

Research

SP-HRNet: single-path neural architecture search for human pose estimation

Mengyu Sun¹ · Yichao Wang¹ · Peng Wang¹ · Wuqi Gao² · Qunli Li¹ · Xiaoyan Li¹ · Ruohai Di¹

Received: 24 August 2024 / Accepted: 2 April 2025

Published online: 14 June 2025

© The Author(s) 2025 **OPEN**

Abstract

In this study, focus on the task of human pose estimation. High-resolution networks (HRNet), while adept at continuously integrating multi-resolution information, tend to compel the network to assimilate redundant information, thereby amplifying computational complexity. The single-path one-shot Neural Architecture Search (NAS) method, an efficacious solution to mitigate this issue. This research introduces a novel NAS method, denoted as SP-HRNet, based on Single-Path One-Step architecture. SP-HRNet meticulously selects the most suitable operator within a high-resolution network and identifies the optimal deep fusion construction methodology. Within SP-HRNet, we propose a hybrid search space incorporating an enhanced search strategy. Unlike the original high-resolution network, instead of modifying the network branches with a fixed decrement or increment, our network structure will be free to choose the most appropriate one among multiple branches. For this purpose, we devise two distinct search spaces labeled as Mixed Operator and Depth Fusion Options and formulate a novel search strategy tailored to this space. A myriad of experiments substantiates the effectiveness of this method. Our proposed network attains outstanding performance on the COCO human pose estimation dataset, achieving heightened accuracy while concurrently reducing the number of model parameters.

Article highlights

- A search space was meticulously designed, incorporating a fusion of lightweight and efficient modules, for the exploration of the backbone network in human pose detection. Experimental findings substantiate the superiority of this combined search space over spaces exclusively comprised of either lightweight or efficient modules.
- By formulating a search space capable of tuning the selection of deep information stacks and fusion layers, and employing NAS searching, the most suitable network structure is successfully found in a complex multibranch network.
- A uniformly sampled single-path neural network structure search is deployed in the field of human pose detection and by replacing the original one-time search with a step-by-step replacement search strategy. Our search methodology facilitates the rapid identification of the optimal subnetwork, and it takes very few training rounds to get the optimal weights.

Keywords Neural network · NAS · Human posture estimation · High-resolution network

✉ Peng Wang, 1065746665@qq.com | ¹School of Electronic Information Engineering, Xi'an Technological University, Xi'an 710021, China. ²School of Computer Science and Engineering, Xi'an Technological University, Xi'an 710021, China.



1 Introduction

Being a pivotal research avenue in computer vision, multi-person pose estimation [1–4] strives to precisely deduce the positions of key joints (e.g., elbows, wrists, eyes) and capture the pose information of the human body within images or videos. Recent literature highlights the remarkable strides made by neural networks in achieving state-of-the-art performance. Many older methodologies involve the transmission of inputs through networks, typically composed of cascaded sub-networks spanning high to low resolutions, followed by an upscaling process. For instance, Hourglass [5] restores the resolution of prominent points through a symmetrical process from low to high, while SimpleBaseline [6] employs transposed convolutional layers to generate high-resolution representations. Furthermore, dilated convolution finds application in disrupting later layers of high to low-resolution networks (such as VGGNet or ResNet) [7]. Most recent studies have used high-resolution network methods [30, 33], they can effectively improve the accuracy of human posture detection. But recent investigations, illuminate that the multi-channel fusion employed in HRnet is redundant. This redundancy has the potential to transform parallel processes into serial processes, thereby diminishing network inference rates.

Three straightforward approaches can be employed to address this issue. Firstly, employing pruning and quantization techniques on HRnet facilitates channel reduction, optimizing the network, diminishing model complexity, and enhancing accuracy. The second approach involves substituting the computationally intensive residual modules in HRnet with lightweight alternatives. The third method entails leveraging Neural Network Architecture Search (NAS) to automatically discover efficient network structures. The first two methods are both handcrafted to create a new network or module and rely on expert experience. Meanwhile, when facing different scenarios in real problems, the generalizability of the network will be more challenged, and often the network performance will be degraded when facing different problems with excellent network structure in one domain. The last effectively mitigates the challenges associated with the high engineering cost of manually crafting compact models and the substantial reliance on human expertise and experience [8]. Networks with varying structures can be obtained directly by searching freshly in the hypernet, drastically reducing the time consumed in re-creating networks when facing different problems.

As a result, this paper adopts a contemporary strategy by embracing neural network architecture search (NAS) as an alternative to manual problem-solving.

This article introduces a novel solution, SP-HRNet, aimed at resolving the aforementioned challenges. SP-HRNet leverages a one-shot Neural Architecture Search (NAS) [9, 40, 41] to explore structures within the backbone network. The devised search space encompasses both modular and structural dimensions. The structural search involves exploring the depth fusion options, determining the number of stacked modules at each network stage, and deciding whether multi-resolution fusion is required for each network stacking search—parameters added to the search space for exploration. A specialized search strategy is tailored to this space, facilitating the separation of module search and structural search during exploration. Following pretraining of the entire search space on the dataset to obtain initial weights, the candidate architecture commences searching by inheriting weights from the supernet. Initially, module selection is locked, and the optimal structure is sought using mAP^{75} as the accuracy indicator with maximum weight (representing the maximum proportion). Subsequently, structure selection is locked, and another metric, mAP^{50} , is utilized as the maximum weight to explore optimal module selection. The combination of these two structures forms the most promising subnet, serving as the parent for subsequent searches utilizing evolutionary algorithms to traverse the search space. This approach enables the simultaneous exploration of both optimal modules and structures, efficiently identifying the optimal subnet within an expansive search space.

The efficacy of SP-HRNet has been thoroughly validated through a series of comprehensive experiments, demonstrating its capability to search for efficient and lightweight human pose estimation networks. The overall search went through only 15,000 epochs and took 3 GPU days. As illustrated in Fig. 1, our method yields distinct network structures depending on the input image size. At an input size of 256×192 , our approach achieves an accuracy of 73.7, showcasing a 1% improvement compared to HRNet, while necessitating merely 13.1 M Params—equivalent to 45.9% of HRNet. Moving to an input size of 354×288 , we identify a subnet with only 35.4% of the parameters of HRNet, coupled with an improved accuracy. Our network exhibits superior fitting effects, requiring a mere 50 epochs for retraining to attain optimal results—merely 23.8% of the 210 epochs typically needed for similar networks.

