electronics

MDPI

Article

Human Pose Estimation via an Ultra-Lightweight Pose Distillation Network

Shihao Zhang [1,2], Baohua Qiang [1], Xianyi Yang [1,*], Xuekai Wei [3,*], Ruidong Chen [1] and Lirui Chen [1]

[1] Guangxi Key Laboratory of Image and Graphic Intelligent Processing, Guilin University of Electronic Technology, Guilin 541004, China; shihaozhang2022@163.com (S.Z.); qiangbh@guet.edu.cn (B.Q.); pgezcrb@163.com (R.C.); lyricschen2022@163.com (L.C.)
[2] School of Information Engineering, Luohe Vocational Technology College, Luohe 462000, China
[3] School of Computer Science, Chongqing University, Chongqing 400044, China
* Correspondence: xianyiyang65@126.com (X.Y.); xuekaiwei2-c@my.cityu.edu.hk (X.W.)

Abstract: Most current pose estimation methods have a high resource cost that makes them unusable in some resource-limited devices. To address this problem, we propose an ultra-lightweight end-to-end pose distillation network, which applies some helpful techniques to suitably balance the number of parameters and predictive accuracy. First, we designed a lightweight one-stage pose estimation network, which learns from an increasingly refined sequential expert network in an online knowledge distillation manner. Then, we constructed an ultra-lightweight re-parameterized pose estimation subnetwork that uses a multi-module design with weight sharing to improve the multi-scale image feature acquisition capability of the single-module design. When training was complete, we used the first re-parameterized module as the deployment network to retain the simple architecture. Finally, extensive experimental results demonstrated the detection precision and low parameters of our method.

Keywords: ultra-lightweight pose estimation; knowledge distillation; re-parameterized module; end-to-end; feature compression

Citation: Zhang, S.; Qiang, B.; Yang, X.; Wei, X.; Chen, R.; Chen, L. Human Pose Estimation via an Ultra-Lightweight Pose Distillation Network. *Electronics* **2023**, *12*, 2593. https://doi.org/10.3390/electronics12122593

Academic Editor: Donghyeon Cho

Received: 26 April 2023
Revised: 2 June 2023
Accepted: 5 June 2023
Published: 8 June 2023

1. Introduction

Human pose estimation has been a research topic of great interest in the field of computer vision for decades. It refers to the recognition and location of the keypoints (e.g., head and shoulder) of each visible human body in pictures or videos captured from image sensors, which plays a significant role in a variety of human-computer interaction applications. Traditional approaches typically use some hand-designed features to detect keypoints, such as tree-structured models [1–4] and graphical models [5–8]. With the rapid development of convolutional neural networks (CNNs) [9–11], the accuracy of human pose estimation based on CNNs has continuously improved. However, most current human pose estimation methods [12–20] have a complex network structure and very high resource costs, which makes them unsuitable for resource-limited devices (e.g., monitoring equipment).

Recently, researchers have conducted several studies [21–27] to achieve good performance and decrease the computational cost of human pose estimation. Cao et al. [21] used a two-branch multi-stage design in which the first branch of each stage generated accurate confidence maps and the second branch of each stage generated helpful part affinity fields, which were then parsed using a greedy inference strategy to generate good multi-person keypoint locations to achieve real-time performance. Kato et al. [22] used the output of the strong teacher model to improve some incomplete labels in the training dataset and designed a label-correction model. Zhang et al. [23] established a compact hourglass network and distilled knowledge of the original state-of-the-art hourglass network to achieve highly cost-effective results, demonstrating the superiority of the knowledge distillation

scheme. Qiang et al. [24] designed a lightweight architecture that used an efficient backbone network composed of modified SqueezeNet and three continuously refined stages to improve detection speed. The above lightweight networks tend to achieve good predictive accuracy, but the degree of reduction in the number of their model parameters have been unsatisfactory. Thus, some researchers have attempted to achieve accurate and lightweight detection results by applying other helpful technologies (e.g., knowledge distillation and re-parameterized technology). Weinzaepfel et al. [25] took advantage of annotated datasets to train some independent teacher models for each part, including body, hand, and face teacher models, and distilled their knowledge into a single deep convolutional network to achieve whole-body 2D-3D pose estimation. Zhong et al. [26] used a lightweight upsampling module and deep supervision pyramid network to enhance the multi-scale image feature representation ability of the model, which resulted in higher detection accuracy and lower computational costs. Wang et al. [27] used a mixed structure that consisted of a multi-branch training network and single-branch deployment network to design an efficient re-parameterized bottleneck block, which resulted in good performance in terms of detection accuracy and detection speed.

Although these methods aim to improve human pose estimation performance using the means of CNNs, the following problems still need to be solved:

(1) Existing CNN-based pose estimation methods often use a complex deployment network that is computationally expensive;

(2) The detection results are unsatisfactory, to a certain extent, if the number of parameters of the pose estimation methods is low.

To address the above problems, we mainly study an ultra-lightweight end-to-end pose distillation network (UEPDN), which applies some helpful techniques to better balance the number of parameters and predictive accuracy of the model. The main contributions of our study are generalized as follows:

- We design a lightweight one-stage pose estimation network, stage 1, which learns from an increasingly refined sequential expert network in an online knowledge distillation manner;
- We construct an ultra-lightweight re-parameterized pose estimation subnetwork that uses a multi-module design with weight-sharing to improve the multi-scale image feature acquisition capability of the single-module design. When training is complete, we use the first re-parameterized module as the deployment network to retain the simple architecture;
- Extensive experimental results demonstrate the superiority of our method on three standard benchmark datasets.

