Fast Human Pose Estimation

Feng Zhang[1]   Xiatian Zhu[2]   Mao Ye[1]

[1]{zhangfengwcy, cvlab.uestc}@gmail.com, School of Computer Science and Engineering,
University of Electronic Science and Technology of China

[2]eddy@visionsemantics.com, Vision Semantics Limited

Abstract

Existing human pose estimation approaches often only consider how to improve the model generalisation performance, but putting aside the significant efficiency problem. This leads to the development of heavy models with poor scalability and cost-effectiveness in practical use. In this work, we investigate the under-studied but practically critical pose model efficiency problem. To this end, we present a new Fast Pose Distillation (FPD) model learning strategy. Specifically, the FPD trains a lightweight pose neural network architecture capable of executing rapidly with low computational cost. It is achieved by effectively transferring the pose structure knowledge of a strong teacher network. Extensive evaluations demonstrate the advantages of our FPD method over a broad range of state-of-the-art pose estimation approaches in terms of model cost-effectiveness on two standard benchmark datasets, MPII Human Pose and Leeds Sports Pose.

1. Introduction

Human pose estimation has gained remarkable progress from the rapid development of various deep CNN models [30, 8, 10]. This is because deep neural networks are strong at approximating complex and non-linear mapping functions from arbitrary person images to the joint locations even at the presence of unconstrained human body appearance, viewing conditions and background noises.

Nevertheless, the model performance advantages come with the cost of training and deploying resource-intensive networks with large depth and width. This causes inefficient model inference, requiring per-image computing cost at tens of FLoating point OPerations (FLOPs) therefore poor scalability particularly on resource-limited devices such as smart phones and robots. There is a recent attempt that binarises the network parameters for model execution speedup [7], which however suffers significantly weak model generalisation capacity.

In this study, we consider the problem of improving the pose estimation efficiency *without* model performance degradation but preserving comparable accuracy results. We observe that the basic CNN building blocks for state-of-the-art human pose networks such as Hourglass [19] are not cost-effective in establishing small networks due to a high number of channels per layer and being more difficult to train. To overcome these barriers, we design a lightweight variant of Hourglass network and propose a more effective training method of small pose networks in a knowledge distillation fashion [13]. We call the proposed method *Fast Pose Distillation* (FPD). Compared with the top-performing alternative pose approaches [32, 10], the proposed FPD approach enables much faster and more cost-effective model inference with extremely smaller model size while simultaneously reaching the same level of human pose prediction performance.

We summarise our *contributions* in follows:

(i) We investigate the under-studied human pose model efficiency problem, opposite to the existing attempts mostly focusing on improving the accuracy performance *alone* at high costs of model inference at deployment. This is a critical problem to be addressed for scaling up the existing deep pose estimation methods to real applications.

(ii) We propose a *Fast Pose Distillation* (FPD) model training method enabling to more effectively train extremely small human pose CNN networks. This is based on an idea of knowledge distillation that have been successfully exploited in inducing object image categorisation deep models. In particular, we derive a pose knowledge distillation learning objective to transfer the latent knowledge from a pre-trained larger teacher model to a tiny target pose model (to be deployed in test time). This aims to pursue the best model performance given very limited computational budgets using only a small fraction (less than 20%) of cost required by similarly strong alternatives.

(iii) We design a lightweight Hourglass network capable of constructing more cost-effective pose estimation CNN

models while retaining sufficient learning capacity for allowing satisfactory accuracy rates. This is achieved by extensively examining the redundancy degree of existing state-of-the-art pose CNN architecture designs.

In the evaluations, we have conducted extensive empirical comparisons to validate the efficacy and superiority of the proposed FPD method over a wide variety of state-of-the-art human pose estimation approaches in the balance of model inference efficiency and prediction performance on two commonly adopted benchmark datasets, MPII Human Pose [1] and Leeds Sports Pose [15].

2. Related Work

Human Pose Estimation The past five years have witnessed a huge progress of human pose estimation in the deep learning regime [30, 28, 6, 31, 19, 11, 32, 20, 22]. Despite the clear performance increases, these prior works focus only on improving the pose estimation accuracy by using complex and computationally expensive models whilst largely ignoring the model inference cost issue. This significantly restricts their scalability and deployability in real-world applications particularly with very limited computing budgets available.

In the literature, there are a few recent works designed to improve model efficiency. For example, Bulat and Tzimiropoulos built parameter binarised CNN models to accommodate resource-limited platforms [7]. But this method leads to dramatic performance drop therefore not satisfied for reliable utilisation. In most cases, high accuracy rates are required. Rafi et al. exploited good general purpose practices to improve model efficiency without presenting a novel algorithm [24]. Further, this method does not provide quantitative evaluation on the trade-off between model efficiency and effectiveness.

In contrast to these previous methods, we systematically study the pose estimation efficiency problem under the condition of preserving the model performance rate so that the resulted model is more usable and reliable in real-world application scenarios.

