NomMer: Nominate Synergistic Context in Vision Transformer for Visual Recognition - Supplementary Materials

Hao Liu[†*] Xinghua Jiang[*] Xin Li Zhimin Bao Deqiang Jiang Bo Ren
Tencent YouTu Lab
{ivanhliu, clarkjiang, fujikoli, zhiminbao, dqiangjiang, timren}@tencent.com

A. Preliminary: Multi-Head Self-Attention

Multi-Head Self-Attention (MHSA) [10] is the core component of the ViT model, on which we build the Synergistic Context Nominator (SCN) and G-NomMer layer in our NomMer. Here, we briefly review this preliminary knowledge. Given queries $\mathbf{Q}$, keys $\mathbf{K}$ and values $\mathbf{V}$, Multi-Head Attention (MHA) is formulated as:

$$\begin{aligned}
&\textit{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \textit{Concat}(\mathbf{H}_1,\mathbf{H}_2,...,\mathbf{H}_h)\mathbf{W}^*,\\
&\mathbf{H}_i = \textit{Attention}(\mathbf{Q}\mathbf{W}_i^Q,\mathbf{K}\mathbf{W}_i^K,\mathbf{V}\mathbf{W}_i^V), i\in\{1,2,...,h\},\\
&\textit{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \textit{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}})\mathbf{V},
\end{aligned}$$

where h is the head number and d_k is the key dimension. $\mathbf{W}_i^Q \in \mathbb{R}^{d_m\times d_k}$, $\mathbf{W}_i^K \in \mathbb{R}^{d_m\times d_k}$, $\mathbf{W}_i^V \in \mathbb{R}^{d_m\times d_v}$ and $\mathbf{W}_i^* \in \mathbb{R}^{hd_v\times d_m}$ are corresponding projection matrices. In particular, the MHA becomes MHSA on the condition that $\mathbf{X}=\mathbf{Q}=\mathbf{K}=\mathbf{V}$, where $\mathbf{X}$ denotes the input. In the G-MHSA (Fig. 3 in main text) and G-NomMer layer (Fig. 2(b) in main text) of our NomMer, the global relations of all tokens are captured by MHSA while the local dependencies of the tokens falling inside the window are built by L-MHSA (Fig. 3 in main text).

B. More Implementation Details

The NomMer architecture is implemented by Pytorch [8] and all experiments are conducted on a workstation with 32 NVIDIA A100-80 GB GPUs. All the reported results are the averaged ones over 10 random seeds. The detailed configurations of NomMer variants and the core pseudo code of Synergistic NomMer block are elaborated as follows.

B.1. Detailed Configurations of NomMer Variants

The configurations of three variants of NomMer are given in Tab. 3, including NomMer-T, NomMer-S and NomMer-B, which refer to tiny, small and base model separately. In the "S-NomMer", the parameters of three types of candidate context aggregators (from top to bottom are "CNN", "L-MHSA", "G-MHSA") are separated by dotted lines.

B.2. Pseudo Code of Synergistic NomMer Block

Algorithm 1 S-NomMer pseudo code.

