ChatGPT

Adapting Foundation Models for Pose Estimation in New Domains (2024–2025)

Overview: Recent works emphasize **fine-tuning large pre-trained pose models** (e.g. Meta's **Sapiens** or the ViTPose family) to **specialized human populations** (infants, elderly, etc.) when ample target-domain data is available. Unlike few-shot or unsupervised approaches, these methods assume sufficient labeled target data and focus on **architectural adaptations, fine-tuning strategies, and domain-specific generalization**. Key themes include replacing or extending output heads, freezing or adapting certain model stages, **supervised domain adaptation** techniques, handling **different body topologies** (e.g. whole-body vs. standard joints, adult vs. infant proportions), and even leveraging additional modalities like depth. Below, we review representative 2024–2025 works from CVPR, ICCV, ECCV that illustrate how large pose models can be adapted to new domains, along with their architectures, training strategies, innovations, and relevance to specialized populations.

Sapiens (Meta, 2024): A Foundation Model for Human-Centric Vision

Model & Architecture: Sapiens is a family of **Vision Transformer (ViT) models** (up to 2B parameters) pre-trained on **Humans-300M**, a curated dataset of 300M in-the-wild human images. Uniquely, Sapiens uses **high-resolution 1024×1024 input patches**, enabling **1K resolution inference** for detailed human analysis. Pre-training is self-supervised via Masked Autoencoder (MAE) on human images, yielding a strong human-specific representation. For downstream tasks, Sapiens adopts a simple **encoder–decoder architecture**: the ViT encoder is initialized from pre-training and a lightweight task-specific decoder head (per task) is added and trained from scratch. **All layers are fine-tuned end-to-end** on the target task data. This design makes it easy to adapt one large pre-trained model to multiple human-centric tasks.

Dense Keypoint Representation: To push pose estimation fidelity, Sapiens introduced a **much richer keypoint definition** with **308 whole-body keypoints** (including 243 facial and 40 hand points) – far more detailed than standard 17 or 68 point sets. A multi-view capture pipeline was used to obtain **high-quality, fine-grained annotations** for these keypoints (Humans-5K dataset). This emphasis on label quality and density helps Sapiens better digitize human pose and shape.

Fine-Tuning Strategy: Sapiens highlights the benefits of **domain-specific pre-training + supervised fine-tuning**. Pre-training on a massive human image corpus (with diverse poses, body shapes, etc.) yields a strong foundation. Then, even with limited but high-quality labels (on the order of a few thousand images), **fine-tuning the entire model** (not just the head) on the target pose task achieves state-of-art accuracy. Notably, Sapiens' entire model (encoder + new head) is trained on the target domain, as opposed to freezing the backbone – this leverages the full capacity of the large model to specialize to new data.

Performance & Generalization: On the Humans-5K test set (whole-body pose), Sapiens dramatically outperforms prior methods. For example, a 0.6B-param Sapiens beat the previous SOTA DWPose-L by **+2.1 AP** (56.2 vs 53.1) on whole-body AP, despite DWPose's complex distillation approach. Scaling up to Sapiens-2B yields **+7 AP** gain over SOTA (61.1 vs 53.1 AP). Even at similar model sizes, Sapiens outperforms ViTPose+ (e.g. Sapiens-0.3B > ViTPose+-L by +0.3 AP), indicating the benefit of human-

focused pre-training and high-res design. Importantly, although Sapiens was fine-tuned on a relatively controlled studio dataset, it shows **strong in-the-wild generalization** – thanks to the diverse pre-training data and high-fidelity annotations, the model handles real-world infant, upper-body, and multi-person images without needing domain-specific tweaks. Sapiens also achieved SOTA on related tasks (body part segmentation, depth, normals) by fine-tuning the same backbone with task-specific heads, underscoring the **broad adaptability** of a well-pretrained human-centric model.

