Efficient, Self-Supervised Human Pose Estimation with Inductive Prior Tuning

Nobline Yoo
Princeton University
nobliney@alumni.princeton.edu

Olga Russakovsky
Princeton University
olgarus@princeton.edu

Abstract

The goal of 2D human pose estimation (HPE) is to localize anatomical landmarks, given an image of a person in a pose. SOTA techniques make use of thousands of labeled figures (finetuning transformers or training deep CNNs), acquired using labor-intensive crowdsourcing. On the other hand, self-supervised methods re-frame the HPE task as a reconstruction problem, enabling them to leverage the vast amount of unlabeled visual data, though at the present cost of accuracy. In this work, we explore ways to improve self-supervised HPE. We (1) analyze the relationship between reconstruction quality and pose estimation accuracy, (2) develop a model pipeline that outperforms the baseline which inspired our work, using less than one-third the amount of training data, and (3) offer a new metric suitable for self-supervised settings that measures the consistency of predicted body part length proportions. We show that a combination of well-engineered reconstruction losses and inductive priors can help coordinate pose learning alongside reconstruction in a self-supervised paradigm.

1. Introduction

Applications of human pose estimation span a wide range, including predicting pedestrian behavior and trajectory on roads with autonomous vehicles [2, 16, 37, 42, 48]. SOTA works [10, 40] have explored larger models or tackled specific failure modes (*e.g.* occlusion). Broadly speaking, some of the latest HPE models address one or more of the following five questions. (1) Can we speed up prediction to enable real-time pose estimation [8, 22]? (2) Can we create lightweight models with smaller memory footprints [22, 25, 28, 39, 46]? (3) Pose estimation models lack robustness under situation X. How can we address this [10, 41, 45]? (4) Can we use transformers [7, 26, 40, 47]? (5) Vision-based pose estimation is unreliable. Can we use other more robust modes of data [9, 20]?

An important caveat is that many of these methods require labor-intensive pose labeling, which limits scalability in training data. To illustrate, COCO's training and validation set alone contain 1.7 million labeled keypoints [23]. MPII Human Pose contains more than 600,000 [1].

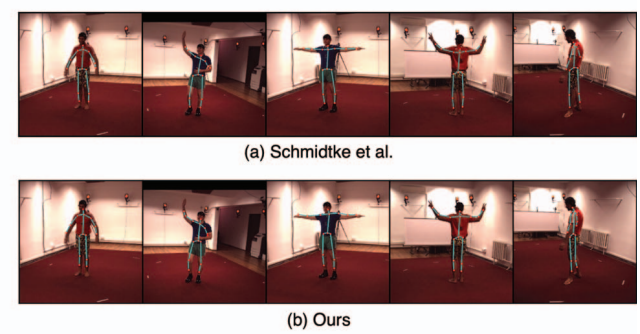

Figure 1. **(a) Baseline predictions** of [29]. **(b) Ours** (with MSE loss, $\boldsymbol{T}_{new}$ template, coarse-to-fine learning). Our predictions more closely follow the outline of the subjects.

To make use of the vast amount of unlabeled data, we look to self-supervised models, which frame classification and regression as reconstruction problems [5, 21, 29, 35], where given some parts of the input space (source), the model reconstructs other parts of the input space (target). These methods are engineered such that classification or regression are necessary to reason about the signals absent in the source but present in the target to be recovered. Hence, rather than optimizing directly for the task, self-supervision learns indirectly by optimizing for reconstruction, yielding representations that surpass the generalizability of those learned via supervised learning [34]. Self-supervised learning holds much potential for autonomous driving, facilitating greater scalability in training data, improving robustness, and enabling lifelong-learning [3, 6]. SOTA methods in this space typically rely on multi-view geometry [4, 12, 15, 36], unpaired pose data [17, 33], or synthetic datasets with pose labels [18].

Recently, Schmidtke et al. [29] use a template-based, self-supervised approach to estimate pose in the Human3.6M dataset [13], without relying on multi-view geometry, unpaired pose data, or synthetic datasets with pose labels. By guiding the model with templates of Gaussians, each representing a body part, the model is able to render

