applied sciences

MDPI

Article

Two-Dimensional Ultra Light-Weight Infant Pose Estimation with Single Branch Network

Viet Dung Nguyen [1,*], Thinh Nguyen-Quang [1], Minh Duc Nguyen [1,2,3,4], Dang Hung Phan [5] and Ngoc Dung Bui [6,*]

[1] Biomedical Engineering Group, School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi 100000, Vietnam; thinh.nq182803@sis.hust.edu.vn (T.N.-Q.); minh.d.nguyen@sydney.edu.au (M.D.N.)
[2] Westmead Applied Research Centre, The University of Sydney, Sydney 2145, Australia
[3] School of Biomedical Engineering, The University of Sydney, Sydney 2006, Australia
[4] Cardiology Department, Westmead Hospital, Westmead 2145, Australia
[5] Center for Information Technology, Hanoi University of Industry, Hanoi 100000, Vietnam; phanhung@haui.edu.vn
[6] Faculty of Information Technology, University of Transport and Communications, Hanoi 100000, Vietnam
* Correspondence: dung.nguyenviet1@hust.edu.vn (V.D.N.); dnbui@utc.edu.vn (N.D.B.); Tel.: +84-913-045-130

Abstract: Motivated by the increasing interest in clinical studies focused on infant movements and poses, this research addresses the limited emphasis on speed and efficiency in existing 2D and 3D pose estimation methods, particularly concerning infant datasets. The scarcity of publicly available infant data poses a significant challenge. In response, we aim to develop a lightweight pose estimation model tailored for edge devices and CPUs. Drawing inspiration from the OpenPose-2016 approach, we refine the algorithm's architecture, focusing on 2D image training. The resulting model, with 4.09 million parameters, features a single-branch structure. During execution, it achieves an algorithmic complexity of 8.97 giga floating-point operations per second (GFLOPS), enabling operation at approximately 23 frames per second on a Core i5-10400f processor.Notably, this approach balances compact dimensions with superior performance on our self-collected infant dataset. We anticipate that this pragmatic methodology establishes a robust foundation, addressing the need for speed and efficiency in infant pose estimation and providing favorable conditions for future research in this application.

Keywords: pose estimation; infant posture; computer vision; lightweight architecture

Citation: Nguyen, V.D.; Nguyen-Quang, T.; Nguyen, M.D.; Phan, D.H.; Bui, N.D. Two-Dimensional Ultra Light-Weight Infant Pose Estimation with Single Branch Network. *Appl. Sci.* **2024**, *14*, 3491. https://doi.org/10.3390/app14083491

Academic Editor: Luis Javier Garcia Villalba

Received: 11 March 2024
Revised: 10 April 2024
Accepted: 12 April 2024
Published: 20 April 2024

1. Introduction

Computer vision and artificial intelligence have become indispensable in numerous domains, providing invaluable assistance to individuals. Within the realm of sports and security, computer vision plays a pivotal role in the monitoring, recognition, and analysis of human behavior and motion. Similarly, the monitoring, tracking, and analysis of infant movements are subjects of great significance, bestowing substantial benefits for the management of young children/infants and yielding medical advantages. This technology enables specialists to scrutinize the movements of infants, and identify early warning signs pertaining to motor development and associated disorders.

In the analysis of infant movements, methods such as live monitoring or video data analysis are being implemented. In addition, the analysis is implemented through the camera with labeling points directed at the body joints. However, manual labeling of joints per data collection is required for the analysis of infant data, which can be time-consuming and challenging to set up. An alternative approach involves the use of artificial intelligence models for pose estimation. This approach is more convenient and allows for more data to be collected from newborns. Recent research based on deep learning

approaches has laid the foundation for computer-supervised motion assessment tools for early intervention and diagnosis of motor neuronal disorders. The utilization of General Movements Assessment (GMA) [1] has proven instrumental in the differentiation of subtle movements from a myriad of smaller movements, offering an early diagnostic avenue for cerebral palsy. In their work, the authors harnessed postural estimation techniques, notably OpenPose, to reconstruct infant posture and meticulously analyze movements for the purpose of cerebral palsy diagnosis. However, the existing methods, while effective, present significant drawbacks in terms of time consumption and computational expenses. The intricate task of developing a novel, compact model is further complicated by the limited availability of public datasets specifically designed for newborns. This scarcity poses a formidable challenge in the pursuit of advancing the field of movement assessment and diagnosis for infants.

In the past decade, impressive progress has been made with the accuracy of pose estimation methods [2–7]. And studies have shifted their focus to developing methods that enhance the speed of these methods. Despite this, the accuracy-focused methods that rely on complex models consume significant computational resources as well as execution equipment resources. Although methods to improve speed have emerged, they often require trade-offs with accuracy. The majority of pose estimation methods can be broadly categorized into two main approaches: top-down and bottom-up. Most of the top-down methods in reference [8] involve detecting the person in the frame first and then estimating the position of each subject. On the other hand, the bottom-up method in this briefing review [9] detects the human keypoints present in the frame and then connects the relevant keypoints to form a human skeleton. The fastest method with high accuracy has been exported in [10], the authors reported this method running at 23 fps with a graphics card for 3 people images and when the number of people if the number of people increases to 20, it will only run at 15 frames per second. Bottom-up is considered to have better stability in the case of a variable number of human objects or error detection (wrong detection of non-human objects).

Pose estimation: The majority of pose estimation methods can be broadly categorized into two main approaches: top-down and bottom-up. The top-down method involves detecting the person in the frame first and then estimating the position of each subject. On the other hand, the bottom-up method detects the human keypoints present in the frame and then connects the relevant keypoints to form a human skeleton. The fastest method with high accuracy was developed in [10]; the authors reported this method running at 23 fps with a graphics card for images of three people, and if the number of people increases to 20, it will only run at 15 frames per second. However, bottom-up approach of Openpose can run stably and immutably when running with a variable number of people. Therefore, we prioritize using the bottom-up approach and improve the model of Openpose to be faster and lighter.