2. Related Work

2.1. Lightweight Pose Estimation Network

Recent studies [28–34] were conducted on lightweight network design to promote human pose estimation applications in resource-limited platforms. For example, Bulat and Tzimiropoulos [28] used binarization technology to design a lightweight pose estimation network for inference acceleration; however, it had low detection accuracy. Xiao et al. [29] built a baseline model, which simply added a few deconvolutional layers to the last convolutional stage of ResNet to directly generate pose heatmaps from image features. Although this method provided some simple and effective model design ideas, its detection performance was not satisfactory. Wang et al. [32] used explicit human estimation regions of interest and relevant 3D directions to directly estimate a 3D pose, which addressed the problems of 2D errors propagating to 3D recovery leading to degenerated results. Li et al. [33] built a multi-branch online knowledge distillation network to simplify the traditional distillation process and improve keypoint detection performance, which they called OKDHP. However, the multi-branch distillation network design increased the training complexity of the model and the accuracy of the model needed improvement. Xiao et al. [34] designed a compact single-stage pose regression method that used a new body representation to

achieve good inference performance for multi-person pose estimation. However, the number of parameters it had was unsatisfactory. Unlike these methods, our method does not need a multi-branch architecture to train a small network; instead, it uses a single-branch iterative pose distillation training network. Simultaneously, we constructed an ultra-lightweight re-parameterized pose estimation subnetwork that uses multi-module design with weight-sharing to improve the multi-scale image feature acquisition capability of the single-module design. This improves the performance of the model while barely increasing the calculational costs. When training was complete, we distributed the weight value to the ultra-lightweight target deployment network through knowledge distillation technology and re-parameterized technology, which maintained good detection accuracy and reduced the model parameters.

2.2. Intermediate Supervision

Intermediate supervision, also known as deep supervision, is popularly applied in multi-stage pose estimation networks (e.g., CPM and OpenPose). It calculates the loss at the prediction location at every stage in the multi-stage network, which has been proven to effectively address the vanishing gradient problem that occurs in the training phase of a deep network and improve keypoint detection performance. Generally, the output of the last stage is used to guarantee accuracy for deployment. We also follow this strategy in our network design, which ensures that the gradient is transferred at all stages and also helps to compress the redundant parameters of the proposed network.

2.3. Structure Optimization

Recently, many lightweight methods based on CNNs, including knowledge distillation, re-parameterized technology, and model pruning, have been proposed to be deployed in resource-limited devices. Li et al. [33] used online knowledge distillation technology to build a small efficient network that distilled the trained knowledge of a multi-branch modified hourglass network into an efficient compact network to decrease the complexity of the traditional two-stage knowledge distillation training process and quantity of model parameters. However, its training costs were unsatisfactory and the accuracy of the model needed improvement. Wang et al. [27] proposed an unbiased lightweight network that consisted of various branch architectures, where the multi-branch architecture, applied in the training stage, would improve detection performance, and the single-branch architecture, used in the deployment stage, would reduce the inference complexity of the model. It used a re-parameterized strategy to implement the conversion of multi-branch parameters to single-branch parameters and showed characteristics of good performance, computational resource savings, and fast inference speed. We also adopted the design concept of the re-parameterized structure in our method. We constructed a re-parameterized structure that introduces the knowledge distillation technique. Our method simultaneously had a low quantity of parameters and good detection accuracy.

3. Proposed Methods

In this study, we developed an ultra-lightweight end-to-end pose estimation network based on online knowledge distillation technology and re-parameterized technology. The structure of the proposed network, including the training network and deployment network, is shown in Figure 1. First, the training images were processed through a re-parameterized network, stage 1, where a modified PeleeNet [35] extracted rich human body features, and re-parameterized pose estimation modules used the multi-module design with weight-sharing to generate increasingly accurate detection results in sequence. Then, stage $s \in \{2, \ldots, S\}$, which included five convolutional blocks that consisted of two 3×3 convolutional layers, a 1×3 convolutional layer, 3×1 convolutional layer, and two 1×1 convolutional layers, predicted increasingly refined keypoint heatmaps using an iterative sequential prediction architecture. The prediction results of the last stage were considered to be the expert model's outputs that were used to teach modules R1, R2,

and stage $s \in \{2, \ldots, S-1\}$. Finally, the re-parameterized module R1 was used as the deployment network to retain the simple architecture.

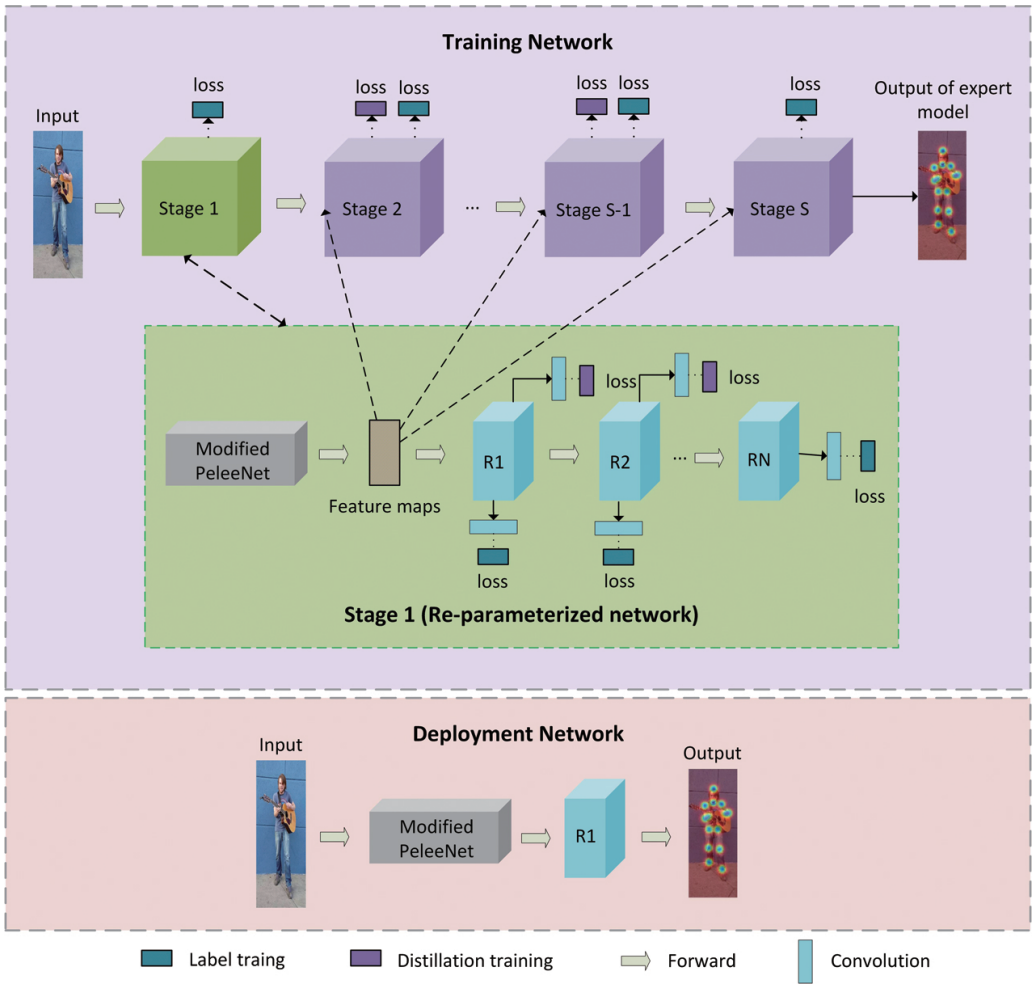

Figure 1. The structure of ultra-lightweight pose distillation network.