Code & Data: Sapiens is a Meta AI project; while code isn't publicly listed in the paper, a model (e.g. Sapiens-1B) is available on HuggingFace. The Humans-300M pre-training dataset is proprietary, but the **Humans-5K** dense keypoint and **Humans-2K** segmentation benchmarks introduced are likely released for evaluation. This work demonstrates how a **foundation model for human pose** can be effectively fine-tuned to a new domain or task with minimal architectural changes – just swapping in a new head and data.

Applicability: For adapting to specialized populations, Sapiens suggests that **simply fine-tuning a sufficiently large, human-pretrained ViT can yield top performance**, provided one has **enough high-quality target data**. For example, one could take Sapiens and fine-tune on an elderly pose dataset or an infant pose dataset (with a suitably adjusted keypoint schema) to obtain a state-of-the-art model for that group. The key is that Sapiens' capacity and pre-training give it a strong prior, so it can **learn subtle domain specifics** (body proportions, common poses, etc.) from a moderate amount of labeled target images, without overfitting. This "pretrain on broad humans, then specialize" paradigm is a recurring theme in recent pose adaptation research.

ViTPose & ViTPose++ (Xu et al., 2022–2023): Plain Transformers Scaled and Adapted to Multiple Pose Domains

Model & Architecture: ViTPose introduced a surprisingly simple yet powerful baseline: a **plain ViT encoder** (non-hierarchical, standard ViT architecture) coupled with a lightweight decoder for pose estimation. Unlike earlier transformer-based pose models that added transformers on top of CNN backbones or crafted task-specific attention modules, ViTPose showed that a vanilla ViT (pre-trained on images) can serve as an effective feature extractor for both top-down and bottom-up 2D pose estimation. The decoder is minimal (two deconvolution layers and a conv to output K heatmaps for K keypoints), relying on the encoder's representational power. ViTPose models were scaled from ~20M to ~1B parameters, demonstrating a **new Pareto frontier** in pose estimation: larger ViT models steadily improved accuracy without special tricks, thanks to the ViT's scalability and high parallelism.

Training & Fine-Tuning: ViTPose leveraged **flexible pre-training and fine-tuning regimes**. Models could be initialized from ImageNet pre-trained weights or self-supervised MAE pre-training, and fine-tuned on pose data. The authors even explored pre-training on **unlabeled pose images** (instead of ImageNet) with MAE, finding it can suffice [1] [2]. This aligns with Sapiens' finding that **domain-specific pretraining boosts downstream pose performance**. For fine-tuning, ViTPose typically trained the whole model on the target pose dataset (e.g. COCO keypoints) in a supervised manner. The approach was **very flexible**: e.g. ViTPose can be fine-tuned in a top-down mode (on person crops) or bottom-up (on full images) by appropriate data preparation, without architecture changes. This showcases a benefit of transformer encoders – they are task-agnostic and can transfer between paradigms and even between human and animal pose tasks.

Performance: ViTPose achieved **state-of-the-art results on mainstream human pose benchmarks** (COCO, MPII, etc.) and its largest model (ViTPose-G, ~1B params) set a new record on COCO test-dev without test-time augmentation. Perhaps more impressively, the same ViTPose architecture proved

effective for **whole-body pose** (COCO-WholeBody) and even for **animal pose** (AP-10K, APT-36K datasets) – domains with different keypoint definitions. However, the original ViTPose was trained per task (e.g. separate models for COCO vs. WholeBody). This raised the question: Can one model handle heterogeneous keypoint schemas **(e.g. 17-point human, 133-point whole-body, 20-point animal)** without sacrificing accuracy?**

ViTPose++ – Knowledge Factorization: To tackle multi-domain adaptation, the follow-up **ViTPose++** introduced a novel architectural adaptation: **knowledge factorization via task-specific feed-forward networks (FFNs)**. In practice, ViTPose++ inserts small task-specific layers or parameters into the ViT such that certain parts of the network are shared across all pose types, while others are specialized for each keypoint set. Specifically, they use **task-agnostic self-attention** but **task-specific FFN parameters** in each transformer layer. This allows a single model to **learn a shared representation (through attention) while tuning the per-task feature transformations** for each keypoint configuration. At inference, the appropriate FFN "expert" is used depending on the pose task. ViTPose++ also adds a **"knowledge token" for distillation**, enabling knowledge transfer from a large model to a smaller one in the same family – effectively a learned token that guides a small student to mimic a big teacher without complicated distillation pipelines.