The remainder of the paper is organized as follows. The related work is described in Sect. 2. Section 3 presents the proposed approach, SP-HRNet, in detail. In Sect. 4, experimental studies are conducted, and compared with state-of-the-art

Search Design (Section 3.2)

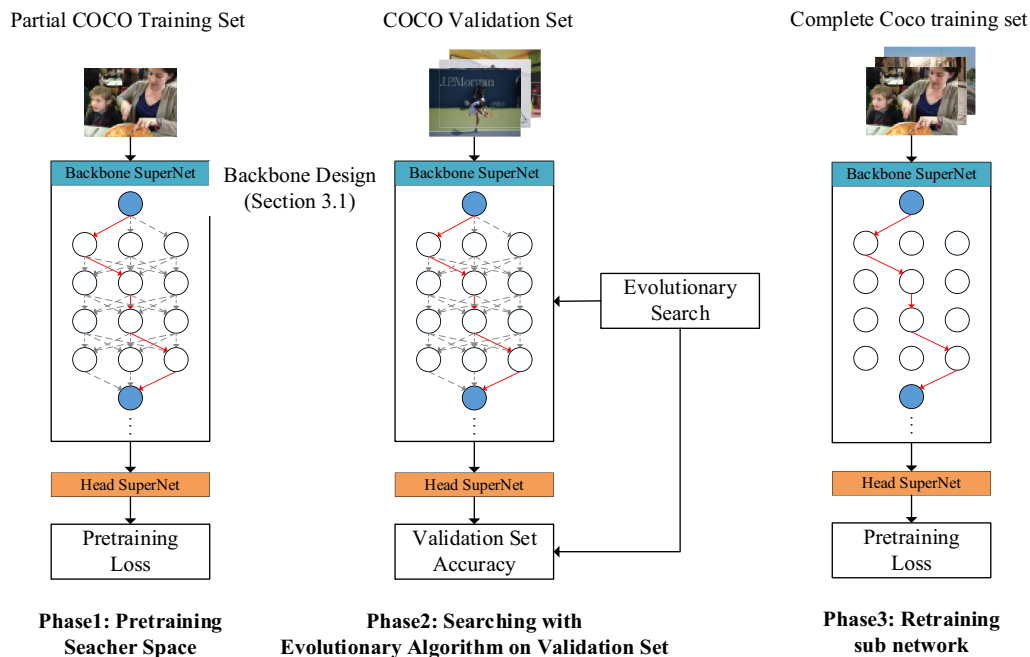


Fig. 1 The search process of this article is divided into three steps: pretraining searcher space, Searching with Evolutionary Algorithm on the Validation Set, and Retraining subnetwork

NAS methods. More importantly, the corresponding experimental results are analyzed and ablation experiments are performed. Finally, Sect. 5 summarizes this paper.

The main creations of this work are:

- A search space was designed, incorporating a fusion of lightweight and efficient modules, for the exploration of the backbone network in human pose detection. Experimental findings substantiate the superiority of this combined search space over spaces exclusively comprised of either lightweight or efficient modules.
- By formulating a search space capable of tuning the selection of deep information stacks and fusion layers, and employing NAS searching, the most suitable network structure is successfully found in a complex multibranch network.
- A uniformly sampled single-path neural network structure search is deployed in the field of human pose detection and by replacing the original one-time search with a step-by-step replacement search strategy. Our search methodology facilitates the rapid identification of the optimal subnetwork, and it takes very few training rounds to get the optimal weights.

2 Related work

2.1 Human pose detection

Traditional solutions for single-person pose detection predominantly rely on probabilistic graphical models or graphical structural models [10], with recent advancements incorporating improved modeling of unary and pairwise energies or emulation of iterative reasoning processes [11] through deep learning. In contemporary scenarios, deep convolutional neural networks emerge as the primary solution [12–15]. Two mainstream methodologies prevail: regressing the position of key points [16], and detecting heatmaps of key points [17], followed by selecting the position with the highest heat value as the key point. Convolutional neural networks utilized for keypoint heatmap detection typically comprise

backbone subnetworks resembling classification networks, generating intermediate resolutions equivalent to the input representation.

Recent advancements in human pose estimation [18–23] center around the design of more robust neural network architectures with increased model capacity to enhance accuracy. Attention mechanisms have proven effective in capturing contextual information for human pose estimation. For instance, Chu et al. [24] introduced multi-context attention to enhance model robustness and accuracy. Su et al. [25] proposed the SCARB module, aiming to augment pyramid features and channel context through spatial enhancement. Various attention modules have been explored extensively, including the Squeeze and Excitation (SE) block [26] for modeling channel relationships and the Global Context (GC) block [27] for capturing global context, as demonstrated in NLNet [28] through additive fusion.

In recent times, the advent of high-resolution networks (HRNet) [29, 30] has garnered attention. However, the network structure of high-resolution networks is deemed excessively complex, encountering challenges in the selection of base modules, determination of module stack numbers, and the rationality of fusion operations within network construction. One approach to address this predicament involves crafting an efficient and compact module. For instance, Kato et al. [31] and M. Ding et al. [32] opt to replace the BasicBlock in HRNet with a novel lightweight module. Additionally, Sota et al. proposed Lite HRNet Plus [33], an end-to-end human pose detection model, to improve on the problem of high computational cost due to fusion blocks connecting feature maps of different resolutions in Lite HRNet by designing a novel fusion block based on channel attention. Yi et al. [30] proposed a lightweight X-HRNet network, which found that a high-resolution representation is necessary to achieve high-performance human pose detection, but leads to high computational complexity. In this paper, the backbone network is constructed utilizing HRNet-like methods, and modifications to network issues are implemented through Neural Network Architecture Search (NAS).

2.2 Neural architecture search

Neural Structure Search (NAS) is centered on the automation of neural network architecture design. Initially applied predominantly to image classification, NAS has recently extended its scope to encompass various domains, including image segmentation and object detection. Concurrently, NAS has made inroads into the realm of human pose, as elucidated by [34, 35], who describe their method in the referenced paper.

In the early stages of Neural Architecture Search (NAS), weight-sharing methods predominantly utilized continuous relaxation to parameterize the search space [36, 37]. The distribution parameters of the architecture undergo joint optimization through gradient-based methods during the training of the supernet, from which the optimal architecture is subsequently sampled. However, this approach is not without its challenges. Firstly, the weights in the hypernetwork are deeply intertwined. Secondly, joint optimization introduces additional coupling between architecture parameters and supernet weights, and the inherent bias in the optimization process based on methods can potentially misguide architecture search. Finally, when sampling and training a large number of architectures for a hypernetwork, the process involves training thousands of architecture candidates from scratch, leading to computationally prohibitive costs.