3.1. Keypoint Feature Extraction

Given the input RGB image $M \in \mathbb{R}^{C \times H \times W}$ of size $H \times W$, we first used a human body detector to obtain human bounding boxes. Then, we cropped every box to 368×368 from the image and sent it to the stage 1 network. Stage 1 is a re-parameterized network that consists of a modified PeleeNet and several re-parameterized modules. We adopted the modified PeleeNet as the backbone network to extract rich human body features. The size of the original feature map extracted from our modified PeleeNet was $46 \times 46 \times 128$. After we passed this result through the first re-parameterized module R1, the size of the human body feature map was adjusted to $46 \times 46 \times 15$. The re-parameterized module $r \in \{2, \ldots, N\}$ continuously generated increasingly accurate detection results. We regarded the prediction results of the last re-parameterized module RN as the stage 1 network's outputs. Then, in the proposed method, we used an iterative sequential prediction architecture in which the keypoint heatmaps with a size of $46 \times 46 \times 15$ pixels generated from the previous adjacent stage, and feature maps with a size of $46 \times 46 \times 128$ pixels generated from the modified PeleeNet, were fused to take abundant characteristic information with learned spatial context features to enhance the network's cognition of multi-scale image features and rich image-dependent spatial features.

3.2. Re-Parameterized Structure

Due to the complex correlation of knowledge transfer from the expert model to the target student model, the final distillation results can be unsatisfactory, to a certain extent, if the student model is just a simplified version of the expert model. To reasonably use information of various scales and high-value information provided by the expert network, we designed a re-parameterized structure that introduced the knowledge distillation technique.

As shown in Figure 2, we assumed that the input feature maps were $x \in \mathbb{R}^{C\times H\times W}$, where C is the number of channels and $H \times W$ represents the size of feature maps. First, we inputted the feature maps into feature space s, and $s(x_i) = W_s x_i$, where W_s is a weight matrix that changes with the intermediate features. We implemented W_s as a 3×3 convolution with a single-channel to obtain spatial information. Second, we processed the intermediate feature $s(x_i)$ using a feature compression module. The feature compression results $f_i \in \mathbb{R}^{C\times H\times W}$ were generated as follows:

$$f_i = LW(s(x_i)) \tag{1}$$

where LW acts on $s(x_i)$, and performs the operations of 1×3 and 3×1 convolutions successively. Then, we inputted the feature compression results f_i into a feature enhancement layer, the outputs of which were $v_i \in \mathbb{R}^{C\times H\times W}$. They were generated as follows:

$$v_i = h(f_i) \tag{2}$$

where h represents the 3×3 convolution operation used to enhance the representation ability of image features. Finally, we used $y \in \mathbb{R}^{C\times H\times W}$ as the final outputs of the re-parameterized module. They were generated as follows:

$$y_i = s(x_i) \oplus v_i \tag{3}$$

where $\oplus$ represents the element addition operation. Then, the re-parameterized module $r \in \{2, \ldots, N\}$ used image features generated from the previous adjacent re-parameterized module to successively enhance the multi-scale image feature acquisition capability of the single-module design.

3.3. Learning in the UEPDN

By reducing the discrepancy between the objective prediction coordinates and given label coordinates, we obtain the optimal mapping between the human image and keypoint coordinates. We apply the l_2 loss to improve the performance of the proposed network:

$$L(p) = \frac{1}{n}\sum_{i=1}^{n}(p - r)^2 \tag{4}$$

where p is the predicted coordinate, r is the real label coordinate, and n is the number of keypoints.

We use two types of loss functions, conventional label loss L_l and specialized distillation loss L_d, to augment training. The overall loss function L can be expressed as follows:

$$L = L_l + L_d, \tag{5}$$

where L_l is the loss between all levels of the prediction coordinates and given label coordinates, and L_d is the loss between the prediction coordinates of the student models and prediction coordinates of the expert model. L_l is calculated as follows:

$$L_l = a \times \left(\frac{1}{n}\sum_{s=1}^{S}\sum_{i=1}^{n}(p_i - r)^2 + \frac{1}{n}\sum_{r=1}^{R}\sum_{i=1}^{n}(p_i - r)^2\right) \tag{6}$$

where a is a hyperparameter, p_i is the predicted coordinate of keypoint i, r is the real label coordinate of corresponding keypoint i, S is the number of stages, and R is the number of re-parameterized modules. L_d is calculated as follows:

$$L_d = b \times \left(\frac{1}{n}\sum_{s=2}^{S-1}\sum_{i=1}^{n}(p_i - p_i^*)^2 + \frac{1}{n}\sum_{r=1}^{2}\sum_{i=1}^{n}(p_i - p_i^*)^2\right) \tag{7}$$

where b is a hyperparameter, p_i is the predicted coordinate of keypoint i generated from stage $s \in \{2, \ldots, S-1\}$ or re-parameterized module $r \in \{1, 2\}$, p_i^* is the predicted coordinate generated from the expert network, S is the number of stages, and r is the number of re-parameterized modules. We can obtain the optimal parameters by minimizing the overall loss function L.

