An Attention-Refined Light-Weight High-Resolution Network for Macaque Monkey Pose Estimation

Sicong Liu †, Qingcheng Fan †, Shanghao Liu, Shuqin Li and Chunjiang Zhao *

Abstract: Macaque monkey is a rare substitute which plays an important role for human beings in relation to psychological and spiritual science research. It is essential for these studies to accurately estimate the pose information of macaque monkeys. Many large-scale models have achieved state-of-the-art results in pose macaque estimation. However, it is difficult to deploy when computing resources are limited. Combining the structure of high-resolution network and the design principle of light-weight network, we propose the attention-refined light-weight high-resolution network for macaque monkey pose estimation (HR-MPE). The multi-branch parallel structure is adopted to maintain high-resolution representation throughout the process. Moreover, a novel basic block is designed by a powerful transformer structure and polarized self-attention, where there is a simple structure and fewer parameters. Two attention refined blocks are added at the end of the parallel structure, which are composed of light-weight asymmetric convolutions and a triplet attention with almost no parameter, obtaining richer representation information. An unbiased data processing method is also utilized to obtain an accurate flipping result. The experiment is conducted on a macaque dataset containing more than 13,000 pictures. Our network has reached a 77.0 AP score, surpassing HRFormer with fewer parameters by 1.8 AP.

Keywords: macaque pose estimation; high-resolution network; HR-MPE; asymmetric convolution

1. Introduction

Macaque has an important position and scientific value in life science research. It is a rare substitute for human beings in the field of psychological science and spiritual science [1]. That has extensive research value in behavior and action psychology. Macaque is a very valuable experimental object and data source in the research of these subjects. The accurate acquisition of macaque pose information is the basis of related subject research [2]. The macaque’s hair is very long and grows rapidly, which makes many sensors unable to be stably worn on the macaque’s body. Furthermore, they are very active and have a wide range of movements. If the sensors are forcibly installed on them, their normal activities will be disturbed, and the data obtained will not reflect the real situation of animals, which is not conducive to subsequent research and analysis [3]. The pose estimation method based on non-wearable devices provides a reliable scheme for capturing the pose of macaques.

The deep learning method has been widely used in human pose estimation based on RGB images and has achieved excellent results [4–6]. People and monkeys are very similar in many ways. Many studies have applied the human pose estimation method to animal pose estimation and achieved good results [7–10]. Rollyn et al. [3] designed a convolutional neural network architecture to estimate the behavior of monkeys in the wild environment. The basic design framework is to add a maximum pool after four convolution layers, which can effectively prevent over-fitting while extracting features. The activation function is RELU. With the down-sampling, the number of layers of the feature map increases and
the resolution decreases. Thus, the convolution layer is a fully connected layer used to
directly obtain the normalized coordinates. In the next work, Rollyn et al. [11] used the
DeepLabCut framework to train a pose estimation network model on the macaque dataset
with 5967 manually annotated monkey images. This model can also detect the monkey
video, but it can only detect a single frame. From this perspective, it is still image detection.
Liu et al. [7] designed a new multi-frame animal pose estimation framework, Optiflex,
which is different from other previous frameworks. This framework can make full use
of the correlation information between frames in the video to analyze the animal pose.
Then, Rollyn et al. [8] created a new macaque monkey pose dataset, called MacaquePose,
with more than 13,000 photos for the study of markerless macaque pose estimation. This
dataset offers convenience and a widely accepted baseline for the subsequent research in
this area. Salvador et al. [12] used ResNet, which has achieved good performance in many
visual tasks, for monkey pose estimation based on MacaquePose. The experimental results
show that the pose estimation accuracy for monkeys has reached the same level as that for
humans. Victoria et al. [10] applied the technique of unlabeled pose estimation to reveal
how primates use vision to guide and correct their actions in real time when capturing their
prey in the wild. This demonstrates the great role of pose estimation in animal behavior.

The essence of the pose estimation problem is to obtain the keypoints of a body from
an input image. This is a challenging basic task of computer vision. The above research
works and many mainstream networks have achieved state-of-the-art performance with
large parameters and huge GFLOPs [6,13]. In many times, the computation source of the
platform is limited, and cannot run the large-scale network smoothly. When it comes to this
issue, many small-scale networks have been proposed. Pose estimation can be performed in
two ways: the first involves using backbones designed for classification tasks directly [14],
and the second considers the characteristics of pose estimation that are sensitive to spatial
information and refines the structure, which could result in better performance [15,16].

In this work, an attention-refined light-weight high-resolution network for macaque
monkey pose estimation (HR-MPE) is developed. We follow the way of the high-resolution
transformer (HRFormer) [16], utilizing the parallel three-stage structure, similar to the
high-resolution network (HRNet) [17], but re-designing basic blocks of the transformer. An
essential point has been proposed: the power of the Vision Transformer (ViT) coming from
the well-designed structure of itself [18]. Based on this critical conclusion, a new basic block
is built via polarized self-attention (PSA) [19] with a simple yet efficient structure to achieve
the purpose of designing a light-weight model. After the three-stage structure, an attention-
refined module is added to improve the performance. An unbiased data processing method
is applied to obtain a more accurate result of the image flipping operation in pre-processing.
Warm-up, Adam, and other training techniques [20] sped up the network convergence. Our
network arrived at a 77.0 AP score, surpassing many state-of-the-art light-weight models.

