Towards Simple and Accurate Human Pose Estimation with Stair Network

Chenru Jiang, Kaizhu Huang*, Shufei Zhang, Xinheng Wang, Jimin Xiao, Zhenxing Niu and Amir Hussain

***Abstract*—In this paper, we focus on tackling the precise keypoint coordinates regression task. Most existing approaches adopt complicated networks with a large number of parameters, leading to a heavy model with poor cost-effectiveness in practice. To overcome this limitation, we develop a small yet discrimicative model called STair Network, which can be simply stacked towards an accurate multi-stage pose estimation system. Specifically, to reduce computational cost, STair Network is composed of novel basic feature extraction blocks which focus on promoting feature diversity and obtaining rich local representations with fewer parameters, enabling a satisfactory balance on efficiency and performance. To further improve the performance, we introduce two mechanisms with negligible computational cost, focusing on feature fusion and replenish. We demonstrate the effectiveness of the STair Network on two standard datasets, e.g., 1-stage STair Network achieves a higher accuracy than HRNet by 5.5% on COCO test dataset with 80% fewer parameters and 68% fewer GFLOPs.** [1]

***Index Terms*—Stair Network, Human Pose Estimation, Feature Diversity.**

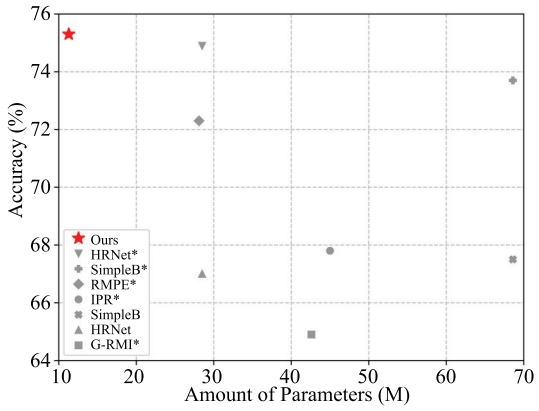


Fig. 1. Performance vs. parameter number of various methods on COCO test data with 384×288 input size. Method in red is 3-stage STair Network, and * means the method adopts pre-training.

I. Introduction

HUMAN pose estimation is one fundamental yet challenging task to estimate precise human joint coordinates (eyes, ears, shoulders, elbows, wrists, knees, etc.). It is one essential component for various high-level visual understanding tasks such as human action recognition [1]–[4], video surveillance [5], and tracking [6], [7]. In recent years, there have been significant advances from single pose estimation [8]–[14] to multiple pose estimation [15]–[20]. These methods can be categorized into bottom-up [15], [17], [19]–[22] and top-down [16], [18], [23]–[27] methods. Top-down methods are gaining more popularity due to their higher accuracy.

In existing top-down methods, multi-stage structures [13], [28], [29] and knowledge distillation [29], [30] are commonly adopted for pose estimation. However, two main drawbacks limit both the efficiency and the accuracy. One defect is that single receptive field for local features is commonly adopted in existing methods. As such, these local features may not be sufficient. To compensate this, current models usually prefer deep and complicated network structures in order to attain good performance. Figure 1 shows a number of existing popular methods which commonly adopt a heavy model with a large number of parameters. Although multi-scale module structure [12], [31] is commonly designed to aggregate information, the feature diversity is still coarse and insufficient for regression tasks. The top part of Figure 2 illustrates a scenario that the single receptive field (the smallest red rectangles) is deficient to distinguish background or different torsos. For knowledge distillation, complicated teacher networks and iterative training process are indispensable. The other defect is that the information loss is inevitable in these structures. We observe that, for localization tasks, the position error will be accumulated and enlarged after the iterative down/upsampling process. In addition, these methods seldom consider the important high-frequency texture representations [32].

To reduce computational cost and still achieve superior performance, we design a novel basic feature extraction block, STair Cell (STC), to simultaneously pursue advantageous feature diversity and high efficiency. In each block, multiple receptive fields structure is introduced to promote local feature diversity and aggregate rich local representations, enabling the block to obtain stronger discriminative capability on multi-scale keypoints or background (as illustrated multiple red rectangles in the top part of Figure 2). Meanwhile, a lightweight

Chenru Jiang is with the Department of Computer Science, University of Liverpool, Liverpool L69 7ZX, U.K. (Email: chenru.jiang@liverpool.ac.uk)

* Corresponding Author: Kaizhu Huang is with Data Science Research Center, Duke Kunshan University, No. 8 Duke Avenue, Kunshan, 215316, China. (Email: kaizhu.huang@dukekunshan.edu.cn)

Shufei Zhang is with the Department of Computer Science, University of Liverpool, Liverpool L69 7ZX, U.K. (Email: shufei.zhang@liverpool.ac.uk)

Xinheng Wang is with the Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China. (Email: xinheng.wang@xjtlu.edu.cn)

Jimin Xiao is with the Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China. (Email: jimin.xiao@xjtlu.edu.cn)

Zhenxing Niu is with School of Electronic Engineering, Xidian University, Xian, 710000, China. (Email: zhenxingniu@gmail.com)

Amir Hussain is with School of Computing, Edinburgh Napier University, Edinburgh, EH11 4BN, UK. (Email: A.Hussain@napier.ac.uk)

[1]The code is released at https://github.com/ssr0512

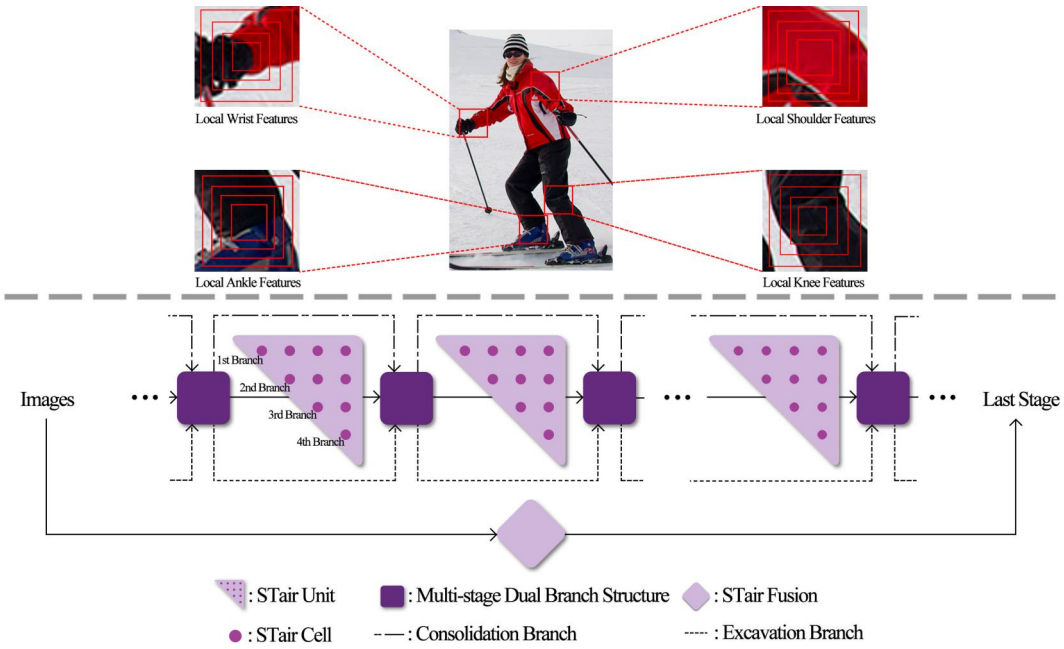


Fig. 2. The upper figure part illustrates the local features of multi-scale keypoints, where the four red boxes on local features are the multiple receptive fields of STair Cell. The lower figure part demonstrates the STair Network structure. The STair Units can be stacked in order to achieve even better accuracy

context attention is embedded to enlarge the features discrimination. Figure 3 visualizes the correlations within STC where multiple receptive fields focus on aggregating different local features and bring rich information difference. For efficiency, the block channel number is gradually halved and depthwise separable mechanism is adopted to attain lower computational cost.