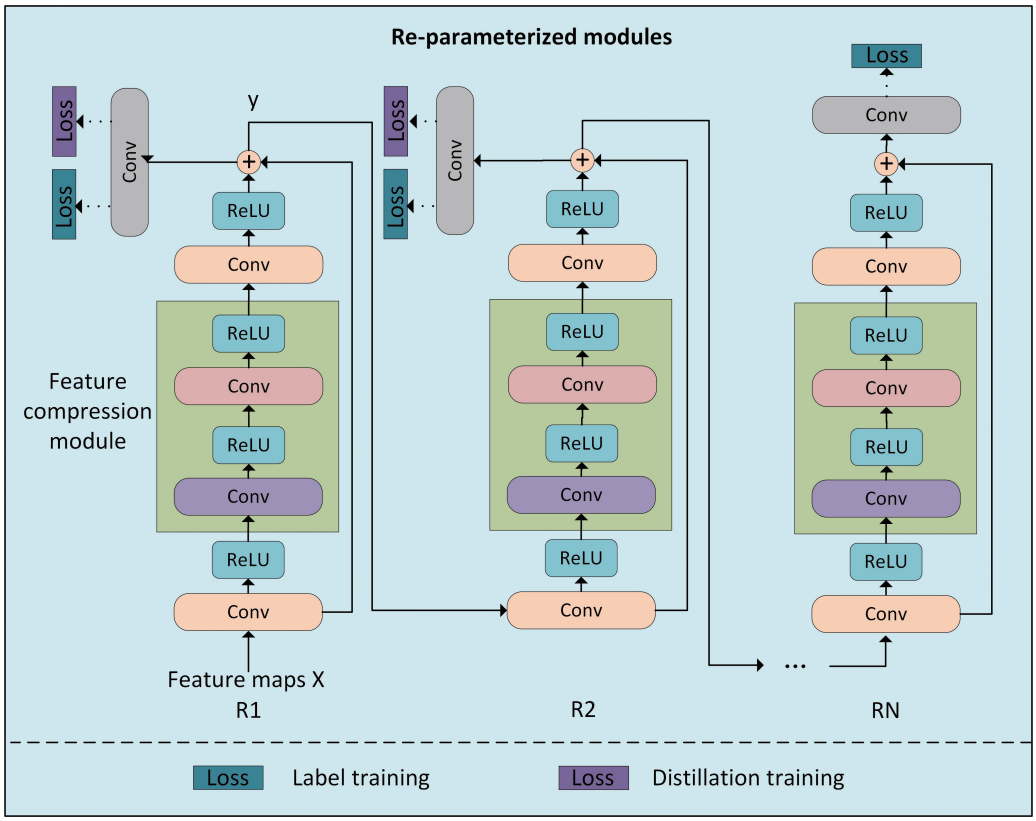

Figure 2. The framework of re-parameterized modules.

3.4. Summary

The complete flow of our method is summarized in Algorithm 1. During the training phase of the proposed model, we obtained result z_i of the re-parameterized module based on the training results of the previous iteration, and inputted the result y_i^{s-1} and feature map f generated from the modified PeleeNet in each subnetwork stage to obtain the result y_i^s of this iteration. We continuously optimized the model parameters by minimizing the overall loss function L. The deployment phase of the proposed model has a simple architecture. First, we inputted the test images into the modified PeleeNet to extract rich human body features, and then we processed these features using the re-parameterized module $R1$ to directly obtain the final detection result p.

In the proposed UEPDN model, we used a re-parameterized structure that introduced the knowledge distillation technique to reasonably use the information of various scales and high-value information provided by the expert network. Simultaneously, we used the efficient overall loss function L, which consisted of conventional label loss L_l and specialized distillation loss L_d, to augment training. Finally, we used the first re-parameterized module

R1 as the deployment network to keep the simple architecture that resulted in good detection performance with high accuracy and fewer model parameters.

Compared with other state-of-the-art lightweight pose estimation algorithms, the proposed method uses an online end-to-end pose distillation architecture and several ultra-lightweight re-parameterized modules with weight-sharing that enhance the multi-scale image feature acquisition capability of the single-module design, while barely increasing the calculational costs, to obtain good detection results.

Algorithm 1 Ultra-lightweight Pose Estimation Algorithm

Input: The human image set $H = \{h_1, h_2, \ldots, h_n\}$ and the corresponding label set $L = \{l_1, l_2, \ldots, l_n\}$.
Output: The predicted keypoints result p.
Let I denote the number of training iterations, R denote the number of the re-parameterized module, z_r denote the output results of re-parameterized module r, S denote the number of the subnetwork stage, and y_s denote the output results of subnetwork stage s;
for $i = 1$ to I **do**
  for $s = 1$ to S **do**
    for $r = 1$ to R **do**
      if $(r == 1)$ **then**
        $z_i^r = RM_r(H)$;
      else
        $z_i^r = RM_r(z_i^{r-1})$;
      end if
    end for
    if $(s == 1)$ **then**
      $y_i^s = RM_r(z_i^R)$;
    else
      $y_i^s = STAGE(y_i^{s-1}, f)$;
    end if
  end for
  Calculate the overall loss function L based on Equation (5) and optimize L;
end for
Deployment: $p = RM_1(image)$.

4. Experimental Results

In this section, we compare the proposed method with excellent pose estimation methods using the MPII [36], LSP [37], and UAV-Human [38] pose estimation benchmarks. Additionally, we conducted extensive ablation experiments to evaluate our method.

4.1. Pose Estimation on the MPII Dataset

4.1.1. Dataset and Performance Metric

The MPII dataset consists of a series of photos of human activities. It contains approximately 25,000 human images, which include 40,000 human instances with 16 labeled keypoints. We used 14 labeled keypoints of each person for our model. We used the same dataset partitioning method as other state-of-the-art pose estimation methods [13,39]. Specifically, we used 25,000 human instances in the training dataset and 3000 human instances in the validation dataset. We used the official evaluation measure, which represents the standard percentage of correct keypoints (PCK) metric, to evaluate the proposed method. It can be given as follows:

$$PCK = \frac{T_p}{T_p + F_p} \tag{8}$$

where T_p represents the number of correct human keypoints predictions and F_p represents the number of incorrect human keypoints predictions. A type of *PCK* is *PCKh@a*, which represents the percentage of keypoints placed at a required distance defined as a of the human head ground-truth length. We used the official evaluation measure, *PCKh*@0.5, to evaluate the proposed method on the MPII dataset. Additionally, we used the quantity of parameters in the entire network (#Params) and floating-point operations (FLOPs) to measure the deployment cost, and used the area under the curve (AUC) to evaluate the authenticity of the proposed method.

4.1.2. Training and Deployment Details

We conducted all the experiments for the proposed method in a server environment based on Ubuntu 16.04, an NVIDIA GTX1080Ti GPU, and Intel Xeon(R) CPU E5-2603 v2. We implemented our method in Caffe [40]. We cropped all the training images based on the ground-truth box and resized them to 368×368. We used the pre-training PeleeNet model at the beginning of training to accelerate the convergence of model training. We used the Adam [41] optimizer to optimize the entire network during the training process. We initialized the learning rate to 8×10^{-5} and weight decay to 5×10^{-4}. We used 200 epochs for the MPII training dataset. When training was complete, we used the first re-parameterized module as the deployment network to retain the simple architecture. We used the universal testing strategy and the ground-truth boxes of people provided in the datasets. Our trained model generated accurate prediction results for every person in the MPII validation dataset.