The main contributions of our work are as follows:

- A multi-branch parallel architecture is built to keep high-resolution representations
  throughout the whole process, which also fuses rich semantic information from
  branches with different resolutions.
- A new basic block of the backbone is designed, relying on the powerful structure
  of the transformer and the PSA that could obtain global attention with a few parameters.
- For striking a balance between the weight and the performance of our model, we
  propose an attention-refined module based on asymmetric convolutions and triplet
  attention with few parameters.

This article is structured as follows. Section 2 discusses related work; Section 3
introduces the internal details of our network; Section 4 demonstrates the dataset used
in the work and describes the experimental settings; Section 5 conducts comparative
experiments and ablation experiments to evaluate the effectiveness of the model; and
Section 6 draws conclusions.
2. Related Work

2.1. Two-Stage Paradigm

There are two main paradigms of pose estimation: bottom-up and top-down. The first one is a kind of end-to-end method, which could directly obtain keypoints of the body. However, the accuracy of it is unsatisfactory in the task. In the training part, this paradigm needs a large size of input image, bringing gigantic pressure in the computation platform, and is time-consuming [4,21]. The second is a kind of two-stage method, which obtains a bounding box of the target body and then detects keypoints of body in the box [22]. Because of the help of the bounding box, the method could achieve a better performance than the first one. Moreover, when the size of body in an image is small, keypoints could also be obtained precisely. In the wild, there are complex backgrounds and environments in pictures of the macaque monkey. The help of the bounding box is particularly important.

2.2. Vision Transformer

The transformer has achieved great success in the field of natural language processing (NLP). The self-attention mechanism in the transformer can effectively deal with the context information of words, and performs well in many NLP tasks. The transformer is also used in the field of computer vision. Thanks to the self-attention mechanism, it can extract richer image information. ViT shows good classification and image segmentation performance on large-scale datasets of ImageNet and COCO [23]. Because there are numerous calculations in the self-attention mechanism, it is not conducive to the training and reasoning of the model. To solve this problem, the Swin Transformer combines the convolution network and designs a local attention mechanism based on a flipped window, which can perform multiple visual tasks well. To achieve better results in resolution-sensitive visual tasks, HRFormer integrates the structure of HRNet and the basic block of the Swin Transformer. The fact that the transformer infrastructure is crucial to computer vision tasks, as noted by Yu et al. [18], is a key inspiration for our work.

2.3. Light-Weight Network

In many application scenarios, powerful computing support cannot be provided, so the research and development of light-weight models is particularly important. There are two main ways of pose estimation. The first way is represented by MobileNet [24] and ShuffleNet [25]. They make good use of depth-wise convolution and point-wise convolution that are classical designs of light-weight network components with high speed and the ability to maintain great results. When compared to previous models, ViT has greater advantages due to its powerful structure and well-designed self-attention block. Based on these characteristics, the model shows remarkable results on a number of widely recognized visual task datasets, such as the Swin Transformer [26] and the pyramid vision transformer (PVT) [27], also including light-weight versions. However, in these networks, as the input image is processed, the resolution of the feature map is continuously compressed, resulting in a low-resolution feature map as the final output. As a result, the estimated position of the body keypoints in the image is inaccurate.

The second way is based on HRNet, maintaining a high-resolution representation throughout the whole process, rather than reviving a high-resolution representation from a series of low-resolution representations by up-sampling. This could prevent information loss and add noise effectively [16]. Using a multi-stage architecture to obtain multiscale representations, high-resolution networks could integrate multiple features with a light-weight component, achieving both good performance and a small scale at the same time [15]. This design provides good guidance for our work.

3. Methods

3.1. Unbiased Data Processing

The difficulty of using the top-down pose estimation paradigm lies in the bias in the data processing. Specifically, numerous repeatable experiments can prove that there are
plenty of inconsistencies between the processing results obtained by common flipping strategies and the original results. At the same time, in the process of training and inference, there are statistical systematic errors in the standard encoder–decoder. These two problems are intertwined, which inevitably reduce the accuracy of pose estimation [28], because the data used for training are biased.

To solve this problem, the unbiased data processing (UDP) [28] method converts data in a continuous space, and pixels are replaced by the unit length (interval between pixels) to measure the image size. The result is consistent with the original image when flipping is performed in inference. Furthermore, a theoretically error-free encoder–decoder method is obtained using the structure of the combined classification and regression pose estimator.

To meet the need of data conversion in this issue, this method can further improve the representation ability of our network. The specific results will be shown in the ablation experiment.

3.2. High-Resolution Network

In this work, the network’s structure is based on the HRNet design and has a three-stage parallel architecture with various resolutions (shown in Figure 1) [29]. The three-stage subnetwork includes one, two, and three different resolution branches ($\frac{1}{4}, \frac{1}{8}, \frac{1}{16}$), consisting of several pre-defined modules (1, 1, 4). Each pre-defined module contains two refined transformer blocks. To integrate information from different branches, there is an information interaction unit in each module, as shown in Figure 2 [30]. Table 1 shows the specific structure of the network.