Knowledge Distillation The objective of knowledge distillation is concerned with information transfer between different neural networks with distinct capacities [5, 13, 3]. For instance, Hinton et al. successfully employed a well trained large network to help train a small network [13]. The rationale is an exploitation of extra supervision from a teacher model, represented in form of class probabilities [13], feature representations [3, 25], or an inter-layer flow [35]. This principle has also been recently applied to accelerate the model training process of large scale distributed neural networks [2], to transfer knowledge between multiple layers [17] or between multiple training states [18]. Beyond the conventional two stage training based offline dis-

tillation, one stage online knowledge distillation has been attempted with added merits of more efficient optimisation [37, 16] and more effective learning [16]. Besides, knowledge distillation has been exploited to distil easy-to-train large networks into harder-to-train small networks [25].

While these past works above transfer category-level discriminative knowledge, our method transfers richer structured information of dense joint confidence maps. A more similar work is the latest radio signals based pose model that also adopts the idea of knowledge distillation [38]. However, this method targets at using wireless sensors to tackle the occlusion problem, rather than the model efficiency issue as we confider here.

3. Fast Human Pose Estimation

Human pose estimation aims to predict the spatial coordinates of human joints in a given image. To train a model in a supervised manner, we often have access to a training dataset $\{I^i, G^i\}_{i=1}^N$ of N person images each labelled with K joints defined in the image space as:

$$G^i = \{g_1^i, .., g_K^i\} \in \mathbb{R}^{K \times 2}, \tag{1}$$

where H and W denotes the image height and width, respectively. Generally, this is a regression problem at the imagery pixel level.

Objective Loss Function For pose model training, we often use the Mean-Squared Error (MSE) based loss function [29, 19]. To represent the ground-truth joint labels, we generate a confidence map m_k for each single joint k ($k \in \{1, \cdots, K\}$) by centring a Gaussian kernel around the labelled position $z_k = (x_k, y_k)$.

More specifically, a Gaussian confidence map m_k for the k-th joint label is written as:

$$m_k(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(\frac{-[(x - x_k)^2 + (y - y_k)^2]}{2\sigma^2}\right) \tag{2}$$

where (x, y) specifies a pixel location and the hyper-parameter σ denotes a pre-fixed spatial variance. The MSE loss function is then obtained as:

$$\mathcal{L}_{\text{mse}} = \frac{1}{K} \sum_{k=1}^{K} \|m_k - \hat{m}_k\|_2^2 \tag{3}$$

where $\hat{m}_k$ refers to the predicted confidence map for the k-th joint. The standard SGD algorithm can then be used to optimise a deep CNN pose model by back-propagating MSE errors on training data in a mini-batch incrementally.

Existing pose methods rely heavily on large deep neural networks for maximising the model performance, whilst neglecting the inference efficiency. We address this limitation for higher scalability by establishing lightweight CNN architectures and proposing an effective model learning strategy detailed below.
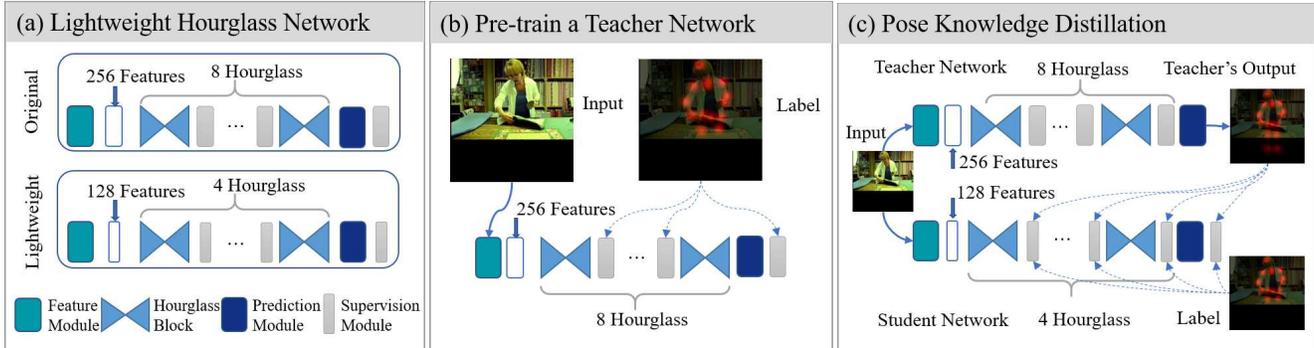
Figure 1. An overview of the proposed Fast Pose Distillation model learning strategy. To establish a highly cost-effective human pose estimation model, We need to build a compact backbone such as **(a)** a lightweight Hourglass network. To more effectively train a small target network, we adopt the principle of knowledge distillation in the pose estimation context. This requires to **(b)** pre-train a strong teacher pose model, such as the state-of-the-art Hourglass network or other existing alternatives. The teacher model is used to provide extra supervision guidance in the **(c)** pose knowledge distillation procedure via the proposed mimicry loss function. At test time, the small target pose model enables a fast and cost-effective deployment. The computationally expensive teacher model is *abandoned* finally, since its discriminative knowledge transferred already into the target model therefore used in deployment (rather than wasted).

