

Article

Lightweight Stacked Hourglass Network for Human Pose Estimation

Seung-Taek Kim and Hyo Jong Lee *

Division of Computer Science and Engineering, Jeonbuk National University, Jeonju 54896, Korea;
kis7279@jbnu.ac.kr

* Correspondence: hlee@jbnu.ac.kr

Received: 17 August 2020; Accepted: 14 September 2020; Published: 17 September 2020

Featured Application: The proposed lightweight hourglass network can be applied as an alternative to existing methods that use the hourglass model as a backbone network.

Abstract: Human pose estimation is a problem that continues to be one of the greatest challenges in the field of computer vision. While the stacked structure of an hourglass network has enabled substantial progress in human pose estimation and key-point detection areas, it is largely used as a backbone network. However, it also requires a relatively large number of parameters and high computational capacity due to the characteristics of its stacked structure. Accordingly, the present work proposes a more lightweight version of the hourglass network, which also improves the human pose estimation performance. The new hourglass network architecture utilizes several additional skip connections, which improve performance with minimal modifications while still maintaining the number of parameters in the network. Additionally, the size of the convolutional receptive field has a decisive effect in learning to detect features of the full human body. Therefore, we propose a multidilated light residual block, which expands the convolutional receptive field while also reducing the computational load. The proposed residual block is also invariant in scale when using multiple dilations. The well-known MPII and LSP human pose datasets were used to evaluate the performance using the proposed method. A variety of experiments were conducted that confirm that our method is more efficient compared to current state-of-the-art hourglass weight-reduction methods.

Keywords: pose estimation; stacked hourglass network; deep learning; convolutional receptive field

1. Introduction

Human pose estimation is a fundamental method for detecting human behavior, and it is applied in virtual cinematography using computer graphics, human behavior recognition, and building security systems. Joint position varies greatly depending on a variety of factors, such as camera angle, clothing, and context. The traditional method estimates or tracks the human pose using additional equipment, such as depth sensors. However, with the advent of convolutional neural networks (CNNs), it is possible to efficiently infer the entire spatial feature from a single image without the need for additional equipment. Accordingly, many studies on human pose estimation using convolutional neural networks are currently underway, and these have achieved great results in terms of their accuracy [1–3].

The stacked hourglass network [2] is one of the best-known methods for resolving performance problems in human pose estimation. It has a stacked structure of hourglass modules composed of residual blocks [4]. Since the hourglass network performs promisingly in resolving the human pose estimation problem, a number of studies have used it as a backbone or modified the original hourglass network to improve performance [5–10]. Ning et al. [11] developed a stacked hourglass

design and inception-resnet module with encoded external features for human pose estimation. Ke et al. [12] introduced the multiscale supervision network with a multiscale regression network using the stacked hourglass module to increase robustness in keypoint localization for complex background and occlusions. Zhang et al. [13] suggested a method to overcome information loss in a repetitive cycle of down-sampling and up-sampling in a feature map. They used a dilated convolution and skip connections applied to a stacked hourglass network to optimize performance while adding extra subnetworks and large parameter sizes.

However, the residual block used in the hourglass network consists of a relatively small kernel with a fixed size. This may not be conducive to extracting the relationship between joints in the entire human body, and performance may significantly deteriorate depending on the size of the person in the input image. Additionally, since the network is stacked, very large memory capacity and computational powers are required. In this paper, we focus on how to reduce the parameter size and achieve the best performance, while others [5–14] deploy an extra subnetwork or layers to the stacked hourglass network. The purpose of this paper is to illustrate an optimized hourglass network with minimized parameter size without sacrificing the quality of the network.

Previous studies investigating human pose estimation [2,6,15] have confirmed that the size of the convolutional receptive field is a major factor in understanding the whole human body. If the receptive field is too small, it is difficult for the network to understand the relationship between each joint. Conversely, if it is too large, information that is not relevant to pose estimation is used for calculation, leading to impaired performance. In addition, if the convolution operation using a large kernel size is used several times to increase the receptive field, the size of the network may be too large.

The goal of this study is to improve the efficiency of the stacked hourglass network [2] in human pose estimation. We propose a method for designing an efficient hourglass network that is lightweight and greatly reduces the number of network parameters. We also propose a residual block that, by expanding the convolutional receptive field of the residual block, enables performance to be maintained on a multiscale basis through multidilation. To verify the performance, we used the MPII dataset [16] and Leeds Sports Poses (LSP) dataset [17], which are widely used human pose estimation datasets, and demonstrated the effectiveness of our approach through various experiments. In summary, our main contributions are threefold:

- A proposed stacked hourglass network structure improves performance in human pose estimation with fewer parameters (Sections 3.1, 3.2.1 and 3.2.2).
- A new structured residual block, known as a multidilated light residual block, which expands the receptive field of convolution operation, effectively represents the relationship of the full human body, and supports multiscale performance through multiple dilations (Section 3.2).
- An additional skip connection in an encoder of the hourglass module that reflects the low-level features of previous stacks on a subsequent stack. This additional skip connection improves performance in pose estimation without increasing the number of parameters. (Section 3.1).

