Fine-Tuning Pose Estimation Models for Specialized Populations

Background and Challenges

Large-scale human pose estimation models (e.g. Meta's **Sapiens** or the ViTPose family) are typically trained on datasets of adult humans in diverse activities. Adapting these **foundation models** to specialized populations – like infants, the elderly, or wheelchair users – poses several challenges. Such populations exhibit **domain shifts** in appearance, body proportions, and typical poses that can confound models trained on adult data. For example, infants have shorter limbs relative to their torso and often lie supine, yielding highly self-occluded poses unfamiliar to models trained on upright adults. Similarly, older adults may have atypical postures or assistive devices (canes, wheelchairs) not seen in standard training data. These differences can lead to poor keypoint localization if models are applied naively across age groups. To bridge this gap, recent research has explored **domain adaptation** and **generalization techniques** to fine-tune pose models for new demographics. Below, we review key approaches focusing on: (1) domain adaptation methods, (2) generalization across age/body variations, (3) keypoint system transferability, and (4) strategies to handle pose and proportion differences, with comparisons of transformer-based vs. CNN-based architectures where relevant.

Domain Adaptation Techniques for New Demographics

Fine-Tuning on Target Data: The most direct approach is to fine-tune a pre-trained model on labeled pose data of the target population. Even a small amount of in-domain data can yield significant accuracy improvements. Jahn *et al.* (2025) found that a generic ViTPose model trained on adults performed best on their infant dataset among off-the-shelf models, and its accuracy improved markedly after fine-tuning on ~4,500 infant images. This underscores that while large pre-trained models have some generalization ability, **targeted fine-tuning** is crucial for optimal performance in a new domain. Foundation models like **Sapiens** are explicitly designed for easy adaptation: Sapiens models pre-trained on 300M human images can be *"simply fine-tuned"* to new tasks or data, exhibiting strong generalization even when the new domain's labeled data is scarce. In practice, fine-tuning typically involves replacing or re-initializing the final keypoint prediction head to match the target keypoint set, then updating the model weights on the new dataset with a smaller learning rate (sometimes freezing early layers to retain generic features).

Adversarial Domain Adaptation: Beyond naive fine-tuning, domain adaptation techniques aim to leverage **unlabeled or synthetic data** to bridge the source-target gap. Ostadabbas and colleagues introduced **FiDIP** (Fine-tuned Domain-adapted Infant Pose) to transfer knowledge from adult poses to infant poses. FiDIP uses an **invariant representation learning** approach: a pose estimation backbone (e.g. HRNet or ResNet) is trained on both adult and infant images, with an additional adversarial loss encouraging the feature embeddings to be **domain-invariant** between adults and infants. They also generated a **Synthetic Infant Pose (SyRIP)** dataset to augment the few real infant images. This combination of synthetic data and adversarial feature alignment led to higher accuracy than straightforward fine-tuning of the same backbones. In fact, FiDIP improved AP by ~1–3 points over plain fine-tuning on multiple CNN backbones, demonstrating the value of explicit domain adaptation.

Unsupervised and Generative Adaptation: When labeled target data is extremely limited or unavailable (a common case due to privacy/ethical issues in infant data), **unsupervised domain adaptation (UDA)** can be applied. Bose *et al.* (2025) proposed **SHIFT**, the first UDA framework for infant pose estimation. SHIFT starts from an adult pose model and adapts it to infants using a *mean teacher* strategy: a student model learns on unlabeled infant images and is guided by a teacher model (an exponential moving average of the student) to produce consistent keypoint predictions under different augmentations. Crucially, SHIFT incorporates infant-specific knowledge in the adaptation process. First, it trains an **infant pose prior** (a generative model of plausible infant skeleton configurations) and penalizes the pose estimator for predicting anatomically implausible poses. Second, to combat infants' frequent self-occlusions, it enforces **visibility consistency**: predicted keypoints must align with the infant's silhouette in the image, via a learned mapping from keypoints to a segmentation mask. These regularizers guide the model to respect infant anatomy and image evidence without needing ground-truth labels. SHIFT significantly outperformed previous UDA methods on infant pose (by ~5% PCK) and even surpassed some fully-supervised infant pose models by 16% – a remarkable result achieved with zero real infant labels.

Another innovative approach uses **generative models for domain adaptation**. Zhou *et al.* (2023) tackle 3D infant pose estimation by leveraging *"generative priors"* learned from large adult pose data. They introduce a diffusion-based pose generator conditioned to produce infant-like 3D poses, effectively **domain-transferring adult poses to the infant domain**. These synthesized infant pose data augment the scarce real data for training a 3D pose model. Coupled with an optimization-based 2D-to-3D lifting method (to sidestep needing huge 3D training sets), their approach achieved state-of-the-art accuracy on infant 3D pose benchmarks (e.g. **MPJPE** 43.6 mm on SyRIP). This highlights the power of *guided data generation* for domain adaptation: by hallucinating target-domain poses (or images), one can fine-tune foundation models without overfitting the few real examples.

Synthetic Data and Style Transfer: Across these methods, a recurring theme is **augmenting the target domain** via synthetic or transferred data. Synthetic generation can take many forms: image rendering (e.g. inserting 3D infant models into backgrounds), geometric warping of adult images to infant proportions, or as above, generating poses or whole images via learned models. Even for elder populations or other scenarios with limited data, researchers suggest using *"synthetic images and/or videos generated by virtual reality"* to expand the training set. For instance, in the context of elderly posture recognition, Bustamante *et al.* (2024) note that domain adaptation and synthetic data are **elegant solutions to data scarcity**. The takeaway is that combining a large pre-trained pose model with target-domain synthetic data and adaptation losses can effectively **transfer** the model to new demographics without requiring prohibitively large real datasets.