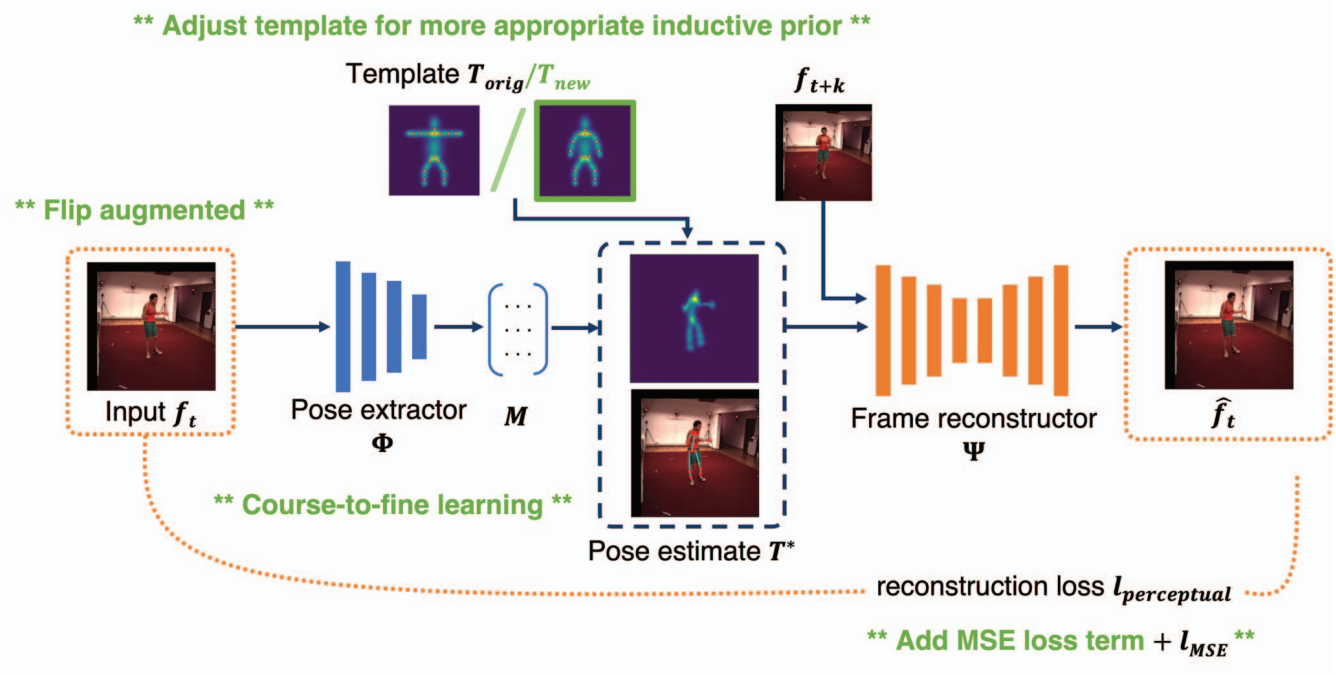

Figure 2. **Baseline model architecture.** Our adjustments to the pipeline are in **green**.

feasible pose estimates. However, without labels, there is no guarantee that predictions will reflect ground truth [31].

In this work, we analyze the relationship between reconstruction optimization and pose estimate accuracy in the absence of ground truth, using a carefully engineered combination of reconstruction loss and inductive prior (in the form of a new template). Using these insights, we develop an efficient model pipeline that exceeds the baseline [29] performance, using a training set that is less than one-third the size of the original, along with a new formulation of matrix transformation that implements coarse-to-fine learning and data augmentation. Further, we propose a metric of consistency for body part lengths that is suitable for self-supervised settings, where ground truth is not available. Our code is available at `https://github.com/princetonvisualai/hpe-inductive-prior-tuning`.

2. Approach

2.1. Baseline

We begin by describing the baseline model [29] which we build off of. It consists of pose extractor Φ and frame reconstructor Ψ (encoder-decoder), and a template T_{orig} of 18 Gaussians (one per body part), which form the inductive prior for viable human shapes (Figure 2). Network Φ is made up of 14 fully convolutional layers, followed by two fully-connected layers. Reconstructor Ψ consists of an encoder and decoder, each of which contain seven fully-convolutional layers.

Network Φ takes input frame f_t, which contains the pose to estimate, and outputs transformation matrices $M = \{M_{1\leq i\leq 18}\}$. MT_{orig} generates pose estimate T^*. Reconstructor Ψ takes T^* along with frame f_{t+k} (a frame with the same background and subject as f_t) to reconstruct f_t by jointly reasoning over the subject style and background information present in f_{t+k} and pose information from T^*.

The model is trained using three objectives: (1) an anchor-point loss to keep adjacent body parts together, (2) a boundary loss to keep the predicted keypoints within the physical 256×256 frame, and (3) a perceptual reconstruction loss with a VGG backbone pretrained on ImageNet to reflect human judgements of similarity [19]. l_{anchor} and $l_{boundary}$ are as defined in [29].

$$l_{recon} = ||VGG(\hat{f}_t) - VGG(f_t)||_1^1 \quad (1)$$

$$l_{tot} = l_{recon} + \lambda_1 l_{anchor} + \lambda_2 l_{boundary} \quad (2)$$

In baseline predictions, the chest, shoulder, and knee keypoints are relatively collapsed, while the general form only loosely follows the subject's outline (Figure 1a). Limb lengths are not always consistent; the second and third frame in Figure 1a depict the same subject from approximately the same camera angle; however, the length of the forearm is significantly reduced in the former, presenting an issue of inconsistency.

2.2. Change 1: MSE reconstruction loss

Since self-supervised HPE uses reconstruction optimization as a proxy for pose learning, we consider how differ-

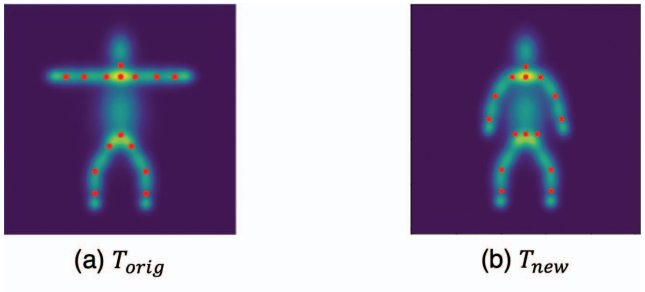

Figure 3. **(a) T_{orig}. (b) T_{new}.** The new template better reflects our data distribution.

ent formulations of reconstruction loss might improve pose estimates. While earlier works measure perceptual against pixel-wise loss [11, 14, 32], recent works in reconstruction use pixel-wise L_1 or L_2 in conjunction with perceptual loss [30, 38, 44]. So, we experiment with adding an MSE term to l_{recon} in Equation 1 to improve reconstruction, with l_{MSE} computed across all pixels and the RGB color channels. The reconstruction loss is now

$$l_{recon} = l_{MSE} + ||VGG(\hat{f}_t) - VGG(f_t)||_1^1 \quad (3)$$

2.3. Change 2: New template T_{new}

We engineer a template T_{new} shown in Figure 3b that better reflects the natural distribution of poses in our dataset, particularly the arms-down pose, providing a more appropriate inductive prior.

2.4. Change 3: Coarse-to-fine learning

Previous works have used multiple networks for coarse-to-fine learning to reconstruct finer details [27, 43]. Instead of using multiple networks, we modify the last layer of network Φ; since matrix M directly influences the model's representation capability, we expand M to hold 20 matrices ($M = \{M_{1 \leq i \leq 20}\}$) to estimate pose in two steps. By expanding the dimensions of matrix M and selectively mapping M_i (in a one-to-one or one-to-many manner), we enable coarse-to-fine learning of pose details without designing a separate network for fine reconstruction. Note that we only apply this two-step procedure to the arms, forearms, and hands, since this is where the +MSE, T_{new} model has the most room for improvement (Figure 12).