Infant Pose Estimation: Presently, approaches to analyzing, recognizing, and clinically researching infant posture heavily rely on visual observation by experts, either in real-time or through recorded sessions [11]. There appear significant differences in body proportions and pose flexibility between infants and adults, creating significant challenges for pose estimation models. Most existing models, often trained on adult datasets due to accessibility, do not exhibit optimal performance on infant datasets. Compounding this issue is the limited availability of infant data, often withheld for confidentiality reasons, posing a major obstacle to training models from scratch. Consequently, only a handful of research efforts have ventured into automating postural monitoring of infants in recorded videos. In a notable attempt to address this challenge, a proposal was presented in [12] to automatically extract key points of infants. The author employed the OpenPose-2019 method [13] for this purpose. However, the accuracy of this approach was compromised due to the scarcity of databases and the inherent complexity of infant movements. A contrasting approach is exemplified by AggPose [14], which boasts high efficiency with minimal computational overhead, adopting a top-down strategy. Nonetheless, it is acknowledged that the top-down approach may exhibit reduced stability compared to the

bottom-up approach, especially in scenarios involving multiple objects within the image. The pursuit of an accurate and efficient solution for infant pose estimation remains a dynamic area of research, balancing the intricacies of infant movements with the limitations of available datasets and computational methodologies.

Navigating the intricate trade-offs between performance and size while optimizing the complexity and parameters of a model poses a challenging endeavor. The study conducted by [15] stands as a noteworthy example where computational costs were successfully curtailed and performance was enhanced. This was achieved through a strategic alteration in the architecture, specifically replacing the original model's 5×5 and 7×7 convolutional layers with two overlapping 3×3 layers. This substitution not only streamlined computational expenses but also yielded an augmentation in the overall efficacy of the method. Building on this paradigm, the techniques outlined in [16] present a broader perspective on improvement methodologies across functions. One consistent approach highlighted in this study involves the systematic replacement of large convolutional layers with overlapping smaller layers, thereby contributing to enhanced efficiency and optimized model performance.

Despite the increasing amount of research related to human pose estimation and real-time improvement in running, current methods have not achieved high accuracy for the infant dataset. Current research has been exclusively conducted on adult datasets, whereas there is a significant difference in body proportions between infants and adults. There need to be multiple foundational components to determine a virtual skeleton (pose estimation/body part localization), such as human object detection, body joint detection, and establishing relationships between the joints. Therefore, current methods often utilize multiple branches or at least two independent branches to determine separate foundations and create a complete framework of a human subject. This paper introduces a new definition of the single-branch model for infant pose estimation. We build and improve models based on modern models with lighter structures consisting of depthwise separable convolution layers and residual blocks. The accuracy and complexity of the proposed method are competitive compared with modern methods on a self-collected dataset.

The summary of the contributions:

- Propose the definition of a single-branch structure that shares convolutional layers and features, which are based on bottom-up approaches with a block-by-block process, in which blocks are applied using the residual block technique.
- Reducing computational complexity of the network by utilizing a single-branch architecture, sharing convolutional layers and features for both confidence map inference and part affinity field inference.
- Propose a self-collected dataset on infant poses.

The structure of this study is as follows: Section 2 outlines the research scope, detailing data collection methods, utilization strategies, as well as implementation and evaluation methodologies. Section 3 delves into the evaluation metrics employed, alongside the experimental setup and results. In Section 4, discussions and explanations are provided regarding the obtained results, accompanied by an exploration of their limitations. Finally, the conclusion is presented in Section 5.

2. Materials and Methods

The application of deep learning to the research and analysis of the movements of newborns, achieved through the automatic detection of their joints and tracking of movement trajectories, holds immense potential benefits. Deep learning, particularly convolutional networks (CNN), stands out as an effective technique in the domain of image detection and recognition. Leveraging this technology for the automatic recognition of the body's joints has the potential to simplify and broaden research on newborn movements, making it more accessible and widespread. This, in turn, facilitates the detection of risks associated with dynamic movement in a diverse range of newborns.

The primary objective of this study is to design a lightweight model suitable for deployment on edge devices while retaining depth and without compromising significantly on accuracy. The proposed model is trained initially with an adult dataset, leveraging the wealth of available data in this domain. Subsequently, the model undergoes specific training using a self-collected dataset focused on infants' postures. This sequential training approach aims to transfer knowledge from the adult dataset to the infant dataset, enhancing the model's ability to accurately detect and track the joints of newborns [17].

In summary, this study bridges the gap between deep learning advancements and infant movement analysis, striving to create a versatile and accurate model that can contribute to the early detection of movement-related risks in newborns. The sequential training methodology, combined with a focus on lightweight design for edge device deployment, underscores the pragmatic and impactful nature of this research endeavor.

2.1. Network Design

The bottom-up method is implemented with 2 parallel inference branches. These branches share features extracted from the backbone; one branch infers the joints of all bodies in the figure, and one branch is used to group joints of the same body together. Instead of using 2 parallel branches like current bottom-up approaches, alternative single-branch blocks with a lighter structure were used to form the proposed model. Specifically, a network architecture was developed that leverages a single branch (Figure 1) for both confidence maps (CMs) and connects the main points of a single human object through the use of part affinity fields (PAFs) [18].

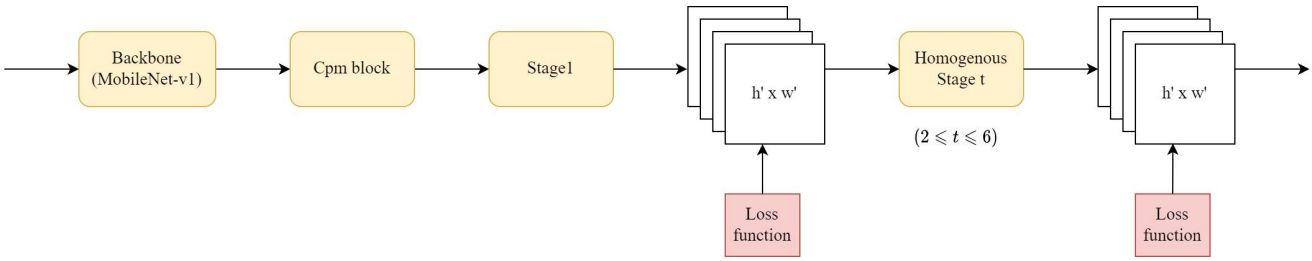

Figure 1. Model pipeline designed with only a single branch. The number of stages can be customized with $2 \leqslant t \leqslant 6$.

Performance comparison tests are conducted on the adult set to evaluate the proposed model. The implementation of this dataset aims to save time and eliminate the need for forwarding reference models with the infant dataset. Once the suitable model has been selected based on the performance comparison tests, the pre-trained model with the adult dataset proceeded with transfer-learning on the self-collected infant dataset. Experiments of the final model for comparison accuracy and performance were conducted on the infant set.

2.1.1. Feature Extractor

Networks with a much lighter structure than VGG-19 have been proposed; they have similar or better classification accuracy than VGG-19 [19–21]. To extract features, the networks from the MobileNet series are highly recommended because of their light structure and the similar accuracy and performance between the SOTA method and VGG-19, VGG-16, AlexNET, DenseNet, DetNet [22,23]. Accuracy and complexity were considered trade-offs before making the decision to use the first 12 layers of MobileNet-v1 [20] as the backbone, and the comparison results and experimental setup are outlined in section IV.