Stage	Building Block
1, 2, 3, 4	Hourglass with 128 channels per layer

Table 1. The structure of a small pose CNN model.

3.1. Compact Pose Network Architecture

Human pose CNN models typically consist of multiple repeated building blocks with the identical structure [8, 31, 19, 11, 32, 20, 22]. Among these, Hourglass is one of the most common building block units [19]. However, we observe that existing designs are not cost-effective, due to deploying a large number of both channels and blocks in the entire architecture therefore leading to a suboptimal trade-off between the representation capability and the computational cost. For example, [19] suggested a CNN architecture of 8 Hourglass stages each having 9 Residual blocks with 256 channels within every layer.

We therefore want to minimise the expense of existing CNN architectures for enabling faster model inference. With careful empirical examination, we surprisingly revealed that a half number of stages (i.e. 4 Hourglass modules) suffice to achieve over 95% model generalisation capacity on the large scale MPII benchmark. Moreover, the per-layer channels are also found highly redundant and reducing a half number (128) only results in less than 1% performance drop (Table 5). Based on these analysis, we construct a very light CNN architecture for pose estimation with only one sixth computational cost of the original design. See Table 1 and Figure 1 for the target CNN architecture specifications.

Remarks Whilst it is attractive to deploy tiny pose networks that run cheaply and fast, it is *empirically* non-trivial to train them although *theoretically* shallow networks have

the similar representation capacities to approximate the target functions as learned by deeper counterparts [3, 26]. A similar problem has been occurred and investigated in object image classification through the knowledge distillation strategy, i.e. let the target small network mimic the prediction of a larger teacher model [13]. However, it remains unclear how well such a similar method will work in addressing structured human pose estimation in dense pixel space. To answer this question, in the following we present a pose structure knowledge distillation method.

3.2. Supervision Enhancement by Pose Distillation

Model Training Pipeline We adopt the generic model training strategy of knowledge distillation:

1. We first train a large teacher pose model. In our experiments, by default we select the original Hourglass model [19] due to its clean design and easy model training. Other stronger models can be considered without any restrictions.

2. We then train a target student model with the assistance of knowledge learned by the teacher model. Knowledge distillation happens in this step. The structure of the student model is presented in Table 1.

An overview of the whole training procedure is depicted in Figure 1. The key to distilling knowledge is to design a proper mimicry loss function that is able to effectively extract and transfer the teacher's knowledge to the training of the student model. The previous distillation function is designed for single-label based softmax cross-entropy loss in the context of object categorisation [3, 13] and unsuitable to transfer the structured pose knowledge in 2D image space.

To address this aforementioned problem, we design a joint confidence map dedicated pose distillation loss function formulated as:

$$\mathcal{L}_{\text{pd}} = \frac{1}{K} \sum_{k=1}^{K} \|\boldsymbol{m}_k^s - \boldsymbol{m}_k^t\|_2^2 \qquad (4)$$

where $\boldsymbol{m}_k^s$ and $\boldsymbol{m}_k^t$ specify the confidence maps for the k-th joint predicted by the pre-trained teacher model and the in-training student target model, respectively. We choose the MSE function as the distillation quantity to measure the divergence between the student and teacher models in order to maximise the comparability with the pose supervised learning loss (Eqn (3)).

Overall Loss Function We formulate the overall FPD loss function for pose structure knowledge distillation during training as:

$$\mathcal{L}_{\text{fpd}} = \alpha \mathcal{L}_{\text{pd}} + (1 - \alpha) \mathcal{L}_{\text{mse}} \qquad (5)$$

where α is the balancing weight between the two loss terms, estimated by cross-validation. As such, the target network learns both to predict the labelled ground-truth annotations of training samples by $\mathcal{L}_{\text{mse}}$ and to match the prediction structure of the stronger teacher model by $\mathcal{L}_{\text{pd}}$.

Further Remarks Why does the proposed pose distillation loss function probably help to train a more generalisable target model, as compared to training *only* on the labelled data? A number of reason may explain this in the context of pose estimation.

1. The body joint labels are likely to be erroneous due to the high difficulty of locating the true positions in the manual annotation process. In such cases, the teacher model may be able to mitigate some errors through statistical learning and reasoning therefore reducing the misleading effect of wrongly labelled training samples (Figure 3 Row (A)).

2. Given difficult training cases say with confusing/cluttered background and random occlusion situations, the teacher prediction may provide softened learning tasks by explained away these hard samples with model inference (Figure 3 Row (B)).