Concretely, previously, transformation matrices $M = \{M_{1 \leq i \leq 18}\}$. Instead, we expand M to $\{M_{1 \leq i \leq 14+6}\}$. In step 1, the first 14 matrices $M_{1 \leq i \leq 14}$ transform T_{new} to coarse estimate $T^{*'}$. In particular, M_{10} and M_{11} are applied to the whole left and right arm, respectively. Effectively, we treat each whole arm as a course unit. In step 2, the last six matrices $M_{15 \leq i \leq 20}$ dictate finer transformations of the individual components of the left and right arms (upper arm, forearm, hands) to get from $T^{*'}$ to final estimate T^*. Specifically, we apply $M_{1 \leq i \leq 20}$ as in Table 1.

M_i	Apply M_i to...
M_1	Core
M_2, M_3	Left/Right hip
M_4, M_5	Left/Right thigh
M_6, M_7	Left/Right shin
M_8, M_9	Left/Right shoulder
M_{10}	*Left upper arm, forearm, hand*
M_{11}	*Right upper arm, forearm, hand*
M_{12}, M_{13}	Left/Right foot
M_{14}	Head
M_{15}, M_{18}	Left/Right upper arm
M_{16}, M_{19}	Left/Right forearm
M_{17}, M_{20}	Left/Right hand

Table 1. **Coarse-to-fine learning.** Steps 1 and 2 in coarse-to-fine pose estimation.

2.5. Change 4: Flip dataset augmentation

We adopt a simple dataset augmentation approach to further improve model training. Concretely, as is standard in image classification, we augment the dataset by flipping the input images across the longitudinal axis to overcome potential discrepancies in distribution between the two sides.

2.6. Change 5: Constraining for consistency

Finally, this brings us to our last contribution of a new metric that helps better constrain the model predictions. Resuming our discussion on issues of inconsistent limb lengths in baseline predictions (Section 2.1), we note that in self-supervised HPE, it becomes important to explicitly code for those things which are automatically coded for in their fully-supervised counterparts via ground truth labels; consistency is one such example. To this day, definitions of "consistency" in HPE have largely been kept to the 3D setting, where it is defined as consistency of 3D representation across camera views for the same subject and pose and is tackled by creating new, reprojection losses [12, 21, 35, 36]. Others have directly addressed inconsistency in limb lengths by learning limb length priors in the 3D supervised setting [24].

We propose a metric for consistency in body part length proportions across frames that can be used in self-supervised settings with no ground truth labels (in the 2D setting, but it can be extended to the 3D setting). We define body part length proportion (`BPLP`) as the proportion of predicted limb length to torso length (Equation 4). The central motivation is that a single subject maintains a level of consistent `BPLP` across different poses. We formulate `BPLP` consistency (`BPLP-C`) as the reciprocal of the standard deviation in `BPLP`s per body part (Equation 5). A

higher BPLP-C is indicative of more consistent predictions.

$$\texttt{BPLP}(\text{limb } i) = \frac{\text{predicted length(limb } i)}{\text{predicted length(torso)}} \tag{4}$$

$$\texttt{BPLP-C} = \frac{1}{\frac{1}{n}\sum_{1 \le i \le n} \sigma_{\texttt{BPLP},i}} \tag{5}$$

$\sigma_{\texttt{BPLP},i}$: std. dev. of BPLP(limb i) over text examples

n: number of limbs

We note how ground truth labels are not needed to calculate BPLP or BPLP-C, making them suitable for evaluating models in the self-supervised setting. To improve BPLP-C, we work on constraining M, since M influences the space of pose estimates. We see this experiment as a proof-of-concept to improve *consistency*, rather than as a method to improve specific performance metrics, like PDJ or L_2 error. We approach the problem as follows. For simplicity, we assume that BPLPs are consistent across all subjects. We recognize this does not hold in the real-world; but for simplicity, we work with this assumption. In concluding, we briefly discuss how we can do away with this assumption in future work by computing subject-specific BPLPs.

The major theoretical constraint we introduce is to apply scaling together on all limbs, since the size of individual limbs does not change drastically and disproportionately across poses, relative to other limbs, for a single subject. Hence, we have a new transformation matrix M_i that is derived as follows:

$$M_i = RLS \tag{6}$$

$$M_i = \underbrace{\begin{bmatrix} cos(\theta) & sin(\theta) & 0 \\ -sin(\theta) & cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{R} \underbrace{\begin{bmatrix} 1 & 0 & \mu \\ 0 & 1 & \delta \\ 0 & 0 & 1 \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} \phi & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{S}$$

$$T^* = MT \tag{7}$$

In this constrained setting, T should be a subject-specific template, since M is not free to scale individual limbs, but in this paper we assume a generic template. R (rotation) and L (localization) are for rotating and repositioning individual limbs, respectively. S (scaling) is for scaling all limbs together.

θ, μ, δ are limb-specific parameters, while ϕ and β are frame-specific parameters, in that for a given frame with a subject in some pose, there will be 3 parameters for each limb and 2 frame-specific parameters that apply to all the limbs. Hence, in the original baseline setting, there would be $(18 \times 3) + 2 = 56$ salient, transformation parameters, whereas in the coarse-to-fine setting, there would be $(20 \times 3) + 2 = 62$ such parameters.

3. Experiments

3.1. Setup

Similar to the baseline [29], we train our models on Human3.6M subjects 1, 5, 6, 7, and 8, evaluate on subjects 9 and 11, and group frames based on subject and background. We downsample frames to 256×256. To encourage training efficiency, we keep the training set relatively small in all experiments—around 180K pairs of frames (f_t, f_{t+k}). Our model is trained over 50 epochs on 1 NVIDIA A100/V100 with a learning rate of 0.001, batch size of 48, $\lambda_1 = 0.5$, and $\lambda_2 = 1$. Training typically takes 34 hours; we evaluate our model on 130K images.