Multi-Task Results: With these adaptations, ViTPose++ achieved **simultaneous state-of-the-art** on a suite of pose estimation tasks. A single ViTPose++ model (with multiple heads/FFNs) reached SOTA on COCO human pose, COCO-WholeBody (133 keypoints), AI Challenger and MPII (different 2D human schemas), and even on AP-10K and APT-36K animal pose **within one unified model**. Remarkably, this multi-task capability came **with no loss in inference speed** compared to task-specific models. The success underscores that **a large transformer can generalize across body topologies** when given minor architectural accommodations for each domain. The code and pre-trained models for ViTPose/ViTPose++ are openly available, making it a foundation that others can fine-tune.

Applicability: For adapting foundation models to specialized domains, ViTPose++ shows the value of **modular architectures**. If one needs to adapt a model like ViTPose to a new domain (say, a new set of keypoints for a specific medical application), one strategy is to **add a task-specific module (e.g. an adapter or FFN branch)** rather than retraining everything from scratch. This preserves knowledge for generic poses while allowing specialization. The knowledge token idea also hints at efficient fine-tuning: instead of heavy full-model training, a small learned token or adapter could imbue new domain knowledge or distill a large model's knowledge – a direction related to parameter-efficient fine-tuning (like LoRA or adapters) [3]. Indeed, very recent work (CVPR 2025) on **LoRA-based adaptation** proposes reusing low-rank adaptation modules for "tuning-free" few-shot transfer of vision models, which could be applied to pose models to avoid full fine-tuning. In summary, ViTPose/ViTPose++ demonstrate both the **baseline strength** of large ViT encoders for pose and the benefit of **inserting task-specific parameters** to handle domain shifts in a supervised multi-task setting.

Insight – Generic Models vs Specialized Data: An independent study by Jahn et al. (Scientific Reports, 2025) evaluated generic vs. specialized infant pose estimators. They found that **ViTPose (trained on adult data) out-of-the-box outperformed dedicated infant pose models** on a new infant movement dataset. This suggests large generic models have strong baseline generalization, likely due to greater capacity and diverse pre-training. However, when they **fine-tuned ViTPose on the infant data, accuracy improved significantly** (PCK increased) – confirming that supervised adaptation with sufficient data yields big gains. Interestingly, the specialized infant models from prior work did not generalize well to this new infant dataset, performing worse than the generic model. This highlights a caution: **models overfit to a narrow "specialized" domain can struggle on even slightly different target distributions**. A foundation model with broad training, on the other hand, may transfer better and serves as a stronger starting point for fine-tuning. The lesson is to leverage the generalization of large models,

then fine-tune on target-domain data – exactly the approach taken by Sapiens and ViTPose. In cases like infant pose, practitioners should be careful to evaluate whether a purported "infant-specific" model actually generalizes to their specific setting. Ensuring diversity in the fine-tuning data (e.g. different camera views – the study noted a top-down camera improved accuracy over the typical diagonal view) is also key for robust adaptation.

Figure: A typical transformer-based pose estimation architecture (as used in ViTPose). A ViT encoder processes image patches and a simple decoder outputs heatmaps for each keypoint [4]. Such plain designs have proven effective when pre-trained at scale, and they can be adapted to new keypoint definitions by modifying the decoder or inserting task-specific layers.