3. The teacher model may provide more complete joint labels than the original annotation therefore not only providing additional more accurate supervision but also mitigating the misleading of missing joint labels (Figure 3 Row (C)).

4. Learning to match the ground-truth confidence map can be harder in comparison to aligning the teacher's prediction. This is because the teacher model has spread some reasoning uncertainty for each training sample either hard or easy to process.

5. On the other hand, the teacher's confidence map encodes the abstract knowledge learned from the entire training dataset in advance, which may be beneficial to be considered in learning every individual training sample during knowledge distillation.

In summary, the proposed model is capable of handling wrong pose joint annotations, e.g. when the pre-trained teacher predicts more accurate joints than manual wrong and missing labels. Due to a joint use of the ground-truth labels and the teacher model's prediction, our model is tolerant to either error but not co-occurring ones. This alleviates the harm of label errors in the training data, in contrast to existing methods that often blindly trust all given labels.

3.3. Model Training and Deployment

The proposed FPD model training method consists of two stages: (i) We train a teacher pose model by the conventional MSE loss (Eqn (3)), and (ii) train a target student model by the proposed loss (Eqn (5)), with the knowledge distillation from the teacher model to the target model being conducted in each mini-batch and throughout the entire training process. At test time, we only use the small target model for efficient and cost-effective deployment whilst throwing away the heavy teacher network. The target model already extracts the teacher's knowledge.

4. Experiments

4.1. Experiment Setup

Datasets We utilised two human pose benchmark datasets, MPII [1] and Leeds Sports Pose (LSP) [15]. The MPII dataset is collected from YouTube videos with a wide range of human activities and events. It has 25K scene images and 40K annotated persons (29K for training and 11K for test). Each person has 16 labelled body joints. We adopted the standard train/valid/test data split [28]. Following [29], we randomly sampled 3K samples from the training set for model validation.

The LSP benchmark contains natural person images from many different sports scenes. Its extended version provides 11K training samples and 1K test samples. Each person in LSP has 14 labelled joints.

Performance Metrics We used the standard Percentage of Correct Keypoints (PCK) measurement that quantifies the fraction of correct predictions within an error threshold τ [34]. Specifically, the quantity τ is normalised against the size of either torso ($\tau = 0.2$ for LSP, i.e. PCK@0.2) or head ($\tau = 0.5$ for MPII, i.e. PCKh@0.5). We measured each individual joint respectively and took their average as an overall metric. Using different τ values, we yielded a PCK curve. Therefore, the Area Under Curve (AUC) can be obtained as a holistic measurement across different de-

Method	Head	Sho.	Elbo.	Wri.	Hip	Knee	Ank.	Mean	AUC	# Param	Deployment Cost
Rafi et al., BMVC'16[24]	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3	57.3	56M	28G
Belagiannis&Zisserman, FG'17[4]	97.7	95.0	88.2	83.0	87.9	82.6	78.4	88.1	58.8	17M	95G
Insafutdinov et al., ECCV'16[14]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5	60.8	66M	286G
Wei et al., CVPR'16[31]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5	61.4	31M	351G
Bulat&Tzimiropoulos, ECCV'16[6]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7	59.6	76M	67G
Newell et al., ECCV'16[19]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9	62.9	26M	55G
Ning et al., TMM'17[21]	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2	63.6	74M	124G
Chu et al., CVPR'17[11]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5	63.8	58M	128G
Peng et al., CVPR'18[22]	98.1	96.6	92.5	88.4	90.7	87.7	83.5	91.5	-	26M	55G
Yang et al., ICCV'17[32]	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0	64.2	28M	46G
Nie et al., CVPR'18[20]	**98.6**	**96.9**	**93.0**	**89.1**	**91.7**	**89.0**	**86.2**	**92.4**	**65.9**	26M	63G
Sekii, ECCV18[27]	-	-	-	-	-	-	-	88.1	-	16M	**6G**
FPD	98.3	96.4	91.5	87.4	90.9	87.1	83.7	91.1	63.5	**3M**	9G

Table 2. PCKh@0.5 and AUC (%) rates on the MPII *test* dataset. M/G: $10^6/10^9$.

Method	Head	Sho.	Elbo.	Wri.	Hip	Knee	Ank.	Mean	AUC	# Param	Deployment Cost
Tompson et al., NIPS'14[29]	90.6	79.2	67.9	63.4	69.5	71.0	64.2	72.3	47.3	-	-
Fan et al., CVPR'15[12]	92.4	75.2	65.3	64.0	75.7	68.3	70.4	73.0	43.2	-	-
Carreira et al., CVPR'16[8]	90.5	81.8	65.8	59.8	81.6	70.6	62.0	73.1	41.5	-	-
Chen&Yuille, NIPS'14[9]	91.8	78.2	71.8	65.5	73.3	70.2	63.4	73.4	40.1	-	-
Yang et al., CVPR'16[33]	90.6	78.1	73.8	68.8	74.8	69.9	58.9	73.6	39.3	-	-
Rafi et al., BMVC'16[24]	95.8	86.2	79.3	75.0	86.6	83.8	79.8	83.8	56.9	56M	28G
Yu et al., ECCV'16[36]	87.2	88.2	82.4	76.3	91.4	85.8	78.7	84.3	55.2	-	-
Peng et al., CVPR'18[22]	**98.6**	**95.3**	**92.8**	**90.0**	**94.8**	**95.3**	**94.5**	**94.5**	-	26M	55G
FPD	97.3	92.3	86.8	84.2	91.9	92.2	90.9	90.8	**64.3**	**3M**	**9G**

