PoseBH: Prototypical Multi-Dataset Training Beyond Human Pose Estimation

Uyoung Jeong[1] Jonathan Freer[2] Seungryul Baek[1] Hyung Jin Chang[2] Kwang In Kim[3]
[1]UNIST [2]University of Birmingham [3]POSTECH

Abstract

We study multi-dataset training (MDT) for pose estimation, where skeletal heterogeneity presents a unique challenge that existing methods have yet to address. In traditional domains, e.g. regression and classification, MDT typically relies on dataset merging or multi-head supervision. However, the diversity of skeleton types and limited cross-dataset supervision complicate integration in pose estimation. To address these challenges, we introduce PoseBH, a new MDT framework that tackles keypoint heterogeneity and limited supervision through two key techniques. First, we propose nonparametric keypoint prototypes that learn within a unified embedding space, enabling seamless integration across skeleton types. Second, we develop a cross-type self-supervision mechanism that aligns keypoint predictions with keypoint embedding prototypes, providing supervision without relying on teacher-student models or additional augmentations. PoseBH substantially improves generalization across whole-body and animal pose datasets, including COCO-WholeBody, AP-10K, and APT-36K, while preserving performance on standard human pose benchmarks (COCO, MPII, and AIC). Furthermore, our learned keypoint embeddings transfer effectively to hand shape estimation (InterHand2.6M) and human body shape estimation (3DPW). The code for PoseBH is available at: https://github.com/uyoung-jeong/PoseBH.

1. Introduction

The demand for human pose estimators has grown substantially in recent years, driving their adoption across diverse applications, including 3D avatar generation [14, 25, 31], motion generation [15, 22, 33], human-robot interaction [20, 44, 59], safety monitoring [7, 18, 39], and VR/AR tracking [12, 43, 47]. These domains pose distinct operational challenges, often necessitating significant domain adaptation for robust pose estimation. To address this, we aim to train a human pose estimator that generalizes effectively across multiple datasets.

However, multi-dataset training (MDT) for pose estimation presents two key challenges. First, keypoints

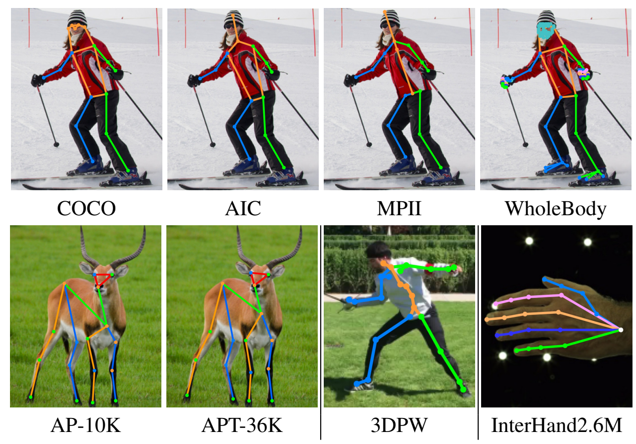

Figure 1. PoseBH unifies diverse skeleton formats, including humans, hands, and animals. The displayed skeletons show pose estimation results from our method, with 3DPW and InterHand2.6M predictions from a transferred model.

are heterogeneous across different datasets. For example, COCO [23] and MPII [1] define 17 and 16 keypoints, respectively, with 9 non-overlapping ones. Even when datasets share keypoints with identical names, their localization characteristics often differ due to domain gaps. For example, while COCO and animal datasets (e.g., AP-10K [58] and APT-36K [56]) use common keypoint names such as eyes, nose, and hips, significant anatomical differences exist (see Fig. 1). As a result, naïvely merging the identically named keypoints is ineffective.

Second, pose estimation MDT suffers from a lack of adequate supervision. In MDT, multiple skeleton types from different datasets are trained for each input instance. However, only a single skeleton-type ground truth is provided per input, leaving others unsupervised. This limitation resembles a semi-supervised learning challenge, constraining performance on smaller datasets. Existing semi-supervised learning methods typically rely on input augmentation or teacher-student knowledge distillation, both of which require substantial extra computational overhead.

Existing approaches fail to fully address these challenges, often relying on a shared backbone network [6] or a mixture of experts [53] to preserve downstream performance. However, these methods tend to focus on mainstream datasets like COCO or MPII, resulting in degraded performance on underrepresented datasets. We propose a method that enhances performance in discrepant domains while maintaining accuracy on standard human pose datasets. A straightforward fine-tuning approach on the target domain often disrupts performance on source datasets, as it fails to balance differences in skeleton structures and dataset distributions. Therefore, effective generalization requires jointly harmonizing skeletal representations and dataset diversity during training.

We achieve this through two primary techniques. First, we introduce prototypes to represent keypoints as embedding vectors, providing a unified representation across datasets. We regress pixel-wise keypoint embeddings from backbone features and match them to corresponding prototypes, automating keypoint mapping across datasets. This enables the integration of diverse keypoint types, including whole-body and animal keypoints, despite significant skeletal differences. Furthermore, our keypoint prototypes are non-learnable parameters, offering adaptability and computational efficiency for domain transfer.

Second, we generate reliable self-supervision signals using the predicted keypoints and their embeddings, a process we term *cross-type self-supervision*. Since keypoint predictions can be obtained via a dot product between embeddings and prototypes, we effectively establish two distinct keypoint prediction modalities. By aligning these predictions, we filter out noisy keypoints and produce reliable labels for supervising unlabeled keypoints. This process enhances the utility of embeddings without requiring additional networks or input augmentation. As a result, we provide effective supervision for underrepresented and discrepant datasets, achieving a higher degree of generalizability.

Our main technical contributions are as follows:

- Keypoint prototypes for learning arbitrary keypoints from multiple datasets ensuring high transferability.
- A cross-type self-supervision mechanism that aligns keypoint regression outputs with keypoint embeddings, enriching supervision for unlabeled keypoints.
- Strong generalization across estimation tasks, including whole-body, animal, and hand pose estimation, as well as human shape estimation.

2. Related works

Multi-dataset training has been widely explored in various computer vision tasks. Relabeling and pseudo-labeling are among the most straightforward yet resource-intensive approaches. For example, MSeg [19] employed relabeling for segmentation, while [63] used offline pseudo-labeling for object detection. However, both methods face scalability challenges, as adding a new dataset requires reprocessing annotations for all previously processed datasets to redefine the class set, leading to a quadratic computational cost.

