Searching for BurgerFormer with Micro-Meso-Macro Space Design

Longxing Yang[1,2,3] Yu Hu[1,2,3] Shun Lu[1,2,3] Zihao Sun[1,2,3] Jilin Mei[1,2,3]

Yinhe Han[1,2,3] Xiaowei Li[2,3]

[1] *Research Center for Intelligent Computing Systems, Institute of Computing Technology, CAS*

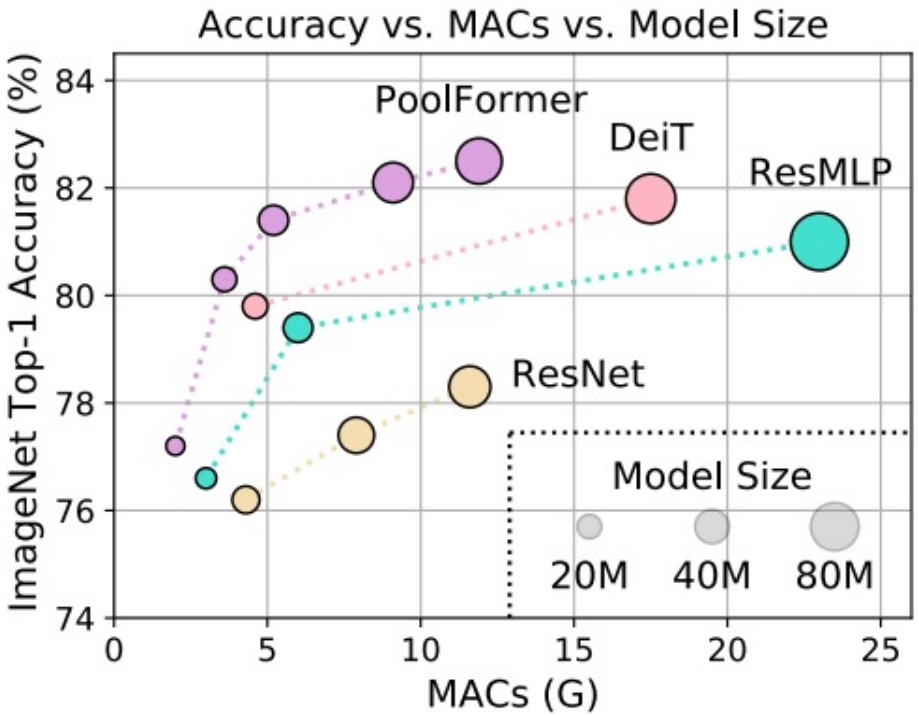
[2] *State Key Laboratory of Computer Architecture, Institute of Computing Technology, CAS*

[3] *School of Computer Science and Technology, University of Chinese Academy of Sciences*

Motivation

Hot Vision Transformers

Simple Pooling can achieve impressive performance

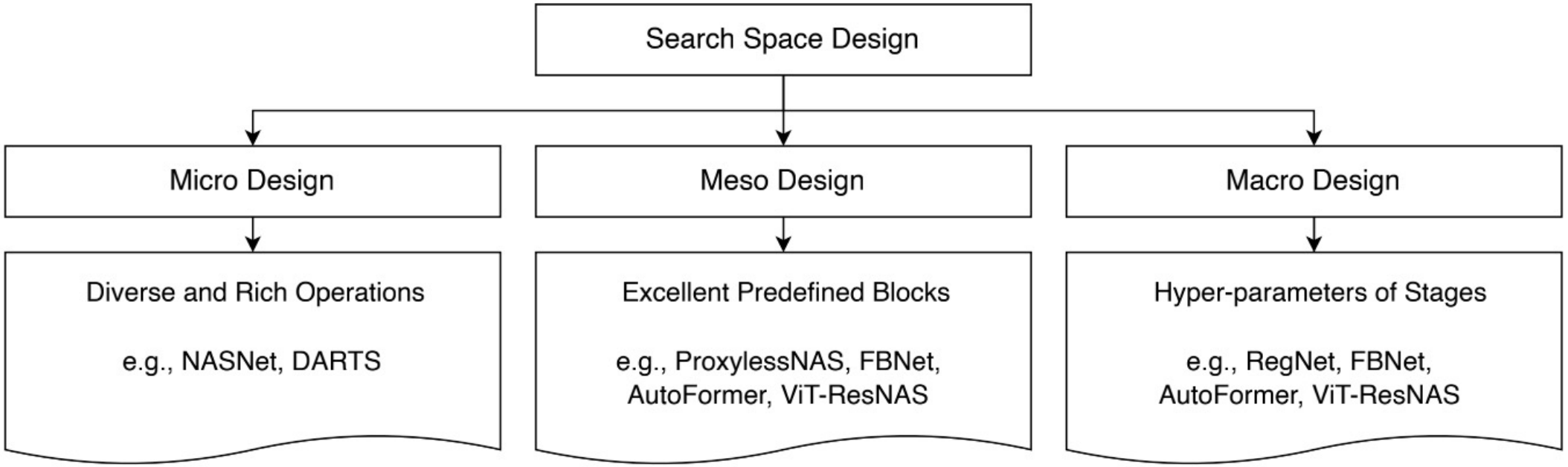


Transformers + Neural Architecture Search

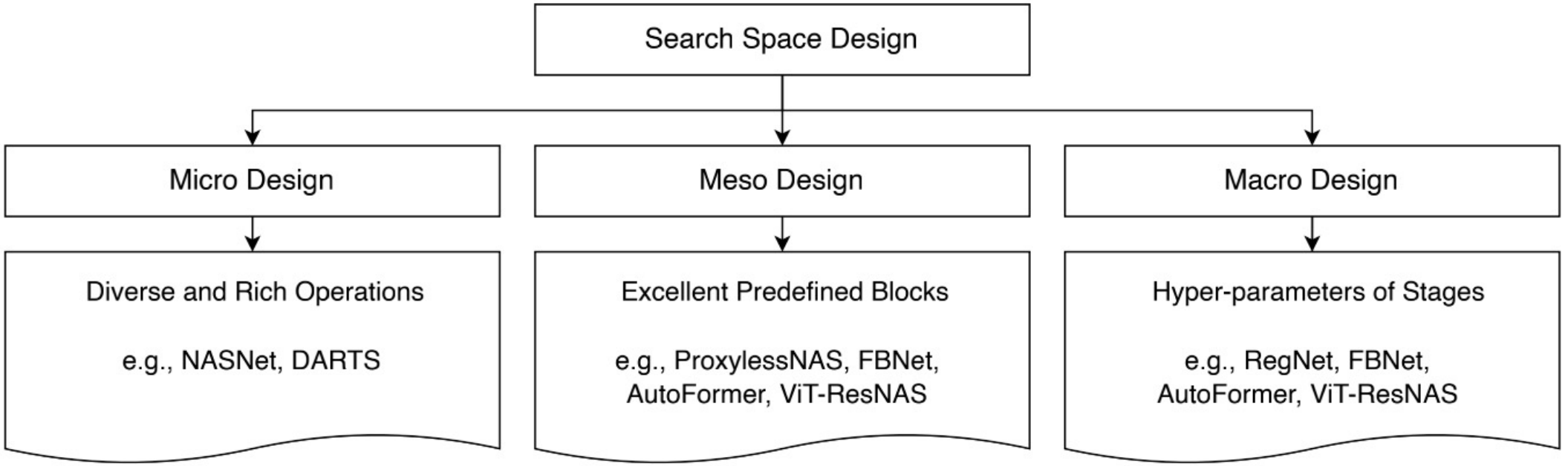
How to design a generic search space to search high-performance Transformer-like architectures ?

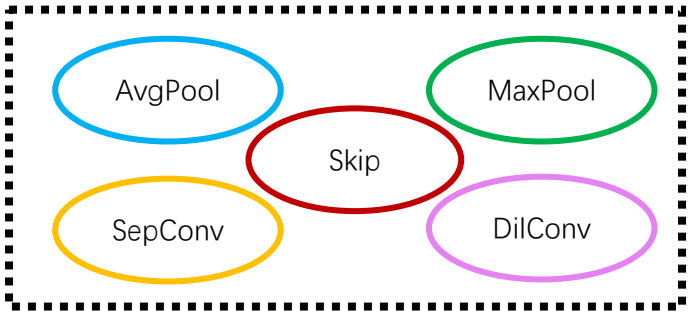
[1] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., and Yan, S. MetaFormer is Actually What You Need for Vision. CVPR 2022 Oral.

