5	41.1
Hierarchical Model	2.7	16.2	79.2	41.9
OverLoCK-XT	2.6	16.4	**80.7**	**43.8**

Ablation Study of ContMix. We conduct a comprehensive comparison of various components within our proposed ContMix framework, as presented in Section 3.2. As listed in Table F, we initially compute $\mathbf{Q}$ and $\mathbf{K}$ using the fused feature map instead of utilizing the channels of $\mathbf{X}$ corresponding to $\mathbf{Z}_i$ and $\mathbf{P}_i$ (the latest *context prior*). This model variant, referred to as "Fusion Affinity", results in a marginal performance decline. Subsequently, we interchange the features used to generate the $\mathbf{Q}$ and $\mathbf{K}$ matrices. This model, denoted as "Reverse QK", also exhibits a decrease in performance. Furthermore, we individually eliminate the Softmax function (referred to as "w/o Softmax"), remove the RepConv (referred to as "w/o RepConv"), and substitute small kernels with large kernels (referred to as "w/o Small Kernel"). These alterations decrease performance on both classification and segmentation tasks.

Table F. Ablation study of ContMix.

Method	# F (G)	# P (M)	Top-1 (%)	mIoU (%)
Baseline	2.6	16.4	**80.8**	**43.8**
Fusion Affinity	2.7	16.6	80.7	43.5
Reverse QK	2.7	16.4	80.6	42.9
w/o Softmax	2.6	16.4	80.5	43.5
w/o RepConv	2.5	16.1	80.6	43.4
w/o Small Kernel	2.8	16.6	80.7	43.3

Table G. A comparison of image classification with 384×384 inputs.

Method	Type	# F (G)	# P (M)	Acc. (%)
Swin-B	T	47.1	88	84.5
MaxViT-B	T	74.2	120	85.7
ConvNeXt-B	C	45.2	88	85.1
InceptionNeXt-B	C	43.6	87	85.2
RDNet-L	C	101.9	186	85.8
PeLK-B-101	C	68.3	90	85.8
OverLock-B	C	50.4	95	**86.2**

Table H. Robustness comparisons of different models.

Models	# F (G)	# P (G)	1K	V2	A	R	Sketch
Swin-T	4.5	28	81.3	69.7	21.1	41.5	29.3
VMamba-T	4.9	29	82.6	72.0	27.0	45.4	32.9
ConvNeXt-T	4.5	29	82.1	72.5	24.2	47.2	33.8
HorNet-T	4.0	22	82.8	72.3	26.6	46.6	34.1
SLaK-T	5.0	30	82.5	72.0	30.0	45.3	32.4
NAT-T	4.3	28	83.2	72.2	33.0	44.9	31.9
RDNet-T	5.0	24	82.8	72.9	27.7	49.0	37.0
UniRepLKNet-T	4.9	25	83.2	72.8	34.8	49.4	36.9
MogaNet-S	5.0	33	83.4	72.6	33.4	49.7	37.8
OverLoCK-T	5.5	33	**84.2**	**74.0**	**39.4**	**53.3**	**40.6**
Swin-S	8.7	50	83.0	72.0	32.5	45.2	32.3
VMamba-S	8.7	50	83.6	73.2	33.2	49.4	37.0
ConvNeXt-S	8.7	50	83.1	72.5	31.3	49.6	37.1
HorNet-S	8.8	50	84.0	73.6	36.2	49.7	36.9
SLaK-S	9.8	55	83.8	73.6	39.3	50.9	37.5
NAT-S	7.8	51	83.7	73.2	37.4	47.3	34.3
RDNet-S	8.7	50	83.7	73.8	33.5	52.8	39.8
UniRepLKNet-S	9.1	56	83.9	73.7	38.3	50.6	36.9
MogaNet-B	9.9	44	84.3	74.3	40.4	50.1	38.6
OverLoCK-S	9.7	56	**84.8**	**74.9**	**45.0**	**57.2**	**45.8**
Swin-B	15.4	88	83.5	72.4	35.4	46.5	32.7
VMamba-B	15.4	89	83.9	73.5	37.2	49.5	38.5
ConvNeXt-B	15.4	89	83.8	73.7	36.7	51.2	38.2
HorNet-B	15.6	87	84.3	73.9	39.9	51.2	38.1
SLaK-B	17.1	95	84.0	74.0	41.6	50.8	38.5
NAT-B	13.7	90	84.3	74.1	41.4	49.7	36.6
RDNet-B	15.4	87	84.4	74.2	38.1	52.7	40.1
MogaNet-L	15.9	83	84.7	74.0	41.0	52.2	39.0
OverLoCK-B	16.7	95	**85.1**	**75.4**	**47.7**	**58.5**	**46.0**