Table 3. PCK@0.2 and AUC (%) rates on the LSP *test* dataset. M/G: $10^6/10^9$.

cision thresholds. To measure the model efficiency both in training and test, we used the FLOPs.

Training Details We implemented all the following experiments in Torch. We cropped all the training and test images according to the provided positions and scales, and resized them to 256×256 in pixels. As typical, random scaling (0.75-1.25), rotating (± 30 degrees) and horizontal flipping were performed to augment the training data. We adopted the RMSProp optimisation algorithm. We set the learning rate to 2.5×10^{-4}, the mini-batch size to 4, and the epoch number to 130 and 70 for MPII and LSP benchmarks, respectively. For the network architecture, we used the original Hourglass as the teacher model and the customised Hourglass with less depth and width (Table 1) as the target model.

4.2. Comparisons to State-Of-The-Art Methods

We evaluated the proposed FPD method by extensively comparing against recent human pose estimation deep methods on MPII and LSP.

Results on MPII Table 2 compares the PCKh@0.5 accuracy results of state-of-the-art methods and the proposed FPD on the *test* dataset of MPII. It is clearly observed that the proposed FPD model is significantly efficient and compact therefore achieving a much cheaper deployment cost. Importantly, this advantage is obtained *without* clearly compromising the model generalisation capability, e.g. achieving as high as 91.1%.

Specifically, compared with the best performer [20], the FPD model only requires 14.3% (9/63) computational cost but gaining 96.4% (63.5/65.9) performance in mean PCKh accuracy. This leads to a 6.7%× (96.4/14.3) cost-effective advantage. When compared to the most efficient alternative competitor [24], our model is 2.9× (26/9) more efficient whilst simultaneously achieving a mean PCKh gain of 4.8% (91.1-86.3). These evidences clearly suggest the cost-effectiveness advantages of our method over other alternative approaches.