Recent Neural Architecture Search (NAS) methods [38–41] have embraced weight-sharing strategies involving the training of super networks. Liao et al. [40] proposed an evolutionary multi-task neural architecture search (EMT-NAS) algorithm that accelerates the search process by transferring architectural knowledge between multiple related tasks. Lu et al. [41] proposed to explicitly minimize the gradient variance of hypergrid training by jointly optimizing the sampling distributions of P_{Path} and D_{Ata} (P_A&D_A). More recent approaches adopt a one-shot weight-sharing strategy to amortize search costs. The fundamental concept involves training a singular hyperparameterized hypernetwork model and subsequently sharing weights across subnets. Uniformly sampling a single path [8] is a representative one-time approach in which only one random path is sampled and a batch of data is utilized to train the path in each iteration to search for the optimal network structure through an evolutionary algorithm. Compared to differentiable NAS this approach can afford a much larger supernet structure, while the uniform sampling approach allows for the design of more complex and sophisticated supernet structures, rather than simply allowing module stacking on one or two branches as in other NAS. The HRNet network structure used in this paper is designed with multiple branches, which also leads to a rapid expansion of the number of operations and model parameters, which makes the uniform sampling single-path approach perform more appropriately. Upon completing the training process, subnets can share weights. The approach followed by SP-HRNet in this article aligns with the one-shot weight-sharing strategy, evaluating each subnetwork through shared

weights to expedite the network search process. Diverging from other one-shot strategies, this paper introduces structured search spaces. Furthermore, the evolutionary algorithm is enhanced for the search space to accelerate the overall search process.

3 Methodology

Our approach endeavors to explore multi-branch search spaces encompassing hybrid lightweight and efficient modules. The general flowchart of our search is illustrated in Fig. 1.

3.1 Search space

3.1.1 Mixed operator

Search space for hybrid lightweight and efficient modules. In earlier algorithms, the construction of search spaces was often task-oriented, specifically designed to consist of either all lightweight or all efficient modules [33, 42]. In some instances, only a single module was retained while altering its channel count or convolution kernel size [35, 42]. However, the inclusion of both lightweight and efficient modules in the search space proves advantageous for Neural Architecture Search (NAS) in uncovering the optimal structure. Experiments reveal that retaining only lightweight modules significantly diminishes accuracy, while using only efficient modules leads to a rapid expansion of the model size. The simultaneous presence of both lightweight and efficient modules not only results in a reduction of model parameters compared to pure precision, due to constraints, but also enhances accuracy in comparison to pure lightweight configurations.

As shown in Fig. 2, we employ the Group Convolution from ShuffleNetV2 [43], Dynamic Convolution with dynamic perceptrons [44] to construct a lightweight module, and a residual module BasicBlock [22] as an efficient module to establish a hybrid search space. All three modules have demonstrated their important role in optimizing the rest of the HRNet structure in previous work. All three modules have good lightweight effect with small number of model parameters and few floating-point operations, and their use in hypernets can effectively reduce the size of the hypernets so that the size of the hypernets can be afforded. However, at the same time, these lightweight modules fully replace the residual modules in HRNet as replacement modules in the original network. This resulted in a poor performance of the detection accuracy of these networks, and the complexity of the high-resolution network structure made it difficult to determine by expert experience where in the network the lightweighting modules should be used without a significant degradation of the network accuracy. Therefore, lightweight modules should be included along with high-precision modules when designing super-networks.

Translated with www.DeepL.com/Translator (free version). This space exhibits diverse composition structures based on the four stages of the network, as outlined in Table 1. In Sect. 4.3, "Ablation and Analysis of Mixed Operator Analysis," we demonstrate the validity of our chosen search space.

As depicted in Table 1, each Block Search in every stage is stacked with 4 building blocks, and the input and output of each building block remain consistent without altering the channel size. The modules in Stage 1 are all built by Bottleneck, including three options with {3, 5, 7} kernel sizes {1, 2, 3} expansion rates. The selection in this stage is mainly used for channel expansion. The remaining stage modules are composed of DYConv, GPConv, and BasicBlock, with an expansion rate and step size of 1. As the stages increase, the number and number of channels gradually increase. For example, in Stage 3, the network has three branches, each with different channel numbers of 32, 64, and 128. Different branches share the same module selection at the same location (Fig. 2).

3.1.2 Depth, fusion options

The construction of the backbone network closely follows the approach of HRNet-W32. We design a multi-branch network as the foundation for the search space encompassing deep information and fusion timing. As illustrated in Fig. 3, the network initiates from the first stage and progressively incorporates a new branch after each stage. In the subsequent stage, additional branches are introduced based on the minimum resolution branch of the current stage, with multi-resolution branches computed in parallel. Owing to the multi-stage nature of the network, it features multiple parallel

Table 1 Search space structure of hybrid lightweight and efficient modules. " $N_{choices}$ " represents the number of choices for the current block, whereas " 4×1 " indicates that a Block Search stack consists of 4 modules with a total of one branch. " Chn " represents the number of channels in each module, and "\/" distinguishes different branches

	Operators	$N_{choices}$	Chn	Padding
Stage 1	3×3 Bottleneck	1×1	256	1
	5×5 Bottleneck	1×1	256	2
	7×7 Bottleneck	1×1	256	3
	DYConv	3×1	256	1
	GPConv	3×1	256	1
	BasicBlock	3×1	256	1
Stage 2	DYConv	4×2	32 / 64	1
	GPConv	4×2	32 / 64	1
	BasicBlock	4×2	32 / 64	1
Stage 3	DYConv	4×3	32 / 64 / 128	1
	GPConv	4×3	32 / 64 / 128	1
	BasicBlock	4×3	32 / 64 / 128	1
Stage 4	DYConv	4×4	32 / 64 / 128 / 256	1
	GPConv	4×4	32 / 64 / 128 / 256	1
	BasicBlock	4×4	32 / 64 / 128 / 256	1