4.1.3. Results on the MPII Dataset

Figure 3 shows the visualized predicted key part (right elbow) heatmaps for various stages and re-parameterized modules. We note that the re-parameterized module $r = 1$ produced initial keypoint heatmaps, and module $r \in \{2, \ldots, 5\}$ and stage $s \in \{1, \ldots, 5\}$ produced increasingly refined keypoint heatmaps. Table 1 shows the *PCKh*@0.5 prediction accuracy, AUC, #Params, and FLOPs of our method and current state-of-the-art methods on the MPII validation dataset. Our proposed network UEPDN achieved a good result: 89.3 mean *PCKh*@0.5 score. The accuracy of UEPDN was slightly lower than that of the top-performing methods (e.g., FPD); however, FPD is a knowledge distillation network that is trained twice, which is cumbersome and not always available. Additionally, the parameter quantities and FLOPs of our method were low. As our model used a re-parameterized structure that barely increased the amount of calculational costs while helping the model to be trained and flexibly deployed, it achieved a good balance between model accuracy and deployment costs. The visualized pose estimation results on the MPII dataset are shown in Figure 4. We clearly observe that the proposed UEPDN model achieved robust and exact detection results in images with various human poses and various complex backgrounds.

Table 1. *PCKh*@0.5, AUC (%) rates, #Params, and FLOPs on the MPII validation dataset.

Methods	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean	AUC	#Params	FLOPs
Hourglass [14]	96.9	95.9	89.5	84.4	88.4	84.5	80.7	89.1	–	25.6 M	55 G
SimCC [42]	97.2	96.0	90.4	85.6	89.5	85.8	81.8	90.0	–	25.7 M	32.9 G
PRTR [43]	97.3	96.0	90.6	84.5	89.7	85.5	79.0	89.5	–	57.2 M	21.6 G
TokenPose [44]	97.1	95.9	91.0	85.8	89.5	86.1	82.7	90.2	–	21.4 M	9.1 G
OKDHP-bran2 [33]	96.7	95.4	89.9	84.1	89.0	84.7	81.1	89.2	–	15.5 M	47 G
OKDHP-bran1 [33]	96.7	95.3	89.2	84.0	87.8	83.9	79.5	88.6	–	13.0 M	41 G
DSPNet-B1 [26]	97.1	96.1	89.7	84.8	89.6	85.5	81.3	89.7	–	12.6 M	1.6 G
DSPNet-B0 [26]	96.7	95.7	88.9	82.6	88.7	84.1	78.7	88.5	–	7.6 M	1.2 G
FPD [23]	–	–	–	–	–	–	–	90.1	62.4	3.0 M	9 G
PCT [45]	97.5	**97.2**	**92.8**	**88.4**	**92.4**	**89.6**	**87.1**	**92.5**	–	221.5 M	15.2 G
Openpose [31]	96.2	95.0	87.5	82.2	87.6	82.7	78.4	87.7	–	–	–
UULPN [27]	96.0	93.6	85.3	78.7	86.2	80.4	75.6	85.7	–	2.8 M	2.23 G
Lite-HRNet-30 [46]	–	–	–	–	–	–	–	87.0	–	**1.8 M**	**0.42 G**
UEPDN-R1 (Ours)	**98.1**	96.7	91.0	84.4	90.3	83.8	76.3	89.3	**64.3**	2.75 M	6.2 G

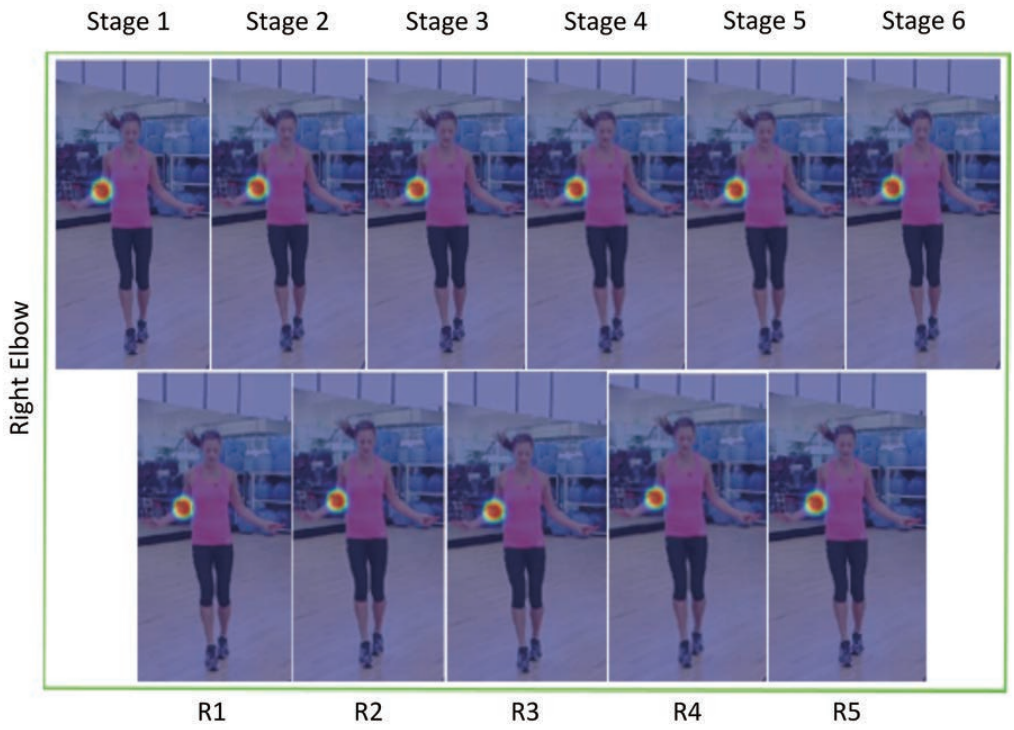

Figure 3. The visualized part (right elbow) heatmaps of six stages and five re-parameterized modules.

Figure 4. Visualized results on the MPII dataset.