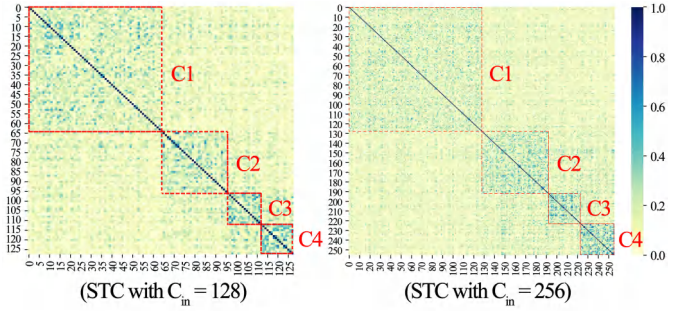


Fig. 3. Branch correlation of STC. Deeper color means stronger correlation. C1, C2, C3 and C4 are corresponding to the four branches of STC in Figure 4. Better viewed in color.

We encapsulate STCs to build a simple multi-branch module termed STair Unit (STU), which is illustrated as triangles in Figure 2. We propose to tackle the information loss problem from two aspects. Within each unit, down/upsampling is not taken in the first branch such that the network can consistently keep the high resolution feature maps to alleviate information loss. Outside each unit, we design a lightweight structure to focus on feature re-usage and re-exploitation among multiple units. Illustrated as squares in Figure 2, such structure can be readily attached after each unit for enhancing feature utilization. Meanwhile, we try to reserve high-frequency texture representations (illustrated as diamonds in Figure 2), to supply these important features at the last unit for precise localization.

STair Units can be simply stacked to form a multi-stage network for the coarse-to-fine estimation. Significantly different from knowledge distillation, by enhancing the feature extraction block while alleviating the information loss problem, the large teacher network is not needed in our network. In addition, our model can be readily combined with other methods to tackle several challenging scenarios. In a nutshell, our contributions are three-fold:

1) To attain a low computational cost and accuracy framework, we propose a lightweight yet effective basic feature extraction block to focus on extracting more diverse local representations, enabling better capability to localize different keypoints with less parameters.
2) To reduce information loss during the training, the first branch within each STC consistently keeps high resolution feature maps. Meanwhile, two efficient mechanisms are proposed to connect STCs to enhance feature utilization and replenish.
3) STNet attains new SOTA performance with lower computational cost on standard benchmarks. Specifically, even 1-stage network can boost the accuracy by 5.5% over previous SOTA network [18] with only 20% parameters when evaluated on COCO test datasets.

II. Related Work

A. Feature Extraction Unit

Residual block is a popular basic feature extraction unit and commonly adopted in existing pose estimation methods [12], [16], [18], [26], [33], [34]. By the efficient bottleneck design with the skip connection, the residual block can be utilized to form much deeper and more complicated structures. Meanwhile, some novel methods can also assist to improve the

residual block capability. For instance, SENet [35] embeds channel attention after each block to reweight each feature map. SKNet [36] increases a convolution with the different kernel size in the block, and adopts soft attention to select useful channels. ResNeXt [37] applies group convolution [38] in order to obtain rich features from different subspace. Currently, RSN [39] proposes to equally split channels into 4 branches in one block, and increases the convolutions and inner connections to fuse features gradually. However, RSN relies heavily on complicated and advanced platforms MegDL, and the performance RSN_18 drops significantly (3.2%) when Pytorch is adopted. As demonstrated in section III-A, the basic feature extraction blocks of this paper consist of four sequential convolutions, where the channel dimension is halved and receptive field is doubled gradually to capture rich diverse local features while maintaining high efficiency.

B. Multi-stage Learning Structure

Multi-stage structure proves suitable to attain accurate localization, which has been widely used in various recent approaches [13], [18], [28], [40] for refining predictions gradually. Hourglass [12] and fast pose [29] take the same sub-module to form multi-stage networks. MSPN [40] follows the hourglass design yet doubles the channel dimension gradually after each downsampling. Among multiple stages, existing methods [29], [41], [42] focus on feature attention, and intermediate supervision [10] to adjust and supervise the sub-module learning process. Differently, as shown in section III-C, we propose a multi-stage dual branch structure which mainly focus on feature re-usage and re-exploitation among stages.

C. Low-frequency Structure Feature Fusion

In the traditional methods [16], [18], [39], multi-scale feature fusion is adopted commonly to extract low-frequency semantic information. Hourglass network [12] proposes a U-shape structure to attain multi-scale feature fusion within each stage. Later works such as cascaded pyramid network [16] execute a multi-scale fusion process between high-to-low and low-to-high structures. HRNet [18] sets up four parallel branches and aggregates branch features iteratively. Based on the HRNet structure, LitePose [43] focuses on simplifying the multi-branch fusion while adopting a larger kernel size to achieve a light-weight model on mobile platforms. Dite-HRNet [44] improves the original structure to extract multi-scale contextual information and long-range spatial dependency of different joints. HigherHRNet [22] utilizes the HRNet structure in the bottom-up learning mechanism for better scale-aware representation learning. Intuitively, current methods all work on low-frequency structure feature fusion. However, high-frequency information is another critical factor [32] for the precise localization tasks as it provides rich texture representations. Unfortunately, high-frequency information is hardly reserved when networks go deeper. Contrary to the conventional approaches, in Section III-D we develop an effective method to directly replenish abundant multi-scale high-frequency representations to the model. On the other hand, our method can be served as a basic model to obtain 2D joints for other high-level applications. For instance, [45] directly utilizes 2D joint locations to regress joints location in 3D space, [46] combine long-wavelength infrared modality and image information to handle in-bed occlusion problem for human behavioral monitoring, [47] explores an effective method which applies the pre-trained 2D pose estimation method on 3D field, and [48] combines Inertial Measurement Unit sensor and video information to handle full-body motion capture.

III. Our Approach

In this section, we first detail our basic component STair Cell. After that, we introduce the multi-scale STair Unit structure consisting of a number of STCs that maintains high resolution feature maps from end to end. Next, we will describe the Multi-stage Dual Branch Structure and STair Fusion mechanism to further enhance the network.

A. STair Cell

a) STair Convolution: For each STair Cell, we propose a sequential stair-shape atrous convolution structure to aggregate significant diverse local features for different scale keypoints through multiple receptive field sizes. In contrast to traditional convolutions, atrous convolution [49] helps to attain larger receptive fields without increasing the computational cost. The kernel sizes of atrous convolution layers are calculated by $K = k + (k-1) \times (R-1)$, where K is the equivalent kernel size, k is the practical kernel size, and R is the dilation rate.

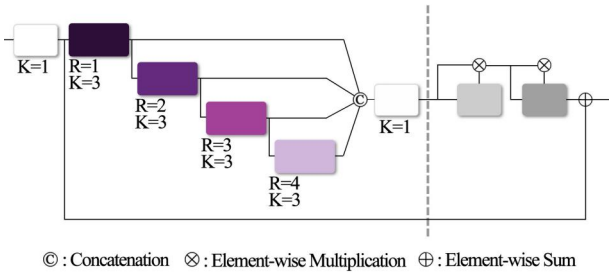


Fig. 4. STair Cell structure. Purple rectangles indicate atrous convolution processes with multiple dilation rates, gray rectangles describe the embedded context attentions.

As shown in Figure 4, one stair shape convolution structure contains four atrous convolution layers with the different dilation rates ($R = 1, 2, 3, 4$). We notice that the atrous convolution has an inherent gridding effect which leads to information inconsistency [50]. To alleviate such gridding effect, we keep the dilation rate as 1 in the top base convolution layer, make it as the same as the normal convolution. In STC, as the receptive field size is increased, the channel number is reduced in half for the efficiency purpose. Concretely, the total computational cost of four-branch structure are: $4T(T = k \times k \times C_{in} \times C_{out} \times H \times W)$, where the computational cost depends multiplication on the number of input channels C_{in}, the number of output channels C_{out}, the kernel size k, and the feature map size $H \times W$. The channel halving

strategy factorizes conventional mechanism into four parts and the amount of parameters are:

$$\frac{1}{2}T + \frac{1}{8}T + \frac{1}{32}T + \frac{1}{64}T = \frac{43}{64}T \quad (1)$$

Thus, the channel halving strategy attains $4/(\frac{43}{64}) \approx 6$ times computation reduction than the conventional structure. Table I demonstrates the computational cost comparison of channel halving strategy. More channels might provide richer diverse features but hinder structure efficiency, which is not pursued in this work. Particularly, we have investigated various structures for STC which are shown in ablation parts. However, the current structure demonstrates the best performance for aggregating diverse local features with high efficiency.

TABLE I
RESULTS OF STC WITH/WITHOUT CHANNEL HALVING STRATEGY.