Typical Search Space Design

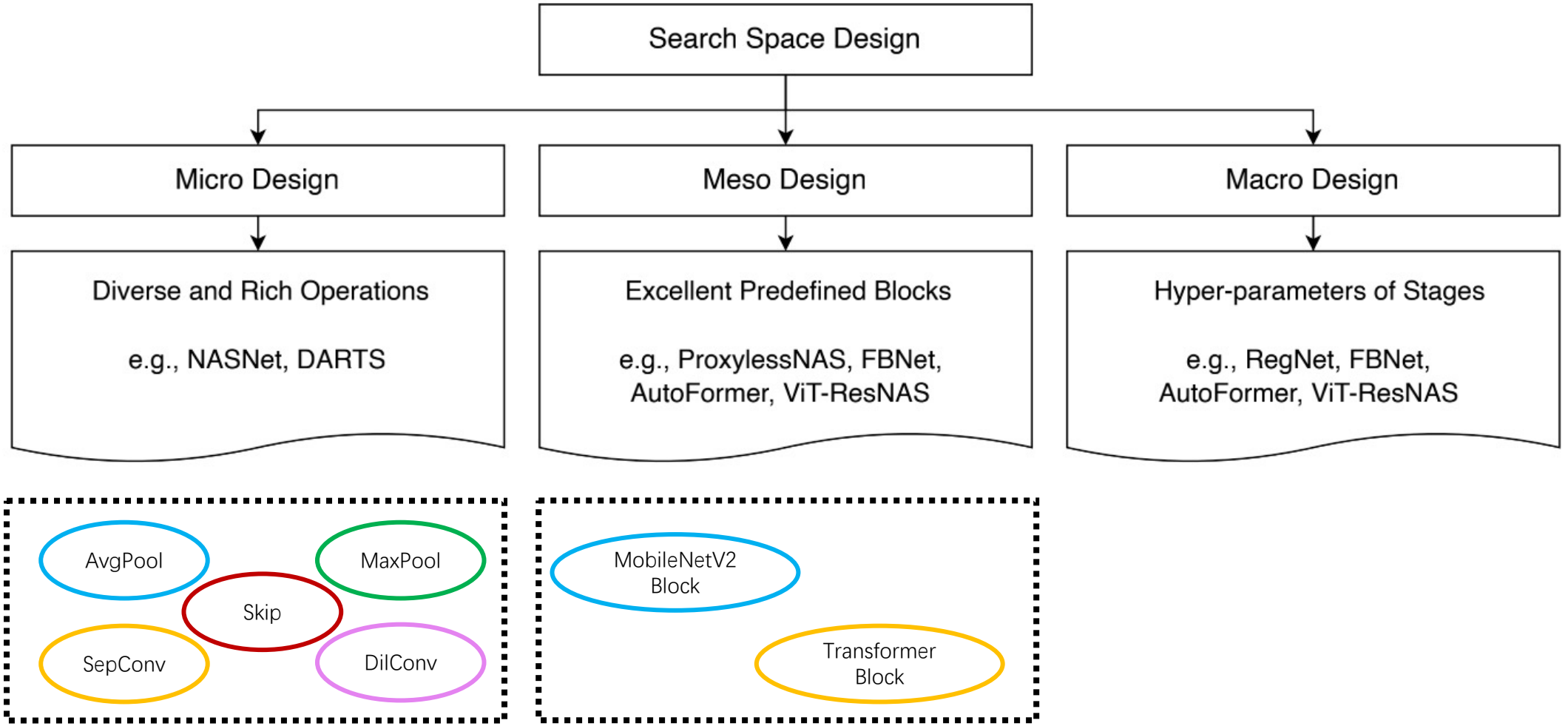


Typical Search Space Design

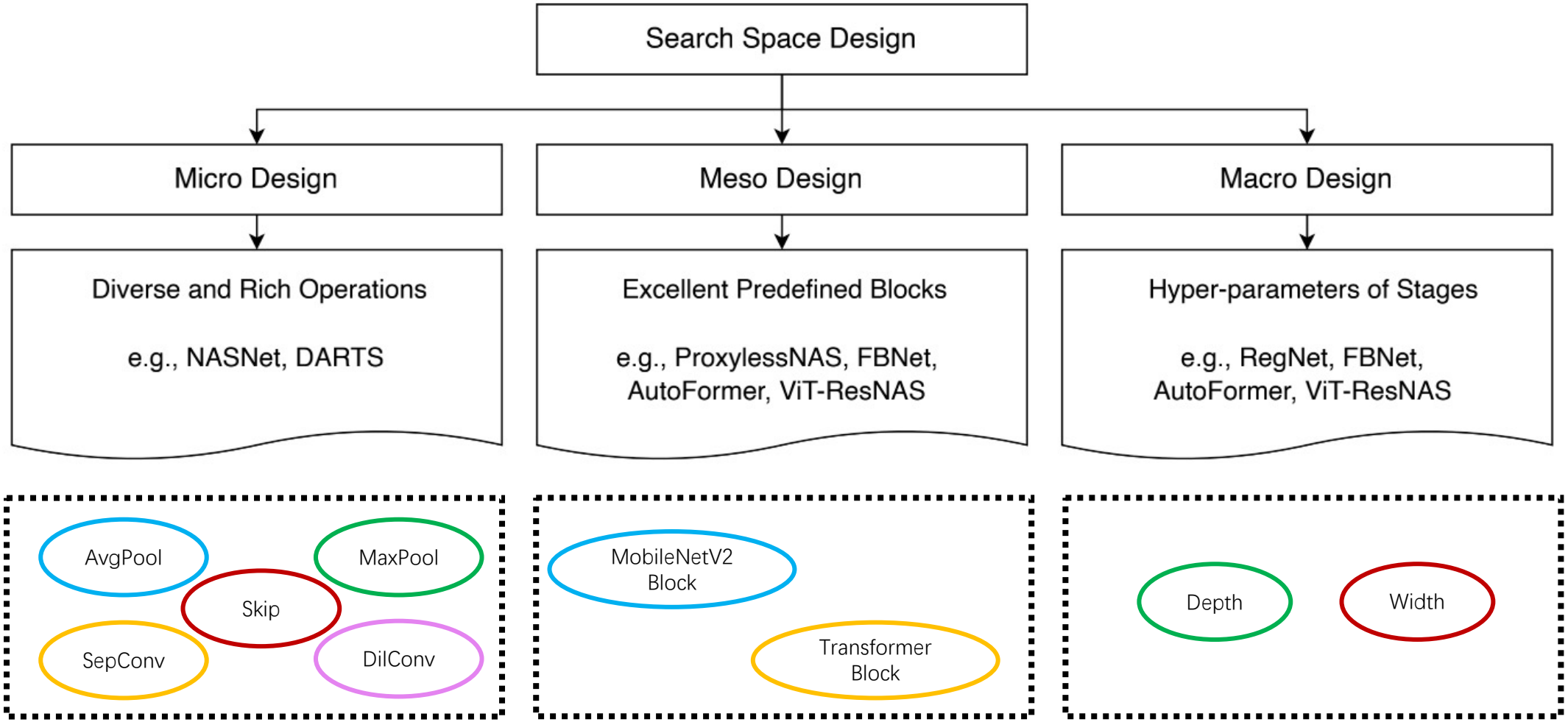




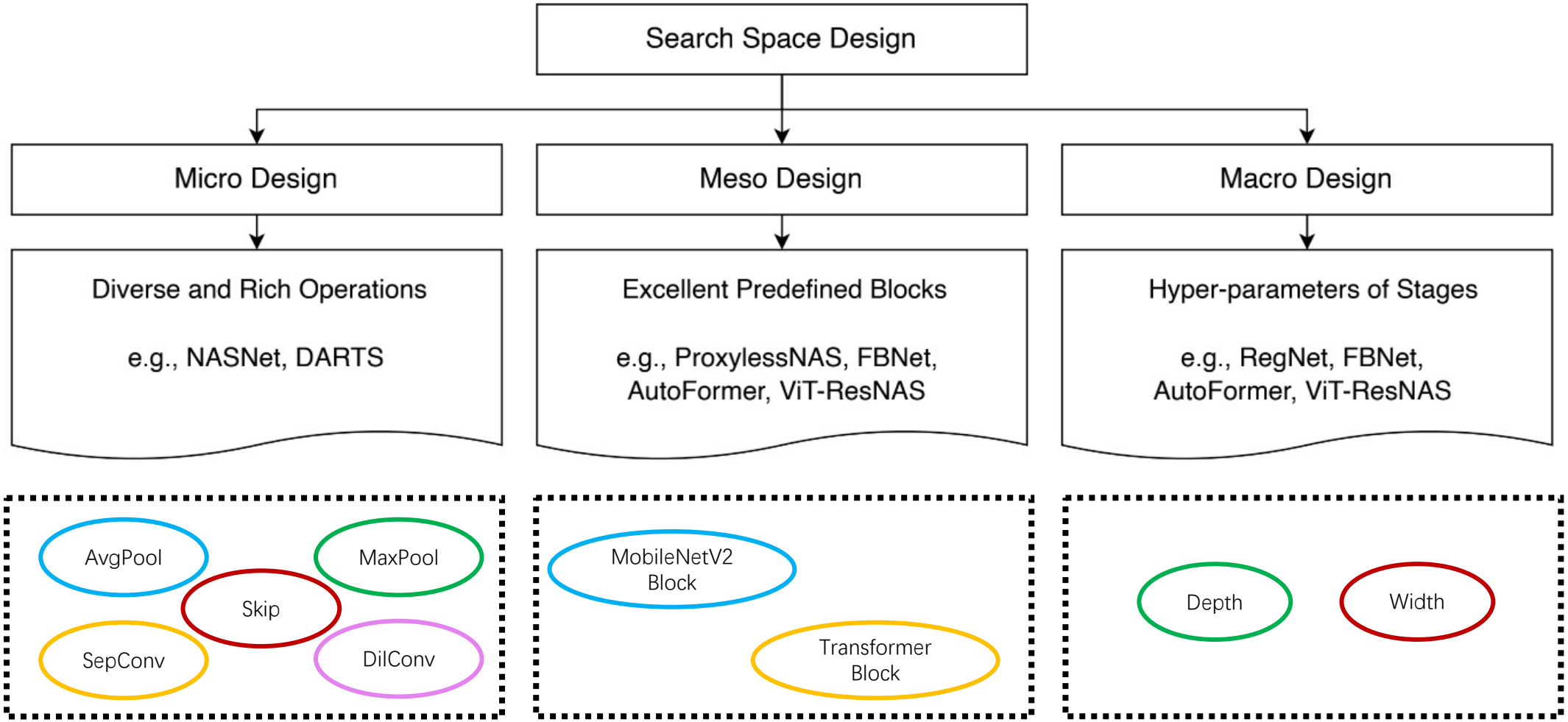
Typical Search Space Design



Typical Search Space Design



Typical Search Space Design

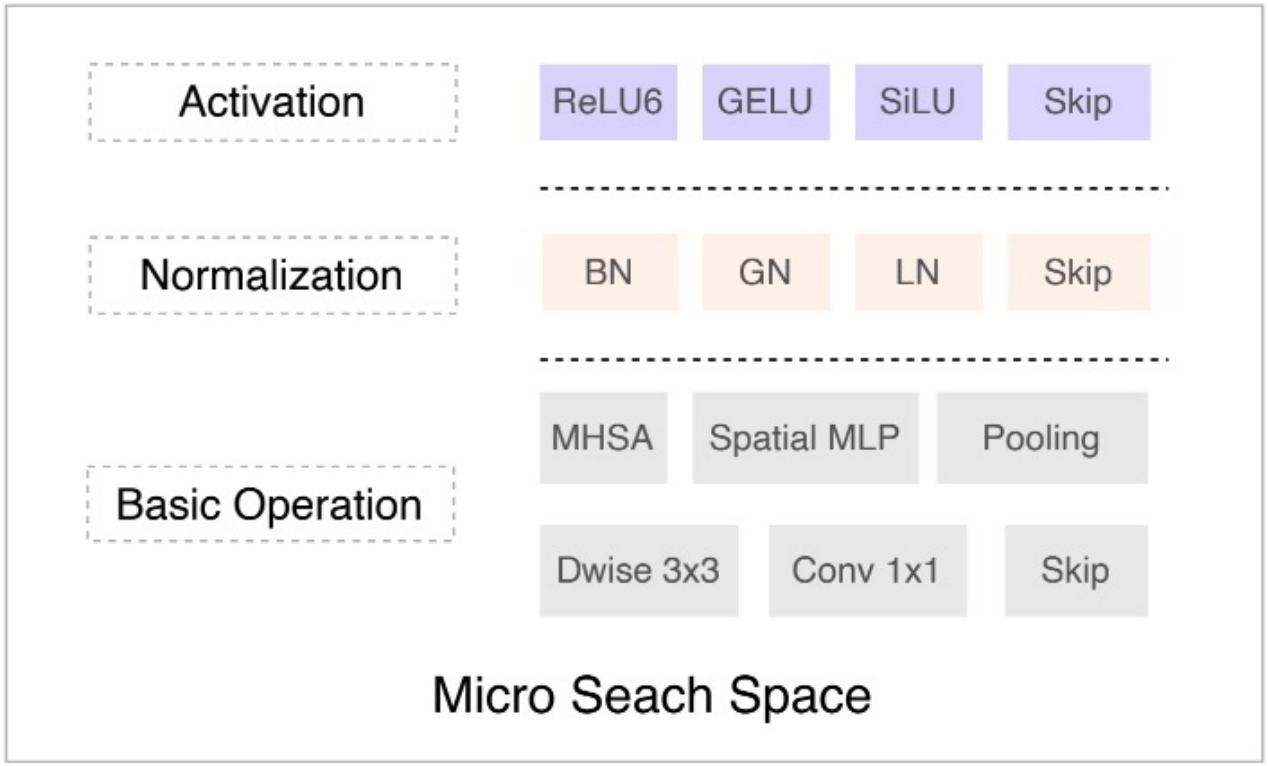