---

**Figure 1.** Illusion of the whole structure of our network.

**Figure 2.** Illusion of the module of fusion.
Table 1. The structure of HR-MPE. FFN-DW: feed-forward network with a 3 \times 3 depth-wise convolution, PSA: polarized self-attention, ARM: attention-refined module, AsyConv: asymmetric convolution, Tri-A: triplet attention, (1, 1, 4, 1): the number of modules, (2, 2, 2, 2): the number of blocks.

| Res. | Stage1 | Stage2 | Stage3 | ARM |
|------|--------|--------|--------|-----|
| \times 4 | [1 \times 1, 64; 3 \times 3, 1 \times 1, 256]_C_1 \times 2 \times 1 | [\text{PSA} \quad \text{FFN-DW}]_C_1 \times 2 \times 1 | [\text{PSA} \quad \text{FFN-DW}]_C_1 \times 2 \times 4 | [\text{AsyConv} \quad \text{Tri-A}]_C_1 \times 2 \times 1 |
| \times 8 | | [\text{PSA} \quad \text{FFN-DW}]_C_2 \times 2 \times 1 | [\text{PSA} \quad \text{FFN-DW}]_C_3 \times 2 \times 4 | |
| \times 16 | | [\text{PSA} \quad \text{FFN-DW}]_C_3 \times 2 \times 4 | | |

In the whole network, three branches use the number of channels with \( C_1, C_2, C_3 \). Because of the aim of building a light-weight network, the number of channels is set as \( (32, 64, 128) \) [16]. Obviously, the number of channels in SimpleBaseline, MobileNet etc. that use the backbones for the classification task is several times as many as ours. HR-MPE maintains a great performance and keeps a reasonable width and parameters of the network at the same time. Benefitting from the well-designed parallel architecture of multiscale representation, high-resolution representation is kept throughout the full process. While down-sampling by a large margin, SimpleBaseline [31], MobileNet [14], etc., lose plenty of spatial information. Additionally, it is challenging to enhance the corresponding expression ability by combining the shallow high-resolution image detail representation with the representation learned from the fuzzy image. The network in our design has three distinct resolution branches. Rich spatial information is maintained in the high-resolution branches, but high-level semantic information, which the low-resolution branches have, is absent. Each stage is completed with a feature fusion module that increases the semantic content of the high-resolution representation.

In the feature fusion unit (shown in Figure 2), there are three situations for information fusion: Firstly, the resolution of the input feature map is the same as that of the output feature map. Direct replication is adopted in the network without any operation, which also ensures that the whole resolution branch has a gradient transmission path. Secondly, the resolution of the input feature map is higher than that of the output feature map. The 3 \times 3 convolution with stride 2 is used as the down-sampling operation in the network, which can reduce the information loss caused by a decline in spatial resolution to a certain extent. Thirdly, the resolution of the input feature map is lower than that of the output feature map. A 1 \times 1 convolution is used to align the number of channels in the network, and then the nearest neighbor interpolation method is used for up-sampling.

3.3. Polarized Self-Attention

Inspired by the polarized light filtering in optics, glare and reflection are always generated in the transverse direction. Therefore, the idea of only allowing light to pass through the normal direction thanks to polarized light filtering can improve the contrast of photos to a certain extent. However, the dynamic range of the filtered light is limited because some light fails to pass through the filter. Therefore, the high dynamic range is used to recover the details. Based on the above simple physical principles, the polarized self-attention (PSA) mechanism is proposed. Firstly, the tensor is folded in a specific dimension to recover the details. Based on the above simple physical principles, the polarized self-attention (PSA) mechanism is proposed. Firstly, the tensor is folded in a specific dimension, and the high-resolution representation is maintained in the normal direction orthogonal to it. After normalization, a sigmoid function is used to expand the dynamic range [19].

The procedure is shown in Algorithm 1 and the structure is shown in Figure 3. Firstly, the input \( X \in R^{C \times H \times W} \) is processed by two standard 1 \times 1 convolutions, achieving \( W_q^{C/2 \times H \times W} \) and \( W_v^{C/2 \times H \times W} \), respectively. Then, the Global Pooling operation is used for \( W_q \) to obtain global information and transfer the dimension into \( R^{1 \times C/2} \). Meanwhile, \( W_v \) only needs a simple reshaping operation and outer-product between \( W_v \) and \( W_q \), which is processed by Softmax, obtaining a Filter\(^{1 \times HW} \). Then, the Filter is handled by Reshape...
and Sigmoid in sequence to obtain a new Filter\(^{1 \times H \times W}\). Finally, there is an inner-product operation between Filter and the input \(X\).

**Figure 3.** The whole architecture of polarized self-attention.

**Algorithm 1: Polarized self-attention.**

```
input : \(X \in \mathbb{R}^{C \times H \times W}\)
output: \(Y \in \mathbb{R}^{C \times H \times W}\)

// The input \(X\) is divided into \(W_q\) and \(W_v\)
1 \(W_q^{C/2 \times H \times W} \leftarrow \text{Conv}_{1 \times 1}(X);\)
2 \(W_q^{C/2 \times 1 \times 1} \leftarrow \text{GlobalPooling}(W_q);\)
3 \(W_q^{1 \times C/2} \leftarrow \text{Reshape}(W_q);\)
4 \(W_v^{C/2 \times H \times W} \leftarrow \text{Conv}_{1 \times 1}(X);\)
5 \(W_v^{C/2 \times HW} \leftarrow \text{Reshape}(W_v);\)
6 Filter\(^{1 \times HW}\) \(\leftarrow \text{Softmax}(W_q) \otimes W_v;\)
7 Filter\(^{1 \times H \times W}\) \(\leftarrow \text{Sigmoid}(\text{Reshape(Filter)});\)
8 \(Y^{C \times H \times W} \leftarrow X \circ \text{Filter}\)
```

Based on the above steps on the description and operation, \(X\) is the input tensor; \(C\) is the number of channels; and \(H\) and \(W\) are the height and width of the characteristic graph, respectively. The time complexity of this method is:

\[
\Omega(\text{Polarized Self-Attention}) = CHW
\]  