2. Related Work

2.1. Network Architecture

The structure of the original hourglass network is shown in Figure 1. The network consists of stacked hourglass modules. An hourglass module is composed of residual blocks [4], and there is a skip connection between each stack. Each module has an encoder–decoder architecture. Loss can be calculated using heat maps obtained from each stack, and the network can perform more stable learning by adjusting to repeated predictions; this process is known as intermediate supervision. The hourglass network has been used as a backbone network for many other studies that have since been shown to perform efficiently to overcome human pose estimation problems. We designed a new residual block to reduce the parameters and improve the performance of the network. In addition, we propose a new network architecture based on an additional skip connection.

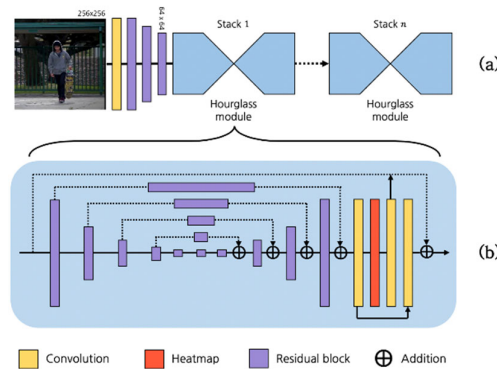


Figure 1. (a) Original hourglass network architecture and (b) hourglass module. The hourglass architecture is composed of hourglass modules stacked n times.

2.2. Residual Block

An hourglass module basically consists of residual blocks [4]. In [8,18], a hierarchical structure of residual blocks was proposed, and the convolution operation was binarized to increase the efficiency of the network. Compared to problems such as object detection, a key problem in human pose estimation is analyzing the full human body by extracting global spatial information. In [9], performance was improved by expanding the receptive field of convolution in residual blocks. Researchers in [5,9] also proposed a multiscale residual block model.

2.3. Human Pose Estimation

It has been well established that convolutional neural networks (CNNs) represent a significant step forward in recognizing spatial features from images. Accordingly, numerous studies have been conducted that use CNNs in human pose estimation where spatial information recognition is critical. The researchers in [19] made the first attempt to use CNNs for human pose estimation problems and showed a dramatic improvement in performance compared to traditional computer vision methods, such as [20,21]. Initial pose estimation methods using CNNs [19,22,23] predicted the coordinates of joints using CNNs and fully connected layers (FCLs). In [3], a method using a heat map generated by a CNN was proposed; this is currently used in most human pose estimation research [1,2,8,15,24–27].

3. Proposed Method

3.1. Network Architecture

Several studies have confirmed that the encoder–decoder architecture makes the network lighter and improves performance [28–30]. The hourglass network [2] used in our study is a productive approach to solving problems in human pose estimation because the network is able to learn more complex features by stacking modules. An encoder first extracts features by reducing the image resolution, while a decoder increases the image resolution and reassembles features. In an hourglass network, the encoder function is connected to the decoder using a skip connection so that the decoder can restore features well. According to recent work [31,32], extracting features is a more important process than simply restoring them. An hourglass network is structured so that input from a previous stack ($n - 1$) is reflected in the current stack (n), in addition to the output of the previous stack ($n - 1$) via the skip connection, as shown in Figure 1. In this structure, only relatively high-level features reconstructed by the decoder are reiterated in the next stack. Since our goal is to resolve this problem while also making the network lighter, we propose a method that enhances feature extraction performance while requiring minimal modification.

Figure 2 shows the proposed hourglass module. A feature extracted by the encoder is transferred to the next stack by the simple addition of parallel skip connections, as shown in Figure 2; this does

not require a significant increase in computation. The proposed structure improves the encoder's extraction performance by transferring features to subsequent stack encoders. This structure produces better performance than the original hourglass network. An architecture with additional skip connections improves performance by increasing the performance of the encoder, even though the size of the network itself remains substantially unchanged.

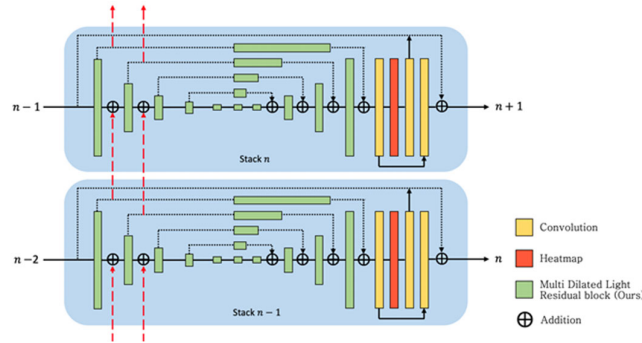


Figure 2. Proposed hourglass architecture, which utilizes an additional skip connection (red dashed arrow in the figure) from the previous stack ($n - 1$) to the current stack (n).