Method	pretrain	Input Size	Halve	#Params	GFLOPs	AP
1-stage	N	256 × 192	✓	5.7M	2.3	72.1
1-stage	N	256 × 192	×	8.7M	3.0	73.0

Table II lists the detail parameter configurations of STC and STU. Since the inputs of each atrous convolution are based on the previous layers, the receptive fields are gradually superimposed. Thus, the fourth branch of the STC contains four receptive fields equivalently. In this regard, STC enables to extract rich diverse local representations for the network. As demonstrated in Figure 3, the correlations between different STC branches (Ci and Cj, $i \neq j$) are low, meaning that multiple branches with various receptive fields focus on aggregating different local features. Consequently, four-branch STair Units can obtain more dense diverse features with STCs, as shown in the second part of Table II.

TABLE II
PARAMETER CONFIGURATION OF STAIR CELLS AND STAIR UNITS. M MEANS THAT THE FEATURE MAP SIZE IS MAINTAINED, AND H MEANS THE FEATURE MAP SIZE IS HALVED BY DOWNSAMPLING.

STC Branch Index	Feature Map Size	Channel Number	Kernel Size
c = 1	M	16	3
c = 2	M	8	3, 5
c = 3	M	4	3, 5, 7
c = 4	M	4	3, 5, 7, 9
STU Branch Index	**Feature Map Size**	**Channel Number**	**Kernel Size**
b = 1	M	32	3, 5, 7, 9
b = 2	H	64	5, 7, 9, 11
b = 3	H	128	7, 9, 11, 13
b = 4	H	256	9, 11, 13, 15

To further reduce the parameters, we adapt the depthwise separable convolution [51] to our STC, which factorizes a standard convolution into a depthwise operation and a pointwise operation. First, each atrous convolution of STC exploits depthwise operation to apply a single filter for each input channel. Then, the pointwise operation applies a 1×1 convolution to combine the depthwise outputs. There is a significant difference on the amounts of network parameters between traditional convolution and depthwise separable convolution. Specifically, the total computational cost of a standard convolution are: T, The depthwise separable convolution splits standard convolution operation into two parts: $T/k^2 + T/C_{out}$. After decomposing convolution as a two-step process of filtering and combining, we can achieve the parameter reduction as below:

$$\frac{T/k^2 + T/C_{out}}{T} = \frac{1}{k^2} + \frac{1}{C_{out}} \quad (2)$$

After the stair-shape convolution, we concatenate four level features and pass them to the lightweight context attention part.

b) Mix Attention: For the postprocessing of STC, we consider separately for different local features so that the block attains a stronger feature diversity and fine-grained recalibration. The proposed mix attention strategy is shown in the right part of Figure 4 which aims to reinforce the extracted diverse features through both channel and spatial dimensions. For calculation efficiency, we introduce soft reduction rate to the channel dimension attention, which is dynamically enlarged as the channels of STU is increased. For spatial dimension attention, STC applies average and max pooling methods to generate only two masks for keeping low computational cost. Table III demonstrates that the mechanism is effective to enhance feature diversity with negligible computational cost. In addition, we maintain skip connection to support gradients propagation of the network. In summary, we invent a lightweight yet effective basic feature extraction block to construct our model. The multiple receptive field design of STC leads to extract more diverse local features which are vital for accurate regression. Meanwhile, we develop the channel halving strategy and mix attention mechanism to attain good balance on effectiveness and efficiency.

TABLE III
COMPARISON OF STC WITH/WITHOUT MIX ATTENTION.

Method	pretrain	Input Size	#Params	GFLOPs	Mix Attention	AP
1-stage	N	256 × 192	5.6M	2.3	×	71.5
1-stage	N	256 × 192	5.7M	2.3	✓	72.1

B. STair Unit

STU can be stacked to compose a multi-stage learning mechanism as demonstrated in Figure 2. Inspired by HRNet [18], each STU adopts a multi-branch structure to attain multi-scale feature fusion. Differently, with the superior STCs, the basic extraction block number of four branches in each STU is: (4,3,2,1), whilst that of the HRNet is: 4x(4,3,2,1). Thus, the computational cost is much less than HRNet. Meanwhile, as seen in the bottom part of Table II, we list the detail kernel size range of the units. More receptive field of STU (3-15) can obtain rich representations than HRNet (3-9). The inner structure of the units is illustrated in Figure 5 where the purple rectangles describe STCs and the gray rectangles are

the multi-scale feature fusion process. To reduce information loss, the sizes of feature maps are reduced in half but branch channels are doubled gradually from top to bottom, and the top branch consistently maintains high resolution featrue map size from end to end. To make the network to obtain rich multi-scale features for precise predictions, we design the four-stage feature fusion in one unit. In feature fusion process, each sub-branch iteratively aggregates the features which are downsampled or upsampled from other parallel sub-branches.

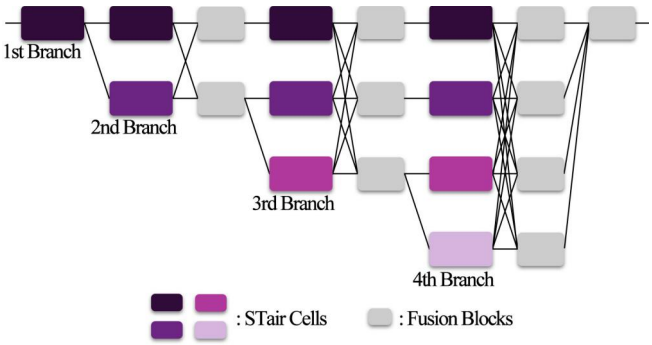


Fig. 5. Structure of each STair Unit. Purple and gray rectangles describe respectively STair Cells and multi-scale feature fusion process.

C. Multi-stage Dual Branch Structure

For feature re-usage, skip connections are widely applied between network stages [16], [34]. However, existing methods pay little attention to feature re-exploration, since the parameters grow quadratically as the densely connected path width increases linearly. In this section, we will explain the limitations of ResNet [33] which focus on feature re-usage, and DenseNet [52] which focus on feature re-exploitation first. We then present the Multi-stage Dual Branch Structure (MDBS), which enjoys the benefits from both path topologies for learning good representations.

a) Revisiting ResNet and DenseNet: Traditional convolution feed-forward networks utilize the output of L^{th} layer as the input for $(L+1)^{th}$ layer, which can be summarized as: $X_L = H_L(X_L - 1)$. (X_L is the output of L^{th} layer, and H_L are the composite operations such as batch normalization, rectified linear unit, or convolution). ResNet proposes a skip connection which bypasses the non-linear transformations with an identity function:

$$X_L = H_L(X_{L-1}) + X_{L-1}. \tag{3}$$

The main contribution of ResNet is that the gradients can propagate directly through the identity function to relieve gradient attenuation and reinforce the feature re-usage in deeper layers. However, ResNet neglects the features re-exploration and the summation process in Equation (3) may impede the information flow in the network. After that, Huang et al. [52] propose DenseNet, where the skip connections are used to concatenate the inputs to the outputs instead of adding operation in order to improve the information flow between layers. Consequently, the L_{th} layer will receive the feature maps of all preceding layers as inputs:

$$X_L = H_L(X_0||X_1||...||X_{L-1}). \tag{4}$$

$X_0||X_1||...||X_{L-1}$ refers to the feature maps concatenation ($||$) of layer $0, \ldots, (L-1)^{th}$. Since the width of the densely connected path and the cost of GPU memory linearly increase as the network goes deeper, building a deeper and wider densenet is substantially restricted.

b) Multi-stage Dual Branch Structure: MDBS combines advantages of both path learning [33], [52], [53], able to reinforce information re-usage and re-exploration among multiple STair Units. In each stage of MDBS, we feed the unit outputs into a simple 1×1 conv to generate two branches: consolidation branch and excavation branch. Similar to DenseNet, we set a constant n as the growth rate in excavation branch, which is half of the unit first branch channel number. The small number of n helps slow the increase on width of branch excavation and the GPU memory occupation. The stage inputs, consolidation branch and excavation branch of MDBS can be expressed as follows:

$$C_i = H_i(X_i^{out})[0:n] + H_{i-1}(X_{i-1}^{out})[0:n], \tag{5}$$

$$E_i = H_i(X_i^{out})[n:end]||...||H_0(X_0^{out})[n:end], \tag{6}$$

$$X_{i+1}^{in} = C_i||E_i \tag{7}$$

X_i^{out} are the MDBS outputs of stage i and X_{i+1}^{in} are the STU inputs of the stage $i+1$. H_i is the shared composite operations that the outputs are equally separated into two parts ($[0:n], [n:end]$) for consolidation and excavation branches. As exhibited in Equation (5) and Equation (6), element-wise summation is applied on consolidation branch for features re-usage, and concatenation operation is applied on excavation branch for the features re-exploration. After that, as listed in Equation (7), two branch features are concatenated together as the inputs for the next STU. As shown in in Figure 2, the method can be readily attached after each unit for enhancing feature utilization. Based on the consolidation and excavation branches design, MDBS can alleviate the position error accumulation problem without incurring obvious computational cost. As shown in Figure 2, MDBS can be readily attached after each unit for performance improvement.