Several approaches have been proposed to improve scalability. In object detection, [4, 65] aim to unify object class labels by merging similar categories. However, they struggle to integrate localization features of the same object due to the absence of spatial information in class logits. UniTrack learns a unified identity embedding through self-supervised learning for multiple object-tracking tasks, including human pose tracking [49]. However, it prioritizes pose propagation over estimation and relies exclusively on the PoseTrack18 [2] skeleton, without incorporating multiple pose datasets.

Another class of methods [48, 51, 54, 62] employs network modules, such as graph convolutional networks (GCN), for cross-dataset adaptation. However, these dataset-specific modules limit domain transferability, as their input and output sizes are dictated by the label structure, making them incompatible with datasets that have different keypoint formats. In contrast, our approach achieves flexible transferability without relying on dataset-specific domain adaptation modules.

Human pose estimation with MDT. In 3D human shape estimation, several methods have used pseudo-ground truths (GTs) generated by the SMPL model for MDT [5, 17, 41, 42]. However, these pseudo-GTs are often inaccurate due to depth ambiguity in 2D images and the limited expressiveness of the parametric 3D model [29, 30]. Similar to [65], Sárándi et al.'s affine-combining autoencoder (ACAE) merges 3D joint coordinates via an autoencoder [36]. However, the autoencoder's learnable parameters are dataset-dependent, limiting its generalizability. Furthermore, ACAE does not preserve pixel-level features, making it unsuitable when spatial information is crucial.

Recent works on category-agnostic pose estimation integrate various objects, such as chairs and vehicles, by combining multiple datasets [13, 52]. These methods follow a relabeling approach, creating a superset of keypoint classes and detecting only the keypoints defined in their relabeled dataset. As a result, they do not address keypoint heterogeneity or label sparsity issues, as training and evaluation occur within a single labeled dataset. In contrast, our method uses multiple skeletons to improve generalization across diverse downstream domains.

In 2D human pose estimation, UniHCP performs multiple human perception tasks using task-specific queries without incorporating task-specific layers inside the backbone [6]. This design limits its ability to generalize to distant domains such as animals and whole-body pose estimation. ViTPose++ [53] and Sapiens [16] employ multi-head architectures for MDT but do not address skeleton heterogeneity or the resulting label sparsity.

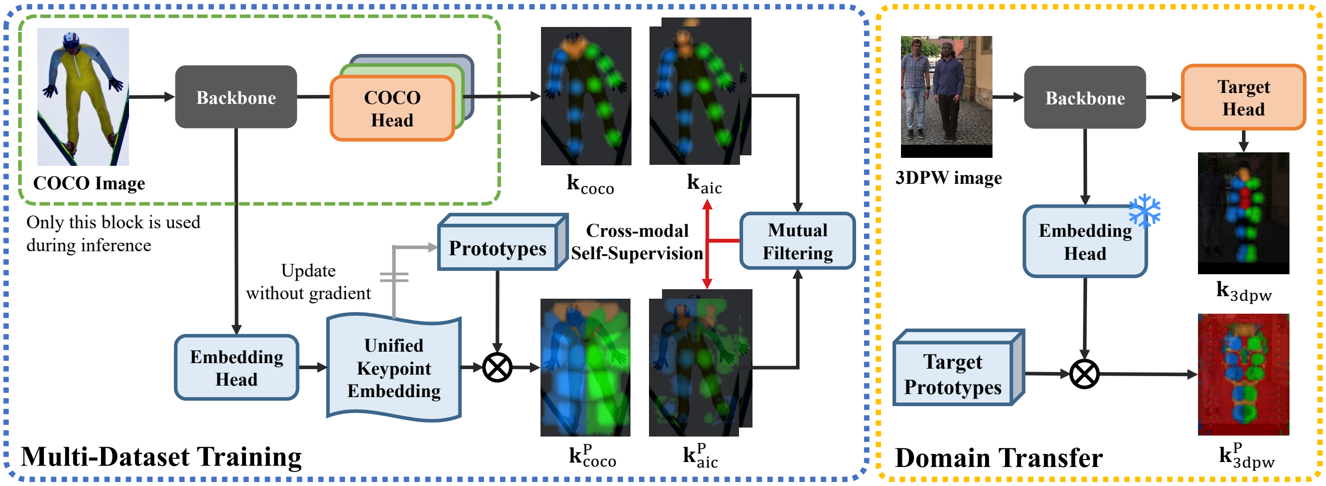

Figure 2. Overview of the PoseBH architecture. During training, the embedding head maps the backbone features into a unified keypoint embedding space. By matching these embeddings with prototypes, dataset-specific keypoint heatmaps $\mathbf{k}^{\text{P}}_{\text{coco}}$ are generated. Prototypes are updated nonparametrically from the embeddings. During inference, the embedding head and the subsequent procedures are removed.

Prototype learning can be interpreted as proxy-based metric learning [35], where the centers of a fixed set of clusters serve as prototypes. This approach has been studied in various domains, including semantic segmentation [8, 10, 64], person re-identification [32], meta-learning [24, 38, 57, 61], and visual interpretability [3, 9, 46].

ProtoSeg [64] introduces a nonparametric approach to semantic segmentation by assigning multiple prototypes per class. ProMotion [27] employs prototypes to unify scene depth estimation and optical flow tasks. While our method draws inspiration from ProtoSeg and aligns with ProMotion in exploiting the benefits of prototypes for MDT, it differs in three key aspects. First, our approach applies keypoint-level supervision, which is inherently sparser than segmentation, depth, or optical flow. Second, we specifically address skeleton heterogeneity and cross-dataset relationships, aspects not considered by ProtoSeg and ProMotion. Third, while ProtoSeg and ProMotion apply fully supervised learning, our method tackles the challenge of unlabeled data, a problem unique to MDT pose estimation.

3. Method

We focus on multi-dataset training (MDT) for 2D human pose estimation. Given D datasets, our model predicts a keypoint heatmap for a person, represented as an array $\mathbf{k}_d \in \mathbb{R}^{J_d \times H \times W}$, where d is the dataset index, J_d is the number of keypoints in the d-th dataset, and H and W denote the height and width of the heatmap, respectively.