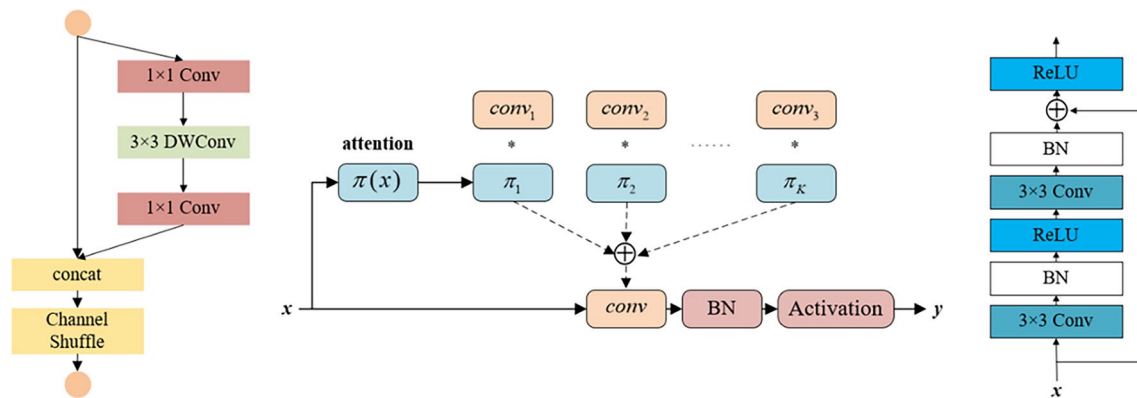
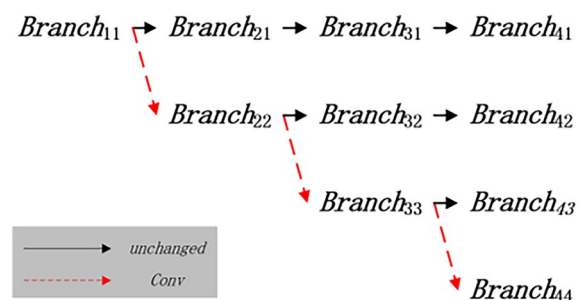


Fig. 2 Schematic structure of the various methods used in this paper. From left to right are GPConv, DYConv, BasicBlock

Fig. 3 An example structure containing a parallel branch network, where $Branch_{xy}$ represents the x branch of the y stage, and $Conv$ represents a convolution operation to add a new branch through the lowest resolution branch



branches with different resolutions, and resolutions from preceding stages persist in subsequent stages. The network comprises two main components: parallel module stacking and fusion modules.

The interplay between module stacking and fusion modules in the network is elucidated in Fig. 4. Each parallel module stack repetitively incorporates the hybrid modules constructed in Sect. 3.1.1 four times within the search space, following which the fusion layer integrates information from different branches.

Depth Information. At each stage of the network, the previously mentioned Depth Blocks are iteratively stacked to augment the network's depth. Enhancing the depth contributes to improved accuracy, but each iteration of stacking Depth Blocks results in a notable increase in the number of model parameters. To address this, this article establishes

Fig. 4 Parallel module stacking and fusion modules. $Sblock_x$ represents the x-th Search Block, and $Fusion$ represents the fusion operation

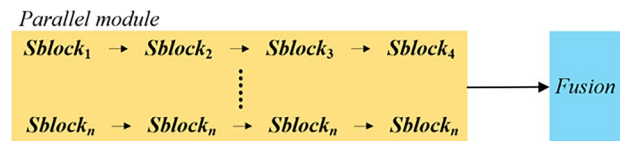
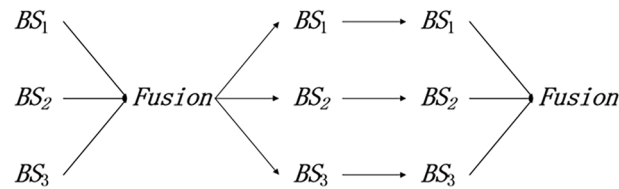


Fig. 5 This is an example of our network search structure, which simulates a situation where the search results are obtained. At this point, the network consists of three branches, constructing three layers of depth. Therefore, the basic module BS_x is repeated three times, while the fusion layer $Fusion$ is used in the first and second depth constructions

a deep information search space. It identifies the optimal depth layer for each stage of the network through architecture search and determines the most suitable depth for the structure. This approach enhances accuracy without causing a significant surge in model parameters, thereby expediting network inference speed.

Fusion Selection. In deep construction, the repetitive stacking of Depth Blocks can result in the network conducting fusion calculations repeatedly. However, fusion calculations significantly impact network performance by transforming serial calculations into parallel calculations, diminishing the inference efficiency of the network. Moreover, fusion calculations entail a substantial amount of floating-point computation. The repetition of fusion, while not essential for enhancing network accuracy, can lead to the learning of redundant information and a decrease in accuracy. To address this concern, a fusion timing search space is formulated to determine through architecture search whether different resolution feature information needs to be fused in various Depth Blocks.

As illustrated in Fig. 5, this serves as an example of our network search structure, simulating a scenario where the search results are obtained. In this instance, the network comprises three branches, building three layers of depth. Consequently, the basic module BS_x is iteratively repeated three times, while the fusion layer $Fusion$ is employed in the first and second depth constructions. A comprehensive network structure diagram is shown in Fig. 6.