D. STair Fusion

In the existing methods [16], [18], [39], multi-scale structures are adopted to focus on low-frequency semantic feature fusion. However, the quantity of high-frequency features is another critical factor for precise localization tasks [32] as the features contain rich texture representations and better discrimination ability. To this end, we present a STair Fusion (STF) mechanism to replenish multi-scale high-frequency features to the network. As shown in Figure 6, STF adopts downsampling to generate multi-scale images. After that, we apply lightweight transformation blocks at each scale to change channels for matching multiple STU branches. The transformation blocks contain just four layers which help to reserve more high-frequency texture features. The reserved high-frequency representations are then concatenated with multi-scale low-frequency features of STU respectively. For the

TABLE IV
COMPARISON RESULTS ON THE COCO TEST-DEV DATASET. #PARAMS ARE TOTAL PARAMETERS OF THE NETWORKS, AND GFLOPS ARE TOTAL COMPUTATIONAL COST OF THE METHODS. * MEANS STC WITHOUT CHANNEL HALVING STRATEGY.

Method	Backbone	Input Size	Pretrain	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
ShuffleNetV2 1× [54]	ShuffleNetV2	384×288	Y	7.6M	2.87	62.9	88.5	69.4	58.9	69.3	68.9
Mask-RCNN [23]	ResNet_50_FPN	-	Y	-	-	63.1	87.3	68.7	57.8	71.4	-
G-RMI [25]	ResNet_101	353×257	Y	42.6M	57.0	64.9	85.5	71.3	62.3	70.0	69.7
HRNet	HRNet_w32	384×288	N	28.5M	16.0	67.0	85.4	74.3	64.8	73.4	78.1
IPR [55]	ResNet_101	256×256	Y	45.0M	11.0	67.8	88.2	74.8	63.9	74.0	-
MobileNetV2 1× [56]	MobileNetV2	384×288	Y	9.8M	3.33	66.8	90.0	74.0	62.6	73.3	72.3
G-RMI + extra data [25]	ResNet_101	353×257	Y	42.6M	57.0	68.5	87.1	75.5	65.8	73.3	73.3
SimpleBaseline [26]	ResNet_152	384×288	N	68.6M	35.5	69.5	90.1	77.0	66.4	75.3	75.5
Lite-HRNet [43]	Lite-HRNet-30	384×288	N	1.8M	0.7	69.7	90.7	77.5	66.9	75.0	75.4
SimpleBaseline	ResNet_50	384×288	N	34.0M	20.2	70.4	90.7	77.5	66.7	76.9	75.8
Dite-HRNet [44]	Dite-HRNet-30	384×288	N	1.8	0.7	70.6	90.8	78.2	67.4	76.1	76.4
SimpleBaseline	ResNet_101	384×288	N	53.0M	27.9	71.9	91.1	79.8	68.7	78.0	77.4
CPN [16]	Resnet_Inception	384×288	Y	-	-	72.1	91.4	80.0	68.7	77.2	78.5
RMPE [57]	PyraNet [13]	320×256	-	28.1M	26.7	72.3	89.2	79.1	68.0	78.6	-
CFN [58]	-	-	Y	-	-	72.6	86.1	69.7	78.3	64.1	-
CPN (ensemble)	ResNet_Inception	384×288	Y	-	-	73.0	91.7	80.9	69.5	78.1	79.0
SimpleBaseline	ResNet_152	384×288	Y	68.6M	35.6	73.7	91.9	81.1	70.3	80.0	79.0
HRNet	HRNet_w32	384×288	Y	28.5M	16.0	74.9	92.5	82.8	71.3	80.9	80.1
HRNet	HRNet_w48	384×288	Y	63.6M	32.9	75.5	92.5	83.3	71.9	81.5	80.5
TokenPose [59]	L/D24	384×288	Y	29.8M	22.1	75.9	92.3	83.4	72.2	82.1	80.8
1-stage	STNet	256×192	N	5.7M	2.3	71.4	91.0	79.0	68.1	77.2	76.8
1-stage	STNet	384×288	N	5.7M	5.2	72.5	91.0	79.7	68.9	78.5	77.8
2-stage	STNet	256×192	N	8.5M	2.9	73.4	91.7	81.1	70.2	78.9	78.7
2-stage	STNet	384×288	N	8.5M	6.5	74.8	92.0	82.1	71.4	80.7	80.0
3-stage	STNet	256×192	N	11.3M	3.5	74.1	91.8	81.8	71.0	79.7	79.4
3-stage	STNet	384×288	N	11.3M	7.8	75.3	92.1	82.7	71.8	81.2	80.4
3-stage*	STNet	384×288	N	20.3M	12.6	75.9	92.3	83.4	72.5	81.8	81.1

fusion stage, we take upsampling and deconvolution methods to gradually enlarge small-scale concatenated feature maps. We then sum four-scale feature maps to attain multi-scale fusion. For efficiency purpose, we apply a reduction factors on the transformation blocks, which is similar to STC context attention. Significantly, we attempt to modify this mechanism into a multi-stage design. The multi-stage design combines high-frequency features with low-frequency features after each STU, where the high-frequency information may still be lost in the next unit, and the results of single STF is better. Therefore, we adopt the single fusion at the end of the network. In a nut shell, STF helps to preserve and replenish essential high-frequency features in the last stage of STNet to tackle the information loss dilemma and refine prediction results further. More experiments are shown in Section IV-C for effectiveness verification.

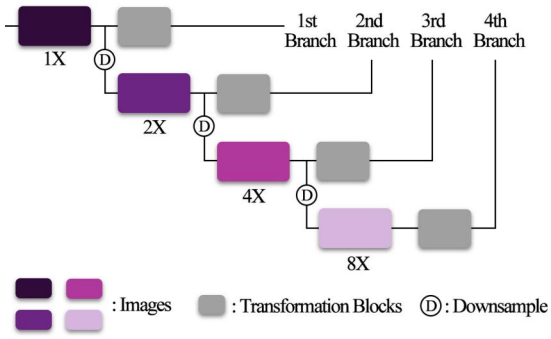


Fig. 6. STair Fusion structure. Purple rectangles are multi-scale original images, and gray rectangles are transformation blocks for matching channel with STU multiple branches.

IV. EXPERIMENTS

STair Network is evaluated on two standard human pose estimation datasets, i.e. the COCO keypoint-2017 dataset [60] and MPII keypoint [61] dataset. We follow the process of [24] to decode the predicted heatmaps. The Adam optimizer [62] is used in training, and we develop STNet on 4 NVIDIA 2080 Ti GPUs.

A. Results on COCO

This dataset contains over 200,000 images and 250,000 human instances labeled with 17 keypoints. The evaluation metric of the dataset is based on *Object Keypoint Similarity (OKS)*. Standard Average Precision (AP) and Average Recall (AR) are used as the metric. AP and AR denotes the mean value of 10 OKS numbers (OKS = 0.50, 0.55, … , 0.95), AP^{50} denotes AP at OKS = 0.5, AP^{M} denotes the AP for medium objects. For this dataset, we follow HRNet and adopt the same human instance detector to provide human bounding boxes for both the validation and test sets. Table IV lists the results of STNet and other competitive methods on the COCO test-dev dataset.The trunk branch channel number of STNet is 32.

As observed, without pre-training, our proposed method achieves very encouraging 75.3% AP score on the 3-stage

TABLE V
COMPARISON RESULTS ON THE COCO VALIDATION SET. OHKM MEANS ONLINE HARD KEYPOINTS MINING [16]. PRETAIN MEANS THE METHOD IS/ISN'T PRETRAINED ON THE IMAGENET CLASSIFICATION TASK. * MEANS STC WITHOUT CHANNEL HALVING STRATEGY.