4.2. Pose Estimation on the LSP Dataset

4.2.1. Dataset and Performance Metric

The LSP dataset consists of a series of images of human sports activities. We evaluated the proposed method on its extended version, the extended Leeds Sports dataset, which includes 12,000 human instances with 14 labeled keypoints. We used the same dataset partitioning method as that of other state-of-the-art pose estimation methods [13,23]. Specifically, we used 11,000 human instances in the training dataset and 1000 human instances in the testing dataset.

We applied *PCK@b*, which represents the percentage of keypoints placed at a required distance defined as *b* of the human trunk ground-truth length, to evaluate the proposed method on the LSP dataset. We used *PCK*@0.2. Additionally, we used #Params, FLOPs, and FPS to measure deployment performance, and used AUC to evaluate the authenticity of the proposed method.

4.2.2. Training and Deployment Details

We used 150 epochs for the LSP training dataset. The other details of the training process were the same as those for the MPII. When training was complete, we used the first re-parameterized module as the deployment network to retain the simple architecture. We also used the universal testing strategy, which used the person boxes provided in the datasets. Our trained model generated accurate prediction results for every person in the LSP test dataset.

4.2.3. Results on the LSP Dataset

Table 2 shows the *PCK*@0.2 prediction results, AUC, #Params, FLOPs, and FPS of our method and other top-performing methods on the LSP test dataset. The proposed network UEPDN-R1 and UEPDN-Stage 1 achieved 87.5 and 91.1 mean *PCK*@0.2 scores, respectively.

Table 2. *PCK*@0.2, AUC (%) rates, #Params, FLOPs, and FPS on the LSP testing dataset.

Methods	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean	AUC	#Params	FLOPs	FPS
CNGM [12]	90.6	79.2	67.9	63.4	69.5	71.0	64.2	72.3	47.3	–	–	–
ECN [47]	95.8	86.2	79.3	75.0	86.6	83.8	79.8	83.8	56.9	56.0 M	28 G	–
CPM [13]	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5	65.4	31.0 M	351 G	3.5
KGDFNN [48]	**98.2**	**94.4**	**91.8**	**89.3**	**94.7**	**95.0**	**93.5**	**93.9**	–	53.1 M	124 G	–
FPD [23]	97.3	92.3	86.8	84.2	91.9	92.2	90.9	90.8	64.3	3.0 M	9 G	–
UEPDN-Stage 1 (Ours)	97.3	92.8	88.8	86.1	91.2	91.5	89.9	91.1	**66.3**	3.8 M	8.4 G	4.0
UEPDN-R1 (Ours)	96.5	91.8	86.0	80.3	88.4	88.4	80.8	87.5	62.9	**2.75 M**	**6.2 G**	**5.3**

At the same time, they had low deployment costs. As our model used a re-parameterized structure and knowledge distillation technology to help the model be efficiently trained and flexibly deployed, we achieved good detection performance and deployment performance while minimally increasing the calculational costs. The visualized pose estimation results on the LSP dataset are shown in Figure 5. We clearly observe that our model obtained robust and exact detection results for images with various human poses and various complex backgrounds.

4.3. Pose Estimation on the UAV-Human Pose Estimation Dataset

4.3.1. Dataset and Performance Metric

The UAV-Human pose estimation dataset contains a total of 22,476 human images. Each image has 17 major body labeled keypoints. Specifically, we used 14 labeled keypoints for our model. At the same time, we used 16,288 human instances from the dataset for training and 6188 human instances for testing.

We applied the mean average precision (mAP) to evaluate the proposed method on this dataset. Additionally, we used #Params and FLOPs to measure the deployment performance of the proposed method.

Figure 5. Visualized results on the LSP dataset.

4.3.2. Training and Deployment Details

We used 180 epochs for the UAV-Human pose estimation training dataset. The other details of the training process were the same as those for the MPII. When training was complete, we used the first re-parameterized module as the deployment network to retain the simple architecture. Our trained model generated good prediction results for every person in the UAV-Human pose estimation testing dataset.

4.3.3. Results on the UAV-Human Pose Estimation Dataset

Table 3 shows the mAP prediction results, #Params, and FLOPs of our method and two prevalent methods on the UAV-Human pose estimation testing dataset. The proposed network UEPDN-R1 and UEPDN-Stage 1 achieved 54.8 and 56.3 mAP scores, respectively. Although the accuracy of our method was slightly lower than that of top-performing methods (e.g., HigherHRNet), our methods had lower deployment costs. At the same time, because our model used a re-parameterized structure and knowledge distillation technology to help the model be efficiently trained and flexibly deployed, it achieved good detection performance and deployment performance while minimally increasing the calculational costs. The visualized pose estimation results on the UAV-Human pose estimation dataset are shown in Figure 6. We clearly observe that our model obtained good detection results for images with various human poses.

Table 3. The mAP, #Params, and FLOPs on the UAV-Human pose estimation testing dataset.

Methods	mAP (%)	#Params	FLOPs
HigherHRNet [18]	56.5	28.6 M	47.9 G
RMPE [15]	**56.9**	59.7 M	–
UEPDN-Stage 1 (Ours)	56.3	3.8 M	8.4 G
UEPDN-R1 (Ours)	54.8	**2.75 M**	**6.2 G**

Figure 6. Visualized results on the UAV-Human pose estimation dataset.

4.4. Ablation Experiments

To illustrate the effectiveness of the proposed ultra-lightweight pose distillation method, we conducted ablation experiments based on the same hardware, software environment, and LSP test dataset used previously in this section.

4.4.1. Effect of Pose Distillation and Re-Parameterized Modules

The effect of using our pose distillation (PD) method and re-parameterized modules (RM) on detection results are displayed in Table 4. It clearly shows that our ultra-lightweight end-to-end pose distillation architecture helped the lightweight re-parameterized modules to achieve good detection performance. The reason for our good detection results were that our proposed pose distillation architecture learned extra helpful image feature information in cases with an incorrect image label and deficient image annotation, making model deployment more flexible. This suggests that the generic theory of knowledge distillation and the re-parameterized technique were effective in their application to the field of structured pose estimation.

Table 4. *PCK*@0.2 of the proposed pose distillation and re-parameterized modules on LSP test dataset.