3.2. Residual Block Design

3.2.1. Dilated Convolution

In human pose estimation, it is important to increase the receptive field so that the network can learn to recognize the features of the full human body. However, if the kernel size is increased to widen the receptive field, the computational cost also increases. Since our goal was to design an efficient hourglass network, we constructed a residual block using dilated convolution [33]. The number of parameters in the standard convolution is shown in Equation (1), where the kernel size is K , the input channel size is C , and the output channel size is M . If the size of the input and output is the same as $H \times W$, the required computational cost is shown by Equation (2):

$$\# \text{ param} = K^2 CM \quad (1)$$

$$\text{Computational Cost} = K^2 CMHW \quad (2)$$

In the case of the dilated convolution, if the kernel size remains the same, the number of parameters and the computational cost remain the same as the standard convolution; however, the receptive field is wider, depending on the dilation size D . For dilated convolutions with a 3×3 kernel size when $D_1 = 2$, the kernel size and computational cost are the same as for the 3×3 standard convolution, but the receptive field is the same as for the 5×5 standard convolution. Additionally, as shown when $D_1 = 2$ and $D_2 = 3$ in Figure 3, dilated convolution has zero padding inside the kernel; thus, the computational cost is slightly lower than it is for standard convolution with the same kernel size.

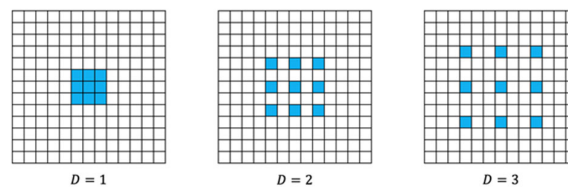


Figure 3. Standard convolution with 3×3 kernel size and dilated convolution. $D = 1$ is the same as the standard convolution. $D = 2, 3$ are calculated by placing zero padding inside the kernel, as shown in the figure.

3.2.2. Depthwise Separable Convolution

We used depthwise separable convolution, as proposed in [34], with dilated convolution for our residual block. Depthwise separable convolution performed pointwise (1×1) convolution after depthwise convolution, which was performed using an independent kernel for each channel. Figure 4 shows the concept of the depthwise separable convolution. This method shows lower performance than it does with standard convolution, but the number of parameters was significantly reduced, and the computation speed was faster. We interpolated the reduced performance due to depth-size separable convolution into a new residual block with dilated convolution.

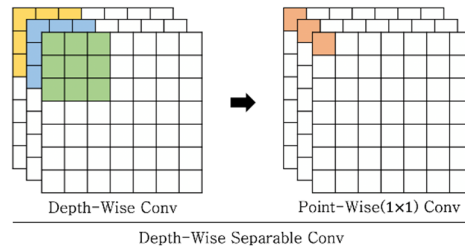
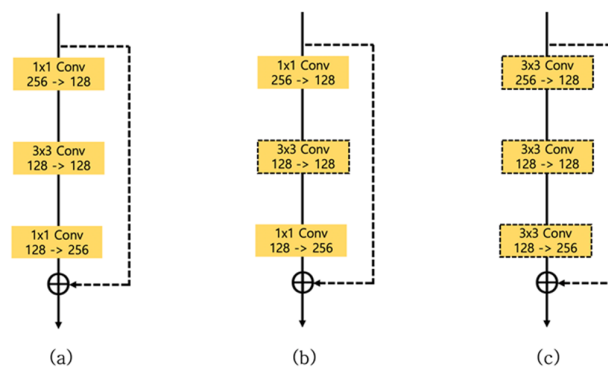


Figure 4. Depthwise separable convolution. Each channel performs convolution using an independent kernel, which is referred to as depthwise convolution. Pointwise convolution is then performed with a 1×1 kernel.

3.2.3. Proposed Multidilated Light Residual Block

The original stacked hourglass network constructed an hourglass module using a preactivated residual block, as shown in Figure 5a. The structure of the preactivated residual block is [ReLU→Batch Normalization→Convolution]; this is unlike that of a conventional residual block, which is designed as [Convolution→Batch Normalization→ReLU]. This structure is advantageous in building a deep network and improves the training speed [35]. However, residual blocks were originally designed for image classification or object detection problems (in which it is important to learn local features) where the convolutional receptive field is relatively small. Moreover, in the deep network architecture, although a bottleneck structure is used to reduce the number of parameters and computation cost, the residual block with a bottleneck structure is still large when designing a stacked multistage network, such as an hourglass network. Therefore, there is a need for a residual block with a new structure that can improve performance in human pose estimation by reducing the size of the network and expanding the receptive field.



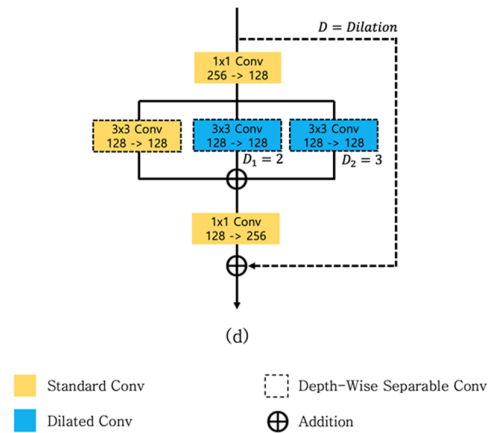


Figure 5. (a) Preactivation residual block used in the vanilla hourglass network. (b) Structure where the 3×3 convolution layer of (a) is changed to a depthwise separable convolution. (c) Structure in which the 1×1 layer of (b) is changed to 3×3 . (d) Our proposed multidilated light residual block.