Method	Backbone	Input Size	Pretrain	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
ShuffleNetV2 1× [54]	ShuffleNetV2	384×288	Y	7.6M	2.87	63.6	86.5	70.5	59.5	70.7	69.7
MobileNetV2 1× [56]	MobileNetV2	384×288	Y	9.8M	3.33	67.3	87.9	74.3	62.8	74.7	72.9
Hourglass [12]	8_Hourglass	256×192	N	25.1M	14.3	66.9	-	-	-	-	-
CPN [16]	ResNet_50	256×192	Y	27.0M	6.2	68.6	-	-	-	-	-
HRNet	HRNet_w32	384×288	N	28.5M	16.0	69.0	84.7	75.8	66.2	77.4	79.0
CPN+OHKM	ResNet_50	256×192	Y	27.0M	6.2	69.4	-	-	-	-	-
SimpleBaseline [26]	ResNet_152	384×288	N	68.6M	35.5	70.2	88.2	77.3	66.8	77.1	76.1
Lite-HRNet [43]	Lite-HRNet-30	384×288	N	1.8M	0.7	70.4	88.7	77.7	67.5	76.3	76.2
SimpleBaseline	ResNet_50	256×192	Y	34.0M	8.9	70.4	88.6	78.3	67.1	77.2	76.3
SimpleBaseline	ResNet_101	256×192	Y	53.0M	27.9	71.4	89.3	79.3	68.1	78.1	77.1
Dite-HRNet [44]	Dite-HRNet-30	384×288	N	1.8M	0.7	71.5	88.9	78.2	68.2	77.7	77.2
HRNet	HRNet_w32	256×192	N	28.5M	7.1	72.1	89.5	78.6	69.5	78.0	78.6
SimpleBaseline	ResNet_152	256×192	Y	68.6M	15.7	72.0	89.3	79.8	68.7	78.9	77.8
HRNet [18]	HRNet_w32	256×192	Y	28.5M	7.1	74.4	90.5	81.9	70.8	81.0	79.8
HRNet	HRNet_w32	384×288	Y	28.5M	16.0	75.8	90.6	82.7	71.9	82.8	81.0
HRNet	HRNet_w48	384×288	Y	63.6M	32.9	76.3	90.8	82.9	72.3	83.4	81.2
1-stage	STNet	256×192	N	5.7M	2.3	72.1	89.1	79.2	68.7	78.7	77.6
1-stage	STNet	384×288	N	5.7M	5.2	73.2	89.1	80.0	69.4	80.1	78.6
2-stage	STNet	256×192	N	8.5M	2.9	73.9	89.7	80.8	70.4	80.6	79.1
2-stage	STNet	384×288	N	8.5M	6.5	75.6	90.0	81.8	71.8	82.5	80.6
3-stage	STNet	256×192	N	11.3M	3.5	74.8	89.8	81.5	71.5	81.3	79.9
3-stage	STNet	384×288	N	11.3M	7.8	76.2	90.4	82.4	72.5	82.9	81.2
3-stage*	STNet	384×288	N	20.3M	12.6	76.8	90.7	83.3	73.3	83.4	81.8

backbone with the 384×288 input size. This result is higher than all the comparison models that are even pre-trained. Specifically, the parameter number of STNet is reduced by over 60% (11.3M) and the GFLOPs are reduced by over 50% (7.8) than those of HRNet_32. Compared with the SimpleBaseline (with ResNet_152 as the backbone), the gain of 3-stage network is 1.6% with the 384×288 input size; moreover, the parameters of STNet are reduced by over 83%, and the reduction on GFLOPs is nearly 80%. It is also noted that, the computational cost of STNet is the least among these popular methods. In contrast to the existing methods, the least reduction on parameters and GFLOPs achieved by 3-stage also attains 60% and 30% respectively. Significantly, with the same 384×288 input size, 1-stage STNet obtains 5.5% improvement with a 80% drop on parameters and 68% drop on GFLOPs when compared with no-pretrained HRNet_32. Furthermore, STNet can achieve higher accuracy (75.9%) once the channel halving strategy is abandoned. Compared with several current small models [43], [44], [54], [56], even 1-stage STNet can lead to the best performance, with a very slight GFLOPs increase. In addition, the performance of STNet can be flexibly enhanced by the stage extension, enabling STNet to obtain a good balance on effectiveness and efficiency on different challenging scenarios. We further report the comparisons results on the COCO validation set in Table V, which also validates the advantages of our proposed model.

B. Results on MPII

This dataset contains 25,000 images with 40,000 human instances labeled with 16 keypoints. The evaluation metric of the dataset is the *Head-normalized Probability of Correct Keypoint (PCKh)* score. For MPII evaluation, we adopt the official testing strategy to use the provided human bounding boxes to estimate joints. We follow the six-scale testing procedure in [13], [41], [63], and the PCKh@0.5 results are reported in Table VI. As the parameters and layer numbers of three SimpleBaseline structures (ResNet_50,101,152) increase gradually, their network capability becomes stronger. The results however show that stacking single receptive field layers tends overfitting without time-consuming pre-training. Table VI shows the comparison results on MPII validation dataset with PCKh@0.5. As observed, STNet outperforms HRNet and SimpleBaseline with much fewer parameters and GFLOPs. Compared to SimpleBaseline (ResNet_152), the gains of three different STNet structures are 3.4%, 4.3% and 4.9% with 92%, 88% and 84% parameters drop. Additionally, we report the comparison results on MPII test set in Table VII. Again, our STNet shows the best performance, which also verifies the superiority of our model.

C. Ablation Analysis

a) STC Structure Exploration: For the STC design, we explore a number of structures as illustrated in Figure 7. For efficiency comparison, we separately list the computational cost of the different STC structures in Equation (8). Apparently, structure E (the final version of STC) attains the minimal computational cost, which is one of the critical factors we pursue in this work. Meanwhile, based on the results of Table VII, we find that the conventional branch separation manner (structure A, B, C, and D) independently extracts features from different receptive fields, which tends to cause semantic features incoherence. Meanwhile, we observe that channel number average separation manner (structure B, C, and D) is not preferable to extract local features.

TABLE VI
COMPARISON RESULTS ON THE MPII VALIDATION SET.

Method	Backbone	Input Size	Pretrain	#Param	GFLOPs	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
SimpleBaseline [26]	ResNet_50	256×256	N	34.0M	12.0	96.5	94.8	87.7	82.2	87.9	82.3	77.1	87.6
SimpleBaseline	ResNet_101	256×256	N	53.0M	16.5	96.1	94.3	86.1	80.3	87.2	81.4	76.2	86.6
SimpleBaseline	ResNet_152	256×256	N	68.6M	21.0	95.7	93.7	84.9	77.9	85.5	79.2	74.3	85.2
HRNet [18]	HRNet_w32	256×256	N	28.5M	9.5	95.6	94.2	88.1	83.7	88.1	83.7	79.3	88.1
HRNet	HRNet_w48	256×256	N	63.6M	19.5	96.0	94.6	88.8	84.1	87.4	83.6	80.4	88.4
TokenPose [59]	L/D24	256×256	Y	28.1M	-	97.1	95.9	90.4	86.0	89.3	87.1	82.5	90.2
1-stage	STNet	256×256	N	5.7M	3.1	96.7	94.9	89.0	83.6	88.4	84.4	79.7	88.6
2-stage	STNet	256×256	N	8.5M	3.9	96.9	95.6	89.8	84.8	89.2	84.7	81.6	89.5
3-stage	STNet	256×256	N	11.3M	4.6	97.0	95.9	90.2	85.2	89.6	86.4	83.1	90.1

TABLE VII
COMPARISON RESULTS ON THE MPII TEST SET.