The preservation of spatial resolution is crucial to avoiding a significant reduction in accuracy when the layers remain the same until the deepest layer matches the output resolution. To address this issue, a convolutional accumulation [17] is incorporated and changed in the backbone block to save spatial resolution and reuse essential weights. By doing so, the number of participants required for the task can be minimized. Specifically, there is a difference in origin to MobileNet-v1; the backbone in Figure 2 shows adjustments

by changing the 7th layer stride to 1, setting the dilated factor of the 8th layer to 2, and padding to 2 to preserve the spatial resolution.

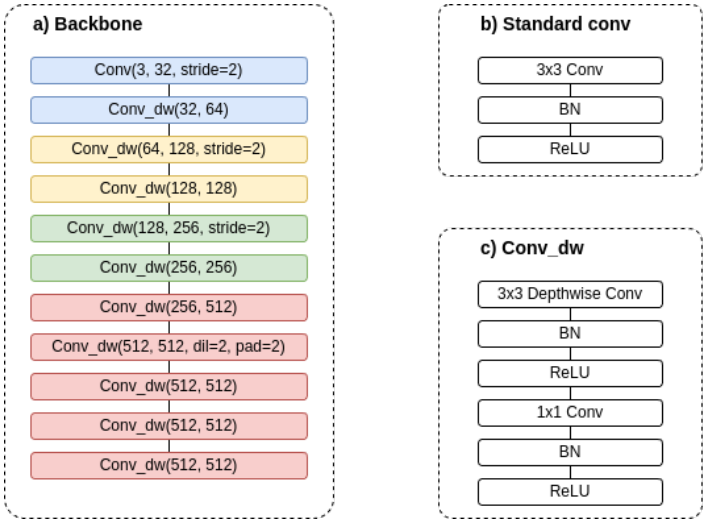

Figure 2. (**a**) First twelve layers of MobileNet-v1 for the backbone; (**b**) standard convolution layer with Batchnorm and ReLU; (**c**) depthwise separable convolution layer with Depthwise, Pointwise, Batchnorm, and ReLU.

A *cpm block* is incorporated after the first 12 layers of modified MobileNetv1 (Figure 3). The first step involves adjusting the number of channels using the initial 1×1 convolutional layer, followed by the passing of the output through a set of 3 *conv_dw_no_bn* layers. Abruptly reducing the number of features from 512 to 19 in CMs features (18 keypoints and 1 background) and from 512 to 38 in PAFs features can result in the loss of important input image information and beneficial features. Therefore, the 3 conv_dw_no_bn layers play a vital role in enhancing and preserving essential characteristics. Finally, the features are integrated using the last 3×3 convolutional layer.

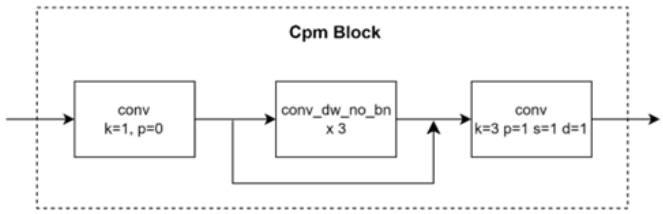

Figure 3. Cpm block add after the first 12 layers of MobileNet-v1 for feature down-sampling.

2.1.2. Stages

Network architectures employing parallel branches are commonly structured to conduct separate inferences concurrently. OpenPose, a prime example of such a design, incorporates two independent inference branches. Specifically, in the OpenPose architecture, convolution layers within stages 2–6 utilize kernels of up to 7×7 in size (Figure 4). This parallelized approach enhances the model's capacity to capture intricate spatial relationships and nuances across various stages of the network, contributing to its robustness in tasks such as pose estimation where capturing detailed features is crucial. The utilization of parallel branches allows the model to process information from multiple perspectives simultaneously, fostering a more comprehensive understanding of the input data.

To generate accurate estimates of heat maps and affinity parts, each stage requires feature extraction from the backbone, along with prior estimates of the two necessary fields. To streamline the operations and expedite the reasoning process, we opted to utilize the same convolutions in the stages, resulting in a single branch (Figure 5). Customizable homogenous stages repeat t times with $2 \leqslant t \leqslant 6$. This stage has a homogenization function

and improves the accuracy, but this improvement is not significant, so these layers can be removed to reduce the computational cost.

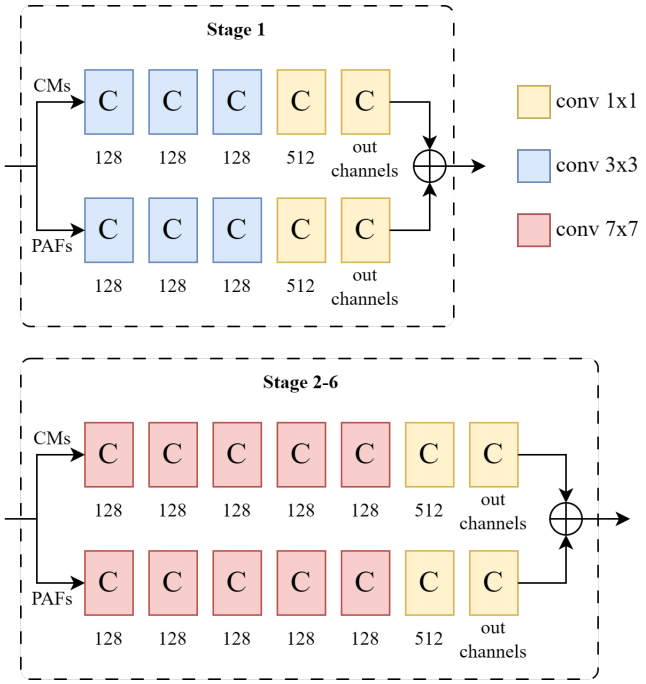

Figure 4. Design of stages in Openpose with a parallel branch.