In pose estimation, an improvement of 0.8% indicates a significant gain particularly on the challenging MPII with varying poses against cluttered background. This boost
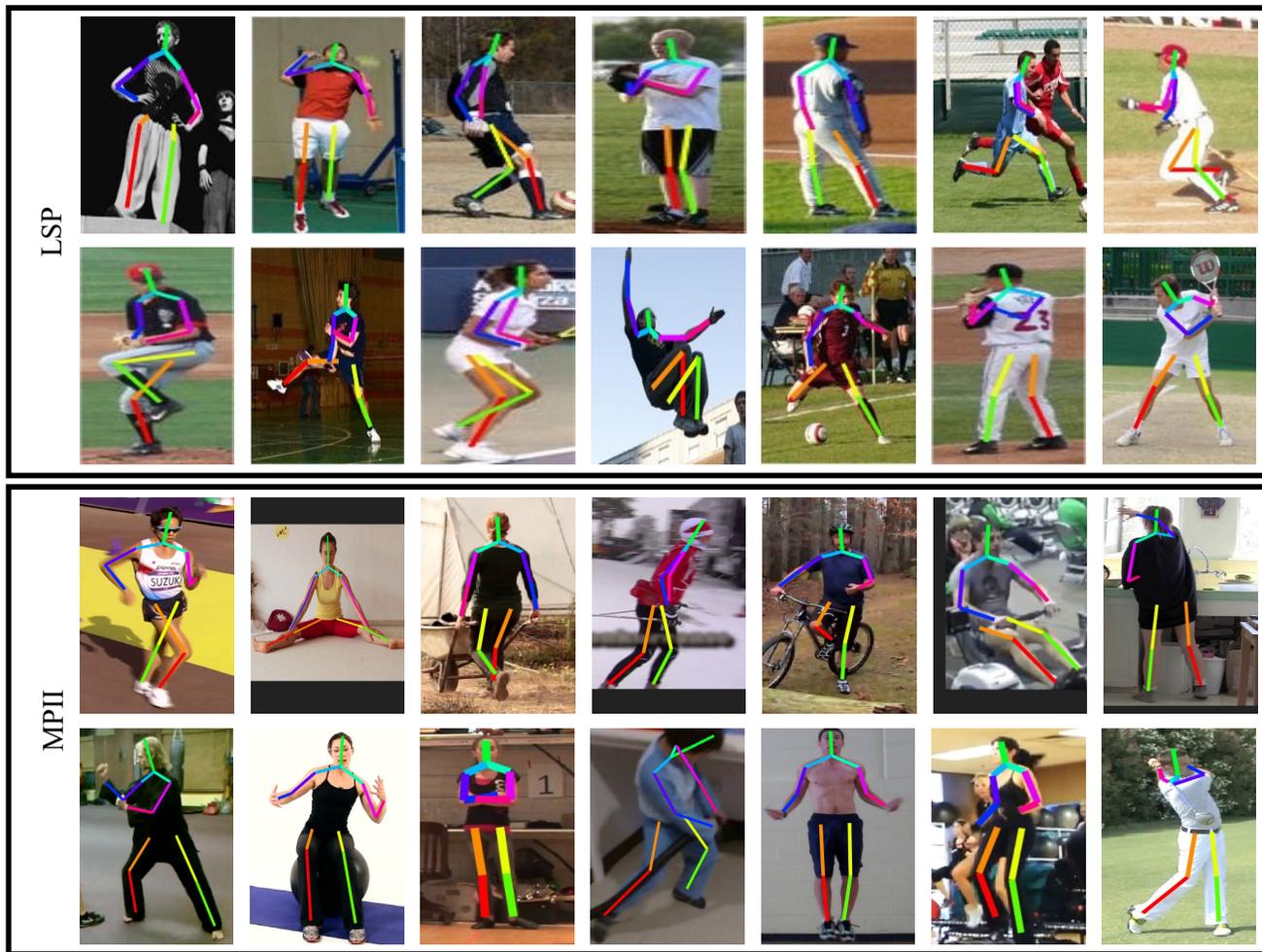
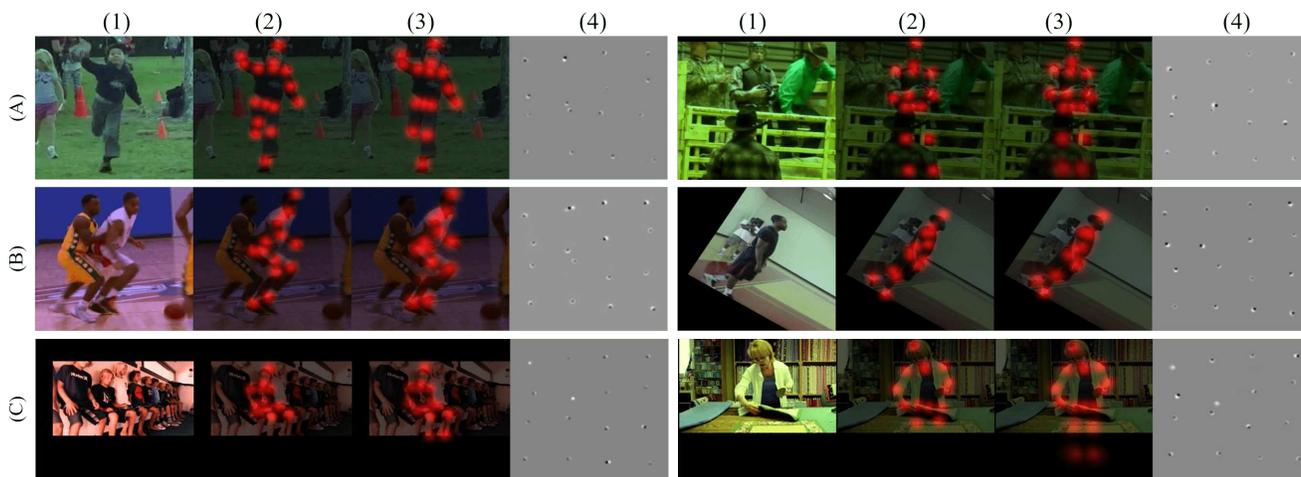Figure 2. Example of human pose estimation on LSP and MPII.

Figure 3. Pose estimation examples on MPII by the proposed FPD model. Column (**1**): The input images. Column (**2**): Ground-truth joint confidence maps. Column (**3**): Joint confidence maps predicted by the teacher model. Column (**4**): The difference between ground-truth and teacher's confidence map. Each row represents a type of pose knowledge transfer. Row (**A**): Error labelling of the right leg ankle in the "ground-truth" annotations, which is corrected by the teacher model. Row (**B**): A softened teacher confidence map with larger uncertainty than the ground-truth due to the highly complex human posture. Row (**C**): Missing joint labels are discovered by the teacher model.

FPD	Head	Sho.	Elbo.	Wri.	Hip	Knee	Ank.	Mean	AUC
✗	97.4	96.0	90.2	85.8	88.2	84.3	80.6	89.4	61.4
✓	**97.5**	**96.3**	**91.4**	**87.3**	**89.4**	**85.6**	**82.0**	**90.4**	**62.4**

Table 4. Generalisation evaluation of the proposed FPD approach. Metric: Mean PCKh@0.5 and AUC.