Method	Backbone	Input Size	Pretrain	#Param	GFLOPs	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
SimpleBaseline	ResNet_152	256×256	Y	68.6M	21.0	98.5	96.6	91.9	87.6	91.1	88.1	84.1	91.5
HRNet	HRNet_w32	256×256	Y	28.5M	9.5	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3
[3]	-	256×256	-	-	-	99.4	94.8	90.1	85.3	94.7	93.0	91.7	92.7
1-stage	STNet	256×256	N	5.7M	3.1	98.3	95.3	90.9	87.5	92.1	89.1	85.3	91.2
2-stage	STNet	256×256	N	8.5M	3.9	98.9	96.0	91.0	88.9	92.8	90.2	88.6	92.3
3-stage	STNet	256×256	N	11.3M	4.6	99.3	96.8	91.5	89.2	93.9	91.2	90.1	93.1

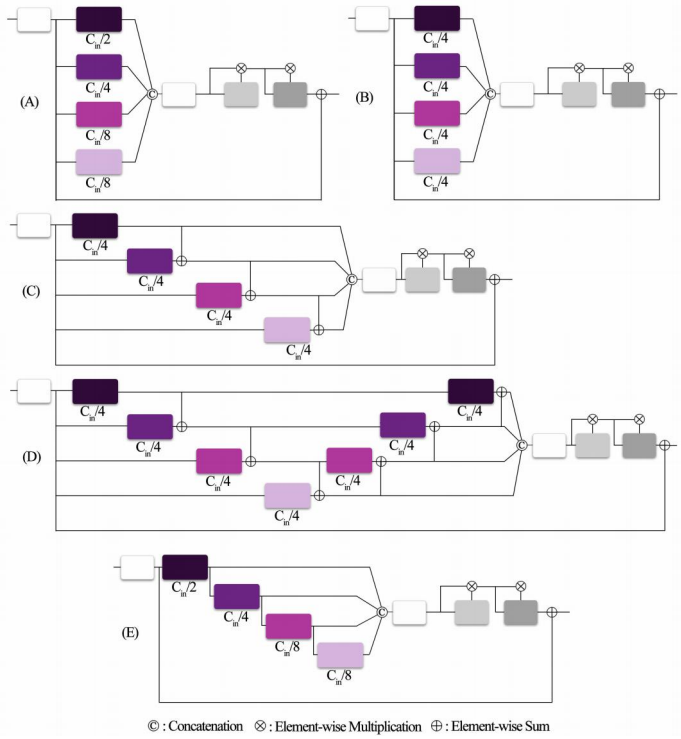


Fig. 7. Different structures investigated in our paper. Structure E is adopted in this work.

TABLE VIII
COMPARISON RESULTS OF MULTIPLE STC STRUCTURES WITH 1-STAGE NETWORK. STRUCTURE E IS ADOPTED IN THIS WORK.

Structure	pretrain	Input Size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
(A)	N	256 × 192	71.5	88.9	78.5	68.0	78.3	77.0
(B)	N	256 × 192	70.6	88.7	77.9	67.0	77.5	76.3
(C)	N	256 × 192	69.7	88.3	76.8	66.4	76.3	75.4
(D)	N	256 × 192	71.5	88.7	78.8	67.8	78.4	77.0
(E)	N	256 × 192	**72.1**	**89.1**	**79.2**	**68.7**	**78.7**	**77.6**

The goal of the bigger receptive field branch is to enhance model discriminative capability on local areas. Thus, more channel numbers (bigger weights) should be assigned to the smaller receptive field branches, which is beneficial for the precise regression process. In addition, the branch separation manner with the sum operation (structure C and D) may cause undesired neutralization on semantic features, which is harmful to final regression precision. By contrast, the stair design of this work (Figure 7 (E)) tackles these problems to coherently aggregate rich local features and attain the best performance with limited parameters and GFLOPs

$$
\begin{aligned}
A: \frac{1}{2}T + \frac{1}{4}T + \frac{1}{8}T + \frac{1}{8}T = T, \\
B: \frac{1}{4}T + \frac{1}{4}T + \frac{1}{4}T + \frac{1}{4}T = T, \\
C: \frac{1}{4}T + \frac{1}{4}T + \frac{1}{4}T + \frac{1}{4}T = T, \\
D: \frac{1}{4}T + \frac{1}{4}T + \frac{1}{4}T + \frac{1}{4}T + \frac{1}{4}T + \frac{1}{4}T + \frac{1}{4}T = \frac{112}{64}T, \\
E: \frac{1}{2}T + \frac{1}{8}T + \frac{1}{16}T + \frac{1}{64}T = \frac{43}{64}T.
\end{aligned}
\tag{8}
$$

b) Ablation Analysis of STC: STC is proposed to focus on extracting multi-scale local representations through aggregating the outputs from four atrous convolution layers with different receptive fields. In this section, we take a closer comparison on STC to evaluate network performance with

the different branch numbers. Table IX shows the results on COCO validation dataset with the 256×192 input size. As demonstrated in Table IX, the performance gain is obviously in different stage networks, and the gain of accuracy becomes slight as we increase branch number more than 4. For 6-branch design, the channel number of the largest kernel size convolution is only 1 which brings a marginal increase. Figure 8 shows the 1-stage STNet inference speed of the network with different branches on COCO validation dataset where the inference speed gradually decreases without obvious accuracy improvement. As such, 4-branch design is adopted in this work.

TABLE IX
ABLATION ANALYSIS OF STC WITH DIFFERENT BRANCH NUMBERS. K MEANS THE KERNEL SIZE.

#Branch	Method	K	AP		AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
c = 1	1-stage	3	70.6		88.4	77.5	67.3	77.2	76.4
c = 2	1-stage	3,5	71.6		89.1	78.8	68.0	78.4	77.1
c = 3	1-stage	3,5,7	71.9		88.9	78.9	68.2	78.7	77.4
c = 4	1-stage	3,5,7,9	72.1	1.5 ↑	89.1	79.2	68.7	78.7	77.6
c = 5	1-stage	3,5,7,9,11	72.1		89.1	79.1	68.7	78.7	77.5
c = 6	1-stage	3,5,7,9,11,13	72.2		89.3	78.8	68.9	78.6	77.6
c = 1	2-stage	3	71.8		88.8	79.2	68.4	78.5	77.3
c = 4	2-stage	3,5,7,9	73.9	2.1 ↑	89.7	80.8	70.4	80.6	79.1
c = 1	3-stage	3	73.2		89.3	80.3	70.0	79.8	78.6
c = 4	3-stage	3,5,7,9	74.8	1.6 ↑	89.8	81.5	71.5	81.3	79.9

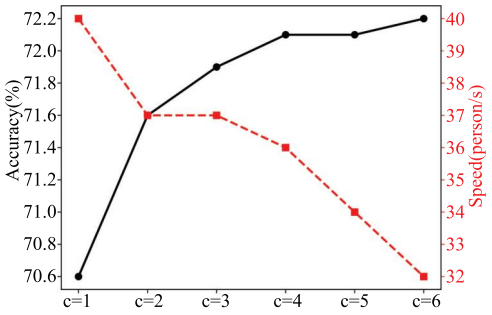


Fig. 8. Accuracy and Speed of STC vs. branch numbers.

c) Ablation Analysis of MDBS and STF: We now examine Multi-stage Dual Branch Structure, and STair Fusion on the COCO validation dataset. For simplicity, we take 1-stage and 2-stage STNets with the 256×192 input size as the illustrative examples. Table X reports the performance when we gradually apply MDBS and STF on the networks. As observed, MDBS can lead to 0.6% improvement than the baseline model on the 1-stage network. On the other hand, when STF is applied, compensation of the high-frequency information can further increase AP by 1.4%. Similarly, MDBS attains 1.7% improvement and STF further achieves 0.6% accuracy increase on the 2-stage network.

d) Ablation Analysis of STU Number: As the basic block of the multi-stage structure, STU can be simply stacked for adjusting the network capability. We perform comparison experiments on multiple architectures from one to three stages with two kinds of input sizes (256×192 and 384×288). We demonstrate the comparison results in Table XI where both the MDBS and STF mechanisms are applied in this section, but pre-training process is not adopted. Table XI demonstrates that the network performance is consistently improved with the increase of unit number.

TABLE X
ABLATION ANALYSIS OF STNET WITH DIFFERENT COMPONENTS.

Method	pretrain	Input Size	#Params	GFLOPs	MDBS	STF	AP
1-stage	N	256×192	3.35M	1.74	×	×	70.1
1-stage	N	256×192	3.37M	1.79	✓	×	70.7
1-stage	N	256×192	5.74M	2.32	✓	✓	72.1
2-stage	N	256×192	6.11M	2.24	×	×	71.6
2-stage	N	256×192	6.17M	2.36	✓	×	73.3
2-stage	N	256×192	8.53M	2.89	✓	✓	73.9

TABLE XI
PERFORMANCE OF STU WITH DIFFERENT UNIT NUMBERS.