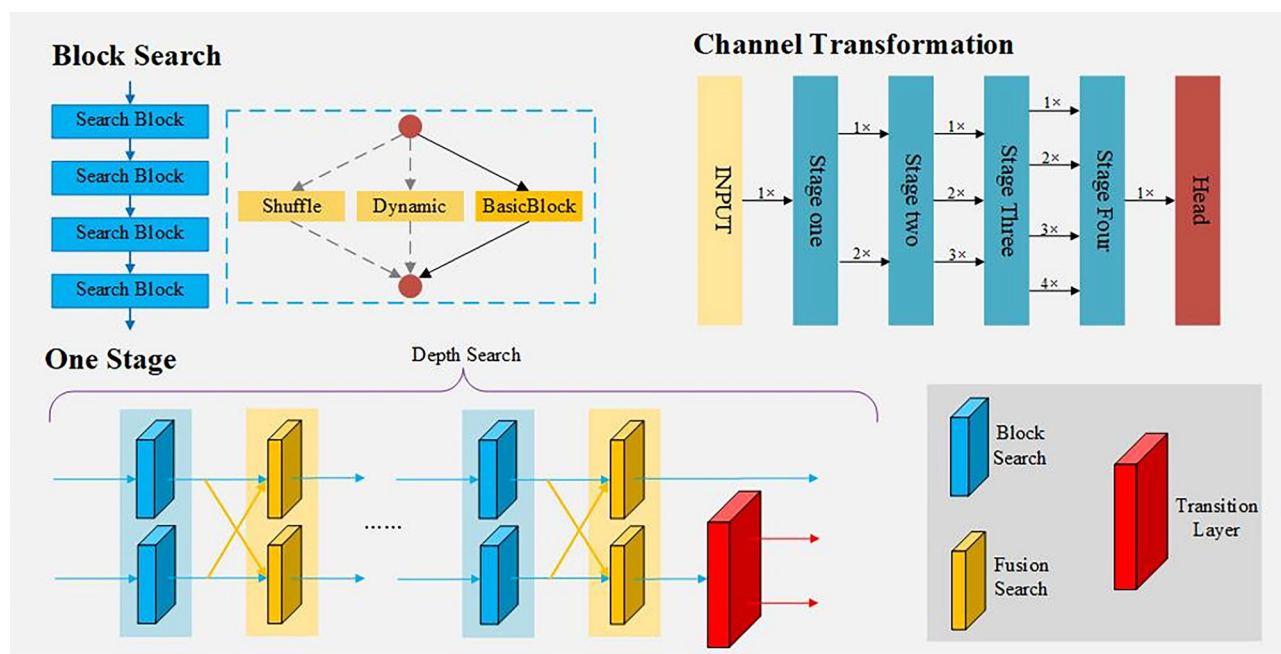


Fig. 6 Backbone design. A diagram showing the complete backbone network

3.2 SPOS NAS

In earlier Neural Architecture Search (NAS) methods, a deep coupling existed between architecture parameters and weights, and the joint optimization of these two factors could impact search time and results. To mitigate this challenge, the single-path one-shot neural network architecture search was introduced. This approach involves designing a comprehensive supernet, wherein all candidate structures are employed as subnets, and the weights of shared subnets are trained. More specifically, the architecture search space The complete flowchart of the search is shown in Fig. 7.

Searching with Evolutionary Algorithm. In the initial phase of training the weights W^p of the supernet, an evolutionary algorithm is employed to search the supernet. To expedite the overall search process, each structure utilizes only a partial validation set to calculate accuracy, without traversing the complete validation set. Through synthesis experiments, we observed that the parallel module stacking and fusion modules in the network structure impact different metrics in the OKS evaluation metrics to varying degrees. Based on this observation, we opted to employ different metrics to evaluate different modules. Firstly, some subnets $\alpha^S \in \mathcal{A}$ are randomly selected from the supernet $\mathcal{A}$ that change only the depth and fusion structure, and the validation set accuracy Oks_{AP75} is calculated based on the supernet weight W^p . The experiment shows that this accuracy is very sensitive to the fusion layer in the network structure. Therefore, the parent generation is sorted using this accuracy first. Then the top-ranked of all deep, fused subnet structures are selected as parents generation $\alpha^{FS} \alpha^{FS}$.

$$\alpha^{FS} = \arg \max_{\alpha^S \in \mathcal{A}} Oks_{AP75}(\mathcal{N}(\alpha^S, W^p(\alpha^S))), \quad (1)$$

Then, the optimal parent α^{FS} is randomly mutated to obtain the offspring α^F . Then, a new round of search will change the selection of modules for the subnetwork structure without changing the depth information and the fusion timing. Parents decide how to mutate to their offspring according to the mutation rate. The mutation rate determines the probability that each node in the parent's structure undergoes a change, and its size ranges from 0.1 to 1. The mutation rate should not be set too large, if it is too large it will cause the network structure to transform too many positions, making it difficult for the offspring to inherit the parent's trained weights efficiently in weight sharing again. You can set the mutation rate larger at the beginning of the search to quickly traverse the search space, and then gradually reduce the mutation rate.

Finally, sort the structures of all parents and children with precision Oks_{AP50} , and select the best-performing structure α^* as the parent for the next search.

$$\alpha^* = \arg \max_{\alpha^{FS} \rightarrow \alpha^F} Oks_{AP50}(\mathcal{N}(\alpha, W^p(\alpha^{FS}, \alpha^F))), \quad (2)$$

Set a threshold $Flops_{max}, Params_{max}$ in the search to filter out the optimal structure α^* that is too large for the model.

$$\begin{aligned} Flops(\alpha^*) &\leq Flops_{max}, \\ Params(\alpha^*) &\leq Params_{max}, \end{aligned} \quad (3)$$

Retraining subnetwork. By searching for the most promising subnet α^* in the second step, inherit the pretraining weight W^p and optimize the subnet weights on the complete training dataset using the minimum loss function $\mathcal{L}_{train}^{com}$, so that the optimal subnet W^* is obtained after training, and evaluate the accuracy Oks_{val} on the validation set.

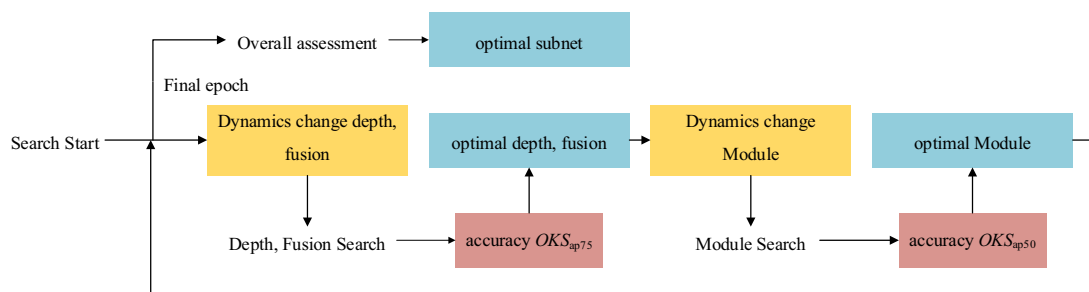


Fig. 7 Search design

$$\begin{aligned}
 W^* &= \arg \min_W \mathcal{L}_{train}^{com}(\mathcal{N}(\alpha^*, W^p)), \\
 Oks_{\alpha^*} &= Oks_{val}(\mathcal{N}(\alpha^*, W^*)),
 \end{aligned}
 \tag{4}$$

4 Experiments

4.1 Implementation details

We evaluate our approach on the COCO dataset [45] and report the comparisons with the latest methods.

Datasets. The MS-COCO dataset includes over 60 k images and 250 k person instances annotated with 17 human joint categories, *e.g.*, left-/right-ear, left-/right-elbow. We train our method on the COCO tran2017 dataset, including 57K images and 150K person instances. The val2017 set and test-dev2017 set are utilized to evaluate our approaches, containing images, respectively.

Pretraining. We do not apply initialization weights and directly use the training dataset to train the supernet. In each round of training, a seed net is randomly selected from the supernet, and each seed net only trains 1024 images. Train 15,000 rounds using the following settings: We extend the human detection boxes in height or width to a fixed aspect ratio: height:width = 4:3, and crop the resized box from the image, which is further rescaled to a fixed size, 256×192 or 384×288 . We randomly augment each person sample with a random scale factor ($[0.75, 1.25]$), random rotation ($[-30^\circ, 30^\circ]$), horizontal flipping and additional half-body augmentation. The network is trained from scratch on NVIDIA RTX4070 with a mini-batch size. For optimization, we adopt SGD optimizer with an initial learning rate.

Search. After the pre-training is completed, we evaluate the subnets on the COCO validation set, using 1024 images per subnet for evaluation. The mutation rate of offspring is set to 0.3. It changes to 0.1 halfway through the search. Image cropping is consistent with pretrain. A total of 30,000 rounds of search were conducted.