PD	RM	R1	R2	R3	R4	(R5) Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6
×	×	–	–	–	–	89.5	91.1	91.8	92.1	92.1	92.1
√	×	–	–	–	–	90.7	91.4	91.8	92.1	92.2	92.1
×	√	85.0	88.3	90.0	90.6	90.9	91.1	91.7	91.9	92.1	92.1
√	√	87.5	88.8	90.2	90.6	91.1	91.2	91.2	91.4	91.6	91.2

√ means that it is used. × means that it is not used.

4.4.2. Effect of Training Stage Size

The effects of the training stage size on detection performance are displayed in Table 5. We selected three teacher models, stage size $s \in \{4, 5, 6\}$, to teach module R1, R2, and stage $s \in \{2, \ldots, S-1\}$, and used the re-parameterized module R1 as the deployment network to retain the simple architecture. We clearly observed that when training stage size $s = 6$, UEPDN obtained good results in terms of its deployment cost and detection accuracy. This suggests that a powerful teacher network substantially helps in training the target student model and obtaining good detection results.

Table 5. *PCK*@0.2 and #Params of module R1 based on different training stage size on the LSP test dataset.

Training Stage Size	Mean	#Params
Stage size s = 6	87.5	2.75 M
Stage size s = 5	87.2	2.75 M
Stage size s = 4	87.1	2.75 M

4.4.3. Effect of Deployment Network

The effects of the deployment network on the detection performance are displayed in Table 6. We set the deployment stage size from 1 to 6 and deployment re-parameterized module number from 1 to 4. We clearly observed that when the re-parameterized module number $r = 1$, UEPDN achieved good performance in terms of its deployment cost and detection accuracy.

Table 6. *PCK*@0.2, #Params, FLOPs, and FPS of different deployment networks on LSP test dataset.

Method	R1	R2	R3	R4	(R5) Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6
Mean	87.5	88.8	90.2	90.6	91.1	91.2	91.2	91.4	91.6	91.2
#Params	2.75 M	3.00 M	3.27 M	3.52 M	3.82 M	5.90 M	7.90 M	10.00 M	12.00 M	14.10 M
FLOPs	6.20 G	6.75 G	7.30 G	7.85 G	8.40 G	12.90 G	17.32 G	21.74 G	26.16 G	30.58 G
FPS	5.3	4.9	4.7	4.5	4.0	3.3	2.6	2.0	1.7	1.4

5. Discussion

Our proposed ultra-lightweight end-to-end pose distillation network architecture explores how to achieve good detection accuracy while compressing the model parameters as much as possible. We summarize the strengths of our approach in three points. First, we designed a lightweight end-to-end pose estimation network that learned from an increasingly refined sequential expert network in an online knowledge distillation manner, which reasonably used high-value information provided by image labels and the expert network to increase the training efficiency of the model. Second, we constructed an ultra-lightweight re-parameterized pose estimation subnetwork that used multi-module design with weight-sharing to improve the multi-scale image feature acquisition capability of the single-module design. Finally, when training was complete, we used the first re-parameterized module as the deployment network to retain the simplest architecture. As our model used a re-parameterized structure, flexible deployment was achieved using various numbers of re-parameterized modules, depending on actual requirements.

Extensive experimental results demonstrated the detection precision and low number of parameters of our method. This suggests that a novel network design based on a re-parameterized structure and online knowledge distillation technique is very helpful for obtaining good detection accuracy, compressing the model parameters, and improving the training efficiency of the model.

Although our ultra-lightweight model achieved good detection performance on three standard benchmark datasets, there were also some limitations to this study. Due to some

occluded keypoints and instances where people were close to each other, a few failures of our model on the MPII, LSP, and UAV-Human datasets occurred, which are displayed in the top, middle, and bottom rows of Figure 7, respectively. For images with complex scenes, such as overlapping people, cluttered backgrounds, and severe occlusion, it may be insufficient to only use the spatial context features extracted from keypoint features. Our proposed method is not in real-time and has not been deployed in realistic resource-limited devices. As real-time performance is necessary for multi-person pose estimation and human pose estimation in the field of video, our method is more suitable for single-person pose estimation and human pose estimation in the field of images captured from image sensors. Additionally, there have been some studies on distilling knowledge from other modalities, and they have achieved good performance. However, for the convenience of training on three standard benchmark datasets consisting of RGB images, our proposed method only focuses on distilling knowledge from RGB modalities to RGB modalities, and cannot deal with other modalities to RGB modalities. In the future, we plan to extend our work on distilling knowledge from other modalities to RGB modalities for good detection performance in complex scenes, and explore applications in realistic resource-limited devices.

Figure 7. Examples of failures on MPII (**top**), LSP (**middle**), and UAV-Human dataset (**bottom**).

6. Conclusions

In this paper, we proposed an ultra-lightweight end-to-end pose distillation network to improve human pose estimation performance. By learning from an increasingly refined sequential expert network in an online knowledge distillation manner, our one-stage lightweight pose estimation network achieved good detection results. We designed an ultra-lightweight re-parameterized pose estimation subnetwork that used multi-module design with weight-sharing to improve the multi-scale image feature acquisition capability of the single-module design. Finally, we used the first re-parameterized module as the deployment network to retain the simple architecture. Extensive experimental results demonstrated the detection precision and low quantity of required parameters of our method. Although the number of parameters was lower for UEPDN than other current lightweight pose estimation methods, we found that the prediction accuracy of the proposed model on primary datasets was slightly lower than that of current state-of-the-art human pose estimation methods. We hope that a re-parameterized structure introducing the knowledge distillation technique can make a contribution to applications in human pose estimation.

Author Contributions: Conceptualization, S.Z.; methodology, S.Z. and X.W.; data curation, L.C.; writing—original draft, S.Z.; project administration, B.Q.; funding acquisition, B.Q.; resources, B.Q.; validation, X.Y.; formal analysis, B.Q.; writing—review and editing, R.C. and L.C.; software, S.Z., X.Y., and X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Natural Science Foundation of Guangxi under grants 2019GXNSFDA185006 and 2019GXNSFDA185007; the National Natural Science Foundation of China under grant 62262006; the Guilin Science and Technology Development Program under grant 20210104-1; and the Guangxi Key Research and Development Program under grants AB17195053 and AD18281002.

Data Availability Statement: Publicly archived datasets used in the study are listed below. MPII: http://human-pose.mpi-inf.mpg.de/ (accessed on 15 February 2023); The Extended Leeds Sports Pose: http://sam.johnson.io/research/lspet.html (accessed on 11 November 2018).