Input: $\mathbf{F}$
Output: $\mathbf{F}^{(S)}$
/* Compressed Global Context Aggregator. */
Function CGCA($\mathbf{F}$)**:**
  /* Discrete Cosine Transform. */
  $\mathbf{f} \leftarrow \textit{DCT}(\mathbf{F})$
  /* Low-Frequency Perceiver. */
  $\hat{\mathbf{f}} \leftarrow \textit{LFP}(\mathbf{f})$
  $\hat{\mathbf{f}}^{(G)} \leftarrow \textit{Conv}(\textit{G-MHSA}(\hat{\mathbf{f}}))$
  /* Inverse Discrete Cosine Transform. */
  $\mathbf{F}^{(G)} \leftarrow \textit{IDCT}(\hat{\mathbf{f}}^{(G)})$
  return $\mathbf{F}^{(G)}$
/* Synergistic Context Nominator. */
Function SCN($\mathbf{F}^{(L)},\mathbf{F}^{(C)},\mathbf{F}^{(G)}$)**:**
  $\mathbf{\Omega} \leftarrow \textit{Conv}(\mathbf{F}^{(L)}+\mathbf{F}^{(C)}+\mathbf{F}^{(G)})$
  $\mathbf{F}^{(S)} \leftarrow \textit{Nominate}(\mathbf{\Omega},\mathbf{F}^{(L)},\mathbf{F}^{(C)},\mathbf{F}^{(G)})$
  return $\mathbf{F}^{(S)}$
/* Synergistic NomMer. */
Function S-NomMer**:**
  $\mathbf{F}^{(L)} \leftarrow \textit{L-MHSA}(\mathbf{F})$
  $\mathbf{F}^{(C)} \leftarrow \textit{CNN}(\mathbf{F})$
  $\mathbf{F}^{(G)} \leftarrow \textit{CGCA}(\mathbf{F})$
  $\mathbf{F}^{(S)} \leftarrow \textit{SCN}(\mathbf{F}^{(L)},\mathbf{F}^{(C)},\mathbf{F}^{(G)})$
  $\mathbf{F}^{(S)} \leftarrow \mathbf{F}^{(S)}+\textit{FFN}(\mathbf{F}^{(S)})$
  return $\mathbf{F}^{(S)}$

C. Image Classification on ImageNet-21K

Experimental setting. We further pre-train NomMer on the larger ImageNet-21K dataset, which contains 14.2M im-

[*]Equal contribution. [†]Contact person.

ages and 21K classes. During pre-training stage, we employ the AdamW [7] optimizer with a weight decay of 10^{-2}, an initial learning rate of 10^{-2}, and a batch size of 4,096 for 90 epochs with the input size of 224^2. In ImageNet-1K finetuning, we train the models for 30 epochs with a batch size of 1,024, a constant learning rate of 10^{-5}, and a weight decay of 10^{-8} on 224^2 and 384^2 input.

ImageNet-21K 224^2 finetuned models			
Method	#param. (M)	FLOPs (G)	Top-1 (%)
Swin-B [6]	88	15.4	85.2
NomMer-B	73	17.6	**85.5**
ImageNet-21K 384^2 finetuned models			
Method	#param. (M)	FLOPs (G)	Top-1 (%)
R-101x3 [4]	388	204.6	84.4
ViT-B/16 [2]	86	55.4	84.0
ViL-B [14]	56	43.7	86.2
Swin-B [6]	88	47.1	86.4
NomMer-B	73	56.2	**86.6**

Table 1. Comparison of different backbones on ImageNet-21K classification.

Performance. The results of pre-training on ImageNet-21K are summarized in Tab. 1, from which we can see that our NomMer-B obtains 85.5% and 86.6% top-1 accuracy under 224^2 and 384^2 input size setting. Compared with the second best method, Swin [6], our method can outperform it by at least 0.2%.

D. Object Detection on COCO with Various Frameworks

Experimental setting. To further verify the effectiveness of our proposed NomMer when working in different detection frame works, we conduct extensive experiments by training four typical object detectors on COCO dataset, including Cascade Mask R-CNN [1], ATSS [15], RepPointsV2 [13] and Sparse R-CNN [9]. For a fair comparison, we utilize the same experimental settings for all four detectors. More concretely, all the models are training with $3\times$ schedule, multi-scale training, AdamW [7] optimizer, initial learning rate of 10^{-4}, weight decay of 5×10^{-2}, and batch size of 16.

Performance. The box mAPs on COCO are reported in Tab. 2. As one can see, the NomMer-T witnesses the substantive improvements on the performance of different detectors, which demonstrates our NomMer architecture is a versatile backbone for various object detection approaches.