Generalization Across Age Groups and Body Proportions

Instead of domain-specific adaptation, another line of work aims for **broader generalization** so that a single model works well across diverse ages, body shapes, and contexts. One strategy is to incorporate **population diversity during pre-training**. The Sapiens model, for example, was pre-trained on Humans-300M, a massive curated image set featuring *"diverse human images"* across varied conditions. This foundation model approach improves robustness: Sapiens showed *"remarkable generalization to in-the-wild data, even when labeled data is scarce or entirely synthetic"*. In practice, such a model would already have exposure to people of different sizes, poses, and possibly ages, making it less prone to fail on, say, an infant or an elderly person if they appeared in-the-wild. Indeed, Sapiens reports state-of-the-art results on multiple human-centric tasks and strong cross-dataset performance by virtue of its scale and diverse pre-training. This suggests that **scale + diversity = generalization**: training on large, varied datasets (and using high-capacity models) helps cover the distributional extremes (tiny infants, stooped elders) that were once out-of-distribution.

Another approach to enhance generalization is through **data augmentation and normalization techniques** that specifically account for body differences. For 3D pose estimation, some works propose augmenting bone lengths and limb proportions during training to teach the model invariance to different physiques. For 2D pose, researchers have used scale jittering, rotation, and even synthetic occlusions to ensure a model doesn't overly rely on a fixed body shape. Hesse *et al.* (2020) introduced the small **MINI-RGBD** infant pose dataset and noted that standard adult-trained models performed poorly until they applied careful augmentation and tuning for the infant's size and lying pose. **Domain generalization** frameworks (e.g. leveraging dual-augmentors or meta-learning) can train a pose network that generalizes to unseen domains without explicit adaptation, though these have been explored more for camera viewpoint or background shifts than age-related shifts.

A practical insight from Jahn *et al.* (2025) is that **multi-angle training data** can help generalization within an age group. They found that incorporating a top-down camera view significantly improved infant pose accuracy compared to the traditional diagonal view used in clinical assessments. This implies that broader coverage of pose appearance (here, different camera perspectives capturing the infant's movement) made the model more reliable. By analogy, a model trained on both upright and bed-ridden people, or both standing and wheelchair-bound postures, is better equipped to generalize. In summary, to improve generalization across age groups one should: (1) train on as **diverse a dataset** as possible (or use a foundation model already pre-trained on such data), and (2) use **targeted augmentations** that simulate the differences in body proportions and pose distributions (e.g. adjusting limb lengths, adding supports like wheelchairs in scenes, varying camera viewpoints).

Keypoint System Transferability

Adapting a pose model to a new population sometimes entails changing the **keypoint definition** itself. For example, infant pose datasets might use a slightly different keypoint set than COCO's 17 joints – perhaps omitting some facial keypoints or adding others relevant to infant movement (such as a different spine reference). Even more dramatically, "whole-body" pose models aim to predict **133 or more keypoints** (including hands and facial landmarks) instead of just the 17 body joints, and animal pose datasets define entirely different joint structures. A key question is how well a foundation model's knowledge can transfer when the **output keypoint system changes**.

Transformer-based architectures have recently shown a elegant solution via **modular heads and knowledge tokens**. Xu *et al.* (2023) extend their ViTPose model to **ViTPose++**, which can handle *"heterogeneous body keypoint categories in different pose estimation tasks"*. ViTPose++ introduces **knowledge factorization** in the transformer encoder: it separates each layer's MLP into a *task-agnostic* part (shared across all keypoint types) and a *task-specific* part that is specialized for a particular keypoint set. During multi-task training, the model learns a robust shared representation while also accounting for differences in keypoint semantics via these task-specific components. This allows a single transformer to output, say, 17 keypoints for a human body or 20 keypoints for an animal, without confusion. Moreover, ViTPose++ uses a **"knowledge token"** to distill information from a large model into a smaller one. In practice, one can pre-train a giant ViT on a superset of keypoints (or multiple datasets), then transfer its knowledge to a smaller model that might focus on a specific keypoint set, achieving efficient adaptation. The result was a *generic* pose estimator that set state-of-the-art on **MS-COCO (17 keypoints)**, **COCO-WholeBody (133 keypoints)**, and even animal pose benchmarks *simultaneously* with one model.

Classic CNN-based approaches have relied on simpler techniques for keypoint transfer. One common method is to use a **shared backbone** (e.g. ResNet or HRNet) and attach multiple **task-specific heads** for different keypoint sets. During fine-tuning, only the head for the target keypoint set is trained (or one can train all heads jointly on respective datasets if doing multi-task learning). For example, the **Sapiens** pose models provide

checkpoints for 17, 133, and 308 keypoint configurations, all derived from the same pre-trained backbone. The 308-keypoint model is fine-tuned for a **densely annotated whole-body schema** (body, hands, feet, face) that Meta introduced to surpass prior datasets. Yet, Sapiens could reuse much of its learned human features – only the final prediction layers differ. The ability to **natively predict 308 keypoints** indicates a high degree of transferability from the original 17-keypoint knowledge to a much richer pose representation. Similarly, for infant adaptation, one might start with an adult model (17 joints) and fine-tune it to predict an infant-specific subset of joints. If some keypoints are not needed (e.g. maybe "neck" was not annotated for infants in a given dataset), those output neurons can be dropped or ignored during training. The main challenge is ensuring the model **understands new keypoints** if introduced. Techniques like **knowledge distillation** can help: e.g. using the predictions of an adult model on overlapping keypoints to guide the infant model (for the shared joints), while learning new joints from scratch.