Table I. Speed comparison among various models. Throughput (Thr.) is tested on a single NVIDIA L40S GPU with a batch size of 128 and an image size of 3×224×224.

Method	# F (G)	# P (M)	Thr. (imgs/s)	Acc. (%)
Swin-T	4.5	28	1324	81.3
Swin-S	8.7	50	812	83.0
Swin-B	15.4	88	544	83.5
MaxViT-T	5.6	31	683	83.7
MaxViT-S	11.7	69	439	84.5
MaxViT-B	24.0	120	241	84.9
NAT-M	2.7	20	1740	81.8
NAT-T	4.3	28	1287	83.2
NAT-S	7.8	51	823	83.7
NAT-B	13.7	90	574	84.3
BiFormer-T	2.2	13	1103	81.4
BiFormer-S	4.5	26	527	83.8
BiFormer-B	9.8	57	341	84.3
VMamba-T	4.9	29	1179	82.6
VMamba-S	8.7	50	596	83.6
VMamba-B	15.4	89	439	83.9
ConvNeXt-T	4.5	29	1507	82.1
ConvNeXt-S	8.7	50	926	83.1
ConvNeXt-B	15.4	89	608	83.8
FocalNet-T	4.5	29	1251	82.3
FocalNet-S	8.7	50	777	83.5
FocalNet-B	15.4	89	481	83.7
SLaK-T	5.0	30	1126	82.5
SLaK-S	9.8	55	747	83.8
SLaK-B	17.1	95	478	83.7
InternImage-T	5.0	30	1084	83.5
InternImage-S	8.0	50	740	84.2
InternImage-B	16.0	97	481	84.9
UniRepLKNet-N	2.8	18	1792	81.6
UniRepLKNet-T	4.9	31	1094	83.2
UniRepLKNet-S	9.1	56	707	83.9
MogaNet-S	5.0	25	766	83.4
MogaNet-B	9.9	44	373	84.3
MogaNet-L	15.9	83	282	84.7
OverLoCK-XT	2.6	16	1672	**82.7**
OverLoCK-T	5.5	33	810	**84.2**
OverLoCK-S	9.7	56	480	**84.8**
OverLoCK-B	16.7	95	306	**85.1**

B. Additional Experiments on Image Classification

B.1. Large Resolution Evaluation

Following previous works [3, 4, 9], we further investigate the image classification performance on the ImageNet-1K dataset at a higher resolution (i.e., 384×384). Specifically, we pre-train the base model on 224×224 inputs and then fine-tune it on 384×384 inputs for 30 epochs. As shown in Table G, our OverLock-B model achieves superior performance under high-resolution input conditions. Notably, OverLock-B surpasses MaxViT-B by 0.5% in Top-1 accuracy while reducing the parameter count by over one-third. Compared to PeLK-B, a large kernel ConvNet, our method also demonstrates significant improvements. These results further validate the robustness of our proposed method in handling large-resolution inputs.