Transformers NAS: Lacking a Search Space Constructed from Three Granularities

Contributions

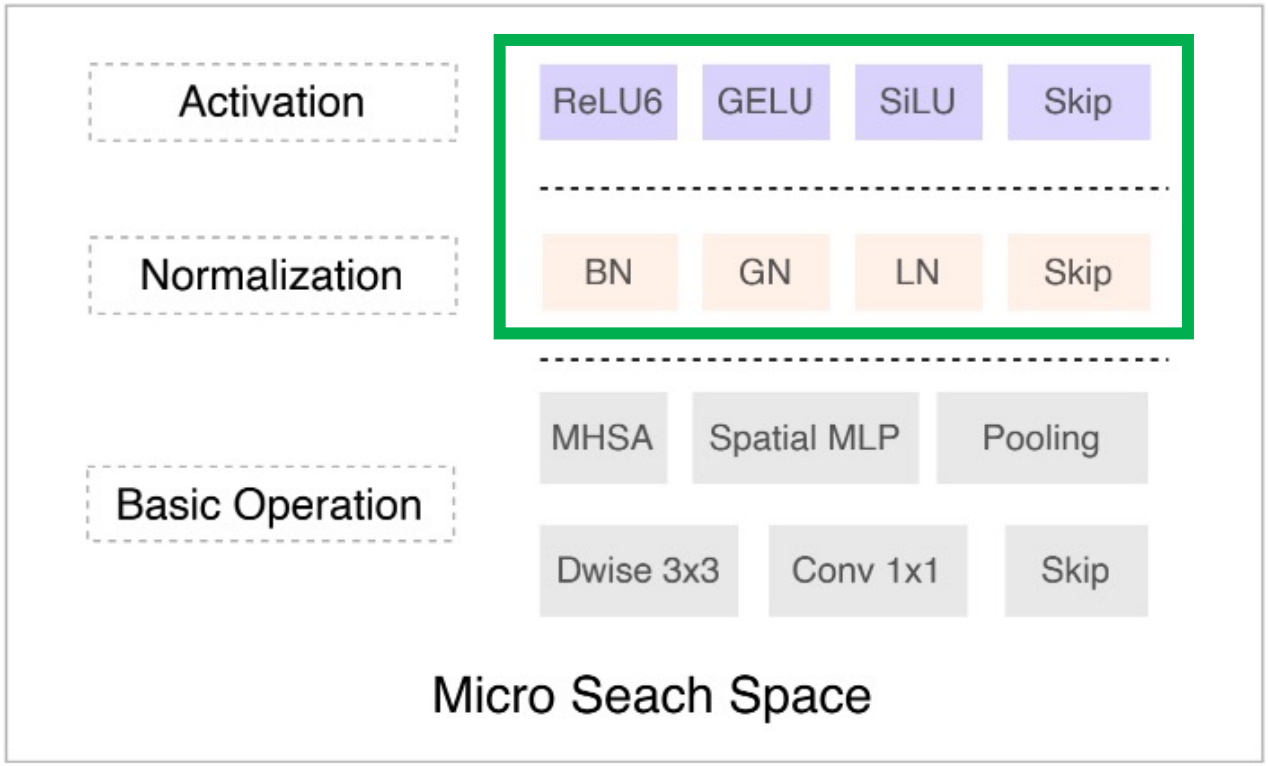
- A Micro-Meso-Macro Search Space for Transformer-Like Architectures
- A Hybrid Sampling Method for Effective One-Shot NAS
- Good Experiment Results on ImageNet and COCO Datasets