3.1. Multi-head baseline

The multi-head baseline model consists of a backbone network that extracts an output backbone feature map from an input RGB image. The backbone network is shared across all datasets, while individual keypoint heads independently regress dataset-wise keypoint heatmaps $\{\mathbf{k}_d\}_{d=1}^{D}$, each trained with its corresponding labels. The keypoint heads share the same architecture, except for the final layer, where the output channel size is the respective number of keypoints. This generic baseline does not share the keypoint output space, limiting the MDT effect up to the backbone.

3.2. Keypoint prototype learning

Overview. Our goal is to learn a unified keypoint space that standardizes keypoint output formats across datasets, scales with dataset size, and retains dataset-specific localization accuracy. To achieve this, we reformulate keypoint regression as a prototype-based distance metric learning problem (see Fig. 2). Our shared keypoint embedding module generates a normalized keypoint embedding map $\mathbf{e} \in \mathbb{R}^{F \times H \times W}$ from an input feature map. The module consists of two deconvolution layers, followed by a residual block and two convolution layers. Additional details are provided in the supplementary document.

Using embeddings and prototypes, we perform keypoint classification by matching embeddings with prototypes $\mathbf{P} \in \mathbb{R}^{J \times M \times F}$, where F is the embedding dimension, J is the total number of keypoints across all datasets, M is the number of in-class prototypes. For each heatmap pixel, we compute the cosine similarity between the prototypes and the embedding vector to obtain keypoint predictions:

$$\mathbf{k}^{\text{P}}_{j,y,x} = \max_{m} \frac{\mathbf{P}_{(j,m,:)} \cdot \mathbf{e}_{(:,y,x)}}{\left\|\mathbf{P}_{(j,m,:)}\right\| \left\|\mathbf{e}_{(:,y,x)}\right\|}, \quad (1)$$

where x, y denote pixel indices in the heatmap, and $\mathbf{P}_{(j,m,:)}$ is a row vector of size F, extracted $\mathbf{P} \in \mathbb{R}^{J \times M \times F}$ along its first two dimensions. For each keypoint class j, $\mathbf{k}^{\text{P}}_{(j,:,:)} \in \mathbb{R}^{H\times W}$ stores the matching score of the most probable prototype among M prototypes. We refer to this as the *prototype keypoint heatmap* $\mathbf{k}^{\text{P}}$ to distinguish it from the multi-head outputs $\mathbf{k}$.

For the remainder of this paper, we use the following notation: $\mathbf{A}_{(:,:,a)}$ denotes the 2D slice of a 3D array $\mathbf{A}$ indexed at a along the third dimension, while $\mathbf{A}_{(:,b,:)}$ represents the 2D slice at index b along the second dimension. Similarly, $\mathbf{A}_{(d:e,:,:)}$ represents the 3D sub-array spanning indices d to e along the first dimension. For a 2D array $\mathbf{B}$, $\mathbf{B}_{(:,c)}$ indicates the c-th column vector, whereas $\mathbf{B}_{(d,c)}$ refers to the element at position (d, c).

Prototype learning. We adopt a nonparametric learning approach [64] to train $\mathbf{P}$. First, we obtain logits $\{\mathbf{l}_j\}_{j=1}^{J} \subset \mathbb{R}^{M\times N}$ and targets $\{\mathbf{t}_j \in \mathbb{R}^{N_j}\}_{j=1}^{J}$, where N_j is the number of foreground samples (i.e., pixels in the ground-truth (GT) heatmap with nonzero values) for the j-th joint, and N is the total number of foreground samples across all joints. We define logits as sampled embedding vectors and targets as keypoint class labels, with each keypoint class selecting an in-class prototype from M candidates.

Since no GT selection is explicitly provided, we assign an in-class prototype for each logit via online clustering using Sinkhorn-Knopp iteration with an equipartition constraint [37], and compute $\mathbf{t}_j$ as follows:

$$\mathbf{l}_j = \mathbf{P}_{(j,:,:)}\bar{\mathbf{e}}, \tag{2}$$

$$\mathbf{t}_j = \arg\max_m \operatorname{diag}(\mathbf{u}) \exp\left(\frac{\mathbf{l}_j}{\kappa}\right) \operatorname{diag}(\mathbf{v}), \tag{3}$$

where $\bar{\mathbf{e}}$ is a 2D array of size $F \times N_j$, obtained by first flattening $\mathbf{e} \in \mathbb{R}^{F\times H\times W}$ along the spatial dimensions into a 2D array of size $F \times (HW)$ and then, selecting N_j foreground objects. $\operatorname{diag}(\mathbf{u})$ is a diagonal matrix derived from the vector $\mathbf{u}$, and $\exp(\cdot)$ is applied element-wise.

Each prototype is then updated with momentum λ:

$$\mathbf{P}^{\text{new}}_{j,m} = \lambda \mathbf{P}^{\text{old}}_{j,m} + (1-\lambda)\frac{1}{N_j}\sum_{n=1}^{N_j} w_{mn}\bar{\mathbf{e}}_{(:,n)}, \tag{4}$$

where w_{mn} is the clustering assignment weight, calculated similarly to Eq. (3), but using $\max$ instead of $\arg\max$.

To optimize the learned embedding with respect to the prototype $\mathbf{P}$, we employ pixel-prototype contrastive learning, using the pixel-prototype contrastive loss (PPC) and pixel-prototype distance (PPD) from [64]:

$$\mathcal{L}_{\text{PPC}}(\mathbf{l}_j, \mathbf{t}_j) = \frac{c_n}{N_j}\sum_{n=1}^{N_j} \text{CE}\left(\mathbf{l}_{j,n}, \mathbf{t}_{j,n}\right), \tag{5}$$

$$\mathcal{L}_{\text{PPD}}(\mathbf{l}_j, \mathbf{t}_j) = \frac{c_n}{N_j}\sum_{n=1}^{N_j} (1 - \mathbf{l}^{m+}_{j,n})^2, \tag{6}$$