In practice, **heterogeneous keypoint transfer** is facilitated by modern frameworks. The ViTPose++ approach of intermixing task-agnostic and task-specific layers is a cutting-edge example that avoids training separate models for each keypoint definition. This is particularly useful if one wants a unified model for, say, full-body pose and whole-body pose – relevant for elder care scenarios where tracking body pose and facial expression together might be important. For specialized populations, adopting a unified skeleton definition can also be beneficial: some works propose adding keypoints that capture clinically significant movements (e.g. distinguishing left/right limb movements in infants) since *"the 17 adult keypoints in COCO do not support infant movement detection well in terms of clinical significance"*. A foundation model can be fine-tuned to this new keypoint schema as long as the differences are accounted for (possibly with a larger decoder or multi-head setup). Overall, the trend is moving toward **flexible pose models** that seamlessly transfer knowledge between keypoint systems, rather than starting over for each new annotation scheme.

Handling Pose Variations and Dataset Biases

Specialized populations often display a narrower or biased set of poses compared to the general population, which can hurt a model's performance. For instance, infants spend most time on their back or stomach and rarely assume the variety of poses adults do. An **adult-trained model may misidentify limbs** or produce implausible poses when confronted with these unusual configurations. Similarly, an elderly subject with a hunched posture might be misinterpreted by a model expecting upright poses. Addressing this requires both data-centric and model-centric strategies:

- **Augmenting Pose Diversity:** As discussed, synthetic data generation is a key strategy. By augmenting the training set with pose variations that the model would otherwise rarely see, one can reduce bias. In infant pose estimation, generating synthetic examples of infants in various positions (curled, stretching, etc.) proved effective – Huang *et al.* (2021) showed that training on a mix of real and synthetic infant data (SyRIP) dramatically raised AP on the test set compared to using a tiny real dataset alone (e.g. AP 69→90 on one backbone). For underrepresented poses in elderly populations (e.g. a senior using a walker), one could similarly generate or simulate data (via pose CGI or manipulating adult data to add walkers) to balance the training distribution.
- **Pose Prior and Constraints:** Incorporating human anatomy knowledge can prevent unrealistic estimations on unfamiliar poses. The SHIFT method's **infant pose manifold prior** is a prime example. It ensured that, even if an infant's legs were curled up (a pose absent in adult data), the predicted joints respected feasible limb lengths and angles for an infant, by penalizing unlikely configurations. Such constraints can be encoded via learned priors or even simple bone-length ratios. In a model for elderly poses, one might include a prior that accounts for limited joint ranges (if known) or common assistive postures (e.g. leaning on a cane, which could be represented as a particular limb orientation pattern). Pose priors act as **regularizers** that keep the model's predictions within the realm of plausible human

poses for that demographic, thereby improving robustness when the visual evidence is ambiguous or biased.

- **Context and Multi-Modality:** Specialized scenarios sometimes benefit from additional context cues. In hospital settings (infants in crib, adults in bed), leveraging the scene context can help the pose model. SHIFT's use of **segmentation masks** to enforce consistency between the estimated pose and the image silhouette is one such technique. If a model predicts an arm position that doesn't align with the visible arm in the segmentation, it can self-correct. For elderly care, one could imagine integrating depth sensors or IMUs for difficult poses (indeed, **WheelPose** (2024) explored using a few low-cost IMUs on wheelchair users to complement vision, greatly improving pose estimation in seated positions). While beyond pure image-based estimation, these strategies highlight that embracing **multi-modal signals or context constraints** can mitigate dataset bias (e.g. a pose estimator that knows a person is sitting can adjust its body pose expectations accordingly).
- **Bias in Annotation:** Lastly, dataset bias can come from how keypoints are annotated. If annotators systematically miss certain poses (e.g. subtle hand movements in infants) or certain demographics are underrepresented in training, the model inherits those biases. **High-quality, diverse annotations** are essential. The Sapiens project addressed this by collecting **denser keypoint annotations** (308 points) to capture fine details across the whole body. By training on such rich data, a model is less likely to, say, ignore a subtle foot movement of an infant or a small head tilt of an elderly person, because the annotation schema itself is more comprehensive. In practice, when extending a model to a new group, one should consider if the standard keypoints are sufficient or if **custom keypoints** would better capture the important movements for that group (e.g. adding keypoints for a wheelchair's contact points or an infant's hands if fine motor assessment is needed).

Transformer vs. CNN Architectures in Adaptation

Transformer-based models (like ViTPose, ViTPose++, Sapiens) have emerged as powerful pose estimators that often **outperform CNN-based models** on both standard benchmarks and specialized domains. In the infant pose study by Jahn *et al.*, the ViT-based ViTPose model achieved the lowest error on infants out of several architectures, beating older CNN approaches (OpenPose, MediaPipe, HRNet) even *without* infant-specific training. This can be attributed to the transformer's ability to capture global context – helpful for understanding occluded or unusual poses – and the benefit of massive pre-training that many ViT models leverage. Transformers also scale well: Xu *et al.* demonstrate scaling ViTPose up to **1 billion parameters** yields a new Pareto frontier of accuracy vs. throughput, and these large models transfer their knowledge impressively to smaller models and new tasks.

That said, **CNN architectures** remain competitive and are often more lightweight. Many domain adaptation works initially built on CNN backbones (ResNet, HRNet) due to their maturity. For example, FiDIP's invariant learning was tested on ResNet-50 and HRNet backbones, showing consistent gains in infant AP in each case. The **HRNet** in particular, with its high-resolution feature stages, was a top performer for pose estimation and still serves as a strong base to fine-tune for new domains. In AggPose (Cao *et al.*, 2022), the authors compared their pure-transformer design against a hybrid CNN-Transformer (HRFormer) and found AggPose slightly outperformed it on both COCO (adults) and a new infant dataset. AggPose's fully transformer design with multi-scale feature aggregation converged faster and was *"easy to transfer from the COCO dataset to [the] infant pose dataset,"* achieving 95.0 AP on infants. This suggests that a well-designed transformer can not only match CNNs on speed/accuracy but also simplify cross-domain transfer.