3.2. Evaluation

We use two overall evaluation metrics on a set J of 15 keypoints: Percentage of Detected Joints (PDJ) (Equation 8) and L_2 error normalized for frame size (Equation 11).

$$\text{PDJ@0.05} = \frac{1}{|J|}\sum_{j \in J} f(j, \hat{j}) \tag{8}$$

$$f(j, \hat{j}) = 1 \text{ if } \text{dist}(j, \hat{j}) \le .05 \times \text{diagonal person length} \tag{9}$$

$$\text{Per-joint accuracy } = \frac{1}{\text{\# test instances}} \sum_{\text{test instances}} f(j, \hat{j}) \tag{10}$$

$$L_2 \text{ error} = \frac{1}{|J|}\sum_{j \in J} \frac{\text{dist}(j, \hat{j})}{\text{image size}} \tag{11}$$

The model detects joint j if estimate $\hat{j}$ is within 0.05 of the diagonal length of the person bounding box. A higher PDJ and lower L_2 error are characteristics of a more accurate model. As with the baseline, we do not predict pose orientation. We hope to address this in future work. In the meantime, we use a frame-centric lens to describe joint-handedness. Before we explore our modifications to the pipeline, we analyze the baseline in the next section.

3.3. Baseline results

First, please note that there are reproducibility issues in the baseline's codebase (which we confirmed with the authors). Going forward, we distinguish between the "baseline" (obtained from the codebase), upon which we implement our proposed changes, and the published checkpoint, against which we ultimately compare our best model.

The baseline yields a PDJ@0.05 of 38.5 and normalized L_2 error of 7.2 (Table 2). Surprisingly, the baseline has a better L_2 error than the published checkpoint, despite a slightly worse PDJ. Figure 4 shows the published checkpoint yields more outliers, with keypoints that overshoot the image frame boundary. Quantitatively, Figure 5, which shows the distribution of L_2 by model, confirms the same.

Model	PDJ	L_2 Error
Published checkpoint	40.8	11.0
Baseline	38.5	7.2
+MSE	26.6	9.2
+T_{new}	33.1	7.5
+MSE, T_{new}	37.2	6.7
+MSE, T_{new}, flip augment	34.3	6.9
+MSE, T_{new}, coarse-to-fine	39.0	7.0
+MSE, T_{new}, coarse-to-fine, flip augment	**42.6**	**6.4**
+MSE, T_{new}, coarse-to-fine, flip augment, constrained M	38.7	7.0

Table 2. **Evaluation metrics by model.** The +MSE, T_{new}, coarse-to-fine, flip augment model yields the best PDJ and L_2.

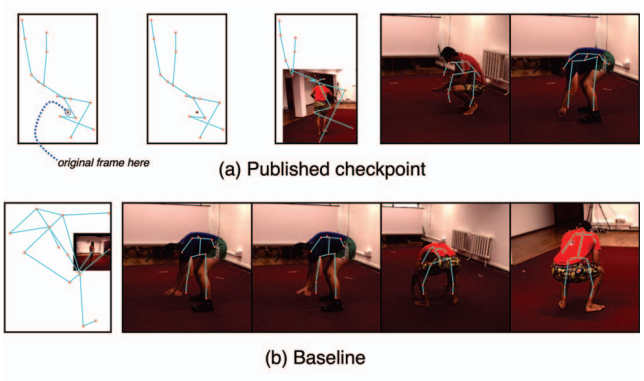

Figure 4. **The bottom 0.5% of worst predictions from the published checkpoint and baseline. (a)** In the first two, the predicted keypoints are so out-of-bounds; the original frame is barely visible. **(b)** The baseline yields less severe outliers.

Next, we explore reconstruction quality in relation to pose estimate accuracy. Figure 6b shows the baseline's reconstruction of f_t after 50 epochs. The torso and legs show higher-quality reconstruction and keypoint grounding. The opposite is true for the elbow and wrists, which show lower-quality reconstruction and keypoint grounding. To improve overall keypoint grounding, we propose modifying the reconstruction loss by adding a pixel-wise loss term: MSE.

One possible explanation for the disparity in reconstruction quality between the arms and legs comes from the distribution of the dataset—the left/right elbow and wrist exhibit the most variation in spatial configuration (Figure 7).

3.4. Results overall

We now briefly compare our full model with [29]'s published checkpoint before diving into a detailed ablation study of the proposed changes. [29] attained a PDJ of 40.8 and L_2 error of 11.0 using 600K images trained over 30 epochs. In comparison, our best model (+MSE, T_{new}, coarse-to-fine, flip augment) reaches a PDJ of 42.6 and L_2 error of 6.4 using only 180K input examples trained over 50 epochs.

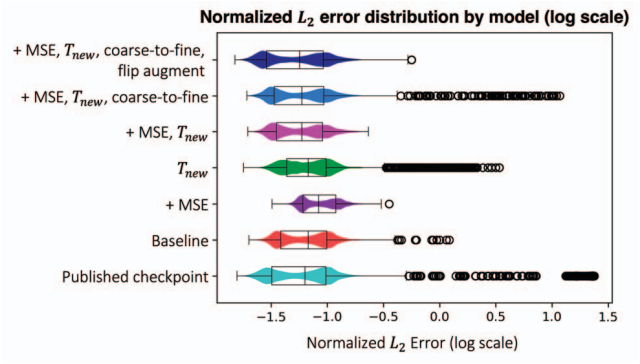

Figure 5. **Error by model.** We compute the normalized L_2 error (on a log scale) per frame and plot its distribution by model.

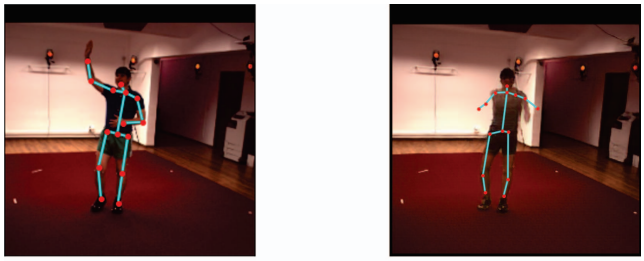

Figure 6. **Reconstruction after 50 epochs.** The model is given only the image during training. **(a)** The ground truth pose shown for illustration purposes. **(b)** Baseline reconstruction of f_t.