Input Size	Pretrain	1-stage	2-stage	3-stage
256×192	N	72.1	73.9	74.8
384×288	N	73.2	75.6	76.2

V. VISUALIZATION ANALYSIS

In this part, we provide more typical comparison images to intuitively demonstrate the superiority of STNet. With the multiple receptive fields structure, STC enables to obtain stronger discriminative capability on multi-scale keypoints or background. As shown in Figure 3, different branches of STC focus on capturing rich feature diversity which is vital to cope with several challenging scenes, such as occlusion, ambiguous pose, and unusual viewpoint. In this comparison, the 3-stage STNet is applied to compare with HRNet_w48 with 384×288 input size. The first and the second columns of Figure 9 show original images and the ground truth location of different human joints. The third column displays the prediction results of HRNet_w48. For clarity, we adopt dotted red circles to denote the failure of HRNet in the third column, and use red circles to highlight the improvements of our method in the fourth column. As illustrated in row (a) and (g), STNet shows robust performance to handle serious occlusions. For some unusual viewpoint scenes like row (c), (d), and (e), there are obvious errors or even no results in HRNet predictions. In contrast, STNet obtains accurate results in these challenging scenarios. In addition, STNet demonstrates finer adjustment on some easy scenes, e.g. row (f), which benefits further accuracy improvements. Without pretraining, HRNet is unstable and tends to generate some serious errors as shown in row (b), where some prediction locations are aggregated together outside the image. However, our method enables to obtain reasonable prediction even without pretraining.

Figure 10 visualizes some successful prediction cases, and the red dotted circles demonstrate some challenging scenes. The small and vague person subjects (5th case of 1st row), serious self-occlusion (3rd and 5th cases of 2nd row), serious

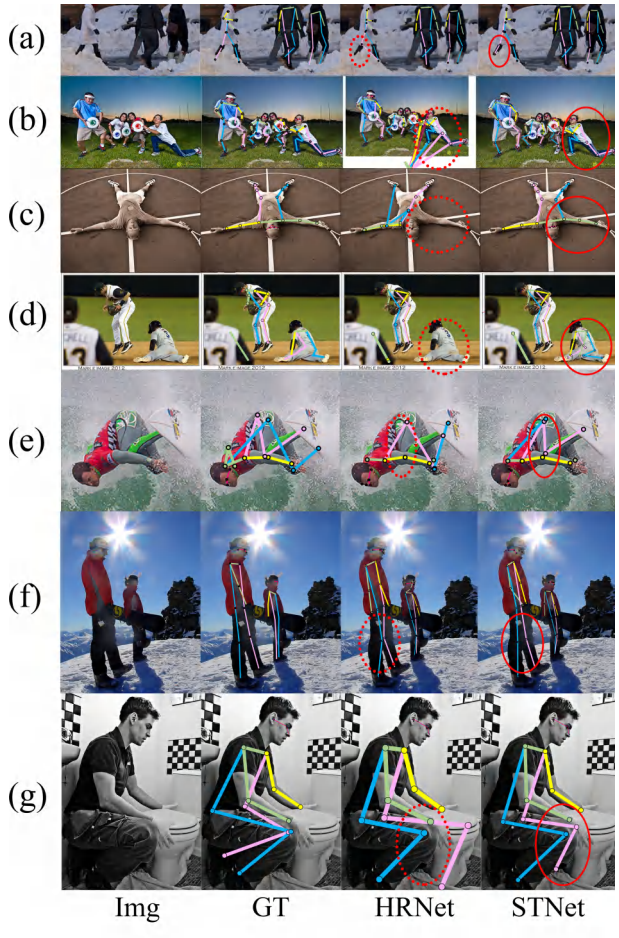


Fig. 9. Visual comparison between HRNet and STNet. The first column contains the original images and the second column details the corresponding groundtruth. The third column shows the results of HRNet and the fourth column shows the results of STNet. The dotted red circles and the red circles denote the failure predictions of HRNet and the improvements of STNet respectively.

occlusion (5th case of 3rd row) and ambiguous post (1st and 2nd cases of 4th row) can be predicted successfully due to the powerful local feature aggregation capability of STNet. In addition, Figure 11 illustrates some failure cases of our method where the red circles point out the error predictions. The ambiguous occlusion and dark illumination are still challenges which we will focus in the future works.

VI. Conclusion

In this work, we present a small yet effective multi-stage network for precise keypoint localization. To reduce the computational cost while maintaining superior performance, we propose a basic feature extraction block to focus on aggregating more diverse local representations through adopting multiple kernel sizes with fewer parameters. We alleviate the information loss problem from two aspects. Within each STair Unit, we keep high resolution feature maps to relieve feature loss. Outside the units, we develop a dual path structure to enhance feature re-usage and re-exploitation with low computational cost. Meanwhile, we design another mechanism to extract high-frequency texture representations. We test the effectiveness of our method through evaluations on standard pose estimation datasets, and the results demonstrate that the STNet's superiority with remarkable efficiency on parameters and GFLOPs.

Acknowledgements

The work was partially supported by the following: National Natural Science Foundation of China under no.61876155; Natural Science Foundation of Jiangsu Province BE2020006-4; Key Program Special Fund in XJTLU under no. KSF-T-06, KSF-E-26.

References

[1] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *CVPR*, 2013.
[2] Z. Liang, X. Wang, R. Huang, and L. Lin, "An expressive deep model for human action parsing from a single image," in *ICME*, 2014.
[3] J. Zhu, W. Zou, Z. Zhu, L. Xu, and G. Huang, "Action machine: Toward person-centric action recognition in videos," *IEEE Signal Processing Letters*, 2019.
[4] C. Wang and X. Song, "Robust head pose estimation via supervised manifold learning," *Neural Networks*, vol. 53, pp. 15–25, 2014.
[5] P. Li, J. Zhang, Z. Zhu, Y. Li, L. Jiang, and G. Huang, "State-aware re-identification feature for multi-target multi-camera tracking," in *CVPR Workshops*, 2019.
[6] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in *CVPR*, 2015.
[7] N.-G. Cho, A. L. Yuille, and S.-W. Lee, "Adaptive occlusion state estimation for human pose tracking under self-occlusions," *Pattern Recognition*, 2013.
[8] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *NIPS*, 2014.
[9] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *CVPR*, 2014.
[10] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016.
[11] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *CVPR*, 2009.
[12] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016.
[13] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *CVPR*, 2017.
[14] P. Witoonchart and P. Chongstitvatana, "Application of structured support vector machine backpropagation to a convolutional neural network for human pose estimation," *Neural Networks*, vol. 92, pp. 39–46, 2017.
[15] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.
[16] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *CVPR*, 2018.
[17] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *ECCV*, 2016.
[18] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019.
[19] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *NIPS*, 2017.
[20] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *CVPR*, 2016.
[21] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy, "Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model," in *ECCV*, 2018.
[22] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *CVPR*, 2020.
[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017.
[24] J. Huang, Z. Zhu, F. Guo, and G. Huang, "The devil is in the details: Delving into unbiased data processing for human pose estimation," in *CVPR*, 2020.

Fig. 10. Successful cases visualization. The dotted circles denote some difficult scenes: small, vague, occlusion, ambiguity.

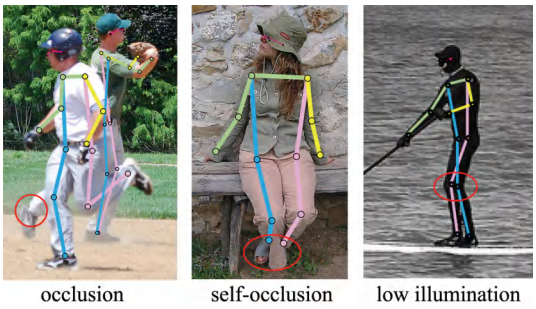


Fig. 11. Several failure cases.