Method	Backbone	#param. (M)	FLOPs (G)	AP^b (%)	AP^b_{50} (%)	AP^b_{75} (%)
Cascade Mask R-CNN [1]	Res50 [3]	82	739	46.3	64.3	50.5
	Swin-T [6]	86	745	50.5	69.3	54.9
	Focal-T [12]	87	770	51.5	70.6	55.9
	NomMer-T	80	755	**51.8**	**70.8**	**56.0**
ATSS [15]	Res50 [3]	32	205	43.5	61.9	47.0
	Swin-T [6]	36	212	47.2	66.5	51.3
	Focal-T [12]	37	239	49.5	**68.8**	53.9
	NomMer-T	30	237	**49.8**	68.6	**54.0**
RepPointsV2 [13]	Res50 [3]	43	431	46.5	64.6	50.3
	Swin-T [6]	44	437	50.0	68.5	54.2
	Focal-T [12]	45	491	51.2	70.4	54.9
	NomMer-T	41	486	**51.6**	**70.7**	**55.1**
Sparse R-CNN [9]	Res50 [3]	106	166	44.5	63.4	48.2
	Swin-T [6]	110	172	47.9	67.3	52.3
	Focal-T [12]	111	196	49.0	69.1	53.2
	NomMer-T	104	195	**49.5**	**69.3**	**53.7**

Table 2. Results on COCO object detection across different object detection methods.

E. More Visual Interpretability

E.1. Nomination Maps on Different Tasks

For the image classification task, we visualize more nomination maps in Fig. 1 of this appendix, which also follows the similar patterns described in the Sec. 4.5 of main text.

Additionally, in Fig. 2 of this appendix, we also visualize the nomination maps from intermediate layers of S-NomMer blocks when semantic segmentation and object detection frameworks adopting NomMer-B as backbones (corresponding to the results in Tab. 2 and Tab. 3 of main text). Compared to the maps on classification task, the nomination maps of dense prediction tasks exhibit different context synergy patterns. Although the CNN context features are also predominant in low-level nomination maps ("Layer 1_1"), the context aggregated by L-MHSA is hardly observed in the first stage layers ("Layer 1_ $\sim$"). Instead, most regions of "Layer 1_3" maps mainly pick up global context accompanied with CNN context focusing on the salient object details, such as outlines of wheels or buildings. The "Layer 1_5" maps further witness the domination of global context. We attribute these phenomena to the larger spatial size of early-stage blocks providing sufficiently precise spatial information, which is indispensable to the dense prediction tasks. This explanation can also be confirmed by the Non-local Neural Networks [11], where the most significant improvement on performance is achieved by inserting the SA-based non-local block in the early stage. In contrast, the "non-local" behavior in our method is implemented in a more graceful way with object

	Output Size	Layer Name	NomMer-T	NomMer-S	NomMer-B
	56 * 56	Patch Embedding	dim 96, conv 4*4	dim 96, conv 4*4	dim 128, 4*4
stage1	56*56	S-NomMer Block	$\begin{bmatrix} \text{dim 96, conv 1*1,} \\ \text{conv 3*3, conv 1*1} \\ \cdots\cdots \\ \text{[dim 96, head 2, wsize 7]} \\ \cdots\cdots \\ \text{[dim 96, head 2, ksize 8]} \end{bmatrix} \times 2$	$\begin{bmatrix} \text{dim 96, conv 1*1,} \\ \text{conv 3*3, conv 1*1} \\ \cdots\cdots \\ \text{[dim 96, head 2, wsize 7]} \\ \cdots\cdots \\ \text{[dim 96, head 2, ksize 8]} \end{bmatrix} \times 6$	$\begin{bmatrix} \text{dim 128, conv 1*1,} \\ \text{conv 3*3, conv 1*1} \\ \cdots\cdots \\ \text{[dim 128, head 2, wsize 7]} \\ \cdots\cdots \\ \text{[dim 128, head 2, ksize 8]} \end{bmatrix} \times 6$
	28*28	Reduction Module	dim 192, conv 3*3, pool /2	dim 192, conv 3*3, pool /2	dim 256, conv 3*3, pool /2
stage2	28*28	S-NomMer Block	$\begin{bmatrix} \text{dim 192, conv 1*1,} \\ \text{conv 3*3, conv 1*1} \\ \cdots\cdots \\ \text{[dim 192, head 2, wsize 7]} \\ \cdots\cdots \\ \text{[dim 192, head 2, ksize 4]} \end{bmatrix} \times 2$	$\begin{bmatrix} \text{dim 192, conv 1*1,} \\ \text{conv 3*3, conv 1*1} \\ \cdots\cdots \\ \text{[dim 192, head 2, wsize 7]} \\ \cdots\cdots \\ \text{[dim 192, head 2, ksize 4]} \end{bmatrix} \times 6$	$\begin{bmatrix} \text{dim 256, conv 1*1,} \\ \text{conv 3*3, conv 1*1} \\ \cdots\cdots \\ \text{[dim 256, head 2, wsize 7]} \\ \cdots\cdots \\ \text{[dim 256, head 2, ksize 4]} \end{bmatrix} \times 6$
	14*14	Reduction Module	dim 384, conv 3*3, pool /2	dim 384, conv 3*3, pool /2	dim 512, conv 3*3, pool /2
stage3	14*14	G-NomMer Block	[dim 384, head 4] $\times$ 8	[dim 384, head 4] $\times$ 16	[dim 512, head 4] $\times$ 16
	7*7	Reduction Module	dim 768, conv 3*3, pool /2	dim 768, conv 3*3, pool /2	dim 1024, conv 3*3, pool /2
stage4	7*7	G-NomMer Block	[dim 768, head 8] $\times$ 2	[dim 768, head 8] $\times$ 4	[dim 1024, head 8] $\times$ 4