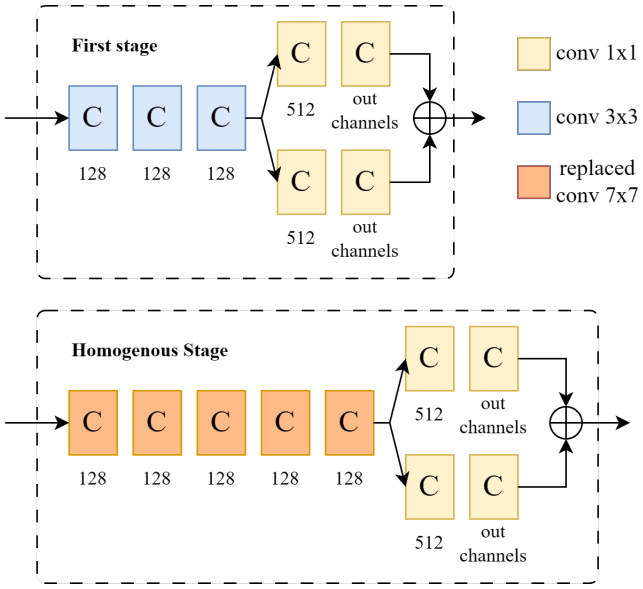

Figure 5. Single prediction branch for the first stage and homogenous stages.

The modification in our research involves the substitution of 7×7 convolutions with a sequence of three consecutive convolutions: 1×1, 3×3, and 3×3 with a dilated factor of 2 (as illustrated in Figure 6).This strategic replacement results in a noteworthy reduction in model complexity by approximately 2.5 times, while still preserving the receptive field, as observed in related work [3]. To further enhance the training performance of the deep network, skip connections, as introduced in [24], are incorporated. Instead of relying on 5 convolutional layers with a kernel size of 7×7, our approach involves the utilization of 15 replaced convolutional layers (organized into 5 replaced blocks). This augmentation proves effective in capturing intricate features and representations, empowering the model to handle complex and hierarchical patterns within the data.

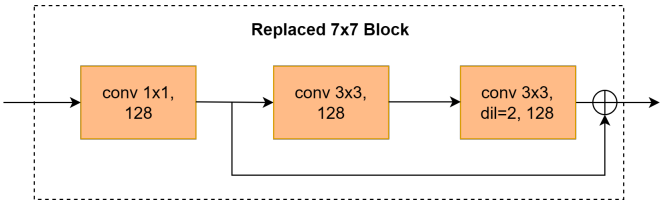

Figure 6. Convolutional block for replacement of 7×7 convolutions in refinement.

2.2. Experiment Setup

Two-Dimensional Pose dataset. The datasets employed in this research are self-collected infant pose data and the MS COCO dataset [25], specifically a collection of infant datasets sourced from YouTube. The MS COCO dataset comprises a diverse range of human activities and postures, featuring 200,000 image data points segregated into train/validation/test sets. The training and validation sets are annotated with up to 1.7 million labeled keypoints. One of the major challenges that we had to contend with was the unavailability of suitable data on infants. Infant data, due to its sensitive nature, is not openly available in existing base datasets. Furthermore, experimental research and studies have been conducted using adult images, and pre-training with adult datasets is believed to be advantageous in carrying over keypoint features, but with larger dimensions. In order to overcome this challenge, we performed a manual data collection of infant data from YouTube and subsequently carried out manual labeling. The dataset that was obtained comprises seven brief video segments, featuring seven distinct infants in the supine position. The videos were formatted in various resolutions and frame rates.

Infant Data Anotation. Infant data were collected by the authors on YouTube. The dataset includes 8 videos with 10 s per video corresponding to 8 distinct newborns/infants (Table 1). They are labeled to include 18 keypoints, and the labeled file is saved in .json format, with occluded body joint points excluded. Following the labeling process using the VaticJS tool, we derived a dataset of 1952 images, each with approximately 35,000 labeled keypoints. The dataset was segregated into train/validation/test sets, with the test set being a distinct dataset that featured a single infant.

Table 1. Summary of the self-collected infant dataset.

	Num. of Videos	FPS	Durations	Num. of Subjects	Ages
Infant-dataset	8	∼24–25	10 s	8	<15 weeks

2.3. Implementation Details

2.3.1. Loss Function

To guide the network in the right direction, repeating the CMs and PAFs prediction operations, the loss function is placed at the end of each stage. The loss function of stage t is the sum of the distances between the prediction and the labeling of both the predicted CMs and PAFs:

$$\mathcal{L}^t = \sum_c \sum_p W_c(p) \cdot \|y_c(p) - \tilde{y_c}(p)\|_2^2 \quad (1)$$

where $y_c(p)$ is the groundtruth corresponding to the prediction c, $c \in [CMs, PAFs]$. Ground truths of $y_c(p)$ for confidence maps and part affinity fields are inspired by [18]. Here, we are faced with the problem of miss-matching labels and incomplete labeling, so to solve this problem, we will use the loss function in space. Specifically, the use of the binary mask $W(p)$, this mask will avoid the penalty on true positive predictions. Placing the loss function helps to maintain and supplement the slope periodically.

$$W(p) = \begin{cases} 0 & \text{where pixel } p \text{ is the missed annotation} \\ 1 & \text{where pixel } p \text{ is annotated} \end{cases} \quad (2)$$

And the overall process object loss function through T stages is:

$$\mathcal{L} = \sum_{t=1}^{T} \mathcal{L}^t. \tag{3}$$

2.3.2. Ground-Truths

Confidence Maps

To compute the loss term $\mathcal{L}^t$ as delineated in Equation (1) within the training regimen, we establish a ground truth for the confidence map $\mathcal{J}^*$ by leveraging the body joint labels associated with the 2D data. This process aims to imbue the confidence map with precise body joint localization information across designated pixels. Specifically, for each human entity discernible within the image, multiple confidence maps $\mathcal{J}^*_{m,n}$ are generated, wherein the ideal manifestation encompasses a solitary confidence peak denoting the corresponding joint match. If multiple individuals are present within the image frame, the expectation is for the generation of distinct confidence peaks $\overline{P}_{m,n}$ within the confidence maps for each visible body part m of each person n.

The confidence map $\mathcal{J}^*_{m,n}$ generated for each person n is constructed such that $\overline{P}_{m,n}$ represents the position of the pixel coinciding with the body joint m of person k in the image. Subsequently, the value of pixel p in the confidence map is defined as,

$$\mathcal{J}^*_{m,n}(p) = \exp\Big(-\frac{\left\|p - \overline{P}_{m,n}\right\|_2^2}{\sigma^2}\Big) \tag{4}$$

The confidence map adopts the structure of a Gaussian normal distribution, where σ represents the radius—a coefficient regulating the spread of the Gaussian peak. The network's output confidence map is a predictive aggregation of the individual confidence maps $\mathcal{J}^*_{m,n}$ via a max operator,

$$\mathcal{J}^*_m(p) = \max_n \mathcal{J}^*_{m,n}(p) \tag{5}$$

The confidence map is synthesized by combining the confidence maps corresponding to each joint by the above equation. This is implemented by taking the maximum value when overlaying the gradient circles (confidence maps) on top of each other. In other words, when the Gaussian distributions corresponding to two joints intersect, we take the value at point p as $max(G_1(p), G_2(p))$.