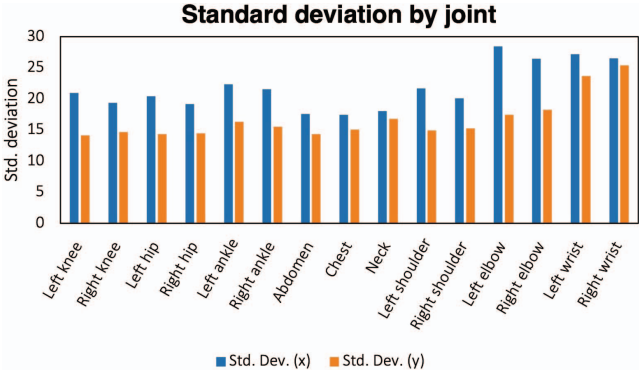

Figure 7. **Standard deviation of (x, y) coordinates of ground truth joints.** The left and right elbow and wrist exhibit the most variation.

Furthermore, by the time our model reaches epoch 10, our L_2 is less than the checkpoint's at 7.3. By epoch 30, our PDJ is comparable with the checkpoint's at 40.4 (using three times fewer training samples) and by epoch 40, surpasses it at 41.8.

Analyzing Figure 1, we observe that (1) our model predictions follow the contour of subject poses cleaner, seeing how the right wrist lies outside the subject in the fifth frame

Model	$\text{Error}_{\text{recon}}$ (L_2)	$\text{Error}_{\text{pose}}$ (L_2)
Baseline	5414.5	7.2
+MSE	4242.0	9.2
+MSE, T_{new}	**3797.9**	**6.7**

Table 3. **Comparison between reconstruction and pose estimate error after 50 epochs. (a) Baseline. (b) +MSE, T_{orig}.** Reconstruction error↓, Pose estimate error↑. **(c) +MSE, T_{new}.** Reconstruction error↓, Pose estimate error↓.

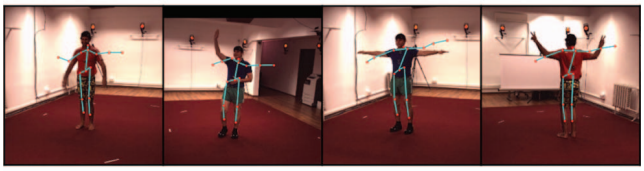

Figure 8. **+MSE model sample predictions.** The model consistently predicts an arms-out pose, regardless of the input frame f_t.

in Figure 1a ([29]'s prediction), (2) our predicted keypoints for the neck and shoulder are less sunken, as observed in the third and fourth frame compared across Figures 1a and 1b, (3) our knee joint estimates are higher up (*i.e.* generally more accurate), as observed in the first four frames compared across Figures 1a and 1b.

3.5. Effect of the new reconstruction loss

We now dive into the individual proposed changes and analyze them one-by-one. We begin by experimenting with the new reconstruction loss described in Section 2.2. Despite the initial hypothesis that improving reconstruction quality would improve pose estimates, we find this is not necessarily the case. Adding an MSE loss term speeds up reconstruction; but the model is worse at grounding keypoints, as shown by the lower reconstruction error accompanied by higher pose estimate error (Table 3). In fact, the +MSE model exhibits worse performance than the baseline (PDJΔ:-11.9 points; $L_2\Delta$:$+2.0$) (Table 2).

Across the four sample predictions in Figure 8, the leg joints are learned reasonably; however, the arms are consistently predicted to be extended outward, resembling the arms-out pose in template T_{orig}, regardless of whether the input frame contains a subject with arms up or down.

Notably, T_{orig} (Figure 3a) is highly unreflective of the pose distribution in the dataset. T_{orig} represents an arms-out pose; however, the test dataset has nearly a $40:1$ ratio of arms-down poses to arms-extended ("arms-out") poses. For simplicity, we assume the test and training set have a similar distribution, since the data was randomly sorted into training/test sets.

Our model is framed as a problem of template-matching (*i.e.* distribution-matching) between estimate T^* and reconstructed frame $\hat{f}_t$. In this context, adding an MSE term to the l_{recon} encourages the model to reconstruct faster, and in the process, to quickly pick up correlations between the transformed template T^* and original frame f_t. However, because the base template is so far from the data distribution, the model learns spurious correlations between T^* and $\hat{f}_t$, resulting in misalignment between reconstruction and pose estimation.

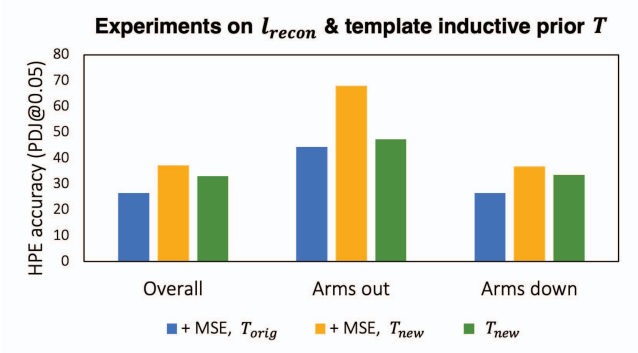

Figure 9. **Effect of adopting T_{new} on pose estimate accuracy.** Adopting T_{new} shows performance improvements when paired with the addition of an MSE loss term.

In pushing the model toward faster, higher-quality reconstruction, it becomes even more important to provide an *appropriate* template that reflects the underlying structure in the data, since the template provides a key inductive prior for learning meaningful relationships between pose estimation and reconstruction.

3.6. Effect of the new template

Adopting the new template T_{new} (Figure 3b) with MSE exhibits a strong lead over T_{orig} with MSE (PDJΔ:$+10.6$ points; $L_2\Delta$:-2.5) and performs comparably with the baseline on PDJ and even outperforms the baseline's L_2 error. T_{new} with MSE also outperforms using T_{new} alone (PDJΔ:$+4.1$ points; $L_2\Delta$:-0.8). When T_{new} is not paired with MSE (*i.e.* the incentive to reconstruct faster), model performance drops overall, but most severely on arms-out poses (relative to T_{new} paired with MSE) (Figure 9).

Furthermore, in Table 3, we see that the combination of MSE with T_{new} reduces both reconstruction and pose estimate error (reconstruction error Δ:-1616.6 points; pose estimation error Δ:-0.5). Engineering an appropriate inductive prior is key to coordinating reconstruction with pose learning. Given the advantage of the MSE, T_{new} combination, we use it as the new local benchmark against which we compare subsequent models. Next, we discuss our experiments updating the model to facilitate coarse-to-fine learning to refine the arm and forearm estimates.