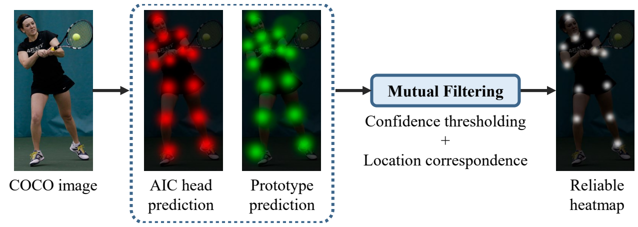

Figure 3. An illustrative example of cross-type self-supervision. Given a COCO image, we jointly refine the AIC head and AIC prototype predictions to produce reliable AIC heatmaps.

where c_n is the confidence value from the GT heatmap, CE denotes the cross-entropy loss, and m^+ indexes the first dimension of $\mathbf{l}_d$. For each dataset d, we collect logits $\mathbf{l}_{j,n} \in \mathbb{R}^{M\times N_d}$ and targets $\mathbf{t}_d \in \mathbb{R}^{N_d}$, where N_d is the number of samples in dataset d. The total prototype loss is then defined as $\mathcal{L}_{\text{Proto}} = \mathcal{L}_{\text{PPC}} + \mathcal{L}_{\text{PPD}}$.

Unlike standard contrastive learning, $\mathcal{L}_{\text{Proto}}$ operates within a single dataset and lacks the push term for differing dataset skeletons. To introduce cross-dataset negatives, we apply K-means clustering to all prototypes, forming K (=96) clusters. This clustering is performed only once during training, while online clustering for in-class prototypes is conducted at each iteration. These clusters serve as negative samples, where the embeddings closest to each centroid are used as logits, with their dataset-specific labels as targets. Finally, we obtain $\mathbf{l}_c$ and $\mathbf{t}_c$ for each cluster and apply Eqs. (5) and (6) to enable cross-dataset contrastive learning.

3.3. Cross-type self-supervision

While our keypoint prototypes enable unified MDT, label sparsity remains a challenge, as each image is annotated for only one dataset, limiting supervision. To address this semi-supervised learning problem, we introduce a self-supervision strategy that bridges keypoint heatmaps and embeddings. By collaboratively filtering out noisy predictions from both the keypoint and embedding heads, we generate reliable keypoint heatmaps for training (see Fig. 3).

To eliminate uncertain keypoint predictions, we apply two filtering conditions. First, for each keypoint class, the confidence scores of both the keypoint and embedding heads must exceed a threshold ($c_{thr} = 0.25$). Second, the root mean square distance between their predictions must be below a threshold ($d_{thr} = 2.1$). The filtered predictions are then combined using a weighted average:

$$\hat{\mathbf{y}}_i = s_i\hat{\mathbf{y}}^{\text{kpt}}_i + (1-s_i)\hat{\mathbf{y}}^{\text{emb}}_i, \tag{7}$$

where $s_i = c^{\text{kpt}}_i/(c^{\text{kpt}}_i + c^{\text{emb}}_i)$, i is the keypoint index, and $\hat{\mathbf{y}}$ and c represent the keypoint prediction and confidence score from each head, respectively.

From $\hat{\mathbf{y}}$, we generate a *reliable heatmap* $\mathbf{k}^{\text{CSS}}$ following the standard GT heatmap generation process. The loss for

Method	GFLOPS	COCO	AIC	MPII	*Avg.*
UniHCP	18.9	76.8	**32.6**	90.9	66.8
ViTPose++	18.5	77.0	31.6	93.1	67.2
Ours	18.5	**77.3**	32.1	**93.2**	**67.5**

Table 1. Comparison of different multi-dataset training methods on general human pose benchmarks, with computational complexity measured in GFLOPS during the evaluation phase.

Method	AP-10K	APT-36K	COCO-W	*Avg.*
UniHCP	56.5	62.0	20.1	46.2
ViTPose++	74.1	76.0	57.1	69.1
Ours	**75.0**	**87.2**	**57.9**	**73.4**

Table 2. Comparison of multi-dataset training methods on whole-body and animal domains. COCO-W refers to COCO-WholeBody. The reported metrics are AP(%).

unlabeled samples is then computed as:

$$\mathcal{L}_{\text{CSS}} = \sum_{d=1}^{D} \zeta \big[\mathcal{L}_{\text{hm}}(\mathbf{k}_{d[\mathbf{u}]}, \mathbf{k}_{d[\mathbf{u}]}^{\text{CSS}}) + \mathcal{L}_{\text{Proto}}(\mathbf{e}_{[\mathbf{u}]}, \mathbf{k}_{d[\mathbf{u}]}, \mathbf{k}_{d[\mathbf{u}]}^{\text{CSS}}) \big], \tag{8}$$

where $\mathbf{u}$ denotes the indices of the unlabeled samples. This cross-type self-supervision (CSS) loss enables self-distillation between the dataset-wise keypoint head and the prototype, mitigating supervision shortage without requiring a teacher model or input duplication.

The final training loss is defined as:

$$\mathcal{L}_{\text{MDT}} = \mathcal{L}_{\text{KPL}} + \mathcal{L}_{\text{CSS}}, \text{ where} \tag{9}$$

$$\mathcal{L}_{\text{KPL}} = \sum_{d=1}^{D} \left[\mathcal{L}_{\text{hm}}(\mathbf{k}_d, \mathbf{k}_d^{\text{gt}}) + \mathcal{L}_{\text{Proto}}(\mathbf{e}, \mathbf{k}_d, \mathbf{k}_d^{\text{gt}}) \right], \tag{10}$$

and $\mathcal{L}_{\text{hm}}$ is the standard JointMSE loss [40].

4. Experiments

4.1. Experimental setup

We follow the experimental setup of ViTPose++ [53], using six human pose datasets for training. (1) COCO consists of 57,000 training and 5,000 validation images, annotated with 17 keypoints [23]. (2) AIChallenger (AIC) contains 210,000 training and 30,000 validation images, annotated with 14 keypoints [50]. (3) MPII includes approximately 15,000 training, 2,729 validation, and 5,700 test images, with 16 keypoints [1]. (4) AP-10K is an animal pose dataset with approximately 7,000 training, 1,000 validation, and 2,000 test images [58]. (5) APT-36K, a video-based animal pose dataset, consists of about 25,000 training, 3,600 validation, and 7,000 test images [56]. Both AP-10K and APT-36K share the same skeletal format with 17 keypoints, largely following COCO. (6) COCO-WholeBody extends COCO with whole-body annotations covering 133 keypoints [50]. We follow standard evaluation protocols, using the average precision (AP) metric based on object keypoint similarity (OKS) for COCO, AIC, COCO-WholeBody, AP-10K, and APT-36K. For MPII, the percentage of correct keypoints (PCK) metric is used [1].