Part Affinity Fields

In the scenario where two gradient circles of confidence maps intersect, we opt to compute the maximum value over the union of the two gradient circles instead of averaging. This approach is warranted as there are still distinct body joints present, and it is imperative to distinctly discern the two vertices corresponding to the respective confidence maps.

In the amalgamation of identified joints into a cohesive virtual skeletal framework emblematic of human posture, the intricate task lies in accurately discerning the association of joints belonging to the same anatomical entity amidst potential misalignments between joints of distinct individuals. This challenge finds resolution through the employment of Part Affinity Fields (PAFs). PAF encapsulates essential positional and directional attributes for each limb, delineating the segment between two successive body joints. The genesis of PAFs stems from the meticulous analysis of supplementary linking cues between adjacent body joints, resulting in a finely-grained 2D vector representation for each limb. Within this representation, each vector is anchored at a pixel denoting a specific limb, collectively portraying the directional orientation from one joint terminus to the other. Thus, each limb is meticulously associated with its corresponding PAF map.

Limb c is defined by 2 body joints, m_1 and m_2, corresponding with ground-truth position $\overline{P}_{m1,n}$ and $\overline{P}_{m2,n}$ of person n in the image. Point p is one of the supplementary

linking cues if it lies on limb c. In this context, the value at the point $L^*_{c,n}(p)$ denotes the vector pointing from joint m_1 to m_2 for a given limb c. All points not associated with limb c are assigned zero.

$$L^*_{c,n}(p) = \begin{cases} \vec{u} & \text{if } p \text{ lies on limb } c, n \\ 0 & \text{otherpoint} \end{cases} \tag{6}$$

Here, vector unit $\vec{u}$ of the direct vector of a limb is defined as:

$$\vec{u} = (\overline{P}_{m2,n} - \overline{P}_{m1,n}) / \left\| \overline{P}_{m2,n} - \overline{P}_{m1,n} \right\|_2 \tag{7}$$

The set of points situated on the limb is characterized by those points lying within the defined region bounded by the distance threshold. Specifically, for a given point p, it is deemed to belong to this set if it satisfies the subsequent condition:

$$0 \leq \vec{u} \cdot (p - \overline{P}_{m1,n}) \leq l_{c,n}$$

and

$$\left| \vec{n} \cdot (p - \overline{P}_{m1,n}) \right| \leq \sigma_l$$

Distance thresholds are in pixels, where σ_l is the width of the limb (ideally the width should be taken in the smallest width of the limb) and $l_{c,n}$ is the length of the limb (Euclidean distance of $\overline{P}_{m2,n}$ and $\overline{P}_{m1,n}$, $\vec{n}$ is the unit vector perpendicular to $\vec{u}$). The ground truth for the part affinity field is constructed by averaging the part affinity fields corresponding to the limbs of the individuals present in the image,

$$L^*_c(p) = \frac{1}{n_{\vec{u}_c}(p)} \sum_n L^*_{c,n}(p) \tag{8}$$

where $n_{\vec{u}_c}$ is the number of $\vec{u}$ of limb c at point p across all n people in the image. Given the study's objective of estimating the pose of a limited number of objects within the image, determining the virtual limb can rely solely on the directional vector, representing the connection between the joints of the limb.

We establish the relationship between joints by computing the integral along the corresponding PAF (Part Affinity Field) over the segment connecting the joint candidate positions. Specifically, for the joint candidate positions d_{j_1} and d_{j_1}, we measure the strong reliability of the relationship between the two joints using the expression:

$$E = \int_{e=0}^{e=1} L_c(p(e)) \cdot \frac{d_{j_1} - d_{j_2}}{\left\| d_{j_1} - d_{j_2} \right\|_2} de \tag{9}$$

Function $p(e)$ is used to interpolate positions along the line segment connecting two candidate joint positions. It is a crucial part of computing the confidence in associating these two positions, $0 \leq e \leq 1$:

$$p(e)) = (1 - e)d_{j_1} + ed_{j_2} \tag{10}$$

- When $e = 0$, $p(e)$ corresponds to the position of the first joint candidate d_{j_1}.
- When $e = 1$, $p(e)$ corresponds to the position of the first joint candidate d_{j_2}.
- When e between 0 and 1, $p(e)$ represents a point on the line segment connecting the two joint candidate positions, calculated by linear interpolation between these two positions.

2.3.3. Implementation Settings

We follow the proposal from the bottom-up approach. The detector is used to estimate the keypoints and the matching link vector fields. COCO val and self-collected infant val sets are used to evaluate the detection results.

Data Augmentation. The diversity of data is enhanced with composite augmentation: hue and saturation change of −30 to 30 out of 255; brightness change of −30 to 30 out of 255; Gaussian noise up to 3% of 255; scaling (−25%, +25%); vertical random flip (50%); random rotation (−30°, +30°). The image size of the input is 368 × 368.

Training. The model is trained on COCO with the CMs supervisor and the PAFs supervisor. To save on computational costs, the first 12 layers of MobileNet are fixed and only cpm is trained—stages in which the learning rate is set at 2 times the base learning rate. The base learning rate and weight decay are set to 4e-5, 5e-4, respectively. With COCO2017, the batchsize is set to 64, and it is trained with 250 epochs. During the learning process, Adam [26] with a multi-step learning rate decay is chosen, which drops at epochs [100, 180, 210]. The learning rate is warmed up with the first 100 epochs, and the learning rate is reduced with $\gamma = 0.333$. The number of homogenous stages is set to 1, so only stage 1 is used to initialize into a homogenous stage for identification and prediction.

Transfer learning. At this phase, the pre-trained model is forward-trained with a self-collected infant dataset. The parameters of this phase are kept as the training phase, only changing the batchsize to 4 and the number of epochs to 230.

3. Results

3.1. Evaluation Metrics

We adhere to the evaluation methodologies of the COCO challenge [25], which defines the object keypoint similarity (OKS) and mean average precision (AP) over ten thresholds as the evaluation value. The di is the Euclidean distance between each corresponding detected keypoint and ground truth, s is defined as the square root of the object segment area, and k_i is substantial for different keypoints. The keypoints on a person's body (shoulders, knees, hips, etc.) tend to have a much larger value than on a person's head (eyes, nose, ears). The OKS is defined as:

$$OKS = \frac{\sum_i \left[\exp\left(-\frac{d_i^2}{2s^2k_i^2} \cdot \delta(v_i > 0)\right)\right]}{\sum_i \left[\delta(v_i > 0)\right]} \tag{11}$$

The number of parameters (params) and computational complexity (GFLOPs) are used to analyze and evaluate the size and performance of the proposed model. Those metrics are compared with current methods and their accuracy to clearly see the trade-offs between performance and accuracy.