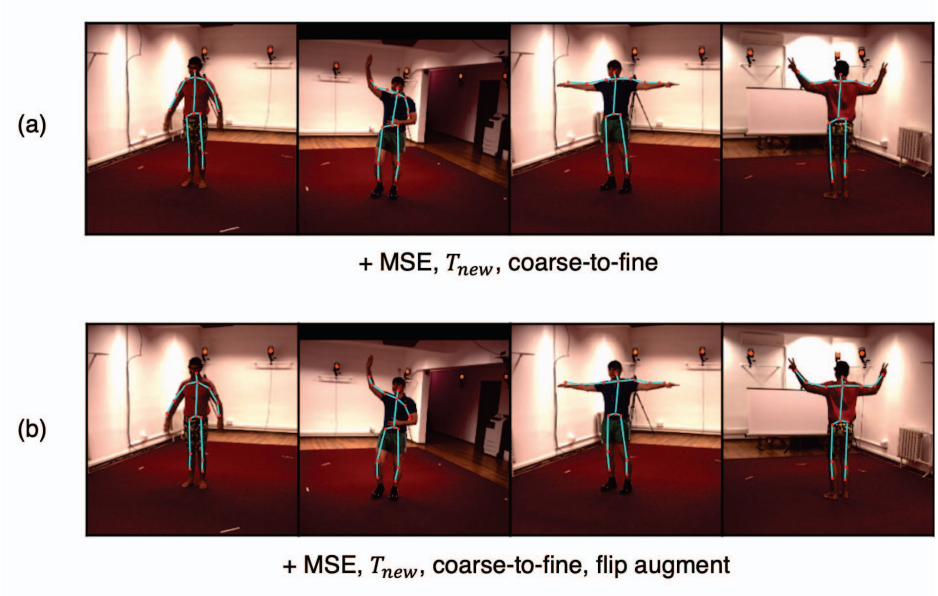

Figure 10. **Sample predictions from models using coarse-to-fine learning, augmentation. (a)** +MSE, T_{new}, coarse-to-fine model. **(b)** +MSE, T_{new}, coarse-to-fine, flip augment model.

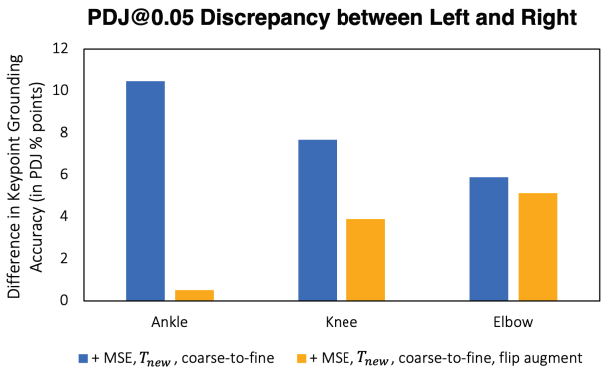

Figure 11. **Difference in PDJ@0.05 between left and right-side ankle, knee, and elbow (the most extreme cases).** The discrepancy decreases after augmentation.

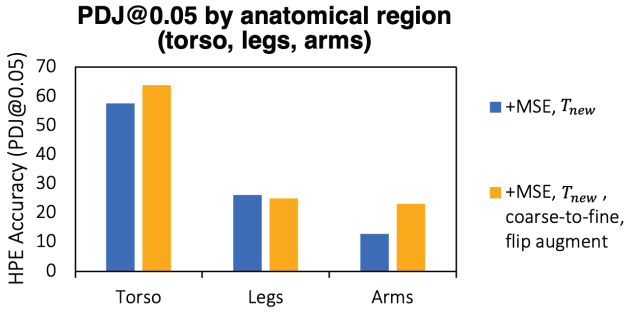

Figure 12. **Model accuracy per anatomical region before/after coarse-to-fine learning and augmentation.** Torso keypoints: abdomen, chest, neck, hips, shoulders. Leg keypoints: knees, ankles. Arm keypoints: elbows, wrists.

3.7. Effect of coarse-to-fine learning

Expanding M and adapting a coarse-to-fine learning strategy described in Section 2.4 yields PDJΔ:+1.8 points and $L_2\Delta$:+0.3 (Table 2). Examining the L_2 distribution, it seems that the slight increase in L_2 error comes from the increase in outliers (Figure 5). On qualitative examples (Figure 10a), coarse-to-fine learning yields pose estimates that closely follow subject contour.

One thing we notice in examining the per-joint accuracy deeper is the discrepancy in keypoint accuracy between left and right-side joints. For the ankle, knee, and elbow (the most extreme cases), the absolute differences are 10.5, 7.7, and 5.9 percentage points (Figure 11). To account for natural differences in distribution between the two sides, we propose data diversification via augmentation.

3.8. Effect of dataset augmentation

We additionally augment the dataset by horizontal flipping as mentioned in Section 2.5. To keep the training size to around 180K pairs, we take approximately half of the frames in the original training set and flip them, yielding a total of 181,728 training pairs (f_t, f_{t+k}), compared to 181,383, pre-augmentation.

After augmenting the dataset, the absolute differences in PDJ for the ankle, knee, and elbow drop to 0.5, 3.9, 5.1 percentage points, respectively (Figure 11). The accuracy on arm keypoints jumps 10.2 percentage points (Figure 12). Overall PDJ jumps to 42.6 (the highest yet) and L_2 error drops to 6.4 (the lowest yet). To analyze whether this jump

Model	BPLP-C
Published checkpoint	2.03
Baseline	9.82
+MSE, T_{new}, coarse-to-fine, flip augment	5.17
+MSE, T_{new}, coarse-to-fine, flip augment, constrained M	12.32

Table 4. **BPLP consistency across four models.** Constraining M yields the highest BPLP-C.