Micro Search Space



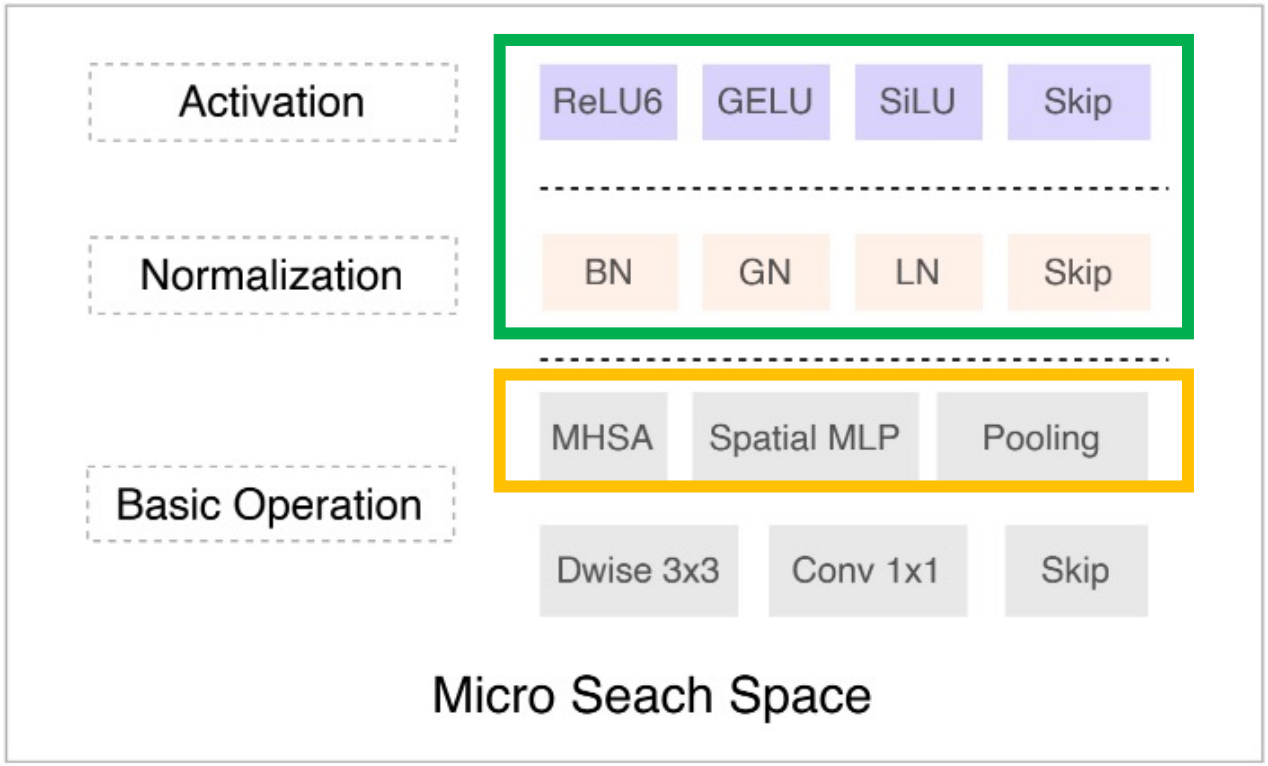
- Transformers and CNN employ different activation and normalization.
- Following Transformer-like architectures, MHSA, Spatial MLP, and Pooling are searched.
- We extend Dwise 3x3, Conv 1x1 for more locality.

Micro Search Space



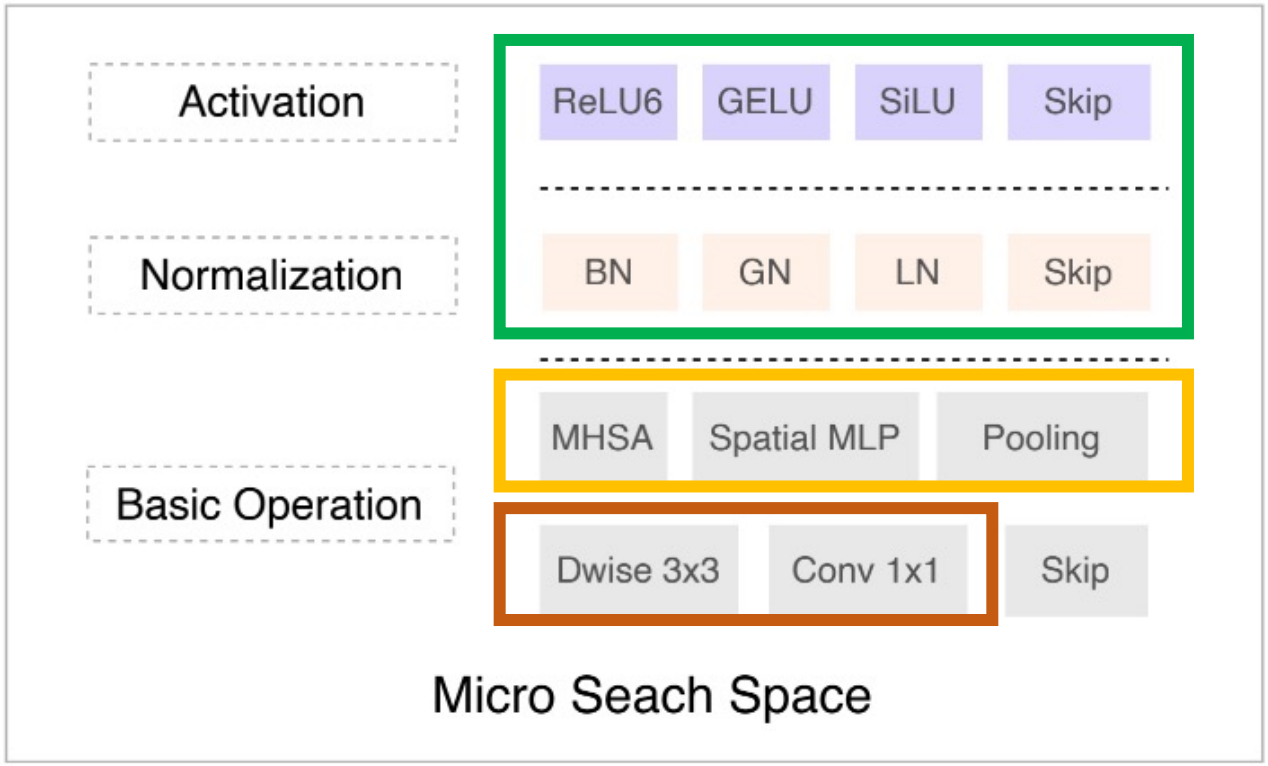
- Transformers and CNN employ different activation and normalization.
- Following Transformer-like architectures, MHSA, Spatial MLP, and Pooling are searched.
- We extend Dwise 3x3, Conv 1x1 for more locality.

Micro Search Space



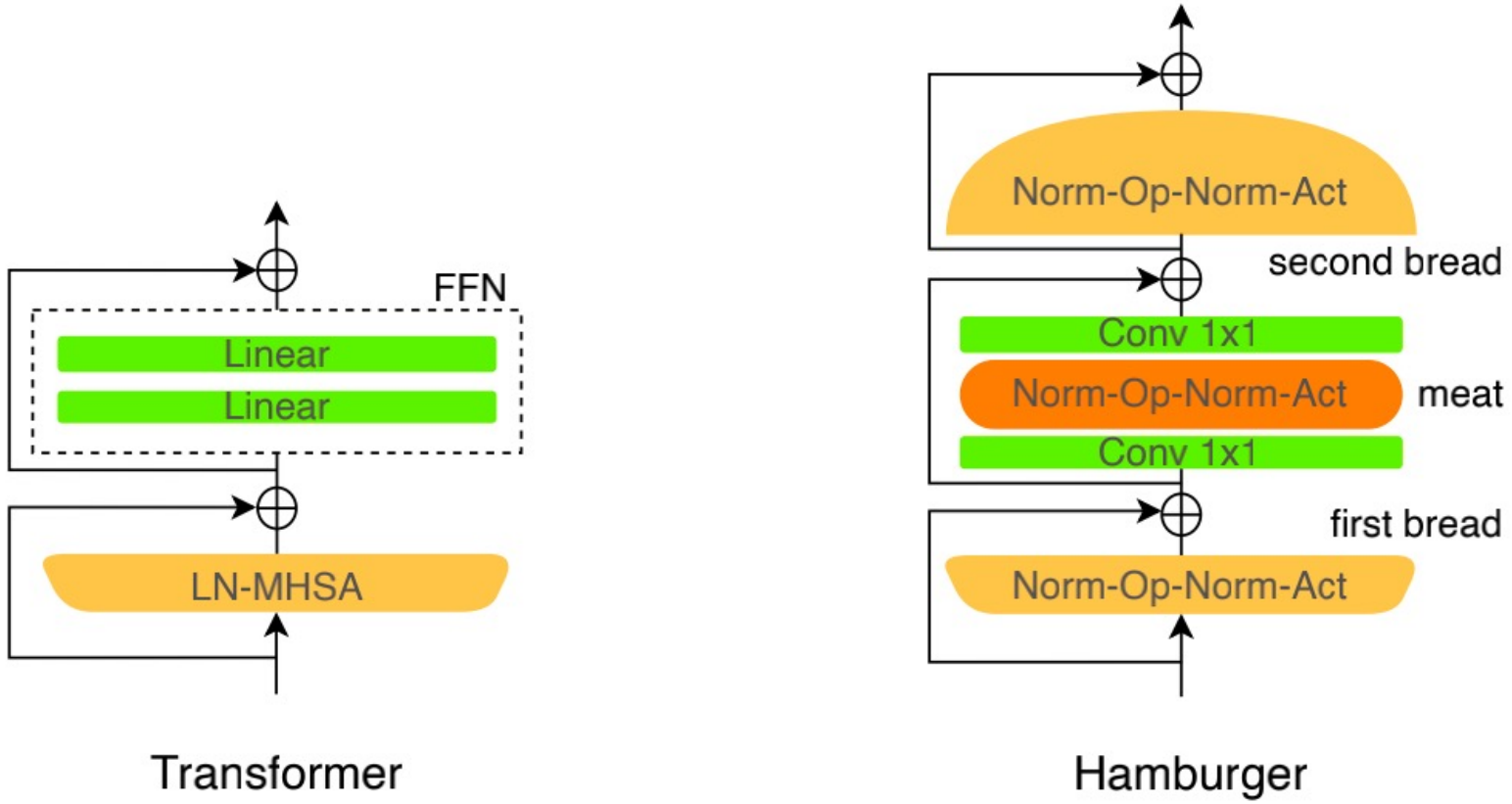
- Transformers and CNN employ different activation and normalization.
- Following Transformer-like architectures, MHSA, Spatial MLP, and Pooling are searched.
- We extend Dwise 3x3, Conv 1x1 for more locality.