The shifted window multi-head self-attention (W-MSA) in the Swin Transformer is as follows [26]:

\[
\Omega(\text{W-MSA}) = 4hwC^2 + 2M^2hwC
\]

where \(hw\) is the number of patch, \(M\) is the window size, and \(C\) is the number of channels. Obviously, the PSA has less parameters and computation than a complex W-MSA. In the ablation experiment, pooling is used as a benchmark to verify the effectiveness of the PSA, as shown in Figure 4a.
3.4. Attention Refined Block

For the purpose of designing a light-weight network, the structure only includes three stages, one stage less than the mainstream networks with high-resolution representations, so the performance may be reduced. In the case of maintaining a few parameters, to make up for this deficiency to some extent, we design the attention-refined block and add it to the end of the parallel structure.

The specific design of the attention-refined block is shown in Figure 4b. Firstly, a $1 \times 3$ and a $3 \times 1$ asymmetric convolution are used to further extract features from the horizontal and vertical based on high-resolution representation. After each convolution, a batch normalization and a RELU activation function are added, respectively. The above components refer to a basic unit. Previously, the PSA in the basic block only extracted spatial attention without acquiring channel information. To obtain more abundant representation information, a triplet attention is added to this block to obtain the spatial attention and the channel attention. This is followed by a same basic unit and a residual link in sequence.

3.4.1. Asymmetric Convolution

Generally, the symmetric convolution block used in the network can obtain a larger receptive field, which is also accompanied by considerable parameters and computation, and it is easy to lead to over-fitting. For the design concept of the light-weight network, we use a set of $1 \times 3$ and $3 \times 1$ convolutions to replace the conventional $3 \times 3$ convolution.

Equation (3) could validate the invariance of the convolution operation:

$$X \ast K^{(1)} + X \ast K^{(2)} = X \ast (K^{(1)} \oplus K^{(2)}),$$

where $X^{C \times H \times W}$ is the input feature map and $K^{(1)}$ is the kernel of convolution with different sizes. In the case of the same shared sliding window, the feature map with the same size...
can be obtained with fewer parameters [32]. To sum up, the operation process of the asymmetric convolution is shown in Equation (4):

\[ y = \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W} K_{j}^{h \times w \times c} S^{h \times w} \times c \]

(4)

where \( S \) is the sliding window on a feature map and \( K_{j} \) is the \( j_{th} \) kernel of a filter.

### 3.4.2. Triplet Attention

The importance of cross-dimensional interactions is captured when calculating attention weights to provide rich feature representation. Cross-dimension interaction occurs through dimension transformation [33]. Channel Pooling is an operator used to concatenate Average Pooling and Max Pooling representations in the zeroth dimension. Subsequently, \( X^{2 \times H \times W} \) is processed by a \( 7 \times 7 \) convolution and a batch normalization layer (shown in Algorithm 2).

**Algorithm 2: Triplet unit.**

```plaintext
input : X ∈ \( R^{C \times H \times W} \)
output : Y ∈ \( R^{C \times H \times W} \)

1 \( X_{\text{original}} \leftarrow X \);
2 \( X^{2 \times H \times W} \leftarrow \text{ChannelPooling}(X) \);
3 \( X^{1 \times H \times W} \leftarrow \text{Conv}_{7 \times 7}(X) \);
4 \( X^{1 \times H \times W} \leftarrow \text{BatchNormalization}(X) \);
5 \( \text{Filter}^{1 \times H \times W} \leftarrow \text{Sigmoid}(X) \);
6 \( Y^{C \times H \times W} \leftarrow X_{\text{original}} \odot \text{Filter} \)
```

The whole process of triplet attention is shown in Algorithm 3 and Figure 5. There are three triplet units in triplet attention, and each unit has the same structure. To rotate the input tensor, we restructure the dimension direction of the tensor by using the permute operation. Finally, we calculate the arithmetic mean value of three units.

**Algorithm 3: Triplet attention.**

```plaintext
input : X ∈ \( R^{C \times H \times W} \)
output : Y ∈ \( R^{C \times H \times W} \)

1 \( X_{\text{perm1}} \leftarrow X_{\text{permute}}(2,1,3) \);
2 \( X_{\text{out1}} \leftarrow \text{TripletUnit}(X_{\text{perm1}}) \) \( X_{\text{out1}} \leftarrow X_{\text{out1} \text{permute}}(2,1,3) \);
3 \( X_{\text{perm2}} \leftarrow X_{\text{permute}}(3,2,1) \);
4 \( X_{\text{out2}} \leftarrow \text{TripletUnit}(X_{\text{perm2}}) \) \( X_{\text{out2}} \leftarrow X_{\text{out2} \text{permute}}(3,2,1) \);
5 \( X_{\text{out3}} \leftarrow \text{TripletUnit}(X) \);
6 \( Y^{C \times H \times W} \leftarrow \frac{X_{\text{out1}} + X_{\text{out2}} + X_{\text{out3}}}{3} \)
```
4. Experiments

In this section, the presented HR-MPE is evaluated on a public macaque monkey dataset, and the results of a series of experiments are demonstrated and analyzed at length. First, the preparation of our experiments is illustrated, containing the dataset, the environment of experiments, the evaluation metrics, and the setting of hyperparameters. Then, HR-MPE is systematically compared with other state-of-the-art deep learning models with great performance. The verification of specific components in our model will be described in the ablation experiment.