Table 3. Detailed architecture specifications of NomMer.

details taken into account. Compared with the "Layer 1_3" maps, the "Layer 1_6" maps are inclined to emphasize CNN contextual features containing details at a more fine-grained level. In the layers of the second stage ("Layer 2_ $\sim$"), the local CNN and L-MHSA contextual features are mostly nominated, where the maps from adjacent layers become complementary, which is different from the synergistic context pattern on classification task.

In summary, Fig. 1 and Fig. 2 vividly show that our NomMer can successfully modulate the synergy behavior of different types of context in terms of specific tasks.

E.2. Representation Structure of NomMer

As aforementioned, the synergy behavior can appear either within or across layers, which has been demonstrated by the Fig. 1 and Fig. 2. To further study the representation structure learned in our NomMer, we plot the Centered Kernel Alignment (CKA) similarities [5] between all pairs of layers across different model architectures in Fig. 3. Given the representations of two layers $\mathbf{X} \in \mathbb{R}^{m \times c_1}$ and $\mathbf{Y} \in \mathbb{R}^{m \times c_2}$, mathematically,

$$CKA(\mathbf{K}, \mathbf{L}) = \frac{HSIC(\mathbf{K}, \mathbf{L})}{\sqrt{HSIC(\mathbf{K}, \mathbf{K}) HSIC(\mathbf{L}, \mathbf{L})}},$$
$$\mathbf{K} = \mathbf{X}\mathbf{X}^\top, \mathbf{L} = \mathbf{Y}\mathbf{Y}^\top,$$

where *HSIC* is the Hilbert-Schmidt Independence Criterion which measures the similarity of centered similarity matrices:

$$HSIC(\mathbf{K}, \mathbf{L}) = \frac{vec(\mathbf{K}') \cdot vec(\mathbf{L}')}{(m-1)^2},$$
$$\mathbf{K}' = \mathbf{HKH}, \mathbf{L}' = \mathbf{HLH},$$

and $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the centering matrix.

From Fig. 3 (b), we can observe that the local self attention-based ViT, Swin-B [6], presents the similarity structure with clear block-like patterns, and the similarity scores are always high within a "block" which contains several adjacent layers while almost become zero outside the block. By introducing global contextual information, Focal-B [12] (Fig. 3 (c)) also exhibits block-like similarity structure but with more smooth edges, indicating that there still have representation similarity between layers with larger interval. Compared with ViT models, in the canonical CNN architecture, ResNet-101 [3] (Fig. 3 (a)), the representation within lower layers present more difference while the similarities tend to be larger between higher layers. Moreover, the similarity scores between lower and higher layers are small.