In practice, the choice may come down to available data and compute. A **large ViT model** pre-trained on ample data (like ViTPose-G or Sapiens-1B) gives a strong starting point; its representations are general enough that only minimal fine-tuning is needed for a new demographic. In contrast, a smaller CNN trained from scratch on a limited infant dataset would likely struggle to generalize beyond that dataset's distribution. However, with techniques like adapters or partial fine-tuning, even CNNs can benefit from large-scale pre-training (e.g. using ImageNet-pretrained weights and then domain-tuning). Some recent works combine the two: a CNN might extract low-level features which are then refined by transformer layers (the *"hybrid"* approach seen in HRFormer, TokenPose, etc.). These can offer a compromise, leveraging CNN's efficiency and Transformer's global reasoning.

To summarize the comparison: **transformers tend to excel in heterogeneous and large-data regimes**, making them well-suited for foundation models that cover many populations. They handle varying body proportions and occlusions gracefully through self-attention across the image. **CNNs,** on the other hand, remain effective for targeted fine-tuning when compute is limited; their inductive biases (like translation invariance and locality) can be advantageous on smaller datasets. The current state-of-the-art sees transformers taking the lead for pose estimation, but often the highest performance is achieved by combining large-scale pre-training (which could be ViT or CNN based) with the right adaptation techniques for the domain at hand.

Summary of Approaches (2019–2025)

To crystallize the insights, the table below compares several recent approaches for adapting pose estimation models to specialized domains:

Method (Year)	Architecture	Target Domain	Adaptation Technique	Key Contributions / Results
Sapiens (Meta, 2024)	ViT-based (0.3–2B params)	Broad (foundation model)	Pre-train on 300M diverse human images; fine-tune lightweight task heads for domain	*Foundation model* for 4 tasks (pose, segmentation, depth, normals). Supports 17, 133, **308 keypoints** in one framework. Fine-tuning yields SOTA on Humans-5K pose with strong generalization.
ViTPose++ (2023)	ViT (plain transformer)	Generic (multi-task: humans, animals)	Knowledge factorization (task-agnostic vs. specific layers); knowledge token distillation	Unified model for **heterogeneous keypoint sets** (human body, whole-body, animal poses). Achieved SOTA on COCO, COCO-WholeBody, and AP-10K (animal) simultaneously without speed loss. Demonstrated transfer of 1B-model knowledge into smaller models.

Method (Year)	Architecture	Target Domain	Adaptation Technique	Key Contributions / Results
AggPose (IJCAI 2022)	ViT (multi-scale transformer)	Infants (clinical poses)	Pre-train on COCO (adults); full fine-tune on new large infant dataset (20k images)	First **large-scale infant pose dataset** + pure ViT model. Outperformed hybrid CNN-Transformer (HRFormer) on infant data by leveraging global self-attention. Reached **95.0 AP** on infant poses and even slightly improved adult COCO pose AP.
FiDIP (FG 2021)	CNN (HRNet/ResNet)	Infants (small-data)	Domain adversarial training for invariant features; **SyRIP** synthetic infant data generation	Pioneered **domain adaptation for infant pose**. Transfers adult pose knowledge to infants via adversarial feature alignment and synthetic training data. Improved AP by ~2–3 points over direct fine-tuning on small infant datasets.
SHIFT (CVPRW 2025)	CNN (HRNet) + custom modules	Infants (unlabeled data)	**Unsupervised DA** with mean-teacher consistency; infant pose prior; silhouette consistency regularization	First UDA method for infant pose (no labels needed). Introduced an infant **manifold pose prior** to enforce plausible poses and a segmentation-guided consistency to handle occlusions. Outperformed supervised infant models by 16% on cross-dataset generalization.
Zhou et al. (WACV 2024)	CNN (optimization-based 3D lifting)	Infants (3D pose)	Diffusion generative model to adapt adult 3D poses into infant-like poses; optimize 3D from 2D with generative prior	**Data augmentation via pose diffusion**. Bridged adult-infant gap in 3D by generating infant-styled poses to augment training. Achieved SOTA 3D infant pose (43.6 mm MPJPE on synthetic+real dataset) with minimal real data. Showed efficacy of generating target-domain examples for training.

Method (Year)	Architecture	Target Domain	Adaptation Technique	Key Contributions / Results
WheelPose (2024) *[Lu et al.]*	ViT (leveraging ViTPose++)	Wheelchair users (seated pose)	Synthetic data (rendered wheelchair scenarios); knowledge factorization for new keypoints (e.g. wheels)	Addressed pose estimation for wheelchair-bound individuals. By synthesizing training images of people in wheelchairs and extending the model's keypoint schema (to include contact points, etc.), improved pose accuracy on wheelchair user datasets. *(Reported to use ViTPose++ with task-specific adaptations; results pending publication.)*

Table: Representative methods for adapting pose models to specialized populations. Transformer-based models (top rows) generally offer strong transfer learning capabilities, while CNN-based methods (bottom rows) have pioneered many domain adaptation techniques.

Notes: The above table illustrates the evolution from early CNN approaches (which needed explicit domain adaptation techniques) to large transformer models that inherently generalize better and handle multi-task learning. However, the approaches are complementary – e.g., one could apply SHIFT's ideas to a ViT model, or use Sapiens as the backbone in FiDIP – combining foundation model power with domain-specific tricks.