PoseBH (Jeong et al., 2025): Multi-Dataset Training with Prototype Representations

Problem: Adapting a pose model to a new domain can lead to **catastrophic forgetting** of the original domain or incompatibility between different keypoint formats. **PoseBH** (arXiv 2505.17475, 2025) tackles this by training a single model on **multiple pose datasets simultaneously**, to achieve broad generalization beyond any one domain. The key challenge addressed is **heterogeneous skeletons**: different datasets define different keypoints (e.g. COCO has 17 body joints, MPII 16 joints, COCO-WholeBody 133 keypoints including face/hand, animal datasets have entirely different anatomy). Naively merging datasets or using multiple output heads can fail because (a) many keypoints are non-overlapping or semantically different, and (b) during training, for any given image only the keypoints of one dataset have labels (others are "missing" labels). This is both a **label heterogeneity** and **semi-supervised learning** issue.

Architecture – Keypoint Prototypes: PoseBH introduces an elegant architectural solution: a **unified keypoint embedding space with learned prototypes**. Instead of having a separate head per dataset, PoseBH has a shared **embedding head** that maps the backbone's feature map into an embedding for any keypoint. It also maintains a set of **prototype vectors** representing each keypoint type across all datasets. During training, for an image from dataset A, the model will produce a heatmap by matching the embedding at each pixel to the prototypes – essentially performing a nearest-prototype classification for keypoint presence. Only the prototypes corresponding to dataset A's keypoints will have ground-truth targets; others remain unsupervised for that image. The prototypes are updated **non-parametrically** (like cluster centers) from the embeddings each batch. Conceptually, this forces the model to **learn a common representation** for semantically similar joints across datasets (e.g. "left eye" in human vs. "left eye" in animals) even if their appearances differ, because the prototype acts as a target for both. It also allows the model to output any dataset's keypoints by selecting the relevant prototype set.

Cross-Type Self-Supervision: To handle the missing label problem (unlabeled keypoints of other types in each image), PoseBH devises a **cross-type self-supervision** mechanism. In essence, it uses the model's own predictions to supervise those parts: for an image from dataset A, the model will predict some "pseudo" keypoints for other datasets' skeletons (via the prototype matching). By enforcing consistency between the embedding-to-prototype predictions and the model's direct multi-head output (or between different prototypes) for those unlabeled parts, it generates a self-training signal. This is done without a separate teacher model or heavy augmentation, making it computationally cheap. Intuitively, if a certain joint (say "wrist") is not labeled in the current dataset, the model can still guess where a "wrist" might be (using its learned prototype for wrist) – and while this guess can't be directly corrected, PoseBH aligns two modalities of guessing to filter out inconsistent predictions, thereby **regularizing the model** to produce plausible outputs for the unlabeled joints.

Training & Data: The authors train PoseBH on a combination of **five datasets**: COCO (17 keypoints), MPII (16), AI Challenger (14), COCO-WholeBody (133), and two animal pose datasets AP-10K & APT-36K (with 17 animal keypoints). The backbone is a ViT (like ViTPose's) shared for all. They compare to baselines like a multi-head network (one head per dataset). A straightforward fine-tuning on a new domain often "forgets" others and yields poor multi-domain performance; PoseBH avoids this by joint training with its unified representation.

Results: PoseBH demonstrates **substantial cross-domain generalization gains**. Notably, it **improves accuracy on under-represented datasets (WholeBody, animals)** while **preserving performance on standard human pose sets**. For example, it outperforms prior multi-dataset approaches on COCO-WholeBody and APT-36K animal poses by a large margin, without hurting COCO or MPII accuracy. This means the model can handle fine-grained keypoints (face, hands) and different anatomical structures in one network. They also show the learned **keypoint embeddings are transferable**: using the PoseBH encoder + prototypes as initialization, they fine-tune on tasks like 3D human shape estimation (3DPW dataset) and 3D hand shape (InterHand2.6M), getting strong results. This suggests the unified embedding space captures semantically meaningful keypoint features that extend to 3D and other modalities.