For the domain transfer experiments, we use InterHand2.6M [28] and 3DPW [45]. InterHand2.6M is a two-hand pose estimation dataset containing approximately 2.6M frames, annotated with 21 keypoints. We adopt the ViTPose++ configuration for frame sampling and dataset splits (training, validation, and test). For evaluation, we use PCK, AUC, and EPE, consistent with ViTPose++. 3DPW is an in-the-wild 3D human shape estimation dataset with 18,000 training and 26,000 test frames, annotated with 24 keypoints. We use COCO-style AP and AR as evaluation metrics. In both InterHand2.6M and 3DPW, the projected 2D keypoints are used as ground-truths.

Further details on the experimental setups, the impact of varying hyperparameters, and failure case analysis can be found in the supplemental document.

4.2. Implementation details

Our multi-head baseline model is ViTPose++ [53]. We use an input image size of 256×192 and apply a flip test in all cases. The model is trained for 100 epochs using the AdamW optimizer with a weight decay factor of 0.1 and a step learning rate scheduler with an initial learning rate of 0.001. The learning rate is reduced by a factor of 0.1 at the 50th and 90th epochs. We use four NVIDIA A100 GPUs for training with the ViT-H backbone and four NVIDIA A6000 GPUs for training with the ViT-B backbone.

We employ a scheduling strategy for both the losses and the learning rate. Initially, we freeze the multi-head model, updating only the embedding module and prototypes for 50 epochs. In the subsequent 40 epochs, we train the multi-heads and the embedding module while keeping the prototypes and the backbone frozen. Finally, during the last 10 epochs, we train the entire network except the prototypes. At this stage, we introduce cross-type self-supervision loss. This progressive optimization approach accelerates training compared to training the entire network from the outset.

4.3. Comparison with existing MDT methods

In Tab. 1, we evaluate the competitiveness of our method against existing MDT approaches. For a fair comparison, all methods use the ViT-Base backbone and are trained under our dataset configuration. While UniHCP performs well on AIC, it achieves the lowest scores on COCO and MPII, resulting in an average of 66.8. We hypothe-

Method	Backbone	Datasets	COCO (AP↑)	COCO (AP^{50} ↑)	COCO (AP^{75} ↑)	COCO (AR↑)	MPII (PCKh↑)	MPII (PCKh@0.1↑)
HRNet	HRNet-W48	1	75.1	90.6	82.2	80.4	90.3†	33.1†
HRFormer	HRFormer-B	1	75.6	90.8	82.8	80.8	-	-
SimCC	HRNet-W48	1	76.1	90.6	82.9	81.2	90.0	36.8
PCT†	Swin-B	1	77.7	91.2	84.7	82.1	92.5	-
PCT†	Swin-H	1	79.3	91.5	85.9	-	-	-
UniHCP	ViT-B	33	76.1	-	-	-	93.2††	-
ViTPose++-B	ViT-B	6	77.0	73.4	84.0	82.6	92.8	39.1
ViTPose++-H	ViT-H	6	79.4	**91.9**	85.7	**84.8**	**94.2**	41.6
Ours	ViT-B	6	77.3	90.8	84.2	82.4	93.2	39.3
Ours	ViT-H	6	**79.5**	**91.9**	**85.8**	84.5	**94.2**	**42.0**

Table 3. Comparison with state-of-the-art methods on COCO and MPII. Methods with † uses 256×256 input size. †† indicates additional fine-tuning on each dataset.

size that UniHCP's parameter-sharing strategy causes well-represented pose datasets to dominate, hindering its generalization across different skeleton structures. ViTPose++ serves as both a multi-head model and the baseline for our method. Our algorithm surpasses ViTPose++ by 0.3 in average score and outperforms it on individual datasets. Given that these benchmark scores are highly saturated and our method maintains the same computation complexity as ViTPose++ during inference, this improvement is significant.

To further assess generalization, we test our method on animal and whole-body pose estimation datasets. Table 2 shows that our method consistently improves performance, despite significant domain discrepancies. UniHCP lags behind with an average score of 27.2, likely due to its parameter sharing strategy, which does not address skeleton heterogeneity. Our method also surpasses ViTPose++ by an average score of 4.3, demonstrating superior generalization beyond human skeletons. Specifically, it achieves gains of 0.9 AP on AP-10K, 11.2 AP on APT-36K, and 0.8 AP on COCO-WholeBody. Unlike existing methods that compromise either mainstream performance or adaptation to distant domains, our approach enhances both synergistically.

4.4. Comparison to state-of-the-art methods

In Tab. 3, we compare our method with state-of-the-art (SOTA) approaches on the COCO and MPII benchmarks, using scores reported in the respective papers. We also evaluate our method against SOTA approaches designed for single datasets, including HRNet [40], HRFormer [60], SimCC [21], and PCT [11]. On the COCO validation set, our method ranks second with the ViT-B backbone and achieves SOTA performance with the ViT-H backbone, reaching 79.5 AP. On the MPII validation set, our ViT-B model performs comparably to the fine-tuned UniHCP and surpasses the SOTA on the PCKh@0.1 metric by 0.4 PCKh.

Method	InterHand2.6M PCK↑	InterHand2.6M AUC↑	InterHand2.6M EPE↓	3DPW AP↑	3DPW AR↑
UniHCP	98.5	87.0	3.72	57.9	61.8
ViTPose++	98.3	86.2	4.02	81.7	85.2
Ours	**98.6**	**87.1**	**3.70**	**83.6**	**87.1**

Table 4. Transfer results on hand (Interhand2.6M) and human shape (3DPW) domains.

These results demonstrate that our method effectively balances adaptability to diverse domains while maintaining competitive performance on standard benchmarks, aligning with our primary objective of improving generalization beyond mainstream tasks.

4.5. Transfer to hands and human shapes

PoseBH has the capability to generalize its learned representations beyond the training dataset. We validate this through domain transfer experiments by fine-tuning the model separately on the InterHand2.6M and 3DPW datasets. InterHand2.6M is a two-hand pose estimation dataset containing 21 MANO keypoints [34], while 3DPW is an in-the-wild human shape estimation dataset with 24 SMPL keypoints [26].