By comparison, unlike the ViT models with heuristic-based design in terms of exploiting local or global-local

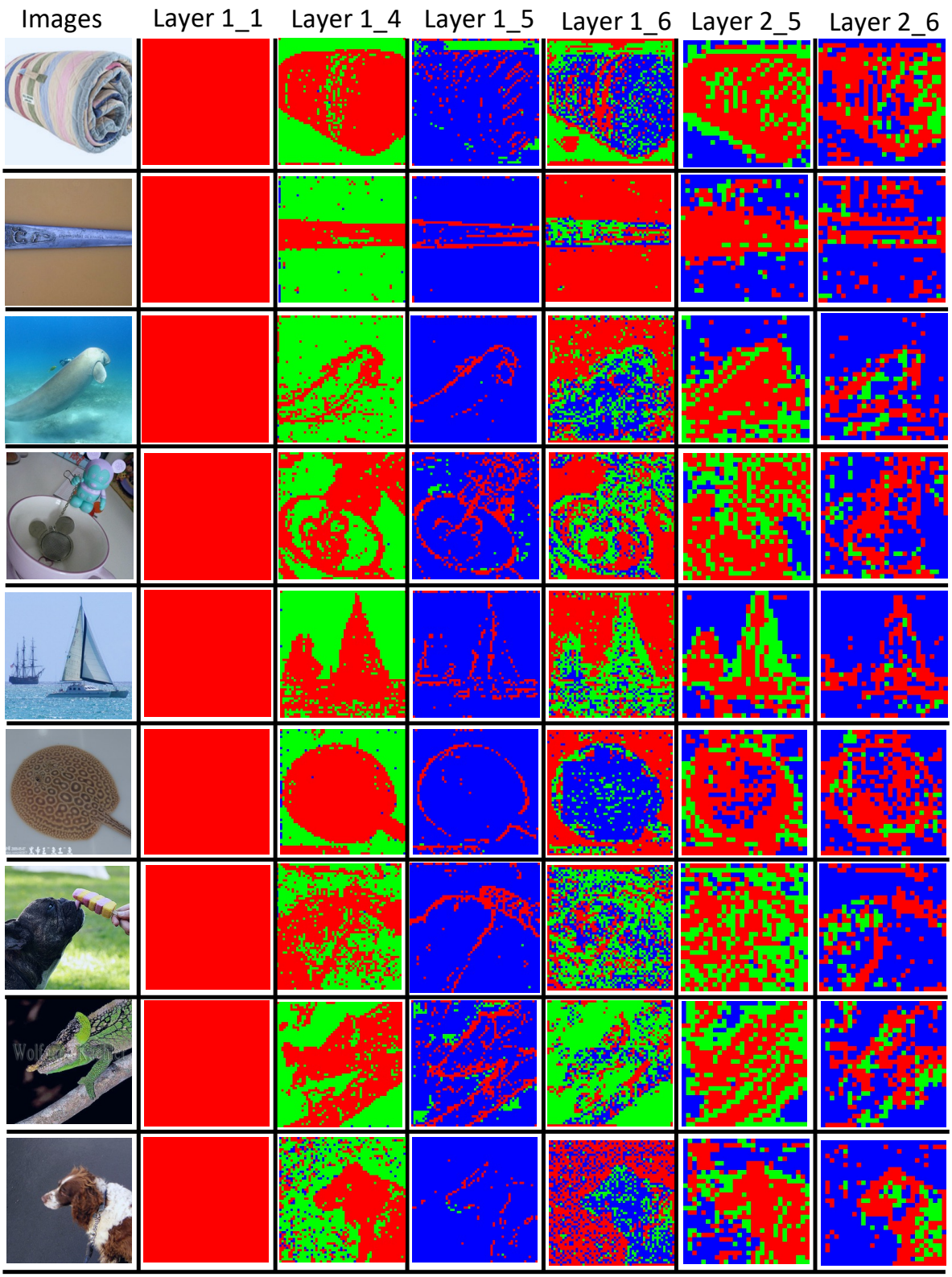


Figure 1. Nomination Maps from NomMer-B on Image Classification task. **Red**: CNN context $\mathbf{F}^{(C)}$. **Green**: Local context $\mathbf{F}^{(L)}$. **Blue**: Compressed global context $\mathbf{F}^{(G)}$. "Layer B_L" stands for that map is from the L-th NomMer layer of NomMer blocks at the B-th stage. Best viewed in color.