4.1. Dataset

In this work, the dataset that we utilize is MacaquePose [8], consisting of 13,083 images of macaque monkeys from zoos (Toyama Municipal Family Park Zoo, Itozu no Mori Zoological Park, Inokashira Park Zoo, and Tobu Zoo), Google Open Images, and the Primate Research Institute of Kyoto University. Additionally, the zoos provide over 78% of images, and images from other ways occupy a tiny part of this dataset. All the images are captured under the condition of natural environment without any human intervention.

The annotations of MacaquePose follow the same format as COCO [34], which consists of 17 keypoints, as shown in Table 2. To maintain the high quality of the dataset, all keypoints are located at the center of joint rotation or organs, and some occluded organs and joints are also annotated. To meet the needs of the experiment, 500 images are randomly selected from the population as the test set, and the rest as the training set, following the way of COCO.
Table 2. Illustrating the definition of macaque joints.

| Keypoint | Definition | Keypoint | Definition |
|----------|------------|----------|------------|
| 1        | Nose       | 10       | Left wrist |
| 2        | Left eye   | 11       | Right wrist|
| 3        | Right eye  | 12       | Left hip   |
| 4        | Left ear   | 13       | Right hip  |
| 5        | Right ear  | 14       | Left knee  |
| 6        | Left shoulder | 15    | Right knee |
| 7        | Right shoulder | 16   | Left ankle |
| 8        | Left elbow | 17       | Right ankle|
| 9        | Right elbow|          |            |

4.2. Evaluation Metrics

The aim of pose estimation is to obtain the spatial position of body keypoints in a two-dimensional or three-dimensional space. Therefore, positioning accuracy becomes an important factor to evaluate the performance of pose estimation. Specifically, the method of evaluation follows COCO.

Evaluation metrics are based on the object keypoint similarity (OKS). The definition of keypoint similarity (KS) is as follows:

\[
ks(\hat{z}_i, z_i) = e^{-\frac{||\hat{z}_i - z_i||_2^2}{2s^2e_i^2}},
\]

where \(\hat{z}_i\) is the predicted coordinate in the \(i_{th}\) keypoint, and \(z_i\) is the ground truth of that. \(s\) is a scale of object, and \(e_i\) represents different constants for different kinds of keypoints. Tolerance of the coordinate bias is controlled by \(s^2e_i^2\). A description of OKS is as follows:

\[
OKS(\hat{z}_i, z_i) = \frac{ks(\hat{z}_i, z_i)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)},
\]

where \(v_i\) describes the visibility of a keypoint.

In this method, average precision (AP) is the most important standard, which refers to the average value of precision when \(OKS = 0.50, 0.55, \ldots, 0.90, 0.95\), respectively. In general, \(OKS \geq 0.75\) is a tiny spatial error range. This requires that the spatial accuracy of the coordinate estimation should be very high, which is the main reason for the high spatial sensitivity of the macaque monkey pose estimation task.

4.3. Experimental Settings

Table 3 presents the settings and hyperparameters of experiments on the macaque monkey dataset. To further illustrate the advantages of HR-MPE, it is compared with other state-of-the-art pose estimation models containing SimpleBaseline [31], MobileNetV2 [14], ShuffleNetV2 [25], HRFormer-Tiny [16], and PVT [27]. To process a fair evaluation, the training and testing datasets keep the same distribution. The best convergent models of each network are chosen. All experiments are executed under the deep learning framework PyTorch 1.8.0 with an NVIDIA GeForce RTX 3090 GPU (24 GB).

Firstly, the macaque body detection frame provided by the dataset is enlarged according to the aspect ratio of 4:3, and then the image in the detection frame is cut out from the original image. In this dataset, the cropped images are scaled to a fixed size of 256 × 192. In data enhancement, random image rotation with an angle change of \([-80^\circ, 80^\circ]\), multiscale enhancement with a zoom range of \([0.5, 1.5]\), and random horizontal image flipping are used in our work.
Table 3. Demonstration of experimental settings and hyperparameters.

| Parameters        | Setting   |
|-------------------|-----------|
| Max epoch         | 210       |
| Base learning rate| $5 \times 10^{-4}$ |
| Ware-up step      | [170, 200] |
| Optimizer         | Adam      |
| Batch size        | 64        |
| Loss function     | Joint MSE Loss |

Adam is used as the optimizer during network training. The initial learning rate is $5 \times 10^{-4}$. In the 170th round (epoch), the learning rate drops to $1 \times 10^{-4}$, and in the 200th round, the learning rate drops to $1 \times 10^{-5}$. The network has trained 210 epochs in total.

5. Results and Discussion

5.1. Experimental Result and Discussion

Table 4 reports the results of the macaque monkey dataset which demonstrated that HR-MPE and other five state-of-the-art backbones are evaluated on six evaluation metrics of $AP$, $AP^{50}$, $AP^{75}$, $AP^M$, $AP^L$, and $AR$ with an input size of $256 \times 192$. It is clear that our HR-MPE ranks the top position on all quotas. It is worth noting that our model exceeds HRFormer by 1.8 $AP$ and 1.1 $AR$, which ranks second, with a smaller number of parameters. SimpleBaseline-18, whose parameters are four times more than our model, lags by a 3.1 $AP$ score and a 3.7 $AR$ score. ShuffleNet and MobileNet rank the first and second from the bottom, respectively, with a difference of 5.4 $AP$ and 7.8 $AP$ from HR-MPE, respectively.