Acknowledgments: We thank the anonymous reviewers whose comments helped improve the quality of this paper.

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

References

1. Felzenszwalb, P.F.; Huttenlocher, D.P. Pictorial structures for object recognition. *Int. J. Comput. Vis.* **2005**, *61*, 55–79. [CrossRef]
2. Andriluka, M.; Roth, S.; Schiele, B. Pictorial structures revisited: People detection and articulated pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
3. Wang, F.; Li, Y. Learning visual symbols for parsing human poses in images. In Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, 3–9 August 2013.
4. Pishchulin, L.; Andriluka, M.; Gehler, P.V.; Schiele, B. Poselet conditioned pictorial structures. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013.
5. Sapp, B.; Toshev, A.; Taskar, B. Cascaded models for articulated pose estimation. In Proceedings of the European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010.
6. Sapp, B.; Taskar, B. Modec: Multimodal decomposable models for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013.
7. Chen, X.; Yuille, A. Articulated pose estimation by a graphical model with image dependent pairwise relations. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014.
8. Cherian, A.; Mairal, J.; Alahari, K.; Schmid, C. Mixing body-part sequences for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
9. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
10. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
11. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.

12. Tompson, J.; Jain, A.; LeCun, Y.; Bregler, C. Joint training of a convolutional network and a graphical model for human pose estimation. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014.
13. Wei, S.E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
14. 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.
15. Fang, H.; Xie, S.; Tai, Y.; Lu, C. RMPE: Regional multi-person pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
16. Nie, X.; Li, Y.; Luo, L.; Zhang, N.; Feng, J. Dynamic kernel distillation for efficient pose estimation in videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
17. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
18. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
19. Zhang, J.; Chen, Z.; Tao, D. Towards high performance human keypoint detection. *Int. J. Comput. Vis.* **2021**, *129*, 2639–2662. [CrossRef]
20. Dong, H.; Wang, G.; Chen, C.; Zhang, X. RefinePose: Towards more refined human pose estimation. *Electronics* **2022**, *11*, 4060. [CrossRef]
21. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
22. Kato, N.; Li, T.; Nishino, K.; Uchida, Y. Improving multi-person pose estimation using label correction. *arXiv* **2018**, arXiv:1811.03331.
23. Zhang, F.; Zhu, X.; Ye, M. Fast human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
24. Qiang, B.; Zhai, Y.; Chen, J.; Xie, W.; Zheng, H.; Wang, X.; Zhang, S. Lightweight human skeleton key point detection model based on improved convolutional pose machines and SqueezeNet. *J. Comput. Appl.* **2020**, *40*, 1806–1811.
25. Weinzaepfel, P.; Brégier, R.; Combaluzier, H.; Leroy, V.; Rogez, G. DOPE: Distillation of part experts for whole-body 3D pose estimation in the wild. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
26. Zhong, F.; Li, M.; Zhang, K.; Hu, J.; Liu, L. DSPNet: A low computational-cost network for human pose estimation. *Neurocomputing* **2021**, *423*, 327–335. [CrossRef]
27. Wang, W.; Zhang, K.; Ren, H.; Wei, D.; Gao, Y.; Liu, J. UULPN: An ultra-lightweight network for human pose estimation based on unbiased data processing. *Neurocomputing* **2022**, 480, 220–233. [CrossRef]
28. Bulat, A.; Tzimiropoulos, G. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
29. Xiao, B.; Wu, H.; Wei, Y. Simple baselines for human pose estimation and tracking. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
30. Martinez, G.H.; Raaj, Y.; Idrees, H.; Xiang, D.; Joo, H.; Simon, T.; Sheikh, Y. Single-network whole-body pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
31. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 172–186. [CrossRef] [PubMed]
32. Wang, J.; Luo, Z. Pointless pose: Part affinity field-based 3D pose estimation without detecting keypoints. *Electronics* **2021**, 10, 929. [CrossRef]
33. Li, Z.; Ye, J.; Song, M.; Huang, Y.; Pan, Z. Online knowledge distillation for efficient pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.
34. Xiao, Y.; Wang, X.; He, M.; Jin, L.; Song, M.; Zhao, J. A compact and powerful single-stage network for multi-person pose estimation. *Electronics* **2023**, *12*, 857. [CrossRef]
35. Wang, J.R.; Li, X.; Ling, C.X. Pelee: A real-time object detection system on mobile devices. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018.
36. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2D human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23-28 June 2014.
37. Johnson, S.; Everingham, M. Clustered pose and nonlinear appearance models for human pose estimation. In Proceedings of the British Machine Vision Conference, Aberystwyth, UK, 31 August–3 September 2010.
38. Li, T.; Liu, J.; Zhang, W.; Ni, Y.; Wang, W.; Li, Z. UAV-Human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 19–25 June 2021.
39. Li, Y.; Shi, Q.; Song, J.; Yang, F. Human pose estimation via dynamic information transfer. *Electronics* **2023**, *12*, 695. [CrossRef]
40. Jia, Y.; Shelhamer, E.; Donahue, J. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014.
41. Kingma, P.D.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

42. Li, Y.; Yang, S.; Liu, P.; Zhang, S.; Wang, Y.; Wang, Z.; Yang, W.; Xia, S.T. SimCC: A Simple Coordinate Classification Perspective for Human Pose Estimation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022.
43. Li, K.; Wang, S.; Zhang, X.; Xu, Y.; Xu, W.; Tu, Z. Pose Recognition with Cascade Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 19–25 June 2021.
44. Li, Y.; Zhang, S.; Wang, Z.; Yang, S.; Yang, W.; Xia, S.T.; Zhou, E. TokenPose: Learning Keypoint Tokens for Human Pose Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.
45. Geng, Z.; Wang, C.; Wei, Y.; Liu, Z.; Li, H.; Hu, H. Human Pose as Compositional Tokens. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 20–22 June 2023.
46. Yu, C.; Xiao, B.; Gao, C.; Yuan, L.; Zhang, L.; Sang, N.; Wang, J. Lite-HRNet: A Lightweight High-Resolution Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 19–25 June 2021.
47. Rafi, U.; Leibe, B.; Gall, J.; Kostrikov, I. An efficient convolutional network for human pose estimation. In Proceedings of the British Machine Vision Conference, York, UK, 19–22 September 2016.
48. Ning, G.; Zhang, Z.; He, Z. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE Trans. Multim.* **2018**, *20*, 1246–1259. [CrossRef]