In this study, to design a residual block with a new structure capable of solving the aforementioned problems, experiments were performed on residual blocks with various structures. Figure 5b shows a structure in which the middle layer of the preactivated residual block shown in Figure 5a has been changed to a depthwise separable convolution. Using this residual block, we carried out experiments to observe the effect of the depthwise separable convolution on the network size and performance in human pose estimation. Since preactivated residual blocks are bottlenecks with the first and last layers having 1×1 convolutions, it did not make sense to reduce the number of parameters by changing the layer to a depthwise separable convolution. The researchers in [34] declared that if a nonlinear function is used between a depthwise convolution and a 1×1 convolution, the performance is significantly reduced. Therefore, all of the depthwise separable convolutions used in this paper consist of a structure that does not use an activation function between the depthwise convolution and 1×1 convolution, such as [ReLU→Batch Normalization→Depthwise Convolution→ 1×1 Convolution].

To evaluate the effect of the bottleneck structure of residual blocks while using a depthwise separable convolution, we designed the new module shown in Figure 5c. Figure 5c is a modified structure of Figure 5b, where we changed the standard convolutions of the first layer [256→128, 1×1] and the last layer [128→256, 1×1] into depthwise separable convolutions of [256→128, 3×3] and [128→256, 3×3], respectively. Figure 5d is our proposed multidilated light residual block, where a residual block of a new structure was used for improving the performance and reducing the number of parameters. Table 1 below shows the detailed structure of the proposed residual block.

Table 1. Proposed multidilated light residual block structure.

	Type	Input Size (C, H, W)	Output Size (C, H, W)	Kernel Size	Dilation Scale (D)
I	INPUT	(256, H, W)	-	-	-
Layer 1					
	ReLU	(256, H, W)	(256, H, W)	-	-
	Batch Norm	(256, H, W)	(256, H, W)	-	-
	Conv	(256, H, W)	(128, H, W)	1×1	1
Layer 2					
B_1	ReLU	(128, H, W)	(128, H, W)	-	-
	Batch Norm	(128, H, W)	(128, H, W)	-	-
	Conv	(128, H, W)	(128, H, W)	3×3	1
B_2	ReLU	(128, H, W)	(128, H, W)	-	-
	Batch Norm	(128, H, W)	(128, H, W)	-	-
	Depthwise Conv	(128, H, W)	(128, H, W)	3×3	2
	Conv	(128, H, W)	(128, H, W)	3×3	1
B_3	ReLU	(128, H, W)	(128, H, W)	-	-
	Batch Norm	(128, H, W)	(128, H, W)	-	-
	Depthwise Conv	(128, H, W)	(128, H, W)	3×3	3
	Conv	(128, H, W)	(128, H, W)	3×3	1
	ADD ($B_1 + B_2 + B_3$)	(128, H, W)	(128, H, W)	-	-
Layer 3					
B_4	ReLU	(128, H, W)	(128, H, W)	-	-
	Batch Norm	(128, H, W)	(128, H, W)	-	-
	Conv	(128, H, W)	(256, H, W)	1×1	1
	ADD ($B_4 + I$)	(256, H, W)	(256, H, W)	-	-
	OUTPUT	-	(256, H, W)	-	-

In this study, the multidilated light residual block was used to greatly lighten the stack hourglass network, while multidilated convolution was used to expand the receptive field to increase the immutability of scale, resulting in improved performance in human pose estimation.

4. Experiments and Results

4.1. Dataset and Evaluation Matrix

The well-known human pose estimation datasets MPII and LSP were used to evaluate the performance of the proposed additional interstack skip connection and multidilated light residual blocks. The MPII dataset contains over 40,000 images of people with joint information, of which around 25,000 images were collected in real-world contexts. For human pose estimation, 16 coordinates for each joint were labeled for each person. In addition, we conducted experiments using the LSP and its extended training datasets [36] for objective evaluation. The LSP dataset contains 12,000 images with challenging athletic poses. In this dataset, each full body is annotated with a total of 14 joints.

To evaluate the performance of our method, we compared the performance with the state-of-the-art lightweight method for the stacked hourglass network [8] with various experiments. As an evaluation method, we used the percentage of correct key-points (PCK) on the LSP datasets and the modified PCK measure, which is the percentage of correct key-points on the head (PCKh) with the MPII dataset, as used in [32]. PCKh@0.5 uses 50% of the ground-truth head segment's length as a threshold. If the error rate is lower than the threshold value when comparing the predicted value with the ground truth, it is determined to be the correct answer.

4.2. Training Details

We followed the same training process as used for the original stacked hourglass network [2] with an input image size of 256×256 . For the data augmentation required for training, rotation ($\pm 30^\circ$), scaling (± 0.25), and flipping were performed. The model used in all experiments was written using PyTorch software [37]. We used the Adam optimizer [38] for training with a batch size of eight. The number of training epochs was 300, and the initial learning rate was 2.5×10^{-4} , which was reduced to 2.5×10^{-5} and 2.5×10^{-6} in the 150th and 220th epochs, respectively. The network was initialized by a normal distribution $\mathcal{N}(m, \sigma^2)$ with mean $m = 0$ and standard deviation $\sigma = 0.001$.