# Stage	# Channel	Mean	AUC	# Param	Deployment Cost
8	256	91.9	63.7	26M	55G
4	256	91.4	63.9	13M	30G
2	256	90.5	63.0	7M	17G
1	256	86.4	58.3	3M	10G
4	256	91.4	63.9	13M	30G
4	128	90.1	62.4	3M	9G
4	64	87.9	59.5	0.95M	4.5G
4	32	83.4	54.9	0.34M	3.1G

Table 5. Cost-effectiveness analysis of the Hourglass model. Metric: PCKh@0.5 and AUC. M/G: $10^6/10^9$.

Pose Distillation	Mean	AUC
✗	90.1	62.4
✓	**90.9**	**63.3**

Table 6. Effect of the proposed pose knowledge distillation. Metric: Mean PCKh@0.5 and AUC (%).

is bigger than other state-of-the-art gains, e.g. +0.3% in 91.2% [21] vs 90.9% [19]; further +0.3% in 91.5% [23]. More specifically, given all 163,814 test joints, each 0.1% gain means correcting 163 joints.

Results on LSP Table 3 compares the PCK@0.2 rates of our FPD model and existing methods with top reported performances on the LSP test data. Compared to MPII, this benchmark has been less evaluated by deep learning models, partly due to a smaller size of training data. Overall, we observed the similar comparisons. For example, our FPD runs more efficiently than the most competitive alternative [24] and consumes much less training energy, in addition to achieving the best pose prediction accuracy rate among all compared methods.

Qualitative Examination To provide visual test, Figure 2 shows qualitative pose estimation evaluations on LSP and MPII. It is observed that such a small FPD model can still achieve reliable and robust pose estimation in arbitrary in-the-wild images with various background clutters, different human poses and viewing conditions.

4.3. Ablation Study

We carried out detailed component analysis and discussion on the *validation* set of MPII.

FPD generalisation evaluation Besides using the state-of-the-art Hourglass as the backbone network, we also

tested the more recent model [32] when integrated into the proposed FPD framework. In particular, we adopted the original network as the teacher model and constructed a lightweight variant as the student (target) model. The lightweight model was constructed similarly as in Table 1 because it is based on the Hourglass design too: reducing the number of stages to 4 and the number of channels in each module to 128. The results in Table 4 show that our FPD approach achieves 1.0% mean PCKh@0.5 gain, similar to the Hourglass case. This suggests the good generalisation capability of the proposed approach in yielding cost-effective pose estimation deep models.

Cost-effectiveness analysis of Hourglass We extensively examined the architecture design of the state-of-the-art Hourglass neural network model [19] in terms of cost-effectiveness. To this end, we tested two dimensions in design: depth (the layer number) and width (the channel number). Interestingly, we revealed in Table 5 that removing half stages (layers) and half channels only leads to quite limited performance degradation. This indicates that the original Hourglass design is highly redundant with poor cost-effectiveness. However, this is largely ignored in previous works due to their typical focus on pursuing the model accuracy performance *alone* whilst overlooking the important model efficiency problem. This series of CNN architecture examinations helps us to properly formulate a lightweight pose CNN architecture with only 16% (9/55) computational cost but obtaining 98% (90.1/91.9) model performance as compared to the state-of-the-art design, laying a good foundation towards building compact yet strong human pose deep models.

Effect of pose knowledge distillation We tested the effect of using our pose knowledge distillation on the lightweight Hourglass network. In contrast to all other

Loss Function	Head	Sho.	Elbo.	Wri.	Hip	Knee	Ank.	Mean	AUC
MSE	**97.7**	**96.4**	**91.8**	87.6	**89.7**	**86.6**	**83.9**	**90.9**	**63.3**
Cross-Entropy	97.6	96.2	91.5	87.6	89.0	86.5	83.6	90.7	63.0

Table 7. Pose knowledge distillation by different types of loss function. Metric: Mean PCKh@0.5 and AUC.

α	0	0.05	0.1	0.5	0.95	0.99
Mean	90.1	90.8	90.8	**90.9**	90.7	90.7
AUC	62.4	63.2	63.2	**63.3**	63.0	63.0

Table 8. Performance analysis of the learning importance parameter of pose distillation. Metric: Mean PCKh@0.5 and AUC (%).

methods, the model [23] additionally benefits from an auxiliary dataset MPII in model training. Table 6 shows that teacher knowledge transfer brings in 0.8% (90.9-90.1) mean PCKh accuracy boost. This suggests that the generic principle of knowledge distillation is also effective in the structured pose estimation context, beyond object categorisation.

To further validate how on earth this happens, we visualise three pose structure transfer examples in Figure 3. It is shown that the proposed mimicry loss against the teacher prediction is likely to pose extra information in cases of error labelling, hard training images, and missing annotation.