During fine-tuning, we freeze the embedding module and nonparametrically learn the prototypes for the target datasets. After convergence, we fine-tune the backbone and head weights alongside the prototype loss. Freezing the embedding module ensures that the previously learned prototypes and those of the target dataset remain aligned within the same embedding space, preserving the learned representations during transfer. Since these experiments do not involve skeleton heterogeneity across datasets, no semi-

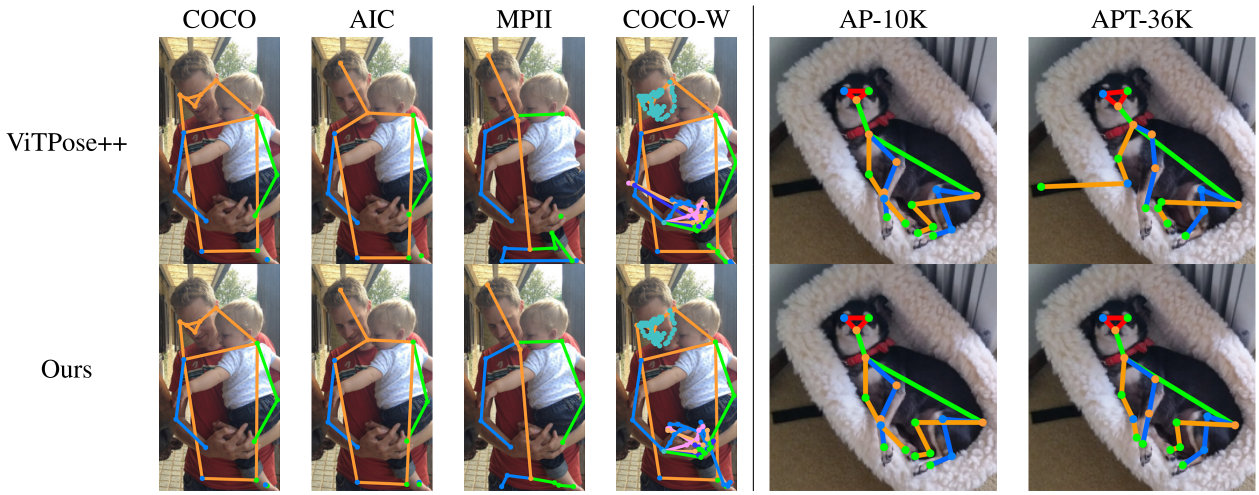

Figure 4. Comparative pose estimation results on human (left) and animal (dog; right).

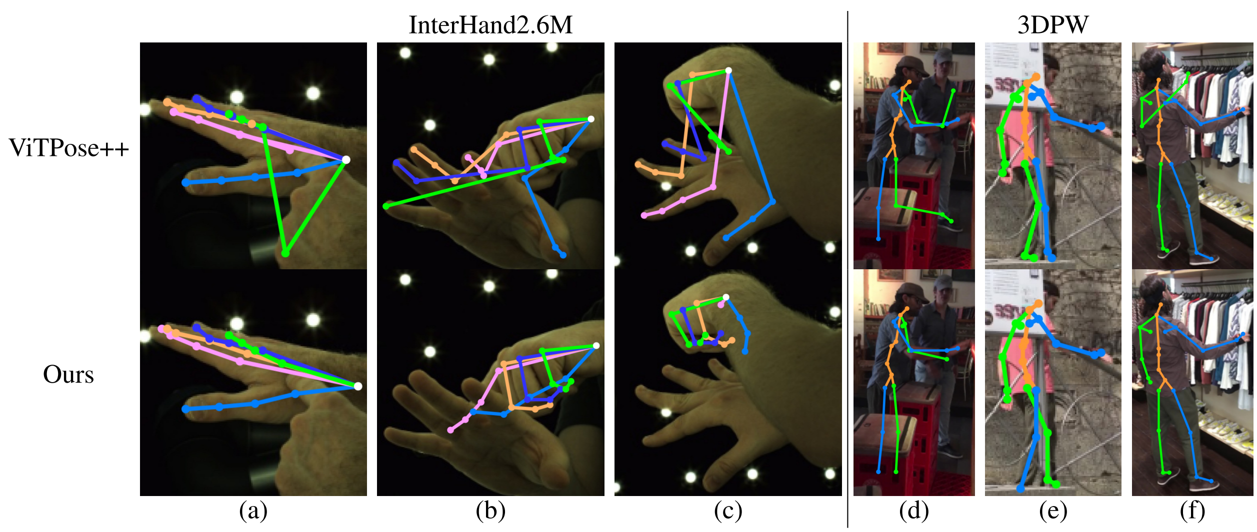

Figure 5. Comparative results on InterHand2.6M (a–c) and 3DPW (e–f).

supervised learning challenges arise, and CSS loss is not applied.

In Tab. 4, we report results on the InterHand2.6M and 3DPW test sets using the ViT-B backbone. Given its large dataset size (∼2.6M instances), InterHand2.6M exhibits a smaller performance gap among different methods. Nonetheless, our method surpasses ViTPose++ by 0.3 PCK, 0.9 AUC, and 0.32 EPE, while also outperforming UniHCP by 0.1 PCK, 0.1 AUC, and 0.02 EPE. For the smaller 3DPW dataset, our method achieves a notable performance gain, surpassing the baseline by 1.9 AP and 1.9 AR, demonstrating strong generalization in keypoint embeddings.

4.6. Ablation studies

In Tab. 5, we present the results of ablation experiments on the COCO (CO), AIC (AI), MPII (MP), AP-10K (AP), APT-36K (APT), and COCO-Wholebody (CW) validation sets using the ViT-Base backbone. The top row represents a multi-head baseline with a shared backbone and dataset-specific heads. This conventional multi-head approach exhibits relatively low performance, particularly on the AP-10K and APT-36K validation sets, due to substantial domain discrepancies. Our keypoint prototype method achieves an average score of 66.3, improving the baseline by 0.8. Additionally, incorporating cross-type self-supervision provides an additional average performance gain of 0.2. Collectively, these enhancements enable our final configuration to outperform the multi-head baseline by an average of 2.4 points across all six datasets.