Conclusion

Adapting human pose estimators to unusual demographics is an active research area bridging **computer vision and biomechanics/clinical needs**. The consensus from recent literature is that large pre-trained models provide an excellent starting point, but **specialized techniques are essential** to reach peak accuracy on domains like infants or elderly. Domain adaptation methods – from fine-tuning with a few real examples to generating synthetic data or enforcing anatomical priors – greatly improve performance on these populations. Generalization across age groups can be enhanced by training on diverse data and using architectures that accommodate different body sizes and keypoint definitions. Notably, the advent of **transformer-based foundation models** (ViTPose++, Sapiens, etc.) has made it easier to carry knowledge across domains and tasks, thanks to their flexibility and capacity. At the same time, targeted solutions (augmented data, domain-invariant features, pose priors) remain crucial to address the specific challenges (like extreme poses or occlusions) that come with specialized populations.

In summary, **fine-tuning large pose models for infants, elderly, or other underrepresented groups is best achieved by a combination of:** (1) leveraging a strong general model (preferably one pre-trained on a broad human dataset) as the baseline, (2) applying domain adaptation techniques (supervised or unsupervised) to account for appearance and pose distribution shifts,, (3) modifying the keypoint output or model head as needed to match the target anatomy, and (4) injecting prior knowledge or synthetic examples of the new domain to guide learning. By following these principles, researchers have attained substantial improvements in pose estimation accuracy for populations that were previously "out-of-scope" for mainstream models – bringing us closer to truly **inclusive** human pose estimation across all ages and conditions.

References: The information and results above are drawn from a range of recent studies, including foundation model descriptions, infant pose estimation benchmarks, domain adaptation frameworks, and multi-task pose model innovations, among others, as cited throughout. These works collectively highlight the state-of-the-art strategies for fine-tuning large pose models to specialized domains in the last five years.

面向特定人群的姿态估计模型微调

背景与挑战

大规模人体姿态估计模型（例如 Meta 的 **Sapiens** 或 ViTPose 系列）通常在包含各种人类活动的成年数据集上训练。将这些**基础模型**适配到如婴幼儿、老年人或轮椅用户等特定人群时，会面临多个挑战。这些人群在外观、身体比例和典型姿态方面存在**领域差异**，这可能会扰乱原本在成年人数据上训练的模型 。

例如，婴儿的四肢相对于躯干更短，而且通常是仰卧的，因此会产生严重的自遮挡现象，这是模型在面对站立成年人的常规姿势时所未遇到的问题 。类似地，老年人可能有非典型的姿态，或使用辅助设备（如拐杖、轮椅），这些情况在标准训练数据中是缺失的。若将这些模型直接应用于其他年龄群体，可能会导致关键点定位效果不佳 。

为了解决这一差距，近期的研究探索了**领域自适应**和**泛化技术**，以对姿态模型进行微调，适应新的群体。下面我们将回顾并对比关键方法，重点关注以下四个方面：

1. 领域自适应方法；
2. 跨年龄与身体变化的泛化方法；
3. 关键点系统迁移策略；
4. 应对姿态变化、身体比例差异与数据集偏差的策略；
5. 我们也将对 Transformer 架构与 CNN 架构在这些任务中的表现进行比较。

针对新群体的领域自适应技术

在目标数据上微调：最直接的方法是，在目标群体的标注姿态数据上对预训练模型进行微调。即使目标领域数据量很小，也可以显著提升精度。Jahn 等人（2025）发现，一个通用的 ViTPose 模型（在成年人上训练）在他们的婴儿数据集上表现最好，而经过在约 4,500 张婴儿图像上的微调后，其准确率显著提高。这表明，虽然大规模预训练模型具有一定的泛化能力，但要在新领域上达到最佳性能，**有针对性的微调**是必不可少的。像 **Sapiens** 这样的基础模型正是为这种易于适配而设计：Sapiens 在 3 亿张人体图像上进行了预训练，可以**“直接微调”**到新的任务或数据上，即使新领域的标注数据很少也能表现良好 。在实际操作中，微调通常包括：替换或重新初始化用于输出关键点的预测头，以适配新的关键点集合，然后以较小的学习率在新数据集上更新模型权重（有时会冻结前面的层以保留通用特征）。

对抗性领域自适应（Adversarial Domain Adaptation）：除了直接微调外，领域自适应方法还旨在利用**无标注或合成数据**来弥合源域与目标域之间的差异。Ostadabbas 及其同事提出了 FiDIP（Fine-tuned Domain-adapted Infant Pose），用于将成年人的姿态知识迁移至婴儿 。FiDIP 使用一种**不变表示学习（invariant representation learning）**方法：在成人与婴儿图像上同时训练一个姿态估计主干网络（如 HRNet 或 ResNet），并引入额外的对抗性损失，使得模型学习的特征在这两个域之间保持一致 。他们还构建了一个**合成婴儿姿态数据集 SyRIP**，用以增强真实婴儿图像数据的稀缺问题 。这种合成数据与对抗特征对齐的结合使得模型准确率提升超过直接微调方法 。实际上，FiDIP 在多个 CNN 主干模型上都比普通微调提高了约 1–3 个百分点的 AP，显示出显式领域自适应方法的价值 。

无监督与生成式自适应：当目标领域缺乏真实标注数据（例如婴儿数据因隐私或伦理问题难以获取）时，可以采用**无监督领域自适应（UDA）**方法。Bose 等人（2025）提出了 SHIFT，这是首个面向婴儿姿态估计的 UDA 框架。SHIFT 从一个成人姿态模型出发，通过*mean teacher* 策略进行自适应：学生模型在无标注婴儿图像上学习，而教师模型是学生的指数滑动平均，指导学生在不同图像增强条件下产生一致的关键点预测。