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^N \sum_{ij} \|H_n(i, j) - \hat{H}_n(i, j)\|^2 \quad (3)$$

The ground-truth heat map $H = \{H_k\}_{k=1}^K$ was generated by applying a Gaussian around k body joints, as shown in [3]. The loss $\mathcal{L}$ between the heat map $\hat{H} = \{\hat{H}_k\}_{k=1}^K$ and H predicted by the network used the mean squared error (MSE). Losses were calculated using the predicted heat maps from each stack and summed up by intermediate supervision. Figure 6a visualizes the loss in the training process, and Figure 6b visualizes the accuracy of PCKh@0.5 in the MPII validation set.

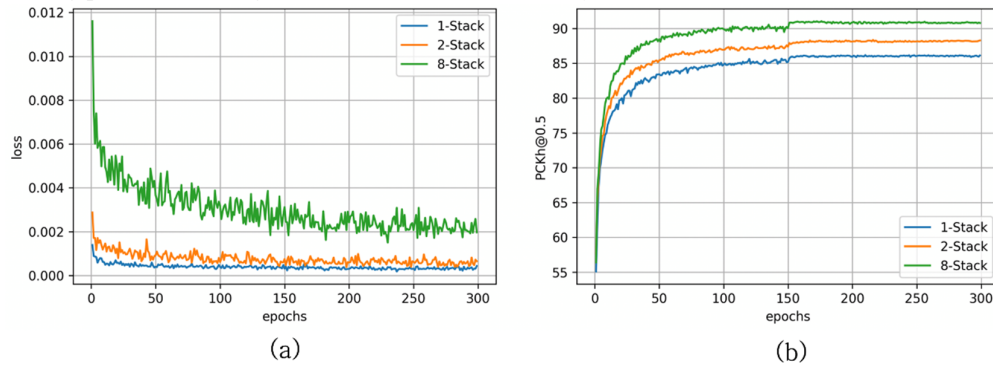


Figure 6. (a) Loss during training and (b) PCKh@0.5 on the MPII validation set.

4.3. Lightweight and Bottleneck Structure

To evaluate the network weight-reduction performance, we used depthwise separable convolution and looked at the effect of the bottleneck structure. Table 2 shows the experimental result obtained by constructing a single-stack hourglass network with each residual block in Figure 5; the number of parameters in the table represents the total number of parameters in a single hourglass network.

In general, in a problem involving localization, such as human pose estimation, performance reduction occurs when using residual blocks in a bottleneck structure that uses 1×1 convolution to reduce the size of the network [18]. However, using 1×1 convolution was inevitable in this study because the network was made lighter by using depthwise separable convolution. Therefore, in order to confirm the effect of the bottleneck structure using 1×1 convolution in this experiment, the 1×1 convolutions of the first and last layers of the original residual block (Figure 5a) used 3×3 kernels. We experimented with a residual block (Figure 5c) that increased the kernel size to 3×3 and applied the depthwise separable convolution. Although the accuracy and parameters increased slightly, our result confirmed that the 1×1 convolution had no significant effect on our experiment.

Table 2. Results for MPII validation sets in a single-stack hourglass network, depending on the type of residual block. (# Params is total number of parameters in a single-stack hourglass network.).

Residual Block	# Params	PCKh@0.5 (Mean)
Bottleneck (Original) (Figure 5a)	3.6M	86.21
Bottleneck (Original) + Depthwise Separable (Figure 5b)	1.4M	85.41
Bottleneck (3×3) + Depthwise Separable (Figure 5c)	1.6M	85.67
Multidilated Light Residual Block (Ours) (Figure 5d)	2.0M	86.12

Through this experiment, we confirmed that the multidilated light residual blocks proposed in this paper showed improved performance in terms of achieving a more lightweight network through a 56% reduction in the number of parameters. In addition, it was confirmed that the PCKh@0.5 performance was reduced by approximately 0.09, despite the large reduction in the number of parameters, as compared to the original residual block (Figure 5a). This slight reduction in accuracy was overcome by using the additional skip connection structure described in Section 3.1.

4.4. Additional Skip Connection

To confirm the effect of the additional skip connection (Section 3.1) on network performance, experiments were conducted on a dual-stack hourglass network (Table 3). When we applied only the proposed method, without using a modified residual block, we observed that the number of parameters remained the same, but the accuracy was greatly increased. The number of parameters in the dual-stack network, using both the proposed network structure (Section 3.1) and the residual block (Figure 5d), was reduced to the level of the original single-stack hourglass network. However, it was confirmed that the accuracy was similar to that of the original dual-stack hourglass network. From this experiment, it was confirmed that the proposed hourglass network using an additional skip connection showed significant results.

Table 3. Comparison of different network architectures with the MPII validation dataset. (# Params is the total number of parameters in a dual-stack hourglass network).