Applicability: PoseBH is particularly relevant when adapting a model to a new domain while wanting to maintain performance on the original domain, or to create one model for many sub-populations. For example, if we have a foundation model on adult poses and we collect a new infant pose dataset, a naïve fine-tuning on infants might degrade adult performance (if we still care about it). PoseBH instead could incorporate the infant data in a multi-domain training and handle both. Its prototype-based approach is a form of **architecture adaptation** that avoids committing to a fixed set of keypoints – useful if the target domain has a different skeleton (e.g. animals, or a clinical marker set). One could initialize prototypes for new keypoints and continue training. This work doesn't assume unlabeled target data (it uses labels per dataset in supervised fashion), but it cleverly uses unlabeled joints within each image to maximize training signal. Overall, PoseBH moves toward a **"universal" pose estimator** that can be adapted to **different body topologies** without manual relabeling or separate models. The code is available, which can help practitioners apply the prototype learning approach to their own combination of datasets.

DWPose (Yang et al., ICCV 2023): Whole-Body Pose Estimation via Two-Stage Distillation

Domain Focus: DWPose ("Distilled Whole-body Pose") addresses adaptation from the standard human pose (body-only joints) to **whole-body pose estimation** – i.e. localizing **body, face, hand, and foot keypoints all together**. Whole-body pose is challenging because it involves **multi-scale features** (e.g. tiny finger joints vs large torso), more keypoints (133 in COCO-WholeBody), and often **data scarcity** for hands/face in full-body images. Rather than designing a new architecture, DWPose focuses on a **training strategy: knowledge distillation (KD)** to effectively **fine-tune and compress a model for whole-body tasks**. This is relevant to domain adaptation in that they start from a high-capacity teacher model and adapt its knowledge to a smaller model that is more efficient for deployment, all while **improving accuracy on the new task (whole-body)**.

Architecture: The base model used is **RTMPose** (by MMPose, 2023) – a CNN-based pose model that had strong performance on COCO-WholeBody after training on a mix of data. DWPose's architecture is essentially RTMPose, but distilled: e.g. using an RTMPose-x (extra-large) as teacher and RTMPose-l (large) as student. The key innovations are in how the distillation is done in two stages:

- **Stage 1 – Full Model Distillation with Visibility-Aware Supervision:** They train the student from scratch using the teacher's intermediate **feature maps** and final **heatmap outputs** as guidance.

Crucially, they do **not discard "invisible" keypoints** in the loss. Normally, if a keypoint is labeled occluded, pose models ignore it for supervision, but DWPose uses the teacher's prediction for even invisible joints as a soft target. This provides additional signal (the teacher can impart a reasonable guess of an occluded joint's position). They also apply a **progressive weight decay to the distillation loss** over training – giving strong guidance early on and then letting the student gradually stand on its own.

- **Stage 2 – Head-Focused Self-Distillation:** After Stage 1, the student's backbone is fairly strong. In Stage 2, they freeze the student's backbone and **fine-tune only the keypoint head** with a novel self-distillation: the student model is cloned; one acts as a fixed "teacher" and the other as the trainable student. Both are initialized from the Stage-1 student. By distilling the outputs (logits) of the frozen model into the trainable one (for the head only), they effectively **refine the pose head** without altering the backbone. This stage is very fast (only 20% of training time) but further boosts accuracy, as the head can learn to better localize keypoints when guided by a copy of itself (this is akin to self-ensembling). It's a **plug-and-play fine-tuning strategy** that could be applied to other dense prediction tasks as well.

Data Augmentation: To address limited data for fine details (like various hand poses, facial expressions), DWPose incorporates an extra dataset called **UBody**. UBody contains a diverse set of upper-body images with rich hand and face annotations. By mixing UBody into training, the model sees more varied hand gestures and facial keypoints, improving its robustness on those parts. This highlights that even with foundation models, data diversity remains crucial – if the target domain lacks certain variations, augmenting with another source (even synthetic or specialized real data) can help.