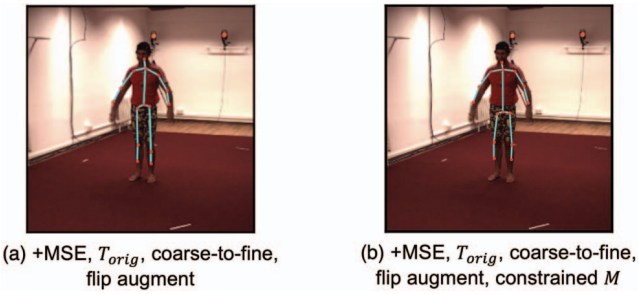

Figure 13. **Sample prediction with/without constrained M. (a)** Without constrained M. **(b)** With a constrained M, left limbs (e.g. left thigh) are relatively more consistent in length with right limbs (e.g. right thigh).

in performance is solely due to the addition of flip augmentation or whether it is a result of the specific combination of flip augmentation with coarse-to-fine learning, we also train a model using MSE, T_{new}, and flip augmentation with the original, unexpanded $M_{1\leq i\leq 18}$. Compared with the +MSE, T_{new} model, the performance drops when only augmentation is added (PDJΔ:-2.9 points; $L_2\Delta$:$+0.2$) (Table 2). Hence, we see that flip augmentation is working together with the expanded M to produce more accurate predictions.

With respect to outliers, we can see from the L_2 distribution (Figure 5) that the combination of MSE, T_{new}, coarse-to-fine learning, and augmentation yields a model with one of the fewest number of outliers.

Despite its success, we notice inconsistent body part length proportions in some of its predictions. Similar to the baseline, the second and third frame in Figure 10b depict significantly different predicted forearm lengths, despite showing the same subject. Next, we offer a new metric, BPLP-C, to measure this type of inconsistency and experiment with constraining M to improve on this metric.

3.9. Effect of constraining M for consistency

Finally, we constrain M for consistency as described in Section 2.6. Constraining M yields a PDJ of 38.7 and L_2 error of 7.0, which is comparable with the MSE, T_{new}, coarse-to-fine model. Table 4 shows BPLP-C by model. We notice that constraining M yields the highest BPLP-C and more consistency between left and right limb lengths (Figure 13b). This serves as a preliminary proof-of-concept that constraining the transformation matrix M along these axes is a viable method to encourage consistent body part length proportions.

4. Conclusion

In this paper, we present a modified self-supervised model pipeline that sets a new benchmark in PDJ@0.05 and normalized L_2 error for 2D human pose estimation that is based on a transformable, Gaussian shape template. Building off [29], we make several new contributions:

1. We analyze the settings in which reconstruction speed-up helps or hurts pose estimation and identify the importance of tuning the inductive prior to reflect some aspect of the data distribution, thereby enabling template-matching and coordinating reconstruction with pose estimation.

2. By proposing ways to combine reconstruction loss, data augmentation, inductive prior tuning, and network-level adjustments, we find a model pipeline that exceeds baseline performance using approximately three times less data.

3. We propose BPLP-C, a metric that can be used to measure consistency in part length proportions in the absence of ground truth and propose one way to influence it—by constraining the transformation matrix.

We consider a few ideas for extensions and future work. To improve our model performance further, it would be a good idea to incorporate some understanding of pose orientation into the model, so that it knows when the subject is facing away or toward the camera. It would be beneficial to test our pipeline on other HPE datasets to check for generalization. To relax the assumption we made about BPLPs being consistent across all subjects, we could create an end-to-end framework that learns not only the poses, but also, subject-specific templates based on a few different poses of the same subject. Furthermore, to measure the value of the BPLP-C metric, we hope to conduct human studies to understand how humans perceive predictions from models with higher BPLP-C in the 2D and 3D setting.

5. Acknowledgements

This work was done as part of NY's undergraduate senior thesis. We are grateful to Princeton Research Computing for providing the compute resources, and to the Princeton SEAS Howard B. Wentz, Jr. Junior Faculty Award (to OR) for enabling the publication and in-person presentation of this research.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014.

[2] Peter Bauer, Arij Bouazizi, Ulrich Kressel, and Fabian B. Flohr. Weakly supervised multi-modal 3d human body pose estimation for autonomous driving. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–7, 2023.

[3] Lars Berscheid, Pascal Meißner, and Torsten Kröger. Self-supervised learning for precise pick-and-place without object model. *IEEE Robotics and Automation Letters*, 5(3):4828–4835, 2020.

[4] Arij Bouazizi, Julian Wiederer, Ulrich Kressel, and Vasileios Belagiannis. Self-supervised 3d human pose estimation with multiple-view geometry. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8, 2021.

[5] Ting Cao, Mohammad Ali Armin, Simon Denman, Lars Petersson, and David Ahmedt-Aristizabal. In-Bed Human Pose Estimation from Unseen and Privacy-Preserving Image Domains. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2022.

[6] Xinke Deng, Yu Xiang, Arsalan Mousavian, Clemens Eppner, Timothy Bretl, and Dieter Fox. Self-supervised 6d object pose estimation for robot manipulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3665–3671, 2020.

[7] Moritz Einfalt, Katja Ludwig, and Rainer Lienhart. Uplift and upsample: Efficient 3d human pose estimation with uplifting transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2903–2913, 2023.

[8] Nicola Garau and Nicola Conci. CapsulePose: A variational CapsNet for real-time end-to-end 3D human pose estimation. *Neurocomputing*, 523:81–91, 2023.

[9] Jiaqi Geng, Dong Huang, and Fernando De la Torre. Densepose from wifi. *arXiv preprint arXiv:2301.00250*, 2022.

[10] Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, and Han Hu. Human pose as compositional tokens. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 660–671, 2023.

[11] Vahid Ghodrati, Jiaxin Shao, Mark Bydder, Ziwu Zhou, Wotao Yin, Kim-Lien Nguyen, Yingli Yang, and Peng Hu. MR image reconstruction using deep learning: evaluation of network structure and loss functions. *Quantitative Imaging in Medicine and Surgery*, 9(9):1516–1527, 2019.

[12] Mohsen Gholami, Ahmad Rezaei, Helge Rhodin, Rabab Ward, and Z Jane Wang. Self-supervised 3d human pose estimation from video. *Neurocomputing*, 488:97–106, 2022.

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Computer Vision – ECCV 2016*, pages 694–711, 2016.

[15] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1077–1086, 2019.

[16] Viktor Kress, Janis Jung, Stefan Zernetsch, Konrad Doll, and Bernhard Sick. Human pose estimation in real traffic scenes. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 518–523, 2018.