4.7. Visual analysis

Figure 4 presents examples of human and animal pose estimation, demonstrating that our algorithm achieves more accurate keypoint localization under complex interactions and

Figure 6. Examples of SMPL fitting using KITRO and with our 2D predictions.

Method	CO	AI	MP	AP	APT	CW	*Avg.*
Baseline	77.0	31.6	93.1	75.3	74.8	57.1	68.2
$+\mathcal{L}_{\text{Proto}}$	77.0	31.8	93.0	76.4	86.4	57.5	70.4
$+\mathcal{L}_{\text{CSS}}$	**77.3**	**32.1**	**93.2**	**76.7**	**86.5**	**57.9**	**70.6**

Table 5. Ablation study of the proposed method. $\mathcal{L}_{\text{Proto}}$: keypoint prototype learning. $\mathcal{L}_{\text{CSS}}$: cross-type self-supervision.

occlusions, as seen in the right-hand keypoints of the standing person (left) and the right front toe of the dog (right).

We further compare the pose estimation results on InterHand2.6M and 3DPW in Fig. 5. In (a), ViTPose++ incorrectly assigns a little finger joint to the opposite hand, whereas our method provides more accurate localization. In (b), ViTPose++ struggles with occlusion from the other hand, producing noisy predictions, while our method correctly separates them. In (c), due to severe self-occlusion, ViTPose++ misestimates occluded keypoints and assigns them to the opposite hand, while our method generalizes well in challenging self-occlusion scenarios. In (d), ViTPose++ produces an implausible foot and hand pose under heavy occlusion, while our method remains robust. In (e), ViTPose++ swaps the left and right legs, while our method correctly distinguishes them. In (f), background clutter causes ViTPose++ to produce noisy predictions for the left hand, while our method correctly identifies the location of the occluded left hand.

To further assess the effectiveness of our 2D SMPL keypoint predictions on the 3DPW test set, we employ a SMPL optimization technique. KITRO [55] optimizes SMPL parameters using 2D keypoint reprojection loss. For a comparative visual analysis, we replace KITRO's 2D keypoint predictions with ours. As shown in Figure 6(a) and (b), KITRO fails to align the mesh with the actual feet, whereas our method achieves more precise alignment. In (c), KITRO mislocalizes the right hand under partial occlusion, while our 2D predictions remain robust. In (d), KITRO exhibits misalignment in both the hands and feet, while replacing its 2D predictions with ours significantly improves the alignment of peripheral body parts.

In the supplemental material, we provide a visualization of the prototypes generated by our algorithm, illustrating how they effectively capture the diversity of representations within the embedding space.

5. Conclusion

This paper presents a new multi-dataset training (MDT) framework for pose estimation across diverse domains and datasets. Traditional approaches encounter challenges with label heterogeneity, dataset-specific parameters, and limited domain generalization. Our approach addresses the limitations with two main contributions. First, to resolve label heterogeneity and dataset-specific parameters, we employ nonparametric prototypes to unify diverse keypoint annotations across multiple datasets. Second, to address the limited multi-dataset supervision, we propose a cross-type self-supervision mechanism. The proposed method has demonstrated substantial improvements over existing MDT across a broad range of datasets, including human pose, animal pose, hand shape, and human shape estimation.

Limitations and future work. A limitation of our approach is that it requires at least a few-shot sample to learn prototypes, making zero-shot inference for unseen skeletons infeasible. Future work should explore extending our method for zero-shot human pose estimation. Future work should also investigate adapting the approach to 3D domains e.g. using axis-angle rotation representations, to facilitate representation learning on 3D manifolds.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant (No. 2021R1A2C2012195) and by the Institute of Information & communications Technology Planning & Evaluation (IITP) grants No. RS-2019-II191906 (AI Graduate School Program, POSTECH), No. RS-2020-II201336 (AI Graduate School Program, UNIST), and RS-2022-II220290, funded by the Korea government (MSIT).

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 5

[2] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. PoseTrack: A benchmark for human pose estimation and tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5167–5176, 2018. 2

[3] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. 3

[4] Yanbei Chen, Manchen Wang, Abhay Mittal, Zhenlin Xu, Paolo Favaro, Joseph Tighe, and Davide Modolo. ScaleDet: A scalable multi-dataset object detector. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7288–7297, 2023. 2

[5] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3D human mesh from in-the-wild crowded scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[6] Yuanzheng Ci, Yizhou Wang, Meilin Chen, Shixiang Tang, Lei Bai, Feng Zhu, Rui Zhao, Fengwei Yu, Donglian Qi, and Wanli Ouyang. UniHCP: A unified model for human-centric perceptions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17840–17852, 2023. 1, 2

[7] Mickael Cormier, Aris Clepe, Andreas Specker, and Jürgen Beyerer. Where are we with human pose estimation in real-world surveillance? In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 591–601, 2022. 1

[8] Anurag Das, Yongqin Xian, Dengxin Dai, and Bernt Schiele. Weakly-supervised domain adaptive semantic segmentation with prototypical contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15434–15443, 2023. 3

[9] Jon Donnelly, Alina Jade Barnett, and Chaofan Chen. Deformable protopnet: An interpretable image classifier using deformable prototypes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10265–10275, 2022. 3

[10] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4320–4329, 2022. 3

[11] Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, and Han Hu. Human pose as compositional tokens. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

[12] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human POSEitioning System (HPS): 3D human pose estimation and self-localization in large scenes from body-mounted sensors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4318–4329, 2021. 1

[13] Or Hirschorn and Shai Avidan. A graph-based approach for category-agnostic pose estimation, 2024. 2

[14] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. NeuMan: Neural human radiance field from a single video. In *European Conference on Computer Vision (ECCV)*, 2022. 1

[15] Roy Kapon, Guy Tevet, Daniel Cohen-Or, and Amit H Bermano. MAS: Multi-view ancestral sampling for 3D motion generation using 2D diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1965–1974, 2024. 1

[16] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569*, 2024. 2

[17] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021. 2

[18] Alberto Lamas, Siham Tabik, Antonio Cano Montes, Francisco Pérez-Hernández, Jorge García, Roberto Olmos, and Francisco Herrera. Human pose estimation for mitigating false negatives in weapon detection in video-surveillance. *Neurocomputing*, 489:488–503, 2022. 1