Network Architecture	Residual Block	# Params	PCKh@0.5 (Mean)
Hourglass (Original)	Bottleneck (Original) (Figure 5a)	6.7M	87.72
Ours	Bottleneck (Original) (Figure 5a)	6.7M	88.68
Ours	Ours (Figure 5d)	3.9M	87.60

4.5. Effect of the Dilation Scale

The dilated convolution in our residual block used zero padding equal to the dilation value D to fit the size of the input and output. Therefore, to check the effect of zero padding on the pose estimation problem according to dilation size, the dilation sizes of $D_1 = 2$ and $D_2 = 3$ and the increased sizes of $D_1 = 3$ and $D_2 = 5$ (proposed in this work) were compared (Table 4).

Table 4. Comparison of different dilation scales with the MPII validation dataset.

Method [$\times$ Stack]	PCKh@0.5 (Mean)
Ours ($D_1 = 2, D_2 = 3$) [$\times 1$]	86.12
Ours ($D_1 = 3, D_2 = 5$) [$\times 1$]	85.67
Ours ($D_1 = 2, D_2 = 3$) [$\times 2$]	87.89
Ours ($D_1 = 3, D_2 = 5$) [$\times 2$]	87.19

The receptive field of the 3×3 dilated convolution extended by $D = 2, 3$, and 5 was the same as the standard convolution using kernels of $5 \times 5, 7 \times 7$, and 11×11 , respectively. In this experiment, optimum dilation enhanced the performance of pose estimation; however, when the dilation size became too large, too much zero padding caused the network to fail to learn the spatial features, resulting in a loss of the ability to localize joints. In the human pose estimation problem, it was confirmed that zero padding due to the size of the receptive field and dilation had a significant effect on performance, while the optimum dilation size was determined to be $D_1 = 2$ and $D_2 = 3$.

4.6. Results and Analysis

To evaluate our method, we compared it with current state-of-the-art lightweight hourglass network methods. The authors of [8] proposed a new hourglass architecture using hierarchical residual blocks and evaluated the performance of network binarization in human pose estimation. Single-stack and an eight-stack networks are implemented for comparison.

As shown in Table 5, our method enhanced performance in pose estimation, despite an approximately 40% reduction in the number of parameters, as compared to state-of-the-art lightweight hourglass networks. It can be seen that the human pose estimation performance using our method was superior.

Table 5. Comparison with state-of-the-art lightweight hourglass methods with the MPII validation dataset.

Method [$\times$ Stack]	[8]-Real [$\times 1$]	[8]-Real [$\times 8$]	Ours [$\times 1$]	Ours [$\times 8$]
Head	96.8	97.4	96.3	98.1
Shoulder	93.8	96.0	94.1	96.2
Elbow	86.4	90.7	85.7	90.9
Wrist	80.3	86.2	80.4	87.2
Hip	87.0	89.6	85.6	89.8
Knee	80.4	86.1	80.3	87.3
Ankle	75.7	83.2	76.0	83.5
# Params	6M	25M	2M	14.8M
PCKh@0.5 (Mean)	85.5	89.8	86.1	90.8

We compared our results with those of existing methods on the MPII and LSP datasets. Table 6 presents the PCKh scores from different methods with the MPII dataset. Table 7 shows the PCK scores with the LSP dataset. The results confirmed that the proposed method shows improved human pose performance compared to the existing methods.

Table 6. Accuracy comparison with existing methods using the MPII validation dataset (PCKh@0.5).

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Pishchulin et al. [39]	74.3	49.0	40.8	34.1	36.5	34.4	35.2	44.1
Tompson et al. [3]	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
Carreira et al. [40]	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Tompson et al. [41]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Hu et al. [42]	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Pishchulin et al. [43]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Lifshitz et al. [44]	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Gkioxary et al. [45]	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Rafi et al. [46]	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Belagiannis et al. [24]	97.7	95.0	88.2	83.0	87.9	82.6	78.4	88.1
Insafutdinov et al. [47]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Wei et al. [15]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Bulat et al. [8]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Ours	98.1	96.2	90.9	87.2	89.8	87.3	83.5	90.8

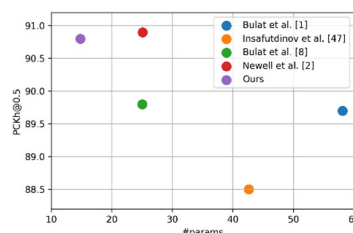
Table 7. Accuracy comparison with existing methods using the LSP validation dataset (PCK@0.2).

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
Lifshitz et al. [44]	96.8	89.0	82.7	79.1	90.9	86.0	82.5	86.7
Pishchulin et al. [43]	97.0	91.0	83.8	78.1	91.0	86.7	82.0	87.1
Insafutdinov et al. [47]	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
Wei et al. [15]	97.8	92.5	87.0	93.9	91.5	90.8	89.9	90.5
Bulat et al. [8]	97.2	82.1	88.1	85.2	92.2	91.4	88.7	90.7
Ours	98.1	93.1	89.2	86.1	92.7	92.8	90.4	91.7

Table 8 shows results from the experiment comparing the number of parameters for the MPII dataset. Figure 7 presents a schematic diagram of Table 8. This experiment confirmed that the method proposed in this paper significantly reduced the number of parameters while enhancing pose estimation performance.