[17] Jogendra Nath Kundu, Siddharth Seth, Varun Jampani, Mugalodi Rakesh, R Venkatesh Babu, and Anirban Chakraborty. Self-supervised 3d human pose estimation via part guided novel image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6152–6162, 2020.

[18] Jogendra Nath Kundu, Siddharth Seth, Pradyumna YM, Varun Jampani, Anirban Chakraborty, and R. Venkatesh Babu. Uncertainty-aware adaptation for self-supervised 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20448–20459, 2022.

[19] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690, 2017.

[20] Shih-Po Lee, Niraj Prakash Kini, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. Hupr: A benchmark for human pose estimation using millimeter wave radar. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5715–5724, 2023.

[21] Yang Li, Kan Li, Shuai Jiang, Ziyue Zhang, Congzhentao Huang, and Richard Yi Da Xu. Geometry-Driven Self-Supervised Method for 3D Human Pose Estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11442–11449, 2020.

[22] Yanping Li, Ruyi Liu, Xiangyang Wang, and Rui Wang. Human pose estimation based on lightweight basicblock. *Machine Vision and Applications*, 34(1):3, 2022.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, 2014.

[24] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Hai Ci, and Yizhou Wang. Context modeling in 3d human pose estimation: A unified perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6238–6247, 2021.

[25] Daniil Osokin. Real-time 2d multi-person pose estimation on cpu: Lightweight openpose. *arXiv preprint arXiv:1811.12004*, 2018.

[26] Xiaoye Qian, Youbao Tang, Ning Zhang, Mei Han, Jing Xiao, Ming-Chun Huang, and Ruei-Sung Lin. Hstformer:

Hierarchical spatial-temporal transformers for 3d human pose estimation. *arXiv preprint arXiv:2301.07322*, 2023.

[27] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning Detailed Face Reconstruction from a Single Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5553–5562, 2017.

[28] Nicholas Santavas, Ioannis Kansizoglou, Loukas Bampis, Evangelos Karakasis, and Antonios Gasteratos. Attention! A Lightweight 2D Hand Pose Estimation Approach. *IEEE Sensors Journal*, 21(10):11488–11496, 2021.

[29] Luca Schmidtke, Athanasios Vlontzos, Simon Ellershaw, Anna Lukens, Tomoki Arichi, and Bernhard Kainz. Unsupervised human pose estimation through transforming shape templates. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2484–2494, 2021.

[30] Prafull Sharma, Ayush Tewari, Yilun Du, Sergey Zakharov, Rares Andrei Ambrus, Adrien Gaidon, William T Freeman, Fredo Durand, Joshua B Tenenbaum, and Vincent Sitzmann. Neural groundplans: Persistent neural scene representations from a single image. In *The Eleventh International Conference on Learning Representations*, 2022.

[31] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-Metric Loss for Self-supervised Learning of Depth and Egomotion. In *Computer Vision – ECCV 2020*, pages 572–588, 2020.

[32] Jake Snell, Karl Ridgeway, Renjie Liao, Brett D. Roads, Michael C. Mozer, and Richard S. Zemel. Learning to generate images with perceptual similarity metrics. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 4277–4281, 2017.

[33] Jose Sosa and David Hogg. Self-supervised 3d human pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4787–4796, 2023.

[34] Atharva Tendle and Mohammad Rashedul Hasan. A study of the generalizability of self-supervised representations. *Machine Learning with Applications*, 6:100124, 2021.

[35] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Self-supervised 3d hand pose estimation through training by fitting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10853–10862, 2019.

[36] Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, and Bodo Rosenhahn. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13294–13304, 2021.

[37] Sijia Wang, Fabian B. Flohr, Hui Xiong, Tuopu Wen, Baofeng Wang, Mengmeng Yang, and Diange Yang. Leverage of Limb Detection in Pose Estimation for Vulnerable Road Users. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 528–534, 2019.

[38] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402 Vol.2, 2003.

[39] Dingning Xu, Lijun Guo, Rong Zhang, Jiangbo Qian, and Shangce Gao. Can relearning local representation help small networks for human pose estimation? *Neurocomputing*, 518:418–430, 2023.

[40] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022.

[41] Cheng-Yen Yang, Jiajia Luo, Lu Xia, Yuyin Sun, Nan Qiao, Ke Zhang, Zhongyu Jiang, Jenq-Neng Hwang, and Cheng-Hao Kuo. Camerapose: Weakly-supervised monocular 3d human pose estimation by leveraging in-the-wild 2d annotations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2924–2933, 2023.

[42] Andrei Zanfir, Mihai Zanfir, Alex Gorban, Jingwei Ji, Yin Zhou, Dragomir Anguelov, and Cristian Sminchisescu. Hum3dil: Semi-supervised multi-modal 3d humanpose estimation for autonomous driving. In *Proceedings of The 6th Conference on Robot Learning*, volume 205, pages 1114–1124, 2023.

[43] Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2315–2324, 2019.

[44] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2737–2746, 2020.

[45] Zhongyang Zhang, Kaidong Chai, Haowen Yu, Ramzi Majaj, Francesca Walsh, Edward Wang, Upal Mahbub, Hava Siegelmann, Donghyun Kim, and Tauhidur Rahman. Neuromorphic high-frequency 3d dancing pose estimation in dynamic environment. *Neurocomputing*, page 126388, 2023.

[46] Zhe Zhang, Jie Tang, and Gangshan Wu. Simple and lightweight human pose estimation. *arXiv preprint arXiv:1911.10346*, 2020.

[47] Shuaitao Zhao, Kun Liu, Yuhang Huang, Qian Bao, Dan Zeng, and Wu Liu. DPIT: Dual-Pipeline Integrated Transformer for Human Pose Estimation. In *Artificial Intelligence*, pages 559–576, 2022.

[48] Jingxiao Zheng, Xinwei Shi, Alexander Gorban, Junhua Mao, Yang Song, Charles R. Qi, Ting Liu, Visesh Chari, Andre Cornman, Yin Zhou, Congcong Li, and Dragomir Anguelov. Multi-modal 3d human pose estimation with 2d weak supervision in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4478–4487, 2022.