Micro Search Space



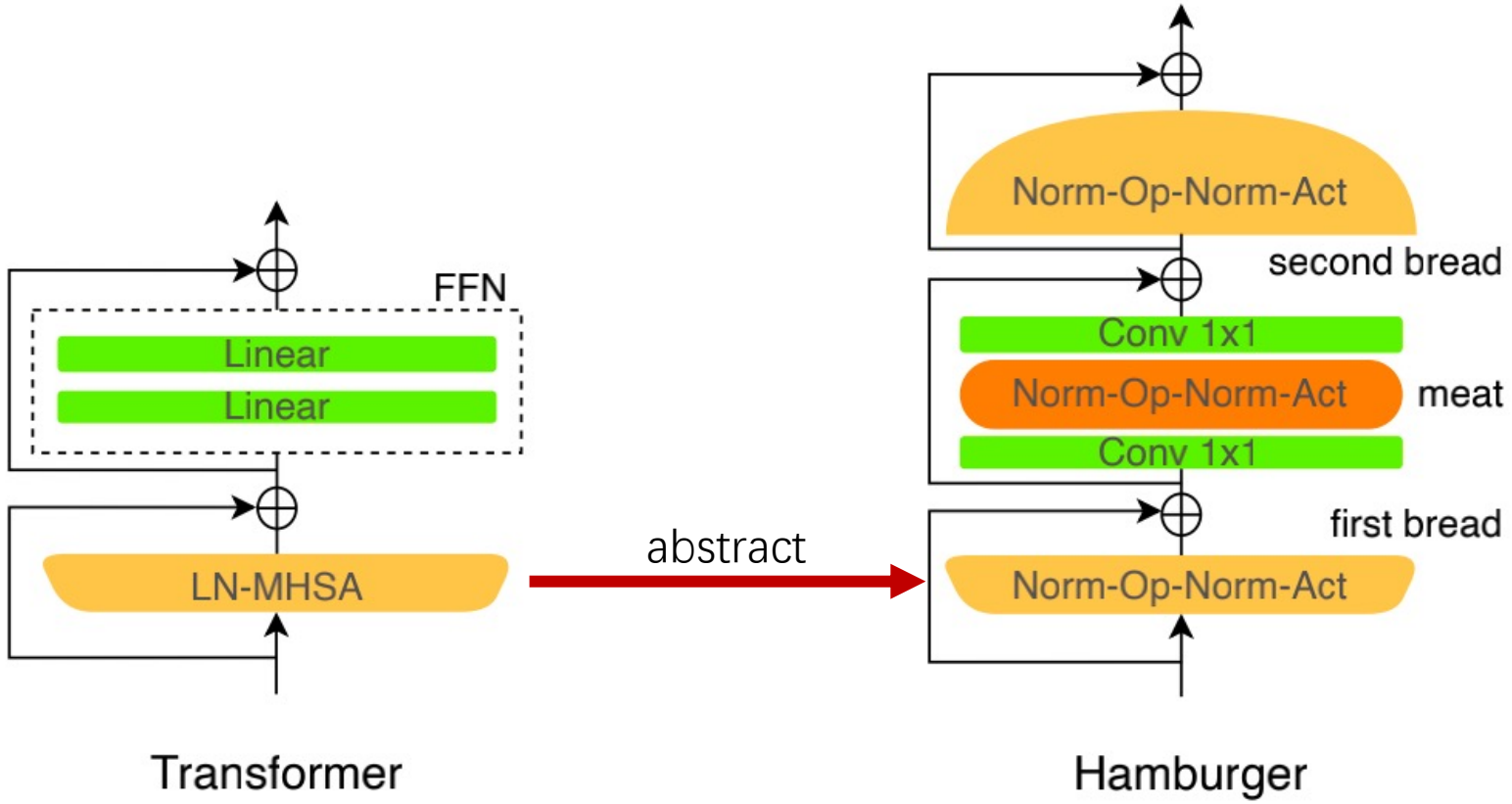
- Transformers and CNN employ different activation and normalization.
- Following Transformer-like architectures, MHSA, Spatial MLP, and Pooling are searched.
- We extend Dwise 3x3, Conv 1x1 for more locality.

Meso Search Space



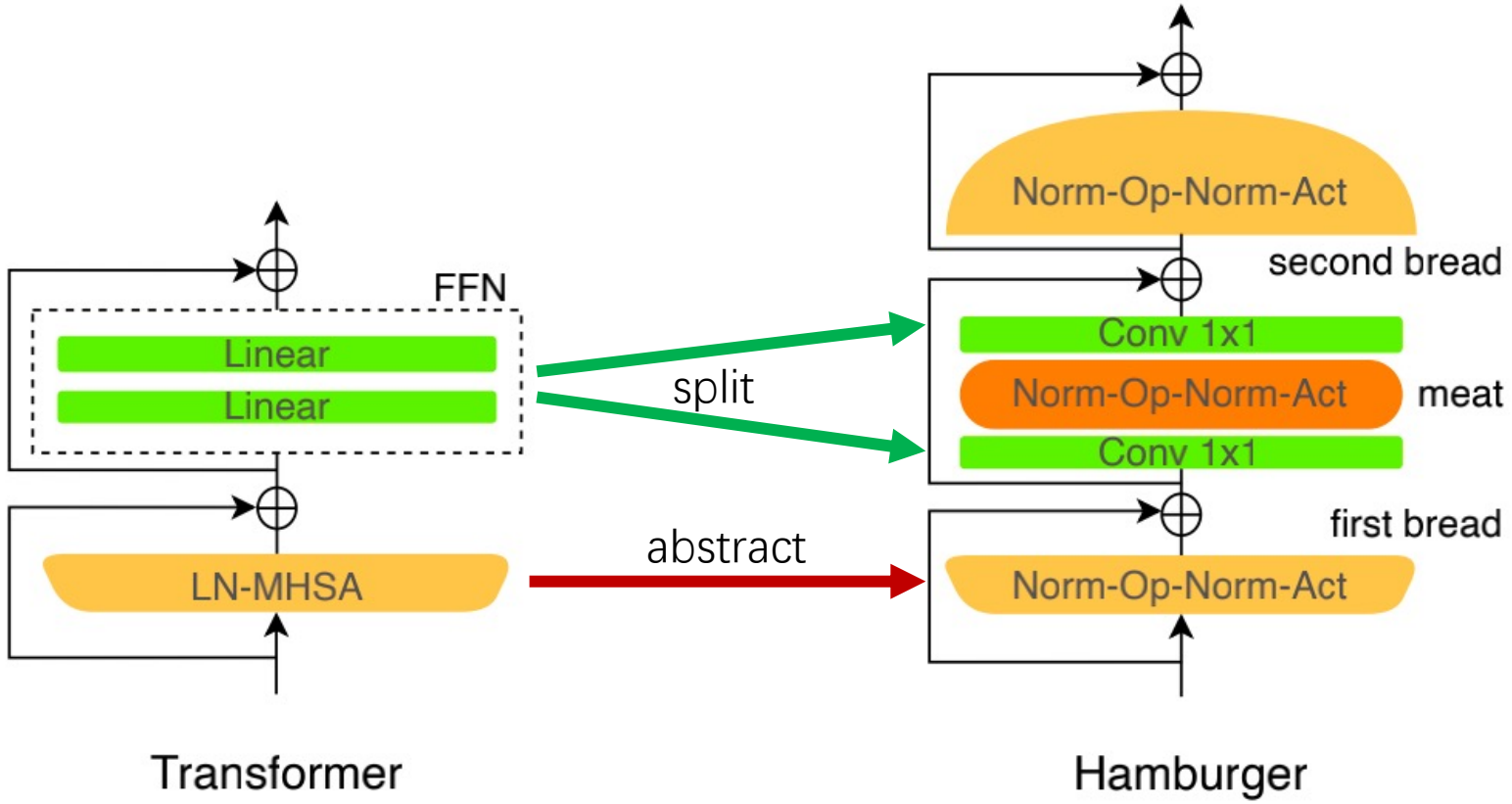
- Abstract the combination of LN and MHSA as a Norm-Op-Norm-Act Structure.
- Split FFN into two Conv 1x1 for keeping inverse bottleneck structures.
- Symmetrical design for more diverse meso blocks.