3.2. Complexity Analysis

Table 2 compares the average precision between Openpose models. The experimental setup involved systematically substituting the backbone one at a time and conducting comparisons to identify the most optimal model, striking a balance between accuracy and computational efficiency. The first 12 layers of MobileNet-v1 are chosen for the backbone.

Table 2. Component selection from the MobileNet series for feature extractor (model's build with 2 branch architecture of Openpose CMU).

	AP, %	GFLOPs
First 6 layers (v1)	37.6	23.3
Dilated first 12 layers (v1)	43.5	27.7
Dilated first 13 layers (v1)	44.7	31.3
First 6 bottle-neck layers (v2)	39.5	27.2

The total number of stages of OpenPose is six, and experiments [13,18] have shown that the accuracy does not increase significantly from stage 2 to stage 6. This is the basis for us to be able to omit stages that are homogeneous and obtain the final network with a structure of only two stages. These two stages have better performances than the six stages of the OpenPose network because of their deeper design.

The OpenPose experiment was set up as a reference for block-to-block comparisons in terms of accuracy/complexity. The input resolution of the network is 368 × 68, which is the same as the input of the first 10 layers of VGG-19. The results suggest that the precision of stages 2–6 are homogenized and have slight increases in accuracy (Table 3). Thus, in terms of insight, we experienced that it is possible to remove a few of the homogenous stages and increase the depth of the remaining homogenous stages for the same accuracy.

Table 3. Accuracy/complexity of OpenPose through stages on the COCO val dataset.

Block	AP%	GFLOPs	GFLOPs Total
Feature extractor	–	40.9	40.9
Stage 1	38.6	2.2	43.1
Homogenous stage 2	53.3	18.6	61.7
Homogenous stage 3	58.1	18.6	80.3
Homogenous stage 4	60.1	18.6	98.9
Homogenous stage 5	61.5	18.6	117.5
Homogenous stage 6	62.6	18.6	136.1

In our proposal, the complexity of the algorithm and the number of parameters of the proposed algorithm have been reduced, as shown in Table 4. For the proposed method, the feature extractor (Backbone + Cpm block) is reduced to 4.28 GFLOPs, equivalent to 10.5%, reduced to 58.2%, with Stage 1 and each homogenous stage block reduced to 18.3%. The above comparisons are compared with experiments on the OpenPose-2016 [18] method. The proposed method exhibits a computational complexity of 8.97 GFLOPs and encompasses approximately 4.09 million parameters. These attributes empower the method with the capability to execute on edge devices in real-time.

Table 4. Complexity and number of parameters (proposed model).

	GFLOPs		Params (M)	
	Block	**Total**	**Block**	**Total**
Backbone	3.72	3.72	1.61	1.61
Cpm block	0.56	4.28	0.27	1.88
Stage1	1.28	5.56	0.6	2.48
Homogenous Stage 2	3.41	**8.97**	1.61	**4.09**

The performed comparisons of the proposed method are implemented on the COCO dataset to evaluate the operation of the method on the adult dataset. The comparison between SOTAs includes both top-down (object detection) and bottom-up (non-object detection: OpenPose-2016, OpenPose-2018, and proposed method) approaches.

Pose estimation on COCO. In addition, using the COCO dataset for comparison is objective because some methods have not been trained on the infant dataset and have low performance on that dataset (Table 5). The proposed method achieves 64.4 AP on COCO with an input size of 368 × 368, which is the method with the best performance among comparing methods with 8.9 GFLOPs. However, the accuracy of the proposed method is similar to top SOTA methods such as HRNet-W48 [27], Transpose [28], HRFormer-B [29], TokenPose-L/D24 [30], OpenPose-2018 [13].

Table 5. Comparisons with SOTA methods on the COCO validation set.

Method	Input Size	Backbone	GFLOPs	*AP*	$AP_{0.5}$	$AP_{0.75}$	AP_M	AP_L	*AR*
SimpleBaseline-Res152 [31]	256 × 192	–	15.7	61.9	74.7	68.1	59.1	67.9	66.9
HRNet-W48 [27]	256 × 192	–	16.0	64.6	77.9	70.7	61.5	70.3	69.1
HRNet-W32 [27]	256 × 192	–	7.1	64.0	77.8	70.4	60.9	69.7	68.6
TransPose [28]	256 × 192	HRNET	21.8	65.2	77.5	70.6	61.8	71.2	69.5
HRFormer-B [29]	256 × 192	HRNET	12.2	65.0	78.1	71.2	61.7	71.0	69.5
TokenPose-L/D24 [30]	256 × 192	HRNET	11.0	65.2	77.7	71.0	62.2	71.1	69.5
Openpose-2016 [18]	368 × 368	VGG19	136.1	62.6	88.2	69.5	68.3	72.2	–
Openpose-2018 [13]	368 × 368	VGG19	–	65.3	85.2	71.3	62.2	70.7	–
Proposed model (12 first layers)	**368 × 368**	**MobileNet**	**8.97**	**64.4**	**81.6**	**72.3**	**61.1**	**71.2**	**69.3**

Pose estimation on infant dataset. In Table 6 shows the comparison of our infant validation set. The proposed method is compared with a representative top-down approach for AggPose [14] and compared with OpenPose, representative of the bottom-up approach. The proposed method archives 82.7 AP with an input size of 368 × 368, which is the lowest complexity method and accuracy ranked only after AggPose-L [14]. The methods were all pre-trained with COCO data and refined and retrained on the infant data domain.

Table 6. Comparison with the infant test set. Comparison of methods with object detection and non-object detection (OpenPose CMU, proposed model). The methods compared in this table are transfer-learning (all weights) and the self-collected infant dataset.

Method	Input Size	GFLOPs	*AP*	*AR*
AggPose-S [14]	256 ×192	9.0	81.3	82.1
AggPose-L [14]	256 × 192	15.0	83.2	83.9
TokenPose-L/D24 [30]	256 × 192	11.0	81.4	82.3
OpenPose-2016 [18]	368 × 368	136.1	79.0	79.8
Pretrained model	**368 × 368**	**8.97**	**41.8**	**46.5**
Proposed model	**368 ×368**	**8.97**	**82.7**	**85.3**

Performance change. First of all, through the experiments of the posture estimation methods, the accuracy results are low when tested on our infant validation set. This experiment demonstrates the existence of differences between adult and pediatric data. And our approach to the infant position estimation problem worked well. For the proposed method, when pretraining with the COCO dataset, the AP results on the Infant validation set only reached 41.8% and increased to 82.7%, AP increases ∼2 times after being trained with the infant dataset (Table 6).