Table 8. Parameter and accuracy comparison with existing methods using the MPII validation dataset (PCKh@0.5).

Method	# Params	PCKh@0.5 (Mean)
Bulat et al. [1]	58.1M	89.7
Insafutdinov et al. [47]	42.6M	88.5
Bulat et al. [8]	25.0M	89.8
Newell et al. [2]	25.1M	90.9
Ours	14.8M	90.8

**Figure 7.** Visualization of performance versus number of parameters among studies using the MPII dataset.

The number of parameters and accuracy according to the number of stacks used are summarized in Table 9. Figure 8 visualizes the pose estimation results for the MPII dataset in the eight-stack network, in which the joints in the areas covered or crossed by the body are correctly estimated. These experiments also confirmed that the proposed method represents an improvement over existing methods in terms of efficiency and performance.

Table 9. Results for MPII validation datasets by number of stacks.

Method	# Params	PCKh@0.5 (Mean)
[$\times 1$]	2.0M	86.12
[$\times 2$]	3.9M	87.89
[$\times 8$]	14.8M	90.78



Figure 8. Prediction results of the proposed method for the MPII dataset.

5. Conclusions

In this paper, we proposed a lightweight stacked hourglass network for human pose estimation. The problem with existing stacked hourglass networks is that they continuously transmit only relatively high-level features from one stack to the next. To solve this problem, we proposed a new hourglass network structure utilizing additional interstack skip connections at the front end of the encoder. These improve the performance by reflecting relatively low-level features extracted by the encoder in the next stack to allow the gradient to flow smoothly during the learning process, even in the case of a deep stack. Moreover, since the skip connection involves a simple elementwise sum operation, performance can be improved without increasing the number of parameters, which assists in constructing a lightweight network.

To maintain accuracy, a multidilated light residual block was also proposed to reduce the number of parameters in the network by about 40% compared to an existing hourglass network. Using a multidilated light residual block improves performance by expanding the receptive field using dilated convolution, significantly reducing both the number of parameters and the computational load by applying depthwise separable convolution. In this paper, a variety of experiments was conducted for objective performance evaluation of the proposed methods, and the results confirmed that our proposed methods demonstrate an effective step forward in meeting the challenges of human pose estimation.

Author Contributions: Conception and design of the proposed method: H.J.L and S.-T.K.; performance of the experiments: S.-T.K.; writing of the paper: S.-T.K.; paper review and editing: H.J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (GR2019R1D1A3A03103736).

Conflicts of Interest: The authors declare no conflict of interests.

References

1. Bulat, A.; Tzimiropoulos, G. Human pose estimation via convolutional part heatmap regression. In Proceedings of the Haptics: Science, Technology, Applications, London, UK, 4–7 July 2016; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 2016; Volume 9911, pp. 717–732.
2. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016.
3. Tompson, J.J.; Jain, A.; LeCun, Y.; Bregler, C. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2014.
4. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2016; pp. 770–778.
5. Wang, R.; Cao, Z.; Wang, X.; Liu, Z.; Zhu, X.; Wang, X. Human pose estimation with deeply learned Multi-scale compositional models. *IEEE Access* **2019**, *7*, 71158–71166, doi:10.1109/access.2019.2919154.
6. Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A.L.; Wang, X. Multi-context Attention for Human Pose Estimation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2017; pp. 5669–5678.
7. Peng, X.; Tang, Z.; Yang, F.; Feris, R.S.; Metaxas, D. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2018; pp. 2226–2234.
8. Bulat, A.; Tzimiropoulos, G. Hierarchical binary CNNs for landmark localization with limited resources. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 343–356, doi:10.1109/tpami.2018.2866051.
9. Yang, W.; Li, S.; Ouyang, W.; Li, H.; Wang, X. Learning feature pyramids for human pose estimation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); Venice, Italy, 22–29 October 2017; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2017; pp. 1290–1299.
10. Tang, W.; Wu, Y. Does learning specific features for related parts help human pose estimation? In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Long Beach, CA, USA, 15–21 June 2019; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2019; pp. 1107–1116.
11. Ning, G.; Zhang, Z.; He, Z. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE Trans. Multimed.* **2018**, *20*, 1246–1259, doi:10.1109/tmm.2017.2762010.
12. Ke, L.; Chang, M.-C.; Qi, H.; Lyu, S. Multi-scale structure-aware network for human pose estimation. In Proceedings of the Haptics: Science, Technology, Applications, Pisa, Italy, 13–16 June 2018; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 2018; pp. 731–746.
13. Zhang, Y.; Liu, J.; Huang, K. Dilated hourglass networks for human pose estimation. *Chin. Autom. Congr.* **2018**, 2597–2602, doi:10.1109/cac.2018.8623582.
14. Artacho, B.; Savakis, A. UniPose: Unified human pose estimation in single images and videos. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2020; pp. 7033–7042.
15. Wei, S.-E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2016; pp. 4724–4732.
16. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2D human pose estimation: New benchmark and state of the art analysis. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2014; pp. 3686–3693.