Table 4. Comparison on the macaque monkey test set. The volume of parameters and FLOPs for the pose estimation network are measured without animal detection and keypoint head.

| Methods        | Input Size | #param. | GFLOPs | $AP$ | $AP^{50}$ | $AP^{75}$ | $AP^L$ | $AR$ |
|----------------|------------|---------|--------|------|-----------|-----------|--------|------|
| ShuffleNetV2   | 256 × 192  | 1.25M   | 0.14   | 69.2 | 89.5      | 80.5      | 69.7   | 73.0 |
| MobileNetV2    | 256 × 192  | 2.22M   | 0.31   | 71.6 | 89.5      | 82.0      | 71.9   | 75.0 |
| SimpleBaseline-18 | 256 × 192 | 11.17M  | 1.78   | 72.9 | 89.6      | 82.9      | 73.4   | 76.3 |
| PVT            | 256 × 192  | 0.86M   | 0.10   | 72.6 | 89.4      | 83.9      | 72.9   | 76.1 |
| Lite-HRNet     | 256 × 192  | 1.76M   | 0.31   | 74.2 | 89.5      | 84.9      | 74.3   | 77.4 |
| HRFormer-Tiny  | 256 × 192  | 2.49M   | 1.39   | 75.2 | 89.4      | 85.8      | 75.7   | 78.9 |
| HR-MPE (ours)  | 256 × 192  | 2.40M   | 1.92   | 77.0 | 89.5      | 86.2      | 77.2   | 80.0 |

From the experimental results, the use of high-resolution representation is conducive to improving the effect of the network. The light-weight high-resolution network (Lite-HRNet), HRFormer-Tiny, and HR-MPE are examples of networks that maintain high-resolution representation throughout the process. SimpleBaseline-18 is an example of a network that recovers high-resolution representation from low-resolution representation (PVT, ShuffleNetV2, and MobileNetV2). This further shows that pose estimation is a resolution-sensitive computer vision task. As we discussed in Section 3.3, our model has clear benefits over the typical convolution network (Lite-HRNet) due to the careful structural design of the transformer and the strong representation capacity of the PSA for spatial information. The use of the PSA, the attention-refined block, and the feature fusion unit enables HR-MPE to achieve better results than HRFormer-Tiny while maintaining a low number of parameters. The triple attention and asymmetric convolution in attention-refined blocks are used for designing light-weight models. Experimental results show that these improvements achieve good performance while maintaining a smaller size.

The annotation results of HR-MPE in the test set are shown in Figure 6. The first column is the collected original image, the second column is the manually annotated pose information, and the third column is the pose information estimated by the model. The
model can accurately predict the pose of macaques, whether it is multiple monkeys or a single monkey. Some occluded parts in the training set are annotated, as described in Section 3.1. The model can effectively learn these features. It is worth noting that certain keypoints in the test set that are not manually annotated are also correctly predicted. Because of body occlusion, the right forelimb of the monkey in the first row, for instance, can not be marked. However, the model can still approximate its location, proving the model’s capacity to be generalized.

Figure 6. Cont.
When analyzing the data, it is noted that the performance of these models on the macaque dataset is better than that of the human pose datasets, such as COCO and MPII [4,5,17,22,35,36]. The dataset size of macaque is much smaller than COCO and MPII. In terms of the characteristics of deep learning, the larger the dataset is, the better the model will be trained. After analysis, we find that the key step of pose estimation is to obtain the keypoints’ information of a body. In the process of data annotation, many keypoints are marked not on the body but on the clothes (such as shoulders and knees). The materials of different clothes are quite unfamiliar, and the corresponding image information is more complex and diverse. However, since monkeys do not wear clothing and there is less variation in monkey fur than there is in clothing, it is easier for the networks to produce better results.

5.2. Ablation Experiment

To verify the effectiveness of each improvement, PoolFormer is used as the baseline. First, an attention-refined block is added to the back of the parallel structure. The results show that although the number of parameters and calculations increases slightly, the effect has been significantly improved by 0.5 AP. This result fulfilled our original design intention. The pooling operation can lead to information loss in the network, and the W-MSA is too complex to be suitable for small-scale models. The PSA mechanism is used to create a balance between complexity and effect. Despite the fact that the PSA parameters are tiny, the results reveal that good results are obtained, which is compatible with the analysis in Section 3.3. As can be seen from Table 5, this change has brought about the improvement of 1.0 AP. In the previous data preprocessing, the inversion operation can bring deviation, which is not conducive to network training. In this work, the UDP method is used for unbiased data processing, which improves the performance by 0.9 AP. This process does not increase the amount of calculation and parameters of the model.
Table 5. Influence of selecting the attention-refined block and polarized self-attention. We show #param, GFLOPs, AP, and AR on the macaque monkey test set.

| No. | Baseline | Attention Refined Block | PSA | UDP | #param. | GFLOPs | AP  | AR  |
|-----|----------|-------------------------|-----|-----|---------|--------|-----|-----|
| 1   | √        |                         |     |     | 2.197   | 1.803  | 74.6| 78.1|
| 2   | √        | √                       |     |     | 2.223   | 1.884  | 75.1| 78.5|
| 3   | √        | √                       | √   | √   | 2.407   | 1.928  | 76.1| 79.4|
| 4   | √        | √                       | √   | √   | 2.407   | 1.928  | 77.0| 80.0|

6. Conclusions

In this work, the attention-refined light-weight high-resolution network based on the top-down paradigm was proposed to obtain a better performance. The well-designed multi-branch parallel architecture could keep high-resolution representation from the start to end. The PSA with a simple structure yet effective achieved global attention. Two light-weight attention-refined blocks were supplemented at the tail of the multi-branch structure, making good use of asymmetric convolutions and triplet attention, almost without parameters, obtaining richer representation information. We utilized the UDP pre-processing method to obtain an unbiased result of flipping. Experiments were conducted on the macaque monkey dataset and the effectiveness of our network was illustrated by a series of experimental results. The model proposed in this work has a great performance and GFLOPs slightly larger than other comparative models. In the future, we will study more efficient convolution and lighter network for macaque monkey pose estimation.