Retaining. After evolutionary search, we use pretrained weights on the COCO training set and retrain the optimal subnet with the same image changes. With initial learning rate $2e^{-3}$. The training process is terminated within 50 epochs, and the learning rate is dropped to $2e^{-4}$ and $2e^{-5}$ at the 40-th and the 45-th epoch, respectively.

Testing. We adopt the top-down paradigm (detect and estimate) with the detection results provided by previous work [46] for a fair comparison. Following the common practice [47–49], we compute the final heatmap by averaging the heatmaps of the original and flipped images. The coordinates of keypoints are decoded from the heatmaps by adjusting the highest response location.

Evaluation metric. The OKS-based mAP metric is adopted to evaluate our approach. The OKS (Object Keypoint Similarity) measures the similarity between ground truth poses and predicted poses. We report standard mean average precision and recall scores: AP (the mean of AP scores with OKS threshold = $[0.5:0.95:0.05]$), AP^{50} , AP^{75} , AP^M (AP score for instances with medium size), AP^L (AP score for instances with large size), AR for a fair comparison with other methods.

4.2 Results and comparisons

COCO val2017. As shown in the Table 2, in order to make a fair comparison with other methods, the input size is scaled to two fixed values of 256×192 and 384×288 for this experiment, and to show the cost of network operations, both GFLOPS and #Params will be compared. Our method achieves higher accuracy while reducing the number of model parameters, resulting in better results compared to the HRNet. Changing the size of the network input image will search different subnets throughout the search process. We set the input image to 256×192 and 384×288 to search two different subnets, SP-HRNet-256 and SP-HRNet-384. When the input size is 256×192 , our SP-HRNet-256 achieves a 73.7 AP score, has higher accuracy. The data in the Table 2 were obtained directly from the paper. HRNet-W32* represents HRNet-W32 replicated by our experimental conditions. our experimental conditions could not reach the data in the paper, but our algorithmic data is still higher than the original HRNet-W32 data. Compared with HRNet, our method achieves a 0.3 AP gain without the need for pretrain, while only requiring 45.9% of parameters and 53.0% of GFLOPs. When the input size is 384×288 , the network accuracy achieved a score of 75.9 AP. When the input size is 384×288 , our SP-HRNet-384 compared to HRNet-W32, it only improved by 0.1 AP. However, the network parameter count is only 35.4% of HRNet-W32 and 52.1% of GFLOPs.

Table 2 Comparisons on the COCO val2017 dataset. SP-HRNet-256 and SP-HRNet-384 represent the two different structures searched by NAS when the input images are 256×192 by 384×288 , respectively

model	pretrain	input size	#Params	GFLOPs	AP	AP ⁵⁰	AP ⁷⁵	AP ^M	AP ^L	AR
<i>Large networks</i>										
8-stage Hourglass	N	256×192	25.1M	14.3	66.9	-	-	-	-	-
CPN	Y	256×192	27.0M	6.20	68.6	-	-	-	-	-
DARK-W48	Y	128×96	63.6M	3.6	71.9	89.1	79.6	69.2	78.0	77.9
HRNet-W32	N	256×192	28.5M	7.10	73.4	89.5	80.7	70.2	80.1	78.9
HRNet-W32*	N	256×192	28.5M	7.10	72.9	89.6	80.7	69.8	77.4	77.5
HRNet-W32	N	384×288	28.5M	16.0	75.8	90.6	82.7	71.9	82.8	81.0
<i>Small networks</i>										
MobileNetV2 1×	Y	256×192	9.6M	1.48	64.6	87.4	72.3	61.1	71.2	70.7
ShuffleNetV2 1×	Y	256×192	7.6M	1.28	59.9	85.4	66.3	56.6	66.2	66.4
Small HRNet-16	N	256×192	1.3M	0.54	55.2	83.7	62.4	52.3	61.0	62.1
Lite-HRNet-30	N	256×192	1.8M	0.31	67.2	88.0	75.0	64.3	73.1	73.3
X-HRNet-30	N	256×192	2.1M	0.30	67.4	87.5	75.4	64.5	73.3	73.5
<i>Ours</i>										
SP-HRNet-256	N	256×192	13.1M	3.76	73.7	92.0	81.4	72.0	77.2	77.6
SP-HRNet-384	N	384×288	10.1M	8.35	75.9	92.8	82.9	72.2	78.4	77.8

COCO test-dev2017. With the input size 384×288 , our method (SP-HRNet-384) achieves 74.9 AP scores, as reported in Table 3. our method maintains the superiority, which demonstrates the robustness of our method. Compared to HRNet-W32, our method achieves comparable performance but requires far fewer FLOPs and parameters. As in Fig. 7, this is a human posture estimation visualization detection result (Figs. 8 and 9).

4.3 Ablation and analysis

Mixed operator Analysis. We evaluate the effects of different operator in our method on COCO val2017 dataset, and report the results in Table 4. We use HRNet-W32 as the backbone network, replacing the modules with various candidate operations in the search space. Input image is 256×192 , with initial learning rate $2e^{-3}$. We're training the network from scratch.

Table 3 Comparisons on the COCO test-dev2017 dataset. The input size is set as 384×288

model	#Params	GFLOPs	AP	AP ⁵⁰	AP ⁷⁵	AR
<i>Large networks</i>						
DARK-W48	63.6M	32.9	76.2	92.5	83.6	81.1
HRNet-W32	28.5M	16.0	74.9	92.5	82.8	80.1
<i>Small networks</i>						
MobileNetV2 1×	9.6M	3.33	66.8	90.0	74.0	72.3
ShuffleNetV2 1×	7.6M	2.87	62.9	88.5	69.4	68.9
Small HRNet-16	1.3M	1.21	55.2	85.8	61.4	61.5
Lite-HRNet-30	1.8M	0.31	67.2	88.0	75.0	73.3
X-HRNet-30	2.1M	0.67	70.0	90.6	77.7	75.5
<i>Ours</i>						
SP-HRNet-384	10.1M	8.35	74.9	93.5	83.1	76.7



Fig. 8 Qualitative results for some example images from the COCO dataset

Fig. 9 Our method is compared with other methods. Efficiency comparisons (i.e., FLOPs) and performance (i.e., APs) are performed for attitude estimation (COCO) comparisons. Each method is represented by a circle, the size of which indicates the number of parameters