Meso Search Space



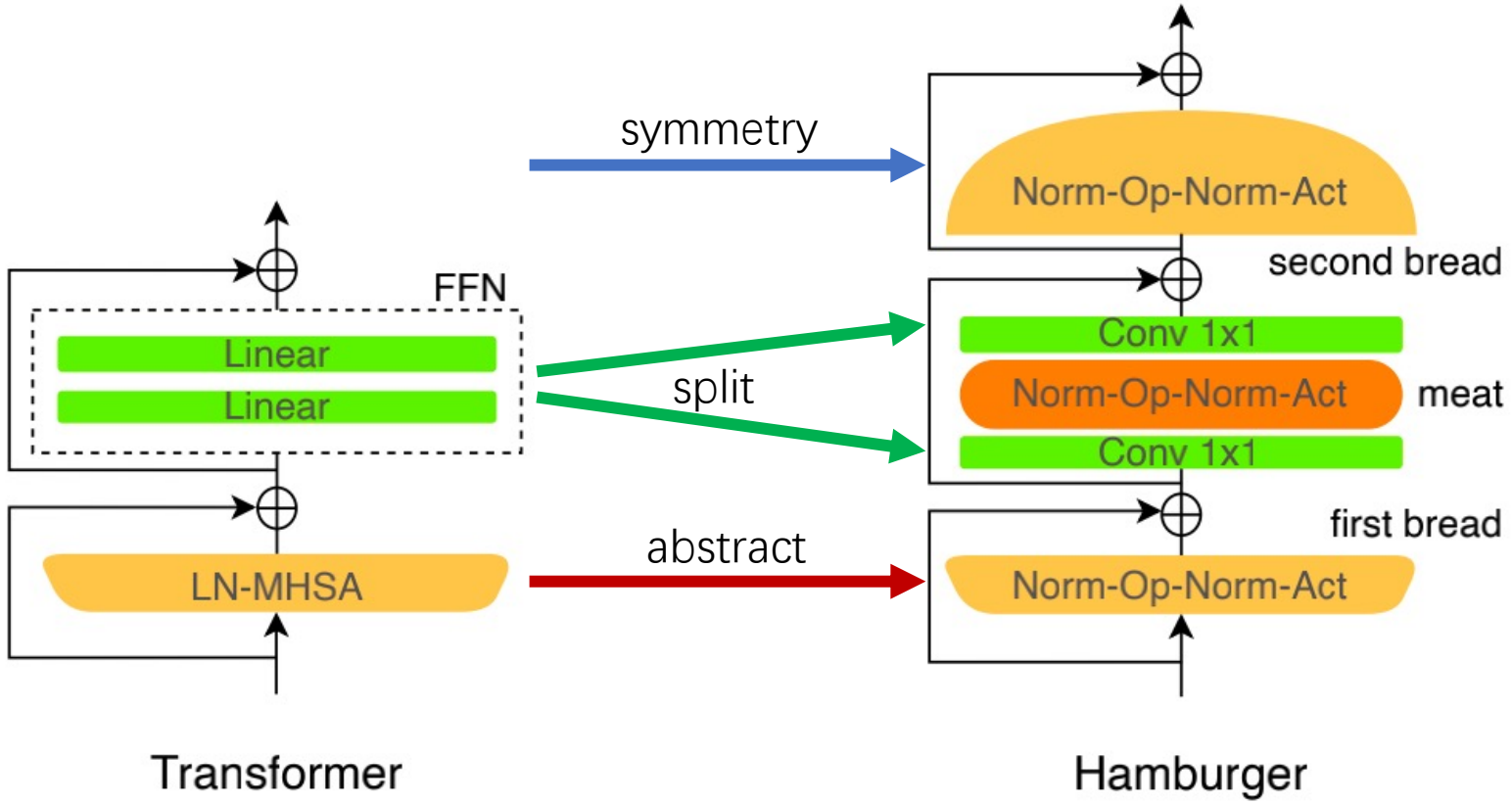
- Abstract the combination of LN and MHSA as a Norm-Op-Norm-Act Structure.
- Split FFN into two Conv 1x1 for keeping inverse bottleneck structures.
- Symmetrical design for more diverse meso blocks.

Meso Search Space



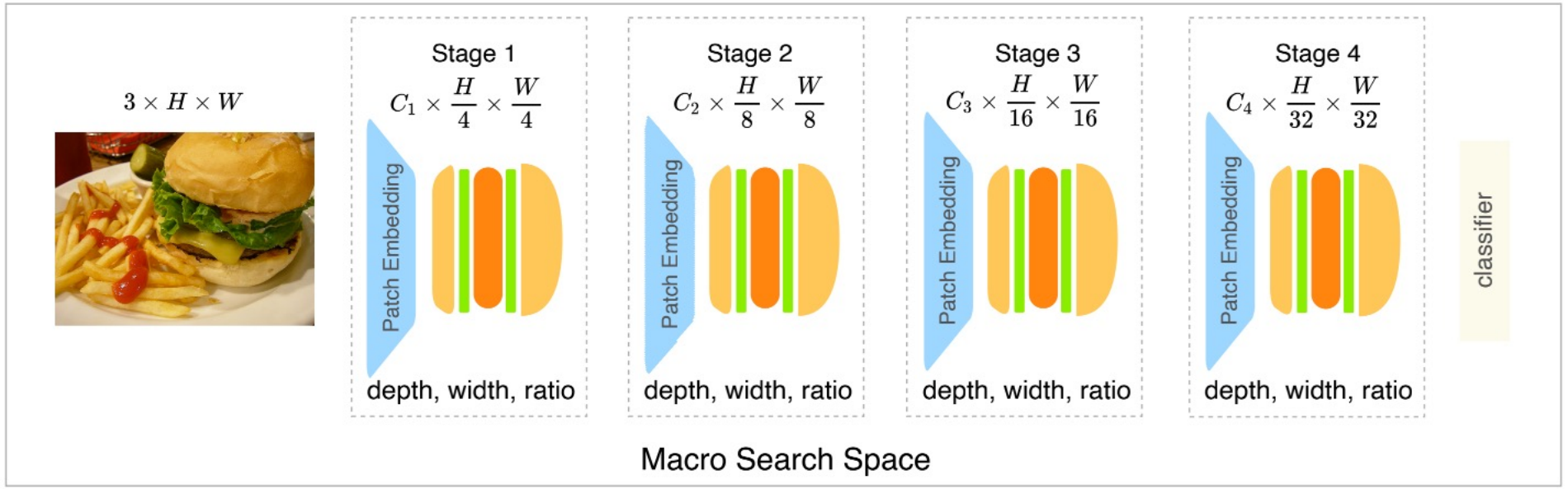
- Abstract the combination of LN and MHSA as a Norm-Op-Norm-Act Structure.
- Split FFN into two Conv 1x1 for keeping inverse bottleneck structures.
- Symmetrical design for more diverse meso blocks.

Meso Search Space



- Abstract the combination of LN and MHSA as a Norm-Op-Norm-Act Structure.
- Split FFN into two Conv 1x1 for keeping inverse bottleneck structures.
- Symmetrical design for more diverse meso blocks.