Author Contributions: Conceptualization, S.L. (Sicong Liu) and Q.F.; methodology, S.L. (Sicong Liu) and Q.F.; software, Q.F., S.L. (Shanghao Liu) and S.L. (Sicong Liu); validation, Q.F. and S.L. (Sicong Liu); data curation, Q.F.; writing—original draft preparation, Q.F., S.L. (Shanghao Liu) and S.L. (Sicong Liu); writing—review and editing, Q.F., S.L. (Shanghao Liu), and S.L. (Sicong Liu); visualization, Q.F.; supervision, S.L. (Shuqin Li) and C.Z.; project administration, S.L. (Shuqin Li). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Thank Rollyn Labuguen of Kyushu Institute of technology in Japan for opening the macaque monkey dataset. This provides great convenience for our research.

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

Abbreviations

The following abbreviations are used in this manuscript:

- UDP: unbiased data processing
- W-MSA: window-based multi-head self-attention
- PVT: pyramid vision transformer
- FFN-DW: feed-forward network with a $3 \times 3$ depth-wise convolution
- PSA: polarized self-attention
- ARM: attention-refined module
- NLP: natural language processing
- HRFormer: high-resolution transformer
- HRNet: high-resolution network
- ViT: vision transformer
- OKS: object keypoint similarity
- KS: keypoint similarity
References

1. Bala, P.C.; Eisenreich, B.R.; Yoo, S.B.M.; Hayden, B.Y.; Park, H.S.; Zimmermann, J. Automated markerless pose estimation in freely moving macaques with OpenMonkeyStudio. Nat. Commun. 2020, 11, 4560. https://doi.org/10.1038/s41467-020-18441-5.

2. Mathis, M.W.; Mathis, A. Deep learning tools for the measurement of animal behavior in neuroscience. Curr. Opin. Neurobiol. 2020, 60, 1–11. https://doi.org/10.1016/j.conb.2019.10.008.

3. Labuguen, P.; Gaurav, V.; Blanco, S.N.; Matsumoto, J.; Inoue, K.; Shibata, T. Monkey Features Location Identification Using Convolutional Neural Networks. bioRxiv 2018, 377895. https://doi.org/10.1101/377895.

4. Bulat, A.; Tzimiropoulos, G. Human Pose Estimation via Convolutional Part Heatmap Regression. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Series Title: Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2016; Volume 9911, pp. 717–732; https://doi.org/10.1007/978-3-319-46478-7_44.

5. Liu, Z.; Chen, H.; Feng, R.; Wu, S.; Ji, S.; Yang, B.; Wang, X. Deep Dual Consecutive Network for Human Pose Estimation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; IEEE: Nashville, TN, USA, 2021; pp. 525–534. https://doi.org/10.1109/CVPR46437.2021.00059.

6. Zhang, F.; Zhu, X.; Ye, M. Fast Human Pose Estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; IEEE: Long Beach, CA, USA, 2019; pp. 3512–3521. https://doi.org/10.1109/CVPR.2019.00363.

7. Liu, X.; Yu, S.Y.; Flierman, N.A.; Loyola, S.; Kammersn, M.; Hoogland, T.M.; De Zeeuw, C.I. OptiFlex: Multi-Frame Animal Pose Estimation Combining Deep Learning With Optical Flow. Front. Cell. Neurosci. 2021, 15, 621525. https://doi.org/10.3389/fncel.2021.621525.

8. Labuguen, R.; Matsumoto, J.; Negrete, S.B.; Nishimaru, H.; Nishijo, H.; Takada, M.; Go, Y.; Inoue, K.I.; Shibata, T. MacaquePose: A Novel “in the Wild” Macaque Monkey Pose Dataset for Markerless Motion Capture. Front. Behav. Neurosci. 2021, 14, 581154. https://doi.org/10.3389/fnbeh.2020.581154.

9. Wenwen, Z.; Yang, X.; Rui, B.; Li, L. Animal Pose Estimation Algorithm Based on the Lightweight Stacked Hourglass Network. Preprints 2022, in Review. https://doi.org/10.21203/rs.3.rs.1627999/v1.

10. Ngo, V.; Gorman, J.C.; Fuente, M.F.D.L.; Souto, A.; Schiel, N.; Miller, C.T. Active vision during prey capture in wild marmosets. Curr. Biol. 2022, 32, 1–6. https://doi.org/10.1016/j.cub.2022.06.028.

11. Labuguen, R.; Bardezoa, D.K.; Negrete, S.B.; Matsumoto, J.; Inoue, K.; Shibata, T. Primate Markerless Pose Estimation and Movement Analysis Using DeepLabCut. In Proceedings of the 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Spokane, WA, USA, 30 May–2 June 2019; IEEE: Spokane, WA, USA, 2019; pp. 297–300. https://doi.org/10.1109/ICIIE.2019.8858533.