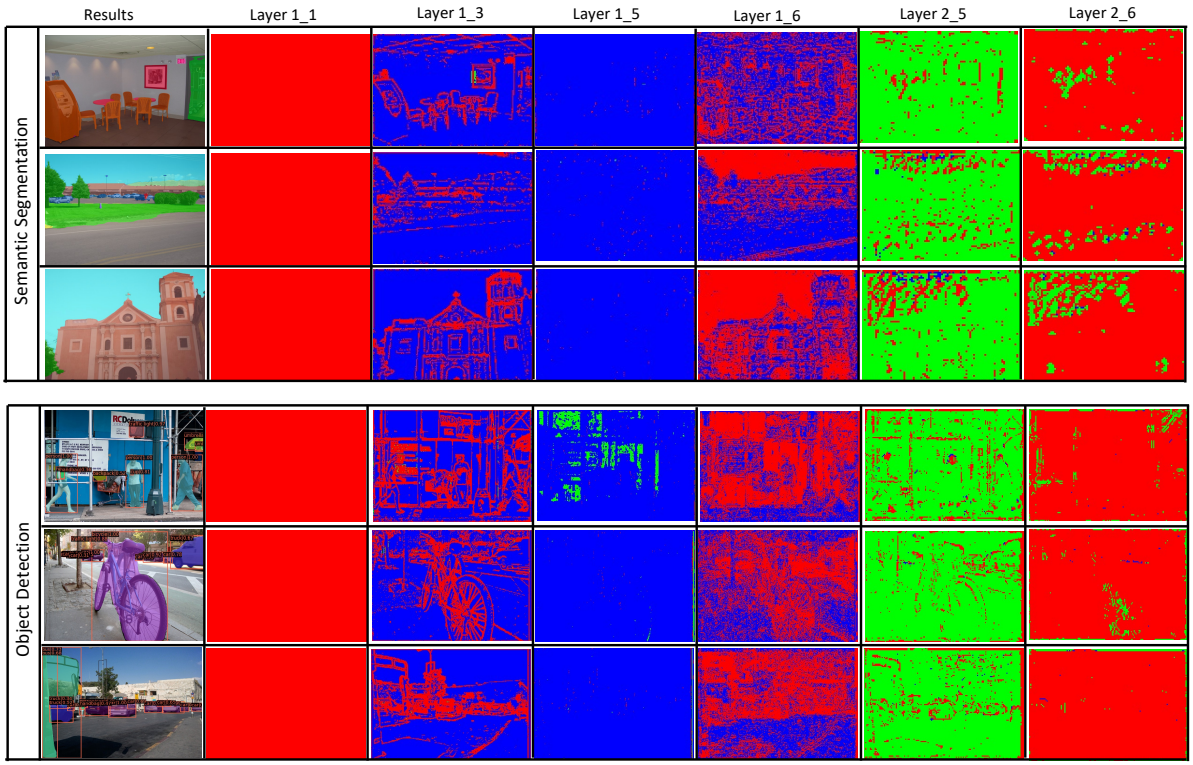
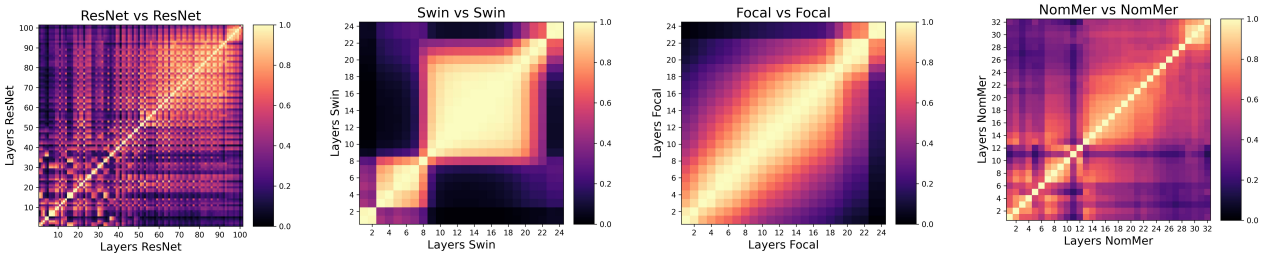