3.3. Visualization Analysis

Qualitative results on the infant set and COCO set are provided in Figures 7 and 8. Despite the difference in body proportions, the infant dataset is labeled in the COCO format, so the training process of the proposed method is quite favorable and shows the relationships between the main points of people well. The proposed method can be used for multi-person recognition in 2D.

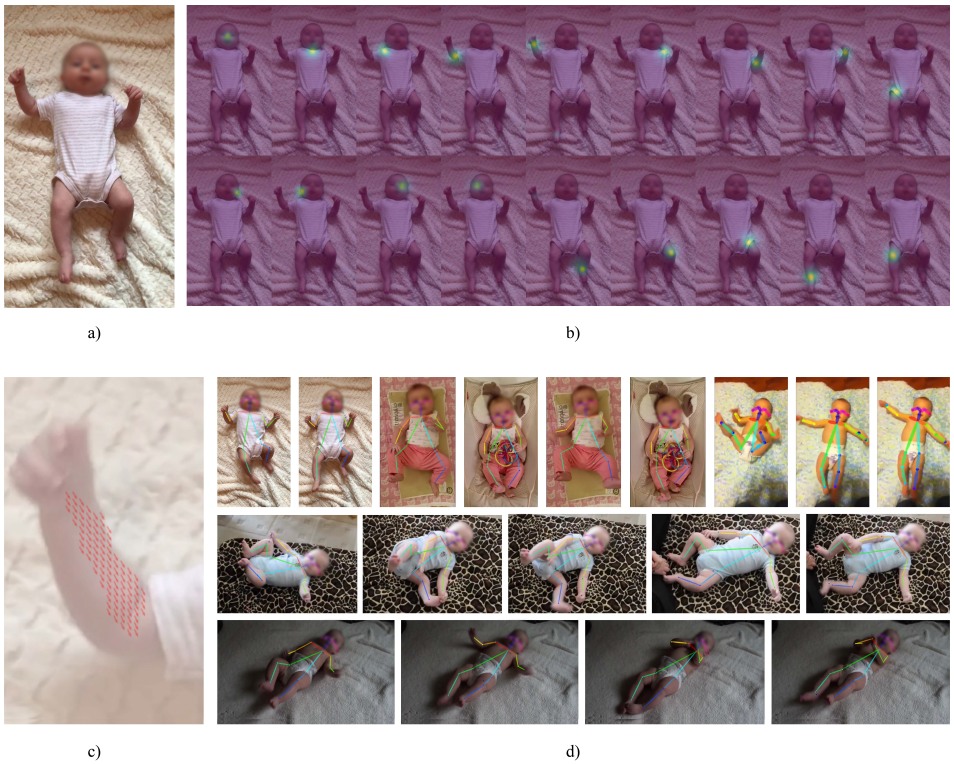

Figure 7. (**a**) Original image. (**b**) Visualization analysis of confidence maps for joint detection. (**c**) Visualization of an example of a part affinity field used for linked body parts. (**d**) Visualization analysis of the pose estimation with the proposed method on the infant test set.

Figure 8. Visualization analysis of pose estimation with the proposed method on the COCO dataset.

4. Discussion

The findings from the research and experimental endeavors provide valuable insights into strategies for diminishing the size and intricacy of CNN models. Experiments are meticulously designed to compare the performance of blocks and layers both before and after reduction and replication. This comparative analysis enables us to discern performance variations and evaluate the effectiveness of each block, thereby facilitating informed decisions in model design.

As mentioned, the first 12 layers of MobileNet-v1 are used for the backbone, and Table 2 compares the average precision between Openpose models instead of backbones and

shows that by using the first 13 layers, MobileNet v1 has the advantage of representations of best AP but higher operations costs; therefore, we accept the option of the first 12 layers of MobileNet v1 to achieve a balance between performance and accuracy. Notably, there exists a substantial contrast in GFLOPs between the initial 12 layers and the subsequent layers in the MobileNet-v1 model. In response, we have intervened to fine-tune the model by using a dilated layer, aiming to capture more expansive features effectively. Dilated layers are helpful in expanding the perception area, broadening the neural network's field of view, thereby enhancing its capacity to observe a wider range of input data. This augmentation facilitates improved recognition of larger objects and offers supplementary contextual information pertaining to the surroundings of the object in focus [32].

In the experiment evaluating the blocks of the Openpose-2016 model, we systematically recreated them to precisely assess the performance of each block. Our findings revealed that stages 2–6 exhibited uniformity with negligible changes in accuracy; these stages shared identical sizes and were computationally intensive. Notably, [13] acknowledged this issue and opted to rebalance the number of stages for confidence maps (CMs) and part affinity fields (PAFs). Their results demonstrated that stages 1 and 2 were particularly effective in yielding satisfactory outcomes. Hence, in our study, we focused on enhancement by exclusively utilizing these two stages. In Openpose-2018, the number of stages remained at 6, with each stage utilized for inference regarding specific segments of the CMs or PAFs results. In our investigation, we endeavored to utilize two stages, with each stage employed for inference for both CMs and PAFs. This approach ensures that the results undergo refinement across two stages, as opposed to just one, thereby potentially enhancing the overall accuracy and quality of the outputs when reducing the complexity of the model.

Training models using adult datasets have proven advantageous due to the substantial knowledge transfer from a larger pool of relevant data [17]. Transfer learning serves as a crucial mechanism for refining the model's weights to facilitate adaptation with newborns' smaller joint ratios.

This research still has several limitations. While we have gathered and contributed data ourselves, the variability in data within consecutive frames of the video remains limited, posing a significant challenge in capturing the diversity of the infants' data for training. Additionally, poses with high complexity are occasionally misinterpreted, a challenge that could potentially be mitigated by augmenting the dataset with more infant-specific data. Moreover, the model's performance tends to degrade in instances where the infant's spine axis is not aligned vertically (rotational angle in the data). To address this issue, we have implemented a strategy where the image data are re-rotated to align the newborn subject's spine axis with the vertical direction [33].