B.2. Robustness Evaluation

We further assess the robustness of our models using the ImageNet out-of-distribution (OOD) benchmarks, including ImageNet-V2 [6], ImageNet-A [2], ImageNet-R [1], and ImageNet-Sketch [8]. As shown in Table H, our method demonstrates excellent robustness on different datasets, outperforming representative ConvNets, Vision Transformers, and Vision Mamba. Notably, although OverLoCK-B improves over MogaNet-L by 0.4% in Top-1 accuracy on ImageNet-1K, it achieves significant gains on OOD datasets, with improvements of 1.4% on ImageNet-V2, 6.7% on ImageNet-A, 6.3% on ImageNet-R, and 6.8% on ImageNet-Sketch. These results showcase the strong robustness of our pure ConvNet.

C. Speed Analysis

We provide a comparison of speed-accuracy trade-off in Figure 1. More details are listed in Table 1, where an OverLoCK variant often achieves faster speed and higher accuracy simultaneously than a larger variant of another network, demonstrating an excellent trade-off between speed and accuracy. For instance, OverLoCK-XT achieves 1672 imgs/s in throughput, improving upon Swin-T by over 300 imgs/s, while significantly enhancing Top-1 accuracy by 1.4%. Also, OverLoCK-T achieves about 200 imgs/s improvement in throughput compared to ConvNeXt-B while achieving better performance at the cost of only around one-third of the FLOPS. When compared to more advanced models, OverLoCK still exhibits significant advantages. For example, OverLoCK-S surpasses MogaNet-B by over 100 imgs/s in throughput while increasing Top-1 accuracy from 84.3% to 84.8%. Likewise, OverLoCK-XT surpasses BiFormer-T by over 600 imgs/s in throughput while remarkably improving Top-1 accuracy by 1.3%.

D. Visualization Analysis

D.1. Effect of Context Guidance

To visually understand the effect of context guidance, we separately visualize the class activation maps generated by

Figure A. Class activation maps of the proposed OverLoCK network. (a), (b), and (c) show the input images, class activation maps of Overview-Net, and class activation maps of Focus-Net, respectively.

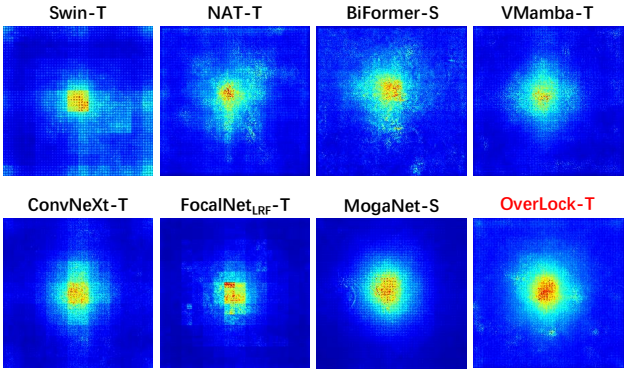

Figure B. Comparison of ERF among various models.

Overview-Net and Focus-Net in OverLoCK-T using Grad-CAM [7] for the ImageNet-1K validation set. As shown in Figure A, Overview-Net first produces a coarse localization of an object, and when this signal is used as the top-down guidance for Focus-Net, the object's location and shape becomes more accurate.

D.2. Effective Receptive Field Analysis

To visually demonstrate the representation capacity of OverLoCK, we compare the Effective Receptive Field (ERF) [5] of our OverLoCK-T with that of other representative models with comparable complexity. The visualizations are generated using over 300 randomly sampled images with a resolution of 224×224 from the ImageNet-1K validation set. As shown in Figure B, our model not only produces global responses but also exhibits significant local sensitivity, indicating that OverLoCK can effectively model both global and local contexts simultaneously.

References

[1] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 3

[2] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021. 3

[3] Donghyun Kim, Byeongho Heo, and Dongyoon Han. Densenets reloaded: Paradigm shift beyond resnets and vits. In *European Conference on Computer Vision*, 2024. 3

[4] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 3

[5] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 29, 2016. 4

[6] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine lLearning*, pages 5389–5400. PMLR, 2019. 3

[7] 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*, 128:336–359, 2020. 4

[8] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[9] Weihao Yu, Pan Zhou, Shuicheng Yan, and Xinchao Wang. Inceptionnext: When inception meets convnext. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[10] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 1, 2