Macro Search Space



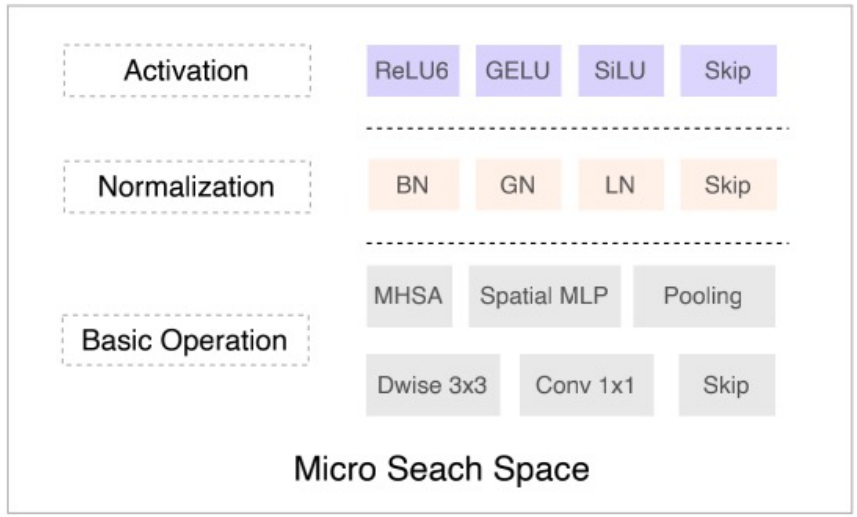
- Multi-Stage Architecture
- Search for Depth, Width, Ratio

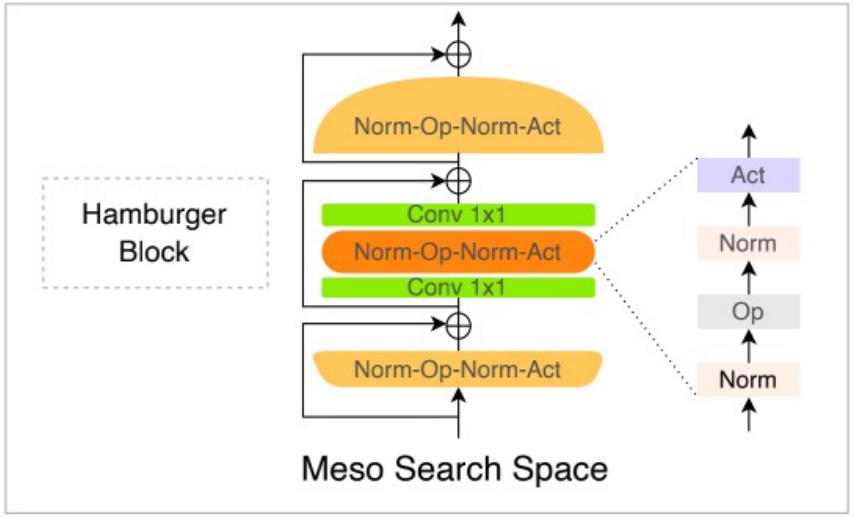
Stage	Depth	Width	Ratio
1	$\{1, 2, 3, 4\}$	$\{32, 64, 96\}$	$\{1, 2, 4, 6\}$
2	$\{1, 2, 3, 4\}$	$\{64, 96, 128\}$	$\{1, 2, 4, 6\}$
3	$\{1, 2, 3, 4, 5, 6, 7, 8\}$	$\{128, 192, 256, 320\}$	$\{1, 2, 4, 6\}$
4	$\{1, 2, 3, 4\}$	$\{128, 256, 384, 512\}$	$\{1, 2, 4, 6\}$

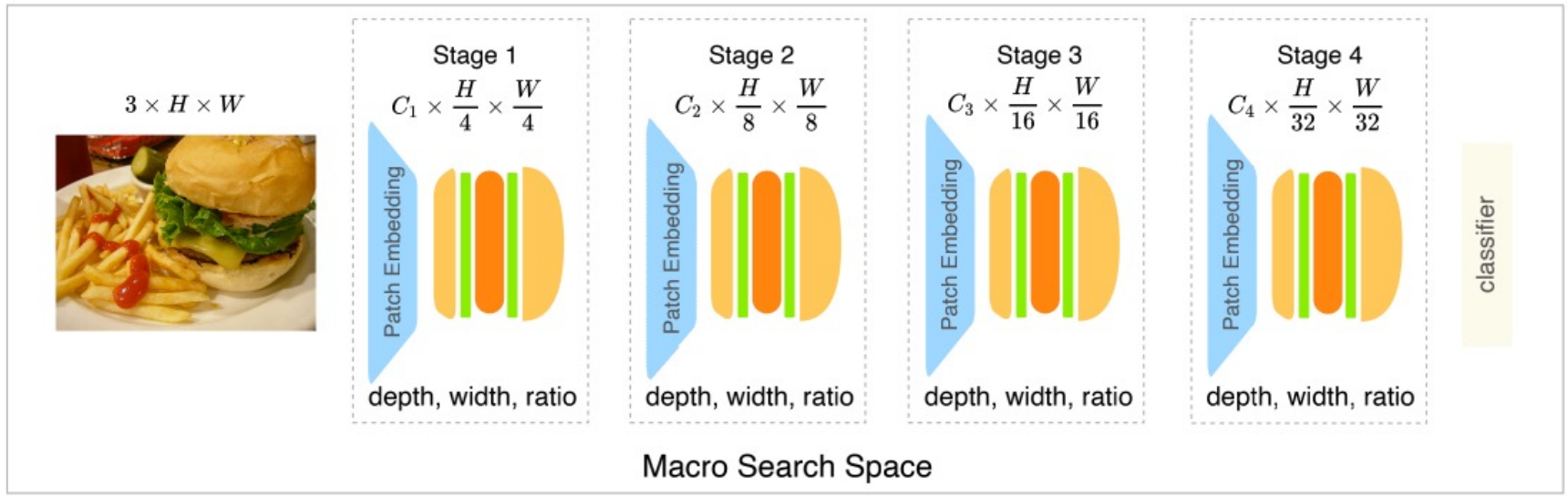
[2] Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. ICCV 2021 Oral.

[3] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. ICCV 2021 Best Paper.

Micro-Meso-Macro Search Space







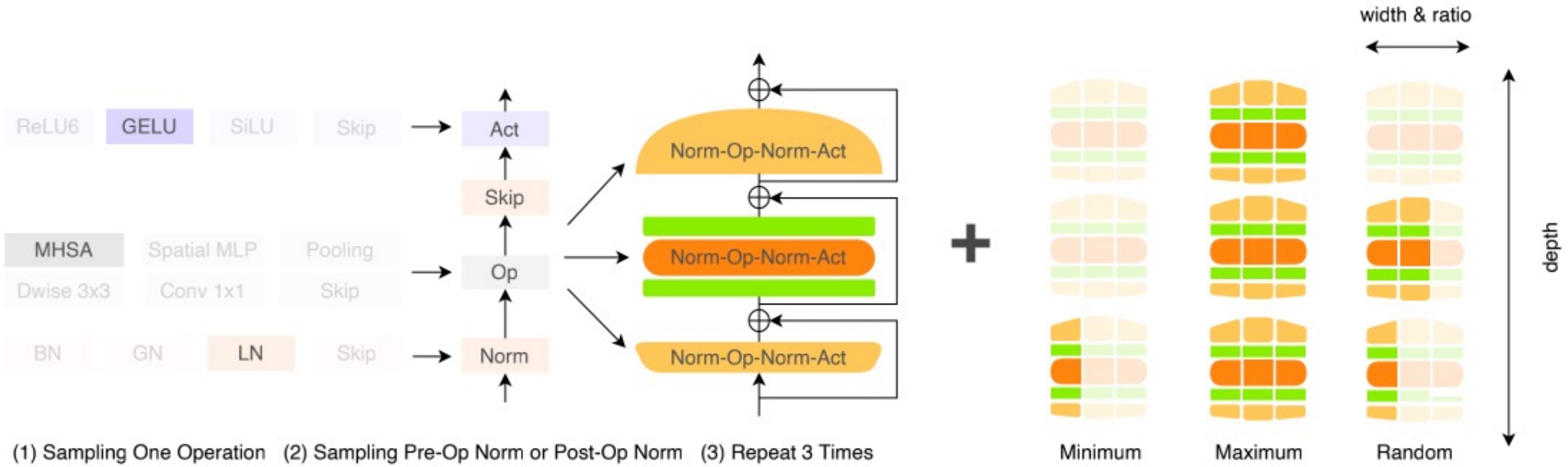
Size: 4.5×10^{28} FLOPs: $0.2G \sim 7.4G$ Parameters: $0.5M \sim 41.6M$

One-Shot NAS & Hybrid Sampling

One-Shot NAS:

$$W_A^* = \arg\min_A L_{train}(N(A, W)),$$
$$\alpha^* = \arg\max_{a \in A} Acc_{val}(N(\alpha, W^*)),$$
$$s.t.\ Resource(N(\alpha, W_{\alpha^*})) \leq C,$$

Hybrid Sampling:



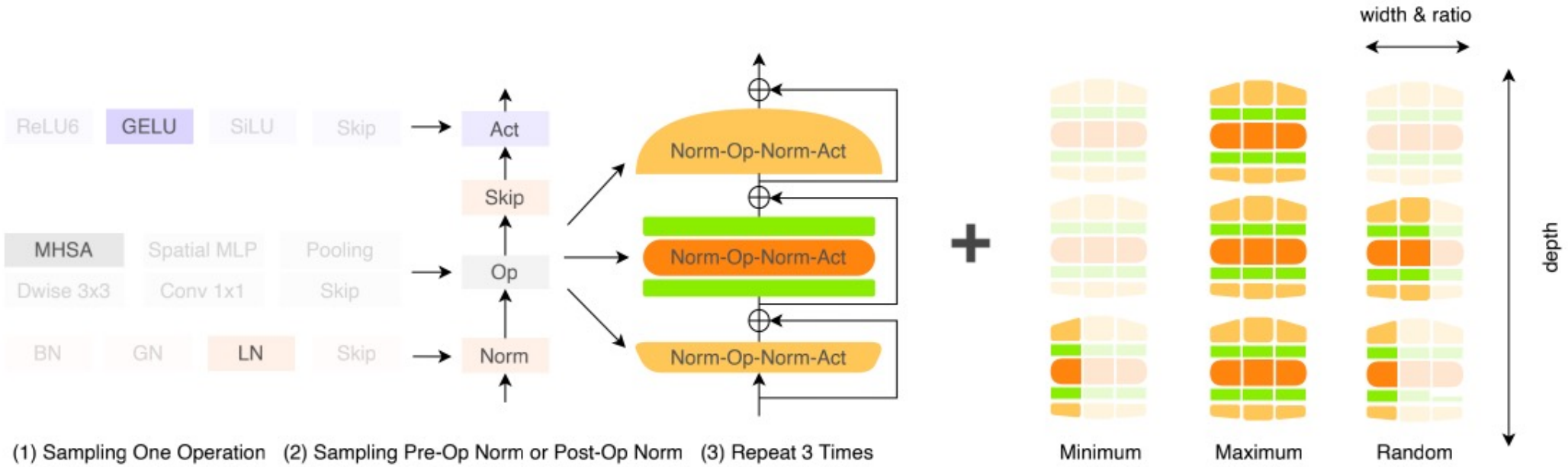
One-Shot NAS & Hybrid Sampling

One-Shot NAS:

$$W_A^* = \arg\min_A L_{train}(N(A, W)),$$
$$\alpha^* = \arg\max_{a \in A} Acc_{val}(N(\alpha, W^*)),$$
$$s.t.\ Resource(N(\alpha, W_{\alpha^*})) \leq C,$$

Methods	FLOPs (G)	Top-1 acc. (%)
Random Search	1.0	75.6
SPOS	1.0	77.0
Sandwich	1.0	77.5
Hybrid Sampling	1.0	78.0

Hybrid Sampling:



ImageNet Results

Model	Params. (M)	FLOPs (G)	Top-1 acc. (%)	Top-5 acc. (%)	Design Type
DeiT-Ti (Touvron et al., 2021b)	6	1.3	72.2	91.1	Manual
TNT-Ti (Han et al., 2021)	6	1.4	73.9	-	Manual
CeiT-T (Yuan et al., 2021)	6	1.2	76.4	-	Manual
AutoFormer-tiny (Chen et al., 2021b)	6	1.3	74.7	92.6	Auto
GLiT-Tiny (Chen et al., 2021a)	7	1.4	76.3	91.1	Auto
ViTAS-DeiT-A (Su et al., 2021)	7	1.4	75.6	92.5	Auto
BurgerFormer-Tiny	10	1.0	78.0	93.7	Auto
ConVi-Ti+ (d'Ascoli et al., 2021)	10	2.0	76.7	93.6	Manual
PVT-Tiny (Wang et al., 2021)	13	1.9	75.1	-	Manual
PoolFormer-S12 (Yu et al., 2022)	12	2.0	77.2		Manual
BurgerFormer-Small	14	2.1	80.4	95.0	Auto
Deit-S (Touvron et al., 2021b)	22	4.7	79.9	-	Manual
Swin-T (Liu et al., 2021)	29	4.5	81.3	-	Manual
CvT-13 (Haiping et al., 2021)	20	4.5	81.6	-	Manual
TNT-S (Han et al., 2021)	24	5.2	81.5	95.7	Manual
PVT-Small (Wang et al., 2021)	25	3.8	79.8	-	Manual
ViL-Small (Pengchuan et al., 2021)	25	4.9	82.0	-	Manual
ResMLP-S12 (Touvron et al., 2021a)	31	6.0	79.4	-	Manual
Twins-PCPVT-S (Xiangxiang et al., 2021)	24	3.8	81.2	-	Manual
PoolFormer-S36 (Yu et al., 2022)	31	5.2	81.4	-	Manual
RegNetY-4G (Ilija et al., 2020)	21	4.0	80.0	-	Auto
AutoFormer-small (Chen et al., 2021b)	23	5.1	81.7	-	Auto
GLiT-Small (Chen et al., 2021a)	25	4.4	80.5	-	Auto
ViTAS-DeiT-B (Su et al., 2021)	23	4.9	80.2	95.1	Auto
S3-T (Minghao et al., 2021)	28	4.7	82.1	95.8	Auto
ViT-ResNAS-Medium (Liao et al., 2021)	97	4.5	82.4	-	Auto
BurgerFormer-Base	26	3.9	82.7	96.2	Auto
Swin-S (Liu et al., 2021)	50	8.7	83.0	-	Manual
Twins-PCPVT-B (Xiangxiang et al., 2021)	44	6.4	82.7	-	Manual
CvT-21 (Haiping et al., 2021)	32	7.1	82.5	-	Manual
BoTNet-S1-59 (A. et al., 2021)	34	7.3	81.7	-	Manual
RegNetY-8G (Ilija et al., 2020)	39	8.0	81.7	-	Auto
BossNet-T1 (Changlin et al., 2021)	-	7.9	82.2	95.8	Auto
BurgerFormer-Large	36	6.5	83.0	96.8	Auto

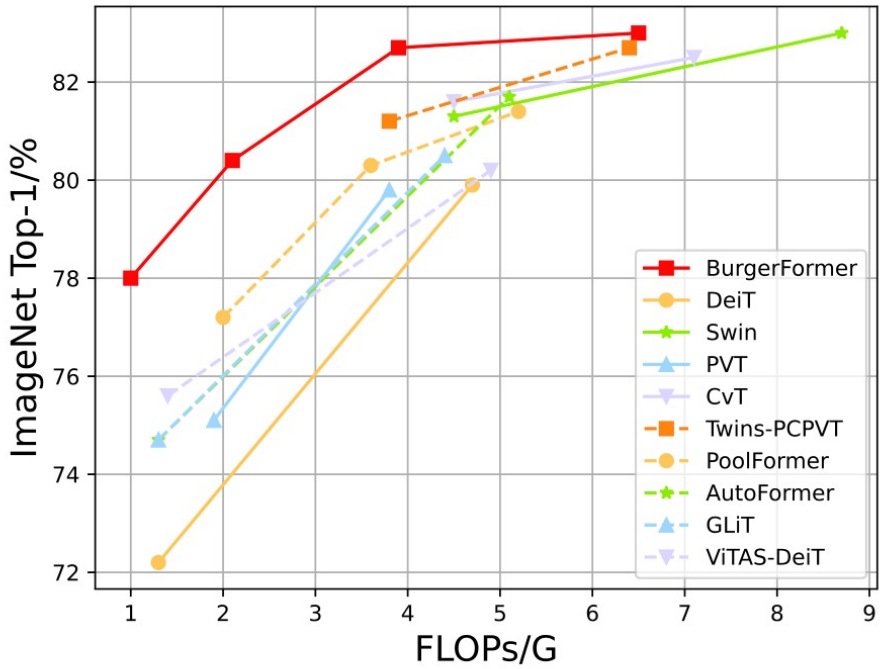


Figure 5. Comparison between BurgerFormer and efficient vision transformers, in terms of ImageNet Top-1 over FLOPs.

COCO Results

BackBone	RetinaNet 1x						
	Param. (M)	AP^b	AP^b_{50}	AP^b_{75}	AP_S	AP_M	AP_L
ResNet50 (He et al., 2016)	37.7	36.3	55.3	38.6	19.3	40.0	48.8
PVT-Small (Wang et al., 2021)	34.2	40.4	61.3	43.0	25.0	42.9	55.7
PoolFormer-S24 (Yu et al., 2022)	31.1	38.9	59.7	41.3	23.3	42.1	51.8
BurgerFormer-Base	35.9	41.2	61.3	43.9	24.2	44.4	55.4
BackBone	**Mask R-CNN 1x**						
	Param. (M)	AP^b	AP^b_{50}	AP^b_{75}	AP^M	AP^M_{50}	AP^M_{75}
ResNet50 (He et al., 2016)	44.2	38.0	58.6	41.4	34.4	55.1	36.7
PVT-Small (Wang et al., 2021)	44.1	40.4	62.9	43.8	37.8	60.1	40.3
PoolFormer-S24 (Yu et al., 2022)	41.0	40.1	62.2	43.4	37.0	59.1	36.9
Swin-T (Liu et al., 2021)	48.0	43.7	66.6	47.7	39.8	63.3	42.7
BurgerFormer-Base	45.9	44.0	65.6	48.1	40.2	62.7	43.4

Table 3. Comparison with state-of-the-art models on COCO.

ICML
International Conference
On Machine Learning

Thank You!

Searching for BurgerFormer with Micro-Meso-Macro Space Design