关键是，SHIFT 在自适应过程中融入了婴儿特有的知识。首先，它训练了一个**婴儿姿态先验模型**（一个可生成合理婴儿骨架结构的模型），并惩罚模型预测不符合解剖结构的姿态。其次，为了解决婴儿常见的自遮挡问题，它引入了**可见性一致性机制**：强制模型预测的关键点应与图像中的轮廓一致，通过一个学习的映射函数将关键点与图像分割掩码对齐。这些正则项在不依赖真实标签的情况下，引导模型学习合理的人体结构。SHIFT 明显优于之前的 UDA 方法，在婴儿姿态任务中提升约 5% 的 PCK，甚至超越部分全监督婴儿姿态模型达 16% ——这在完全无标注条件下取得如此成果，十分惊人。

另一种创新方式是使用**生成式模型进行领域自适应**。Zhou 等人（2023）处理的是 3D 婴儿姿态估计任务，他们利用从成人大规模数据中学习到的 *"生成式先验"*。他们引入了一个基于扩散的姿态生成器，该生成器被训练以生成具有婴儿特征的 3D 姿态，实质上将成年人的姿态转换为婴儿域的姿态。这些合成的婴儿姿态用于补充稀缺的真实数据，从而训练一个 3D 姿态模型。该方法还采用了优化式的 2D 到 3D 姿态提升方案（避免了对海量 3D 标注数据的需求），最终在婴儿 3D 姿态评估中取得了 SOTA 表现（如 SyRIP 数据集上的 MPJPE 为 43.6 毫米）。这表明，通过**生成目标域样本**（无论是姿态还是图像），可以有效微调基础模型而不会过拟合极少量的真实样本。

合成数据与风格迁移：贯穿上述方法的共同点是通过合成或迁移方式**扩展目标领域的数据**。合成方式多种多样：可以是图像渲染（例如将 3D 婴儿模型嵌入背景中），也可以是对成人图像进行几何变换以匹配婴儿体型，或者如前所述，通过生成模型生成完整图像或姿态。即便是面向老年人或其他小众人群的研究，学者们也建议使用 *"通过虚拟现实生成的图像或视频"* 来扩充训练集。例如，在老年人姿态识别的场景中，Bustamante 等人（2024）指出，领域适应与合成数据是应对数据稀缺问题的**优雅解决方案**。总结来说，结合大规模预训练姿态模型与目标域的合成数据和自适应损失项，可以**有效迁移模型到新的群体**，而无需大量真实数据。

跨年龄群体与身体比例的泛化能力

与专门针对某一领域的自适应方法不同，另一研究方向致力于实现**更广泛的泛化能力**，即让单一模型能在不同年龄群体、身体形态和应用场景中都能保持良好性能。其中一种策略是，在预训练阶段引入**群体多样性**。例如，Sapiens 模型是在 Humans-300M 这一超大规模的人体图像集合上预训练的，该集合包含了*"在各种条件下的多样人体图像"*。这种基础模型策略可提升模型的鲁棒性：Sapiens 显示出*"即便标注数据稀缺或完全为合成数据，仍可在真实世界中具备显著泛化能力"*。实践中，这种模型已在多个以人为中心的任务中取得了 SOTA 表现，并在跨数据集评估中展现出优秀迁移能力，其核心正是其大规模与多样化的预训练。这意味着，**规模 + 多样性 = 泛化能力**：在大数据与多样样本上进行训练，能够覆盖以往被认为是"分布外"的极端人群（如婴儿、驼背老人）。

另一种提升泛化能力的方法是使用**数据增强与归一化技术**，使模型能够适应不同身体特征。例如，在 3D 姿态估计任务中，有研究提出通过训练时改变骨长与四肢比例，来让模型学会在不同体型下保持姿态判断能力。而在 2D 姿态估计中，研究者使用尺度扰动、旋转变换，甚至模拟遮挡等方式，使模型不依赖于固定的身体结构。Hesse 等人（2020）提出了一个名为 MINI-RGBD 的小型婴儿姿态数据集，他们指出，标准成人模型在该数据集上的表现较差，除非应用针对婴儿比例和姿态的增强方法。**领域泛化（Domain Generalization）**框架也可用于训练不依赖特定领域的模型，例如利用双重数据增强器 或元学习策略，这些方法主要被用于解决摄像机视角或背景变化问题，但也可被扩展至年龄与身体差异建模。

Jahn 等人（2025）提出一个实用的观点：**训练中引入多视角数据**有助于提升模型在特定群体内部的泛化能力。他们发现，相较于传统的临床对角视角，加入从顶部拍摄的图像可以显著提升婴儿姿态估计的准确性。这说明更丰富的外观变化覆盖（如不同摄像头视角）使得模型更加可靠。类比而言，如果一个模型在训练时同时接触到站立者与卧躺者、健步者与轮椅使用者，它将更容易泛化至广泛姿态状态。

总之，要想提高模型在年龄群体间的泛化能力，推荐做法包括：（1）使用**尽可能多样化的数据集**进行训练（或使用已在此类数据上预训练的基础模型），（2）应用**针对性的数据增强策略**，以模拟目标人群常见的身体比例与姿态分布差异（如调整肢体长度、加入轮椅等支持物、变化视角等）。

关键点系统的可迁移性

将姿态估计模型适配到新的人群时，有时意味着需要更改**关键点定义系统**本身。例如，婴儿姿态数据集可能采用与 COCO 数据集中 17 个关节点不同的关键点集合——可能省略部分面部关键点，或添加其他对婴儿动作更有意义的点（如用于反映脊柱中线的附加点） 。差异更显著的情况还包括"全身"姿态估计任务，它要求预测**133 个以上的关键点**（包括手部与面部关键点），而动物姿态数据集中定义的骨架系统则完全不同。一个核心问题是：当输出关键点集合发生变化时，一个基础模型的知识能否**良好迁移**。