5. Conclusions

In this paper, we have introduced a new definition of the single-branch model in the bottom-up approach and used this method to estimate the infant's posture by training the model on a self-collected infant dataset. The proposed method is an improved model based on the model of the OpenPose-2016 method. The tests showed promising results from the proposed method with an AP level of 82.7% and computational complexity of 6.6% (8.97 GFLOPs) compared to the OpenPose-2016 method. The proposed method has a backbone architecture of the first 12 layers of MobileNet-v1 and uses 1×1 and 3×3 convolutional layers to minimize the number of operations and parameters. In addition, the proposed method is not effective with complex body shape data, a number of overlapping joints, such as supine posture, and complex limbs, and the proposed method is time-consuming to train. This is due to the fact that the model has a large depth that slows down the backpropagation process.

With the proposed number of parameters and computational complexity, deploying on mobile/edge devices to widely disseminate support resources for the needs of infant pose estimation applications is promising. However, we acknowledge that there is a long way to go to develop a model for mobile/edge applications that analyze infant movements.

Author Contributions: Conceptualization, V.D.N.; methodology, V.D.N and T.N.-Q.; software, T.N.-Q., M.D.N. and N.D.B.; validation, V.D.N., T.N.-Q. and N.D.B.; formal analysis, V.D.N.; investigation, D.H.P. and N.D.B.; resources, T.N.-Q. and N.D.B.; data curation, T.N.-Q.; writing—original draft preparation, V.D.N. and T.N.-Q.; writing—review and editing, T.N.-Q., M.D.N. and N.D.B.; visualization, T.N.-Q., N.D.B., M.D.N. and D.H.P.; supervision, V.D.N. and N.D.B.; project administration, V.D.N.; funding acquisition, D.H.P. 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: Data are contained within the article.

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

References

1. Silva, N.; Zhang, D.; Kulvicius, T.; Gail, A.; Barreiros, C.; Lindstaedt, S.; Kraft, M.; Bölte, S.; Poustka, L.; Nielsen-Saines, K.; et al. The future of General Movement Assessment: The role of computer vision and machine learning—A scoping review. *Res. Dev. Disabil.* **2021**, *110*, 103854. [CrossRef] [PubMed]
2. Toshev, A.; Szegedy, C. DeepPose: Human Pose Estimation via Deep Neural Networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1653–1660. [CrossRef]
3. Wei, S.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional Pose Machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4724–4732.
4. Yang, W.; Li, S.; Ouyang, W.; Li, H.; Wang, X. Learning Feature Pyramids for Human Pose Estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1281–1290.
5. Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A.L.; Wang, X. Multi-context Attention for Human Pose Estimation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5669–5678. [CrossRef]
6. Nie, X.; Feng, J.; Zuo, Y.; Yan, S. Human Pose Estimation with Parsing Induced Learner. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2100–2108. [CrossRef]
7. Bulat, A.; Tzimiropoulos, G. Human pose estimation via Convolutional Part Heatmap Regression. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 717–732.
8. Nguyen, T.D.; Kresovic, M. A survey of top-down approaches for human pose estimation. *arXiv* **2022**, arXiv:2202.02656.
9. Kresovic, M.; Nguyen, T.D. Bottom-up approaches for multi-person pose estimation and it's applications: A brief review. *arXiv* **2021**, arXiv:2112.11834.
10. Kocabas, M.; Karagoz, S.; Akbas, E. MultiPoseNet: Fast Multi-Person Pose Estimation using Pose Residual Network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 417–433.
11. Airaksinen, M.; Räsänen, O.; Ilén, E.; Häyrinen, T.; Kivi, A.; Marchi, V.; Gallen, A.; Blom, S.; Varhe, A.; Kaartinen, N.; et al. Automatic Posture and Movement Tracking of Infants with Wearable Movement Sensors. *Sci. Rep.* **2020**, *10*, 169. [CrossRef] [PubMed]
12. McCay, K.D.; Ho, E.S.L.; Marcroft, C.; Embleton, N.D. Establishing Pose Based Features Using Histograms for the Detection of Abnormal Infant Movements. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 5469–5472. [CrossRef]
13. 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]
14. Cao, X.; Li, X.; Ma, L.; Huang, Y.; Feng, X.; Chen, Z.; Zeng, H.; Cao, J. AggPose: Deep Aggregation Vision Transformer for Infant Pose Estimation. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, Vienna, Austria, 23–29 July 2022. [CrossRef]
15. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
16. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.S.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
17. Lee, J.H.; Kvinge, H.J.; Howland, S.; New, Z.; Buckheit, J.; Phillips, L.A.; Skomski, E.; Hibler, J.; Corley, C.D.; Hodas, N.O. Adaptive Transfer Learning: A simple but effective transfer learning. *arXiv* **2021**, arXiv:2111.10937.
18. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In Proceedings of the European Conference on Computer Vision (ECCV), Las Vegas, NV, USA, 27–30 June 2016.
19. Hong, S.; Roh, B.; Kim, K.; Cheon, Y.; Park, M. PVANet: Lightweight Deep Neural Networks for Real-time Object Detection. *arXiv* **2016** arXiv:1611.08588.

20. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017** arXiv:1704.04861.
21. 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; pp. 4510–4520. [CrossRef]
22. Elharrouss, O.; Akbari, Y.; Almaadeed, N.; Al-Maadeed, S. Backbones-Review: Feature Extraction Networks for Deep Learning and Deep Reinforcement Learning Approaches. *arXiv* **2022**, arXiv:2206.08016.
23. Bo, Y.; Wu, J.; Hattori, G. Face Mask aware Robust Facial Expression Recognition during the COVID-19 Pandemic. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021. [CrossRef]
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
25. Lin, T.; Maire, M.; Belongie, S.J.; Bourdev, L.D.; Girshick, R.B.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
26. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
27. 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. [CrossRef] [PubMed]
28. 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, 10–17 October 2021; pp. 11782–11792. [CrossRef]
29. Yuan, Y.; Fu, R.; Huang, L.; Lin, W.; Zhang, C.; Chen, X.; Wang, J. HRFormer: High-Resolution Transformer for Dense Prediction. *arXiv* **2021**, arXiv:2110.09408.
30. 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 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 11293–11302. [CrossRef]
31. Ye, S.; Zhang, Y.; Hu, J.; Cao, L.; Zhang, S.; Shen, L.; Wang, J.; Ding, S.; Ji, R. DistilPose: Tokenized Pose Regression with Heatmap Distillation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 2163–2172. [CrossRef]
32. Lei, X.; Pan, H.; Huang, X. A Dilated CNN Model for Image Classification. *IEEE Access* **2019**, *7*, 124087–124095. [CrossRef]
33. Zhao, X.; Takata, S.; Fukumori, K.; Tanaka, T. Infant Posture Assessment Based on Rotational Keypoint Detection. In Proceedings of the 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC),Tokyo, Japan, 14–17 December 2021; pp. 1546–1550.