17. Johnson, S.; Everingham, M. Clustered pose and nonlinear appearance models for human pose estimation. In Proceedings of the British Machine Vision Conference 2010, Aberystwyth, UK, August 31–September 3; British Machine Vision Association and Society for Pattern Recognition: Durham, UK, 2010.
18. Bulat, A.; Tzimiropoulos, G. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2017; pp. 3726–3734.
19. Toshev, A.; Szegedy, D. DeepPose: Human pose estimation via deep neural networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2014; pp. 1653–1660.
20. Yang, Y.; Ramanan, D. Articulated pose estimation with flexible mixtures-of-parts. In Proceedings of the CVPR 2011, Providence, RI, USA, 20–25 June 2011; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA 2011; pp. 1385–1392.
21. Ferrari, V.; Marín-Jiménez, M.J.; Zisserman, A. Progressive search space reduction for human pose estimation. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, USA, 24–26 June 2008; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA 2008; pp. 1–8.
22. Li, S.; Liu, Z.; Chan, A.B. Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. *Int. J. Comput. Vis.* **2014**, *113*, 19–36, doi:10.1007/s11263-014-0767-8.
23. Jain, A.; Tompson, J.; Andriluka, M.; Taylor, G.W.; Bregler, C. Learning human pose estimation features with convolutional networks. *arXiv* **2013**, arXiv:1312.7302.
24. Belagiannis, V.; Zisserman, A. Recurrent human pose estimation. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2017; pp. 468–475.
25. Chu, X.; Ouyang, W.; Li, H.; Wang, X. Structured feature learning for pose estimation. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2016; pp. 4715–4723.
26. Yang, W.; Ouyang, W.; Li, H.; Wang, X. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2016; pp. 3073–3082.
27. Cao, Z.; Simon, T.; Wei, S.-E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA 21–26 July 2017; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2017; pp. 1302–1310.
28. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1520–1528.
29. Badrinarayanan, V.; Badrinarayanan, V.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495, doi:10.1109/tpami.2016.2644615.
30. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: New York, NY, USA, 2015; pp. 234–241.
31. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–21 June 2019; pp. 5686–5696.
32. Xiao, B.; Wu, H.; Wei, Y. Simple Baselines for Human Pose Estimation and Tracking. In Proceedings of the Haptics: Science, Technology, Applications, Munich, Germany, 8–14 September 2018; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 2018; pp. 472–487.
33. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.

34. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. In Proceedings of the Haptics: Science, Technology, Applications, London, UK, 4–7 July 2016; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 2016; Volume 9908, pp. 630–645.
36. Johnson, S.; Everingham, M. Learning effective human pose estimation from inaccurate annotation. In Proceedings of the CVPR 2011, Providence, RI, USA, 20–25 June 2011; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2011; pp. 1465–1472.
37. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in pytorch. In Proceedings of the NeurIPS Workshop, Long Beach, CA, USA, 4–9 December 2017.
38. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
39. Pishchulin, L.; Andriluka, M.; Gehler, P.; Schiele, B. Strong appearance and expressive spatial models for human pose estimation. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3487–3494.
40. Carreira, J.; Agrawal, P.; Fragkiadaki, K.; Malik, J. Human pose estimation with iterative error feedback. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4733–4742.
41. Tompson, J.; Goroshin, R.; Jain, A.; LeCun, Y.; Bregler, C. Efficient object localization using convolutional NETWORKS. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 648–656.
42. Hu, P.; Ramanan, D. Bottom-Up and Top-Down Reasoning with Hierarchical Rectified Gaussians. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5600–5609.
43. Pishchulin, L.; Insafutdinov, E.; Tang, S.; Andres, B.; Andriluka, M.; Gehler, P.; Schiele, B. Deepcut: Joint subset partition and labeling for multi person pose estimation. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4929–4937.
44. Lifshitz, I.; Fetaya, E.; Ullman, S. Human pose estimation using deep consensus voting. In *European Conference on Computer Vision*; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 2016; Volume 9906, pp. 246–260.
45. Gkioxari, G.; Toshev, A.; Jaitly, N. Chained Predictions Using Convolutional Neural Networks. In Proceedings of the Haptics: Science, Technology, Applications, London, UK, 4–7 July 2016; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 2016; Volume 9908, pp. 728–743.
46. Rafi, U.; Leibe, B.; Gall, J.; Kostrikov, I.; Wilson, R.C.; Hancock, E.R.; Smith, W.A.P.; Pears, N.E.; Bors, A.G. An efficient convolutional network for human pose estimation. In Proceedings of the British Machine Vision Conference 2016, York, UK, 19–22 September 2016; British Machine Vision Association and Society for Pattern Recognition: Durham, UK, 2016.
47. Insafutdinov, E.; Pishchulin, L.; Andres, B.; Andriluka, M.; Schiele, B. Deepcut: A deeper, stronger, and faster multi-person pose estimation model. In Proceedings of the Haptics: Science, Technology, Applications, London, UK, 4–7 July 2016; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 2016; Volume 9910, pp. 34–50.