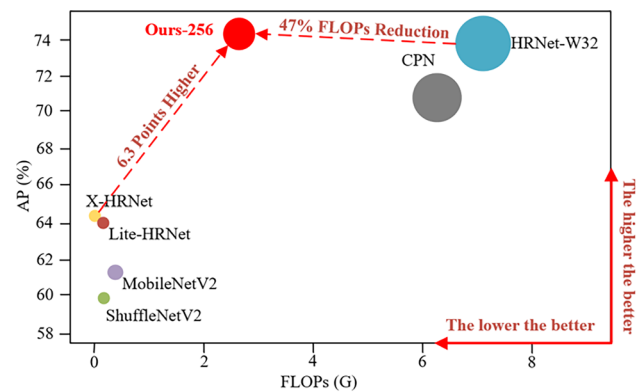


Table 4 Params, GFLOPs and AP of different operators

operator	#Params	GFLOPs	AP
DYConv	3.4M	0.34	63.1
GPConv	3.2M	0.37	64.8
BasicBlock	28.5M	7.10	73.4
BasicBlock*	28.5M	7.10	72.9

Table 5 Impact of change depth and fusion on networks

networks	#Params	GFLOPs	AP
<i>ours(BasicBlock)</i>	19.1M	4.85	73.5
HRNet-W32	28.5M	7.10	73.4
BasicBlock*	28.5M	7.10	72.9

The training process is terminated within 60 epochs, and the learning rate is dropped to $2e^{-4}$ and $2e^{-5}$ at the 40-th and the 50-th epoch, respectively. As reported in Table 4, The performance of building a backbone network separately for each operator in the search space is not as good as our method.

Impact of deep information and fusion timing. We discussed the effectiveness of deep information and fusion timing on networks. We use the searched network (SP-HRNet-256) and change the operator selection. Compare with HRNet-W32 under the same training conditions, terminated within 210 epochs, and the learning rate is dropped to and at the 170-th and the 200-th epoch. As reported in Table 5, Changing deep information and fusion timing can maintain similar accuracy to HRNet-W32 while reducing the number of model parameters and GFLOPs.

Analysis of Searched Architecture. We observe interesting phenomena. We decouple the operators and structures in the search and rank them using different accuracy metrics. Compared to other evolutionary algorithms, our method is specifically designed to adjust the high-resolution network structure. Our method separates the selection of operators and structures, and searches for their respective optimal subnets. This method can quickly search for potential subnets, and the entire search time takes 2 days. On the experimental surface, the fusion layer has a greater impact on stricter accuracy. Our search strategy first determines the fusion layer and then determines the operator, which results in a looser accuracy metric of the network achieving better results, while stricter metrics cannot achieve it.

5 Conclusion

This paper attempts to design a human pose estimation network through neural structure search. The proposed method, i.e., SP-HRNet redesigns the one-shot NAS specialized for high-resolution networks and introduces two effective search spaces. Extensive experiments on the COCO dataset have shown that the SP-HRNet achieves more advanced performance

than HRNet while using fewer triggers and parameters. In addition, SP-HRNet can use less search and training time, significantly reducing the network construction cost.

Author's contribution Mengyu Sun: Writing, Software—review & editing. Yichao Wang: Conceptualization, Methodology, Writing—original draft. Peng Wang: Supervision, Validation. Wuqi Gao: Writing—review & editing. Qunli Li: Writing—review & editing. Xiaoyan Li: Writing—review & editing. Ruohai Di: Writing—review & editing.

Funding This work was supported by National Natural Science Foundation of China (62171360), Key research and development program of Shaanxi Province (2024GX-YBXM-048), Shaanxi Science Fund for Distinguished Young Scholars (2024JC-JCQN-57), 2022 Shaanxi University Youth Innovation Team Project, 2023 Shaanxi University Engineering Research Center, Engineering Technology Research Center of Shaanxi Province for Intelligent Testing and Reliability Evaluation of Electronic Equipments.

Data availability Data is provided within the manuscript information files.

Declarations

Ethics approval and consent to participate Authors' adherence to ethical principles and the implementation of informed consent are essential requirements for protecting individuals' right to privacy and ensuring fairness and transparency in the use of data.

Consent for publication I (We), the undersigned author(s), hereby declare that the manuscript titled **"SP-HRNet: Single-Path Neural Architecture Search for Human Pose Estimation"** is an original work created by me (us) and does not infringe upon any third-party rights, including but not limited to intellectual property rights. I (We) agree to submit this manuscript to *Discover Applied Sciences* for review and potential publication. I (We) confirm that this manuscript has not been published previously, nor is it under consideration for publication elsewhere. Upon acceptance, I (we) grant *Discover Applied Sciences* the right to edit, format, and publish the manuscript in accordance with its publication policies. I (We) also agree to comply with the ethical and copyright regulations of the journal. All listed authors have read and approved this statement and confirm the authenticity and academic contribution of the manuscript.

Competing Interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. All authors disclosed no relevant relationships.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

1. Kocabas M, Karagoz S, Akbas E. Multiposenet: Fast multi-person pose estimation using pose residual network. *Lecture Notes in Computer Science*. Cham: Springer; 2018. p. 437–53.
2. Sekii T. Pose proposal networks. In ECCV. 2018. https://doi.org/10.1007/978-3-030-01261-8_21.
3. Nie X, Feng J, Xing J, Yan S. Pose partition networks for multi-person pose estimation. In ECCV. 2018.
4. Papandreou G, Zhu T, Chen L-C, Gidaris S, Tompson J, Murphy K. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In ECCV. 2018.
5. Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation. In ECCV. 2016. pp. 483–499. https://doi.org/10.1007/978-3-319-46484-8_29.
6. Xiao B, Wu H, Wei Y. Simple baselines for human pose estimation and tracking. In ECCV. 2018. pp. 472–487. https://doi.org/10.1007/978-3-030-01231-1_29.
7. Elsken T, Metzen JH, Hutter F. Neural architecture search: a survey. *J Mach Learn Res*. 2019;20:1–21.
8. Guo Z, Zhang X, Mu H, Heng W, Liu Z, Wei Y, et al. Single path one-shot neural architecture search with uniform sampling. In ECCV. 2020. https://doi.org/10.1007/978-3-030-58517-4_32.
9. Pishchulin L, Andriluka M, Gehler PV, Schiele B. Poselet conditioned pictorial structures. In CVPR. 2013. pp. 588–595. <https://doi.org/10.1109/CVPR.2013.82>.
10. Chu X, Ouyang W, Li H, Wang X. Structured feature learning for pose estimation. In CVPR. 2016. pp. 4715–4723. <https://doi.org/10.1109/CVPR.2016.510>.
11. Tang W, Yu P, Wu Y. Deeply learned compositional models for human pose estimation. In ECCV. 2018. https://doi.org/10.1007/978-3-030-01219-9_12.
12. Daneshfar F, Bartani A, Lotfi P. Image captioning by diffusion models: a survey. *Eng Appl Artif Intell*. 2024;138: 109288.