[25] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in *CVPR*, 2017.
[26] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *ECCV*, 2018.
[27] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," in *CVPR*, 2020.
[28] H. Zhang, H. Ouyang, S. Liu, X. Qi, X. Shen, R. Yang, and J. Jia, "Human pose estimation with spatial contextual information," *arXiv preprint arXiv:1901.01760*, 2019.
[29] F. Zhang, X. Zhu, and M. Ye, "Fast human pose estimation," in *CVPR*, 2019.
[30] Z. Li, J. Ye, M. Song, Y. Huang, and Z. Pan, "Online knowledge distillation for efficient pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 740–11 750.
[31] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017.
[32] H. Wang, X. Wu, Z. Huang, and E. P. Xing, "High-frequency component helps explain the generalization of convolutional neural networks," in *CVPR*, 2020.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[34] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018.
[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.
[36] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *CVPR*, 2019.
[37] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *CVPR*, 2017.
[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.
[39] Y. Cai, Z. Wang, Z. Luo, B. Yin, A. Du, H. Wang, X. Zhou, E. Zhou, X. Zhang, and J. Sun, "Learning delicate local representations for multi-person pose estimation," *arXiv preprint arXiv:2003.04030*, 2020.
[40] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, and J. Sun, "Rethinking on multi-stage networks for human pose estimation," *arXiv preprint arXiv:1901.00148*, 2019.
[41] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *CVPR*, 2017.
[42] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *ECCV*, 2018.

[43] Y. Wang, M. Li, H. Cai, W.-M. Chen, and S. Han, "Lite pose: Efficient architecture design for 2d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 126–13 136.
[44] Q. Li, Z. Zhang, F. Xiao, F. Zhang, and B. Bhanu, "Dite-hrnet: Dynamic lightweight high-resolution network for human pose estimation," *arXiv preprint arXiv:2204.10762*, 2022.
[45] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2640–2649.
[46] M. Afham, U. Haputhanthri, J. Pradeepkumar, M. Anandakumar, A. De Silva, and C. U. Edussooriya, "Towards accurate cross-domain in-bed human pose estimation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2664–2668.
[47] B. Zhang, Y. Xiao, F. Xiong, C. Wu, Z. Cao, P. Liu, and J. T. Zhou, "3d human pose estimation with cross-modality training and multi-scale local refinement," *Applied Soft Computing*, vol. 122, p. 108950, 2022.
[48] T. Von Marcard, G. Pons-Moll, and B. Rosenhahn, "Human pose estimation from video and imus," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1533–1547, 2016.
[49] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *TPAMI*, 2017.
[50] Y. Fang, Y. Li, X. Tu, T. Tan, and X. Wang, "Face completion with hybrid dilated convolution," *Signal Processing: Image Communication*, 2020.
[51] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
[52] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.
[53] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," in *NIPS*, 2017.
[54] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.
[55] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *ECCV*, 2018.
[56] A. Howard, A. Zhmoginov, L.-C. Chen, M. Sandler, and M. Zhu, "Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation," 2018.
[57] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *ICCV*, 2017.
[58] S. Huang, M. Gong, and D. Tao, "A coarse-fine network for keypoint localization," in *ICCV*, 2017.
[59] Y. Li, S. Zhang, Z. Wang, S. Yang, W. Yang, S.-T. Xia, and E. Zhou, "Tokenpose: Learning keypoint tokens for human pose estimation," 2021.
[60] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.
[61] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014.
[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[63] W. Tang, P. Yu, and Y. Wu, "Deeply learned compositional models for human pose estimation," in *ECCV*, 2018.

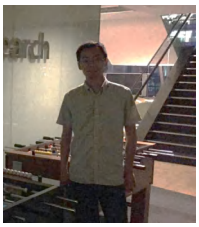
Chenru Jiang is currently a PhD student at University of Liverpool (UoL), UK. He received the B.E. degree in digital media technology from Xi'an Jiaotong-Liverpool University (XJTLU), Suzhou, China, in 2015, and the M.Sc. degree in multimedia telecommunications from UoL in 2017. His research interests include computer vision, and deep learning.

Kaizhu Huang is currently a Professor in ECE, Duke Kunshan University, China. He was a Professor as well as associate dean of research in School of Advanced Technology, Xi'an Jiaotong-Liverpool University. Prof. Huang obtained the PhD degree from Chinese University of Hong Kong (CUHK), the Master degree from Institute of Automation, Chinese Academy of Sciences (CASIA), and the Bachelor degree from Xi'an Jiaotong University. He worked in Fujitsu Research Centre, CUHK, University of Bristol, CASIA from 2004 to 2012. Prof. Huang has been working in pattern recognition, machine learning, and neural information processing. He was the recipient of 2011 Asia Pacific Neural Network Society Young Researcher Award. He received best paper or book award six times. He serves as associated editors/advisory board members in a number of journals and book series. He was invited as keynote speaker in more than 30 international conferences or workshops.

Shufei Zhang is currently a PhD student at University of Liverpool (UoL), UK. He obtained his Master degree from Informatics school of University of Edinburgh in 2015. He worked as intern in Alibaba DAMO Academy from 2017 to 2018. He has been working in machine learning, neural information processing, and pattern recognition.

Xinheng Wang received the B.E. and M.Sc. degrees in electrical engineering from Xi'an Jiaotong University, Xian, China, in 1991 and 1994, respectively, and the Ph.D. degree in electronics and computer engineering from Brunel University, Uxbridge, U.K., in 2001. He is currently a Professor with the School of Advanced Technology and the Head of Department of Mechatronics and Robotics, Xian Jiaotong-Liverpool University (XJTLU), Suzhou 215123, China. Prior to joining XJTLU, he was a professor with different universities in the UK. He has been an Investigator or Co-Investigator of nearly 30 research projects sponsored from EU, UK EPSRC, Innovate UK, China NSFC, and industry. He has authored or coauthored over 180 referred papers. He holds 15 granted patents, including 1 US, 1 Japan, 4 South Korea and 9 China patents. His current research interests include Tactile Internet, indoor positioning, Internet of Things (IoT), acoustic localization, communications and sensing, and big data analytics for intelligent services, where he has developed the world's first smart trolley with Chigoo Interactive Technology Co. Ltd. His research has led to a few commercial products in condition monitoring, wireless mesh networks, and user-centric routing and navigation for group users (www.ciptechnology.co.uk).

Jimin Xiao received the B.S. and M.E. degrees in telecommunication engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2004 and 2007, respectively, and the Ph.D. degree in electrical engineering and electronics from the University of Liverpool, Liverpool, U.K., in 2013. From 2013 to 2014, he was a Senior Researcher with the Department of Signal Processing, Tampere University of Technology, Tampere, Finland, and an External Researcher with the Nokia Research Center, Tampere. Since 2014, he has been a Faculty Member with Xi'an Jiaotong-Liverpool University, Suzhou, China. His research interests include image and video processing, computer vision, and deep learning.



Zhenxing Niu is a Associate Professor of School of Electronic Engineering at Xidian University. He received his Ph.D. degree in Electronic Engineering from Xidian University in 2012. His current research interests include computer vision, deep learning and machine learning. During 2011 and 2014, he was a research intern with Dr. Gang Hua and Dr. Qi Tian. During 2013 and 2014, he was a visiting researcher in Department of Computer Science at the University of Texas at San Antonio. He has published papers on T-PAMITIPPR, CVIU, IEEE CVPR, IEEE ICCV, and ACM Multimedia. He served as a reviewer for international journals such as TIP, TCSVT, PR, etc and international conferences such as IEEE CVPR, ICCV and ACM Multimedia.

Amir Hussain received his B.Eng (highest 1st Class Honours with distinction) and Ph.D degrees, from the University of Strathclyde, Glasgow, U.K., in 1992 and 1997, respectively. He is a member of the member of the UK Computing Research Committee (UKCRC) - the Expert Panel of the Institution of Engineering and Technology (IET) and the BCS, The Chartered Institute for IT, for computing research in the UK. He has been invited Advisor/Consultant for various international Governments and organisations, including at: Kuwait Institute for Scientific Research (KISR), Kuwait Government; and the National Centre of Big Data Cloud Computing (NCBC), Higher Education Commission, Pakistan Government. He acts as a Consultant for various global companies and is co-founder/Advisor for a number of successful spin-out/start-up companies, including SenticNet, Smart Big Data Solutions Ltd. and AiGenics. He has been appointed invited Associate Editor/Editorial Board member for a number of prestigious journals, including: the IEEE Transactions on Artificial Intelligence (AI), the IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Systems, Man, and Cybernetics: Systems, Information Fusion, AI Review, IEEE Computational Intelligence Magazine, and the IEEE Transactions on Emerging Topics in Computational Intelligence.