[19] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. MSeg: A composite dataset for multi-domain semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[20] Duy Tho Le, Chenhui Gou, Stavya Datta, Hengcan Shi, Ian Reid, Jianfei Cai, and Hamid Rezatofighi. JRDB-PanoTrack: An open-world panoptic segmentation and tracking robotic dataset in crowded human environments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22325–22334, 2024. 1

[21] Yanjie Li, Sen Yang, Peidong Liu, Shoukui Zhang, Yunxiao Wang, Zhicheng Wang, Wankou Yang, and Shu-Tao Xia. SimCC: A simple coordinate classification perspective for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 89–106. Springer, 2022. 6

[22] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-X: A large-scale 3D expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1

[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 *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 1, 5

[24] Sun-Ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, and Ting Yao. Learning orthogonal prototypes for generalized few-shot semantic segmentation. In

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 3

[25] Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. HumanGaussian: Text-driven 3D human generation with gaussian splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6646–6657, 2024. 1

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 6

[27] Yawen Lu, Dongfang Liu, Qifan Wang, Cheng Han, Yiming Cui, Zhiwen Cao, Xueling Zhang, Yingjie Victor Chen, and Heng Fan. ProMotion: Prototypes as motion learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28109–28119, 2024. 3

[28] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2020. 5

[29] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. NeuralAnnot: Neural annotator for 3D human mesh training sets. In *Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2022. 2

[30] Gyeongsik Moon, Hongsuk Choi, Sanghyuk Chun, Jiyoung Lee, and Sangdoo Yun. Three Recipes for Better 3D Pseudo-GTs of 3D Human Mesh Estimation in the Wild. In *Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2023. 2

[31] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3D gaussian avatar. In *European Conference on Computer Vision (ECCV)*, 2024. 1

[32] Haocong Rao and Chunyan Miao. TranSG: Transformer-based skeleton graph prototype contrastive learning with structure-trajectory prompted reconstruction for person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22118–22128, 2023. 3

[33] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. HuMoR: 3D human motion model for robust pose estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[34] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 6

[35] Karsten Roth, Oriol Vinyals, and Zeynep Akata. Non-isotropy regularization for proxy-based deep metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7420–7430, 2022. 3

[36] István Sárándi, Alexander Hermans, and Bastian Leibe. Learning 3D human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023. 2

[37] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrics. *Pacific Journal of Mathematics*, 21(2):343–348, 1967. 4

[38] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 3

[39] Markus D Solbach and John K Tsotsos. Vision-based fallen person detection for the elderly. In *IEEE/CVF international conference on computer vision workshops (ICCVW)*, pages 1433–1442, 2017. 1

[40] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5, 6

[41] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3D people. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[42] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3D people in depth. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[43] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xR-EgoPose: Egocentric 3D human pose from an hmd camera. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7728–7738, 2019. 1

[44] Edward Vendrow, Duy Tho Le, Jianfei Cai, and Hamid Rezatofighi. JRDB-Pose: A large-scale dataset for multi-person pose estimation and tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[45] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, 2018. 5

[46] Chong Wang, Yuyuan Liu, Yuanhong Chen, Fengbei Liu, Yu Tian, Davis McCarthy, Helen Frazer, and Gustavo Carneiro. Learning support and trivial prototypes for interpretable image classification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2062–2072, 2023. 3

[47] Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. Scene-aware egocentric 3D human pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13031–13040, 2023. 1

[48] Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. Towards universal object detection by domain attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7289–7298, 2019. 2

[49] Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip Torr, and Luca Bertinetto. Do different tracking tasks require different appearance models? *Advances in Neural Information Processing Systems (NeurIPS)*, 34:726–738, 2021. 2

[50] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. AI Challenger: A large-scale dataset for

going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017. 5

[51] Hang Xu, Linpu Fang, Xiaodan Liang, Wenxiong Kang, and Zhenguo Li. Universal-RCNN: Universal object detector via transferable graph r-cnn. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 12492–12499, 2020. 2

[52] Lumin Xu, Sheng Jin, Wang Zeng, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Pose for everything: Towards category-agnostic pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 398–416. Springer, 2022. 2

[53] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 5

[54] Di Yang, Yaohui Wang, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. UNIK: A unified framework for real-world skeleton-based action recognition. *British Machine Visison Conference (BMVC)*, 2021. 2

[55] Fengyuan Yang, Kerui Gu, and Angela Yao. KITRO: Refining human mesh by 2D clues and kinematic-tree rotation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1052–1061, 2024. 8

[56] Yuxiang Yang, Junjie Yang, Yufei Xu, Jing Zhang, Long Lan, and Dacheng Tao. APT-36K: A large-scale benchmark for animal pose estimation and tracking. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:17301–17313, 2022. 1, 5

[57] Sung Whan Yoon, Jun Seo, and Jaekyun Moon. TapNet: Neural network augmented with task-adaptive projection for few-shot learning. In *International conference on machine learning*, pages 7115–7123. PMLR, 2019. 3

[58] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. AP-10K: A benchmark for animal pose estimation in the wild. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 1, 5

[59] Xinyi Yu, Xin Zhang, Chengjun Xu, and Linlin Ou. Human–robot collaborative interaction with human perception and action recognition. *Neurocomputing*, 563:126827, 2024. 1

[60] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. HRFormer: High-resolution transformer for dense prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 6

[61] Baoquan Zhang, Xutao Li, Yunming Ye, Zhichao Huang, and Lisai Zhang. Prototype completion with primitive knowledge for few-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3754–3762, 2021. 3

[62] Bo Zhang, Jiakang Yuan, Botian Shi, Tao Chen, Yikang Li, and Yu Qiao. Uni3D: A unified baseline for multi-dataset 3D object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9253–9262, 2023. 2

[63] Xiangyun Zhao, Samuel Schulter, Gaurav Sharma, Yi-Hsuan Tsai, Manmohan Chandraker, and Ying Wu. Object detection with a unified label space from multiple datasets. In *European Conference on Computer Vision (ECCV)*, pages 178–193. Springer, 2020. 2

[64] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking Semantic Segmentation: A Prototype View. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2582–2593, 2022. 3, 4

[65] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7571–7580, 2022. 2