Results: DWPose achieved a new **state-of-the-art on the COCO-WholeBody benchmark**, with a whole-body AP of **66.5** for the student model – notably **surpassing its teacher's AP (65.3)**. In other words, the distilled large model outperforms the original extra-large model, which is a compelling result of knowledge transfer. It improved the baseline student (without KD) by a sizable margin (+1.7 AP, from 64.8 to 66.5). Qualitatively, it better localizes fine keypoints like fingers and facial landmarks. The authors released a **series of models (tiny through large)** distilled with this method, allowing practitioners to choose a model size that fits their efficiency needs without much sacrifice in accuracy. Code is available on GitHub via the MMPose project.

Fine-Tuning Insights: DWPose exemplifies **stage-wise fine-tuning** and **teacher–student adaptation** for a domain with more keypoints. In the context of foundation models, one could imagine using a similar two-stage approach to adapt a huge model to a specialized domain and then compress it. For instance, if we have a 2B-parameter Sapiens model for general humans, and we want a smaller model optimized for an AR device tracking an elderly person's full body including face, we could: (1) fine-tune Sapiens (teacher) on an elderly whole-body dataset, (2) distill to a smaller ViT or CNN student model using DWPose's strategies. The visibility-inclusive loss is particularly relevant to domains like infants or elderly in which occlusions (self-occlusion, medical equipment, etc.) are common – using a teacher's guess for occluded joints could yield more anatomically plausible predictions. Also, freezing the backbone and only training the head (Stage 2) is a form of **partial model fine-tuning** that is efficient and can be seen as analogous to only updating an adapter or head in a foundation model. It aligns with the idea that the backbone provides general features, while the head can be cheaply refit to domain specifics. This strategy can help avoid over-tuning the entire model and preserve general features, which might be useful if the target domain data, while "sufficient," is still smaller than the pretraining data.

Adapting to Specialized Populations: Additional Notes

The above works primarily focus on broad human pose tasks, but their techniques are **directly applicable to specific populations** like infants, patients, or the elderly:

- **When target data is plentiful and labeled (supervised adaptation):** Fine-tuning a large model (ViTPose, Sapiens) on the target data is very effective. It's often as simple as replacing the output head to match the target's keypoint schema (if different) and then updating all weights on the new dataset. As seen, the large models' prior knowledge helps; e.g., a strong adult-trained model can jump-start infant pose training, achieving higher accuracy than training from scratch or using smaller models. The key is to ensure the target dataset's label definition aligns or is mapped to the model's output. If not (e.g. infant pose might use slightly different joint definitions), one may use multi-dataset methods like PoseBH or simply define a new keypoint head and train the model end-to-end on the new labels.

- **Architectural tweaks for body differences:** Infants have different body proportions (head larger relative to body, often lying down), and elderly subjects might have different pose priors (e.g. less limb flexibility). While a generic model can handle these to an extent, some works incorporate domain knowledge: for example, the **SHIFT** framework (CVPR 2025 workshop) added an **infant pose prior manifold** to ensure predicted infant poses are physically plausible [5] [6]. SHIFT was an unsupervised domain adaptation approach (adult to infant) that assumed no infant labels and thus used a mean-teacher setup with consistency losses. It's somewhat outside our "sufficient data" assumption, but its findings are useful: accounting for **anatomical differences via a prior** and **occlusions via segmentation cues** led to big improvements in infant pose estimates. Once we do have labels, incorporating such priors (e.g. as a regularizer during fine-tuning) could further improve performance on specialized domains. For instance, one could fine-tune Sapiens on an infant dataset while adding a loss term that penalizes unlikely infant joint configurations (perhaps learned from infant motion data) – to avoid the model inadvertently predicting adult-like poses.