13. Nie X, Feng J, Zuo Y, Yan S. Human pose estimation with parsing induced learner. In CVPR. 2018. <https://doi.org/10.1109/CVPR.2018.00224>.
14. Peng X, Tang Z, Yang F, Feris RS, Metaxas D. Jointly optimize data augmentation and network training: adversarial data augmentation in human pose estimation. In CVPR. 2018.
15. Carreira J, Agrawal P, Fragkiadaki K, Malik J. Human pose estimation with iterative error feedback. In CVPR. 2016. pp. 4733–4742. <https://doi.org/10.1109/CVPR.2016.512>.
16. Chu X, Yang W, Ouyang W, Ma C, Yuille AL, Wang X. Multi-context attention for human pose estimation. In CVPR. 2017. pp. 5669–5678. <https://doi.org/10.1109/CVPR.2017.601>.
17. Cheng B, Xiao B, Wang J, Shi H, Huang TS, Zhang L. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020. pp. 5386–5395.
18. Duan H, Lin KY, Jin S, Liu W, Qian C, Ouyang W. Trb: a novel triplet representation for understanding 2d human body. In Proceedings of the IEEE/CVF international conference on computer vision. 2019. pp. 9479–9488.
19. Jin S, Liu W, Xie E, Wang W, Qian C, Ouyang W, et al. Differentiable hierarchical graph grouping for multi-person pose estimation. In European Conference, Computer Vision. 2020. https://doi.org/10.1007/978-3-030-58571-6_42.
20. Jin S, Xu L, Xu J, Wang C, Liu W, Qian C, Ouyang W, Luo P. Whole-body human pose estimation in the wild. In European Conference on Computer Vision. 2020. pp. 196–214.
21. Li J, Wang C, Zhu H, Mao Y, Fang HS, Lu C. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. pp. 10863–10872.
22. Sun K, Xiao B, Liu D, Wang J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. pp. 5693–5703.
23. Chu X, Yang W, Ouyang W, Ma C, Yuille AL, Wang X. Multi-context attention for human pose estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. pp. 1831–1840.
24. Su K, Yu D, Xu Z, Geng X, Wang C. Multi-person pose estimation with enhanced channel-wise and spatial information. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. pp. 5674–5682.
25. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. pp. 7132–7141.
26. Cao Y, Xu J, Lin S, Wei F, Hu H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE/CVF international conference on computer vision workshops. 2019.
27. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. pp. 7794–7803.
28. Sun K, Xiao B, Liu D, Wang J. Deep high-resolution representation learning for human pose estimation. In CVPR. 2019. <https://doi.org/10.1109/CVPR.2019.00584>.
29. Wang J, Sun K, Cheng T, Jiang B, Deng C, Zhao Y, Liu D, Mu Y, Tan M, Wang X, Liu W, Xiao B. Deep high-resolution representation learning for visual recognition. CoRR. 2019. abs/1908.07919.
30. Zhou Y, Wang X, Xu X, Zhao L, Song J. X-HRNet: Towards lightweight human pose estimation with spatially unidimensional self-attention. In 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan. 2022. pp. 01–06. <https://doi.org/10.1109/ICME52920.2022.9859751>.
31. Kato S, et al. Lite-HRNet Plus: Fast and accurate facial landmark detection. In 2023 IEEE International Conference on Image Processing (ICIP). 2023. pp. 1500–1504.
32. Ding M, et al. HR-NAS: Searching efficient high-resolution neural architectures with lightweight transformers. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA. 2021. pp. 2981–2991.
33. Kato S, Hotta K, Hatakeyama Y, Konishi Y. Lite-HRNet Plus: Fast and Accurate Facial Landmark Detection. In 2023 IEEE International Conference on Image Processing (ICIP), Kuala Lumpur, Malaysia. 2023. pp. 1500–1504. IEEE. <https://doi.org/10.1109/ICIP49359.2023.10221969>.
34. Cai H, Zhu L, Han S. Proxylessnas: Direct neural architecture search on target task and hardware. arXiv preprint arXiv:1812.00332. 2018.
35. Cai H, Gan C, Wang T, Zhang Z, Han S. Once-for-all: Train one network and specialize it for efficient deployment. In Int. Conf. Learn. Represent. 2019.
36. Cai H, Zhu L, Han S. Proxylessnas: Direct neural architecture search on target task and hardware. In Int Conf Learn Represent. 2018.
37. Li X, Lin C, Li C, Sun M, Wu W, Yan J, Ouyang W. Improving one-shot nas by suppressing the posterior fading. In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. 2020. pp. 13836–13845.
38. Liu J, Li C, Liang F, Lin C, Sun M, Yan J, et al. Inception convolution with efficient dilation search. arXiv preprint arXiv:2012.13587. 2020.
39. Liu J, Zhou S, Wu Y, Chen K, Ouyang W, Xu D. Block proposal neural architecture search. IEEE Trans Image Process. 2020;30:15–25.
40. Liao P, Jin Y, Du W. EMT-NAS: Transferring architectural knowledge between tasks from different datasets. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. pp. 3643–3653.
41. Lu S, Hu Y, Yang L, et al. Pa&da: jointly sampling path and data for consistent NAS. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. pp. 11940–11949.
42. Yan B, Peng H, Wu K, Wang D, Fu J, Lu H. LightTrack: finding lightweight neural networks for object tracking via one-shot architecture search. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA. 2021. pp. 15175–15184.
43. Ma N, Zhang X, Zheng HT, Sun J. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer Vision – ECCV 2018. Cham: Springer; 2018.
44. Zhao Z, Dong M. Channel-spatial dynamic convolution: An exquisite omni-dimensional dynamic convolution. In 2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China. 2023. pp. 1707–1711. <https://doi.org/10.1109/ICSP58490.2023.10248781>.
45. Lin T-Y, et al. Microsoft COCO: common objects in context. In ECCV, Springer; 2014. pp. 740–755.
46. Xiao B, et al. Simple baselines for human pose estimation and tracking. In ECCV. 2018. pp. 466–481. https://doi.org/10.1007/978-3-030-01231-1_29.

47. Sun K, et al. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. pp. 5693–5703.
48. Wang J, et al. Deep high-resolution representation learning for visual recognition. Trans. PAMI. 2020.
49. Li J, et al. Simple pose: rethinking and improving a bottom-up approach for multi-person pose estimation. Proc AAAI Conf Artif Intell. 2020;34:11354–61. <https://doi.org/10.1609/aaai.v34i07.6797>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.