Transformer 架构近期提供了优雅的解决方案，其核心是**模块化的预测头（heads）与知识 token 表示**。Xu 等人（2023）将 ViTPose 模型扩展为 **ViTPose++**，该模型可处理 *"在不同姿态估计任务中存在异构关键点类别"* 的情况 。ViTPose++ 在 Transformer 编码器中引入了**知识因子分解**机制：它将每一层的 MLP（多层感知机）划分为一个 *任务无关*部分（在所有关键点类型之间共享），以及一个 *任务特定*部分，用于针对特定关键点集合进行优化 。在多任务训练过程中，模型学习到一个稳健的共享表示，同时又通过这些任务特定的分量适配不同关键点语义。这允许单一 Transformer 模型输出例如人类 17 个关键点或动物 20 个关键点，而不会产生混淆。此外，ViTPose++ 还引入了**知识 token**机制，可用于将大模型的能力蒸馏到小模型中 。实践中，研究者可以先在包含关键点超集（或多个数据集）的超大模型上进行预训练，然后将其知识迁移到关注特定关键点集的小模型上，从而实现高效的模型适配。最终结果是一种 *通用* 姿态估计器，可同时在 **MS-COCO（17点）**、**COCO-WholeBody（133点）**，甚至动物姿态基准任务上达到 SOTA 表现 。

传统的基于 CNN 的方法则采用更直接的策略来实现关键点的转移。常见做法是使用一个**共享主干网络**（如 ResNet 或 HRNet），并连接多个**任务特定的输出头（heads）**，分别用于不同关键点集合的回归。在微调过程中，只训练目标任务对应的输出头（或同时训练所有任务的 heads，以实现多任务学习）。例如，**Sapiens** 模型就提供了用于 17、133 和 308 个关键点预测的多个版本，这些版本都是基于相同的主干网络，通过不同的解码头微调得到的 。其 308 关键点模型专为一种**密集注释的全身骨架系统**而训练，该系统由 Meta 团队提出，以超越以往的数据集定义 。尽管关键点数量变化很大，Sapiens 模型仍能重用大量原有的人体表示——仅有预测输出层不同。这种能够**原生支持 308 个关键点预测**的能力表明，模型已成功将原本仅支持 17 点的知识迁移到更丰富的姿态表示上 。

在面向婴儿的微调过程中，研究者通常会从一个成人模型出发（17 关键点），再将其微调以预测适合婴儿的数据集关键点子集。如果某些关键点在目标任务中不再需要（例如婴儿数据集中未标注"脖子"），这些输出节点可以在训练中丢弃或忽略。挑战在于：**当引入新关键点时，模型是否能理解并正确学习它们**。为解决这个问题，可以使用**知识蒸馏**等技术：例如，利用成人模型在共有关键点上的预测，引导婴儿模型进行学习，而对于新加入的关键点则从头学习。

在现代的开源框架支持下，**异构关键点迁移**变得更加便捷。例如 ViTPose++ 的"任务无关+任务特定"混合结构就是一项前沿设计，它避免了每个关键点定义都训练一个独立模型的繁琐步骤 。这对实际部署尤为重要——如老年人护理场景中，可能需要同时监控身体姿态与面部表情，此时，使用统一模型预测全身姿态与面部关键点就显得尤为有价值。对于某些特殊人群，采用统一的骨架定义也有助于数据一致性与可迁移性：有研究者提出可添加关键点以捕捉更具临床意义的动作（例如在婴儿评估中区分左右肢体移动），因为 *"COCO 中的 17 个成年关键点在婴儿运动检测中不具备足够临床意义"* 。一个基础模型可以通过微调扩展至这种新的骨架系统，只要差异部分通过更大的预测头或多任务结构加以适配即可。

总之，趋势正在转向构建**灵活的姿态模型**，它们可以无缝迁移知识、支持不同关键点定义，而无需为每一个新注释体系重新训练整个模型。

应对姿态变化与数据集偏差的策略

特定人群通常会表现出比一般人群**更窄或更偏态的姿态分布，**这可能会影响模型性能。例如，婴儿大多数时间处于仰卧或俯卧状态，几乎不会呈现成年人多样化的动态姿态 。一个在成人数据上训练的模型在面对这些不常见的配置时，可能会**错误识别肢体位置**或预测出不合常理的姿态 。同样地，姿态佝偻的老年人也可能被模型误判为姿态异常。要解决这类问题，需结合数据层面与模型结构层面的策略：