Figure 2. Nomination Maps from NomMer-B on dense prediction tasks including semantic segmentation and object detection. **Red**: CNN context $\mathbf{F}^{(C)}$. **Green**: Local context $\mathbf{F}^{(L)}$. **Blue**: Compressed global context $\mathbf{F}^{(G)}$. "Layer B_L" stands for that map is from the L-th NomMer layer of NomMer block at the B-th stage. Best viewed in color.



(a) ResNet-101 (CNN) (b) Swin-B (Local ViT) (c) Focal-B (Global-local ViT) (d) NomMer-B (ours)

Figure 3. CKA similarities between all pairs of layers across different model architectures trained on ImageNet-1k. The results are shown as heatmaps in which the horizontal and vertical axes indexing the layers from input to output. Best viewed in color and zoom in.

context, the representation structure of our NomMer-B has somewhat similarity with that of ResNet-101 while the "block" patterns in ViT are also preserved. This more sophisticated structure can be attributed to the "dynamic nomination" of NomMer, which effectively integrate the contributions of local and global context.

E.3. More Qualitative Analysis on NomMer

This part will illustrate more evidence on indispensable design of nominator and G-NomMer block in our method.

Hard sampling vs. soft sampling. By replacing the Gumbel-softmax in SCN of NomMer with canonical softmax, we find that, in contrast to the hard version (Fig. 4(b)), the local-attention features (green) present dominant in the learned maps of soft version (Fig. 4(a)). Correspondingly, the classification activation maps of soft version become more unstable than hard version. As a result, the top1 accuracies of different NomMer variants on image classification task all drop round 0.4 on image classification task. We attribute it to the redundancy not well reduced by soft sam-

pling, where local and global features could have negative impact on each other.

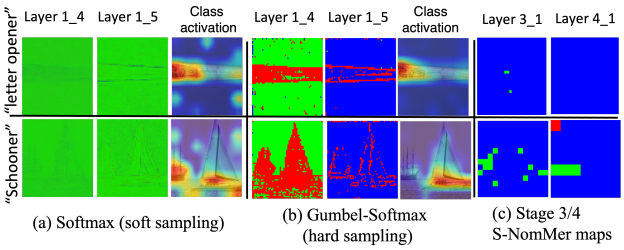


Figure 4. Nomination maps (soft sampling (a) vs. hard sampling (b)) and class activation attention maps from NomMer-B on image classification task. (c) Nomination maps at stage 3 and 4.

G-NomMer block. As shown in Fig. 4 (c), if the S-NomMer blocks applied to all stages, the global SA features (blue) would be dominant at both stage 3 (Layer3_1) and stage 4 (Layer4_1). It is probably because the global context and local context could become homogeneous when feature size become small at higher-level stage 3 and 4, as claimed in main text. Therefore, we adopt the G-NomMer block only equipped with canonical global SA.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 2

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 2

[3] Kaiming He, 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*, pages 770–778, 2016. 2, 3

[4] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020. 2

[5] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019. 3

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 2, 3

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 1

[9] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14454–14463, 2021. 2

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

[11] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2

[12] Jianwei Yang, 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. 2, 3

[13] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9657–9666, 2019. 2

[14] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. *arXiv preprint arXiv:2103.15358*, 2021. 2

[15] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9759–9768, 2020. 2