A Search Space

A.1 Detailed Description of the Once-For-All Approach

The Once-For-All (OFA) methodology incorporates several key dimensions for searching optimal configurations in Convolutional Neural Networks (CNNs), specifically focusing on depth, input resolution, width, and kernel size. This approach is inspired by and follows the settings outlined in MSuNAS [26].

We structure the CNN into five distinct stages, with each stage undergoing thorough parameter exploration for depth, width, and kernel size as illustrated in Figure 7. The resolution parameter is adjusted between 192 to 256, with incremental steps of 4, offering 17 unique settings. Depth levels considered are 2, 3, or 4, while the expansion ratio is chosen from the set {3, 4, 6}, and the kernel size options are {3, 5, 7}. Notably, a fixed-length encoding scheme is employed for representing the neural network architectures. When the depth parameter does not equal 4, zero padding is applied to extend the block's encoding to a consistent length of 8, ensuring uniformity in representation across different configurations.

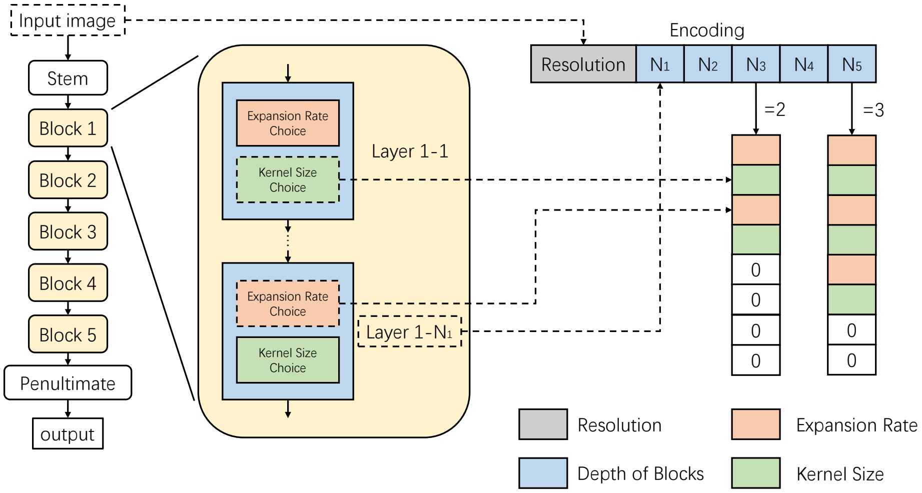

Fig. 7. OFA-based Search Space: The candidate architecture in the search space contains five blocks. Each block, denoted as b^{th} block, comprises N_b layers with a specific expansion rate and kernel size.

A.2 Detailed Description of the AutoFormer-based Search Space

In our approach, the AutoFormer-based search space is used to decompose the Vision Transformer (ViT) architecture into several fundamental dimensions, specifically Embedding Dimension, MLP Ratio, Head Number, and Depth as Table 3.

The Depth and Embedding Dimension parameters are crucial as they directly influence the overall complexity of the ViT model. For each block stacked vertically in the model's depth, the MLP ratio and the Head Number parameters are explored independently. To standardize the configuration representation across different models, zero-padding is utilized to transform the parameter encoding into a fixed-length decision vector, as depicted in Figure 8.

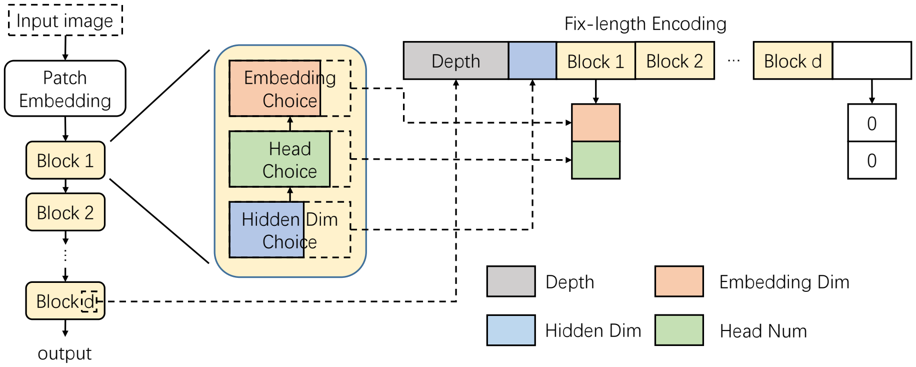

Fig. 8. AutoFormer Search Space: In this search space, the Vision Transformer architecture is conceptualized as a series of stacked blocks. Each block features a specific embedding dimension, while the number of attention heads and the hidden dimensions of the multi-layer perceptron within each block can be adjusted variably.

Table 3. Search space of AutoFormer. Tuples of three values represent the lowest value, the highest value, and steps.

	Supernet-tiny	Supernet-small	Supernet-base
Embed Dim	(192, 240, 24)	(320, 448, 64)	(528, 624, 48)
MLP Ratio	(3.5, 4, 0.5)	(3, 4, 0.5)	(3, 4, 0.5)
Head Num	(3, 4, 1)	(5, 7, 1)	(8, 10, 1)
Depth Num	(12, 14, 1)	(12, 14, 1)	(14, 16, 1)

B Parameter Restriction in AutoFormer

We replicated AutoFormer based on the source code and pre-trained networks provided at https://github.com/microsoft/Cream/tree/main/Autoformer. The neural architecture search (NAS) process within AutoFormer is conducted by constraining the search parameter ranges and employing an evolutionary algorithm (EA). We set up seven different parameter restrictions based on the figures provided in the AutoFormer literature and descriptio in the project's GitHub issue https://github.com/microsoft/Cream/issues/74 for completing the replication process. The detailed settings of these seven parameter groups are shown in Table 4 and correspond one-to-one with Figure 5.

Table 4. AutoFormer Parameter Restriction

Supernet-Tiny		Supernet-Small		Supernet-Base	
Code	#Params	Code	#Params	Code	#Params
1	[5M, 6M]	2	[16M, 18M]	5	[41M, 44M]
		3	[22M, 24M]	6	[52M, 54M]
		4	[17M, 29M]	7	[65M, 68M]