12. Blanco Negrete, S.; Labuguen, R.; Matsumoto, J.; Go, Y.; Inoue, K.I.; Shibata, T. Multiple Monkey Pose Estimation Using OpenPose. Neurosci. 2021. https://doi.org/10.1101/2021.01.28.428726.

13. Mao, W.; Ge, Y.; Shen, C.; Tian, Z.; Wang, X.; Wang, Z. TFPose: Direct Human Pose Estimation with Transformers. arXiv 2021, arXiv:2103.15320.

14. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE: Salt Lake City, UT, USA, 2018; pp. 4510–4520. https://doi.org/10.1109/CVPR.2018.00474.

15. Wang, Y.; Li, M.; Cai, H.; Chen, W.M.; Han, S. Lite Pose: Efficient Architecture Design for 2D Human Pose Estimation. arXiv 2022, arXiv:2205.01271.

16. Yuan, Y.; Fu, R.; Huang, L.; Lin, W.; Zhang, C.; Chen, X.; Wang, J. HRFormer: High-Resolution Vision Transformer for Dense Predict. In Proceedings of the NeurIPS 2021, Virtual, 6–14 December 2021.

17. Wang, X.; Tong, J.; Wang, R. Attention Refined Network for Human Pose Estimation. Neural Process. Lett. 2021, 53, 2853–2872. https://doi.org/10.1007/s11063-021-10523-9.

18. Yu, W.; Luo, M.; Zhou, P.; Si, C.; Zhou, Y.; Wang, X.; Feng, J.; Yan, S. MetaFormer is Actually What You Need for Vision. arXiv 2021, 2101.11482.

19. Liu, H.; Liu, F.; Fan, X.; Huang, D. Polarized Self-Attention: Towards High-quality Pixel-wise Regression. arXiv 2021, arXiv:2107.00782.

20. Zhang, Y.; Wa, S.; Sun, P.; Wang, Y. Peer Defect Detection Method Based on ResNet and DCGAN. Information 2021, 12, 397. https://doi.org/10.3390/info12100397.

21. 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; pp. 468–475.

22. Kreiss, S.; Berton, L.; Alahi, A. PiPaf: Composite Fields for Human Pose Estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; IEEE: Long Beach, CA, USA, 2019; pp. 11969–11978. https://doi.org/10.1109/CVPR.2019.01225.

23. Intarat, K.; Rakwatin, P.; Panboonyuen, T. Enhanced Feature Pyramid Vision Transformer for Semantic Segmentation on Thailand Landsat-8 Corpus. Information 2022, 13, 259. https://doi.org/10.3390/info13050259.
24. Howard, A.; Sandler, M.; Chen, B.; Wang, W.; Chen, L.C.; Tan, M.; Chu, G.; Vasudevan, V.; Zhu, Y.; Pang, R.; et al. Searching for MobileNetV3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; IEEE: Seoul, Korea, 2019; pp. 1314–1324. https://doi.org/10.1109/ICCV.2019.00140.

25. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Series Title: Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11218, pp. 122–138; https://doi.org/10.1007/978-3-030-01264-9_8.

26. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.

27. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 568–578.

28. Huang, J.; Zhu, Z.; Guo, F.; Huang, G. The Devil Is in the Details: Delving Into Unbiased Data Processing for Human Pose Estimation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; IEEE: Seattle, WA, USA, 2020; pp. 5699–5708. https://doi.org/10.1109/CVPR42600.2020.00574.

29. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 2021, 43, 3349–3364. https://doi.org/10.1109/TPAMI.2020.2983686.

30. Li, Y.; Wang, C.; Cao, Y.; Liu, B.; Luo, Y.; Zhang, H. A-HRNet: Attention Based High Resolution Network for Human pose estimation. In Proceedings of the 2020 Second International Conference on Transdisciplinary AI (TransAI), Irvine, CA, USA, 21–23 September 2020; IEEE: Irvine, CA, USA, 2020; pp. 75–79. https://doi.org/10.1109/TransAI49837.2020.00016.

31. Xiao, B.; Wu, H.; Wei, Y. Simple Baselines for Human Pose Estimation and Tracking. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Series Title: Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11210, pp. 472–487; https://doi.org/10.1007/978-3-030-01251-1_29.

32. Ding, X.; Guo, Y.; Ding, G.; Han, J. ACNet: Strengthening the Kernel Skeletons for Powerful CNN via Asymmetric Convolution Blocks. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27–28 October 2019; IEEE: Seoul, Korea, 2019; pp. 1911–1920. https://doi.org/10.1109/ICCV.2019.00200.

33. Misra, D.; Nalamada, T.; Arasanipalai, A.U.; Hou, Q. Rotate to Attend: Convolutional Triplet Attention Module. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; IEEE: Waikoloa, HI, USA, 2021; pp. 3138–3147. https://doi.org/10.1109/WACV48630.2021.00318.

34. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; Zitnick, L. Microsoft COCO: Common Objects in Context. In Proceedings of the ECCV European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.

35. 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; IEEE: Columbus, OH, USA, 2014; pp. 3686–3693. https://doi.org/10.1109/CVPR.2014.471.

36. Yang, S.; Quan, Z.; Nie, M.; Yang, W. TransPose: Keypoint Localization via Transformer. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; IEEE: Montreal, QC, Canada, 2021; pp. 11782–11792. https://doi.org/10.1109/ICCV48922.2021.01159.