Pose distillation loss function We finally evaluated the effect of loss function choice for pose knowledge distillation. To that end, we further tested a Cross-Entropy measurement based loss. Specifically, we first normalise the entire confidence map so that the sum of all pixel confidence scores is equal to 1, i.e. L1 normalisation. We then measure the divergence between the predicted and ground-truth confidence maps using the Cross-Entropy criterion. The results in Table 7 show that the MSE is a better choice in comparison to Cross-Entropy. The plausible reason is that MSE is also the formulation of the conventional supervision loss (Eqn (3)) therefore more compatible.

Parameter analysis of loss balance We evaluated the balance importance between the conventional MSE loss and the proposed pose knowledge distillation loss, as controlled by α in Eqn (5). Table 8 shows that equal importance (when $\alpha = 0.5$) is the optimal setting. This suggests that the two loss terms are similarly significant with the same numerical scale. On the other hand, we found that this parameter setting is not sensitive with a wide range of satisfactory values. This indicates that the teacher signal is not far away from the ground-truth labels (see Figure 3 Column (4)), possibly providing an alternative supervision as a replacement of the original joint confidence map labels.

5. Conclusion

In this work, we present a novel *Fast Pose Distillation* (FPD) learning strategy. In contrast to most existing human pose estimation methods, the FPD aims to address the under-studied and practically significant model cost-effectiveness quality in order to scale the human pose estimation models to large deployments in reality. This is made possible by developing a lightweight human pose CNN architecture and designing an effective pose structure knowledge distillation method from a large teacher model to a lightweight student model. Compared with existing model compression techniques such as network parameter binarisation, the proposed method achieves highly efficient human pose models *without* accuracy performance compromise. We have carried out extensive comparative evaluations on two human pose benchmarking datasets. The results suggests the superiority of our FPD approach in comparison to a wide spectrum of state-of-the-art alternative methods. Moreover, we have also conducted a sequence of ablation study on model components to provide detailed analysis and insight about the gains in model cost-effectiveness.

6. Acknowledgement

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[2] Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert Ormandi, George E Dahl, and Geoffrey E Hinton. Large scale distributed neural network training through online distillation. In *International Conference on Learning Representations*, 2018.

[3] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems*, 2014.

[4] Vasileios Belagiannis and Andrew Zisserman. Recurrent human pose estimation. 2017.

[5] Cristian Bucilua, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[6] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision*, 2016.

[7] Adrian Bulat and Georgios Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *IEEE International Conference on Computer Vision*, 2017.

[8] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[9] Xianjie Chen and Alan L Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in Neural Information Processing Systems*, 2014.

[10] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *IEEE International Conference on Computer Vision*, 2017.

[11] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L. Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[12] Xiaochuan Fan, Kang Zheng, Yuewei Lin, and Song Wang. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv*, 2015.

[14] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. *European Conference on Computer Vision*, 2016.

[15] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference*, 2010.

[16] Xu Lan, Xiatian Zhu, and Shaogang Gong. Knowledge distillation by on-the-fly native ensemble. In *Advances in Neural Information Processing Systems*, pages 7528–7538, 2018.

[17] Xu Lan, Xiatian Zhu, and Shaogang Gong. Person search by multi-scale matching. In *European Conference on Computer Vision*, 2018.

[18] Xu Lan, Xiatian Zhu, and Shaogang Gong. Self-referenced deep learning. In *Asian Conference on Computer Vision*, 2018.

[19] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, 2016.

[20] Xuecheng Nie, Jiashi Feng, Yiming Zuo, and Shuicheng Yan. Human pose estimation with parsing induced learner. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[21] G. Ning, Z. Zhang, and Z. He. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE Transactions on Multimedia*, PP(99):1–1, 2017.

[22] Xi Peng, Zhiqiang Tang, Fei Yang, Rogerio S. Feris, and Dimitris Metaxas. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[23] Xi Peng, Zhiqiang Tang, Fei Yang, Rogerio S Feris, and Dimitris Metaxas. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2226–2234, 2018.

[24] Umer Rafi, Bastian Leibe, Juergen Gall, and Ilya Kostrikov. An efficient convolutional network for human pose estimation. In *British Machine Vision Conference*, 2016.

[25] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv e-print*, 2014.

[26] Frank Seide, Gang Li, and Dong Yu. Conversational speech transcription using context-dependent deep neural networks. In *Twelfth annual conference of the international speech communication association*, 2011.

[27] Taiki Sekii. Pose proposal networks. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[28] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[29] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems*, 2014.

[30] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[31] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[32] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation. In *IEEE International Conference on Computer Vision*, 2017.

[33] Wei Yang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[34] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2878–2890, 2013.

[35] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[36] Xiang Yu, Feng Zhou, and Manmohan Chandraker. Deep deformation network for object landmark localization. In *European Conference on Computer Vision*, pages 52–70. Springer, 2016.

[37] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.

[38] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.