- **Use of Depth/3D Data:** Some specialized applications have access to depth cameras or multi-view setups (e.g. hospital settings). While most 2D pose models ignore depth, having depth can help disambiguate occlusions and improve robustness. Approaches like the infant 3D pose framework by Soualmi et al. (2024) used **stereo camera images** and fine-tuned an adult 2D pose network to infer 3D infant pose. They achieved a mean error of 1.72 cm in 3D by training on a large stereo infant dataset (88k images). This underscores that **if sufficient 2D+depth data is available, fine-tuning a model with a multi-view or 2.5D approach can greatly improve accuracy**. Depth can be incorporated by extending the input channels (for a depth map) or by multi-task training (estimating 2D pose and depth jointly). In foundation models, one could imagine extending Sapiens (which already has a depth prediction head) to also take depth input for pose, or using depth as an intermediate supervision to aid domain adaptation (as some 3D UDA methods do by reconstructing 2D images). In general, additional modalities that remain consistent across domains can serve as bridges for adaptation.

- **Evaluation and Generalization:** Finally, an important practical aspect when fine-tuning on a specialized domain is to test generalization on related but unseen scenarios. As noted, an infant model trained on one dataset often fails on another without adaptation. To avoid this, one should incorporate diversity in training (different ages, environments, camera angles for infants, etc., or even use data augmentation/style transfer). Multi-domain training (like PoseBH) can be a powerful way to maintain a model's utility across sub-populations. For instance, a single model could be trained on adult, infant, and elderly data together using a unified output (perhaps using

the union of keypoints or separate heads). This would ensure the **foundation model remains strong overall, not overfitted to one group**. The downside is it requires collecting all such data and possibly more complex training. A compromise could be sequential fine-tuning with caution: e.g., start with a foundation model, fine-tune on infant data, but also periodically mix in some adult data (with lower weight) to retain adult capability (a form of continued pre-training).

In summary, the recent literature converges on the view that **large, pre-trained pose models are highly adaptable to new domains through supervised fine-tuning**, often outperforming bespoke models. By leveraging strategies like architectural modularity (task-specific heads or FFNs), knowledge distillation, and multi-dataset learning, one can **specialize a pose estimator to a particular population (or multiple) without assuming data scarcity**. The availability of open-source foundation models (ViTPose, etc.) and frameworks (MMPose, DeepLabCut, etc.) makes it feasible for practitioners to apply these findings: for example, fine-tuning ViTPose on a custom dataset of patients with Parkinson's, or using PoseBH's prototype method to train one model on humans and pets together. The works from CVPR/ICCV/ECCV 2024–2025 provide both the **tools (code, models)** and the **insights (importance of data quality, simple architectures, and careful adaptation losses)** to guide such domain adaptation efforts.

Sources:

- Xu et al., "ViTPose++: Vision Transformer for Generic Body Pose Estimation," 2023.
- Khirodkar et al., "Sapiens: Foundation for Human Vision Models," 2024.
- Jeong et al., "PoseBH: Prototypical Multi-Dataset Training Beyond Human Pose Estimation," 2025.
- Yang et al., "Effective Whole-body Pose Estimation with Two-stage Distillation (DWPose)," ICCV 2023.
- Jahn et al., "Comparison of generic vs. specialized infant pose estimators for GMA," Sci. Reports 2025.
- Bose et al., "SHIFT: Unsupervised Infant Pose Estimation via Synthetic Adult Data," CVPR W 2025 [6].
- Soualmi et al., "3D pose estimation for preterm infants in NICU," MTAP 2024.

---

[1] [2] [3] ViTPose++.pdf
file://file-FYU2me95W3R5fB3UEwRo4j

[4] Sapiens: Foundation for Human Vision Models by Meta
https://learnopencv.com/sapiens-human-vision-models/

[5] [6] Leveraging Synthetic Adult Datasets for Unsupervised Infant Pose Estimation
https://arxiv.org/html/2504.05789v1