- **增强姿态多样性**：如前所述，合成数据生成是一项关键策略。通过为模型补充那些原本训练集中很少或未见的姿态，可以有效减少偏差。在婴儿姿态估计任务中，通过生成婴儿处于各种位置（如蜷缩、伸展等）的合成图像，研究者显著提高了性能——Huang 等人（2021）指出，在结合真实与合成数据训练时（如 SyRIP 数据集），相较仅使用少量真实数据，模型的 AP 可从 69 提升至 90 。同理，对于在训练中难以覆盖的老年人姿态（如使用助行器时的姿势），可通过姿态图像合成或在成人图像中添加助行器模拟，以平衡训练样本的分布。
- **姿态先验与结构约束机制**：引入人体解剖知识有助于避免模型在遇到不熟悉姿势时生成不合常理的预测。SHIFT 方法中的**婴儿姿态流形先验**即为典型例子 。即使婴儿双腿蜷缩（成年数据中未出现该姿态），该机制依然能限制模型输出符合婴儿解剖结构，通过惩罚不合理的骨长与关节角度，确保预测结果生理合理 。类似地，在老年人姿态建模中，也可以引入有关关节活动范围受限或常见依赖姿态（如靠拐杖站立等）的结构约束。姿态先验在视觉证据不确定或数据偏见严重时充当**正则器**角色，提升模型鲁棒性。
- **利用上下文与多模态信息**：在特定应用中，增加额外上下文信息能有效提升模型效果。例如在医院环境（如婴儿处于婴儿床中、成人卧床）中，场景上下文可辅助模型判断姿态。SHIFT 使用**分割掩码一致性约束**就是一个例子：如果模型预测的手臂位置与图像中的可见区域（分割轮廓）不符，它会自动调整预测。在老年人护理场景中，研究者也提出结合深度传感器或 IMU（惯性测量单元）以解决姿态模糊问题——例如 **WheelPose（2024）** 就研究在轮椅用户身上佩戴少量低成本 IMU，以补充视觉数据，从而提升坐姿状态下的姿态估计性能 。虽然这超出了单纯图像模型的范畴，但这些策略说明，借助**多模态信号或上下文信息**可有效缓解数据偏差带来的模型性能下降（例如，知道对象处于坐姿状态后，模型可动态调整其身体姿态的先验分布）。
- **标注偏差与骨架定义调整**：数据集偏差也可能来源于关键点注释的不一致。如果注释者系统性地忽略某些姿态（如婴儿手部的微动作），或某些群体在训练中严重欠采样，模型将不可避免地继承这些偏差。此时，构建**高质量且多样化的注释体系**极为关键 。Sapiens 项目为解决这一问题，收集并定义了**密集的关键点标注体系**（共 308 个点），以覆盖身体各部分的精细细节 。训练于这种丰富标注数据上的模型，不容易忽视婴儿脚趾的微动或老年人头部微微歪斜等细节动作。
  在实践中，当将模型扩展到新群体时，应评估现有关键点定义是否足够全面，必要时可引入**自定义关键点系统**，以更好捕捉目标人群的关键动作（如轮椅接触点，或婴儿用于神经运动评估的手部动作等）。

微调中的 Transformer 与 CNN 架构比较

基于 Transformer 的模型（如 ViTPose、ViTPose++、Sapiens）近年来已经成为强大的姿态估计器，在标准基准和特定领域任务中通常**超越基于 CNN 的模型**。在 Jahn 等人对婴儿姿态的研究中，基于 ViT 的 ViTPose 模型在未经过婴儿特定训练的情况下，表现优于多个传统架构（OpenPose、MediaPipe、HRNet）。这种表现可归因于 Transformer 模型擅长捕捉全局上下文信息——这对于理解遮挡严重或非典型姿态尤为关键——同时还得益于许多 ViT 模型所使用的**大规模预训练机制**。Transformer 也具有良好的可扩展性：Xu 等人展示了 ViTPose 可扩展至 **10 亿参数**，构建了一个准确率与推理速度双重领先的新边界前沿，而这些大模型的知识也能出色地迁移到更小模型和新任务上 。

尽管如此，**基于 CNN 的架构**依然具备竞争力，特别是在对资源有限或对模型轻量化有需求的任务中。许多领域适应方法最初都构建于 CNN 主干之上（如 ResNet、HRNet），因为它们结构成熟、工程实现稳定。例如，FiDIP 所提出的不变特征学习框架就在 ResNet-50 和 HRNet 上进行测试，在两种主干下都带来了婴儿姿态 AP 的提升 。其中 **HRNet** 尤其出色，其多分辨率高保真特征表达使其长期以来是姿态估计的主力架构，在微调到新领域任务时仍然表现优异。在 AggPose（Cao 等人，2022）中，作者将他们提出的纯 Transformer 设计与一种 CNN-Transformer 混合架构（HRFormer）进行比较，发现 AggPose 在 COCO 成人数据与婴儿数据集上都稍有领先 。AggPose 采用多尺度特征聚合的 Transformer 架构，训练收敛更快，且 *“从 COCO 数据集迁移到婴儿姿态数据集过程十分平滑”*，最终在婴儿数据集上取得了 **95.0 AP** 的性能 。

这表明，一个设计良好的 Transformer 不仅可以在精度上超越传统 CNN 架构，还可以在迁移学习中带来更大的便利性。

在实际应用中，模型架构的选择往往取决于可用数据量与计算资源。一个**预训练良好的大型 ViT 模型**（如 ViTPose-G 或 Sapiens-1B）可作为极具表现力的起点；其通用特征表达足够强大，仅需最小幅度的微调即可适配新的人群或任务。相比之下，一个在小型婴儿数据集上从头训练的 CNN 模型则很难实现跨分布泛化【16†L81-L89adapter)"或部分微调策略，CNN 模型也能从大规模预训练中受益（如使用 ImageNet 权重初始化后进行领域微调）。

此外，近年来一些研究探索将两者结合：CNN 负责提取低层特征，Transformer 负责建模全局语义，这类**混合架构**（如 HRFormer、TokenPose 等）被证明在许多任务中取得了平衡的效果 。这类架构在效率和建模能力之间提供了良好折中，逐渐成为主流趋势。

总结比较如下：**Transformer 架构在任务异质性大或数据丰富场景下表现最佳**，非常适合用作基础模型；其自注意力机制在处理不同体型与遮挡问题时表现优越。相比之下，**CNN 架构在低计算预算或部署端轻量需求下更具实用性**，且其局部感受野与位移不变性对于小样本学习任务仍具优势。当前 SOTA 架构往往是在**大规模预训练模型基础上，结合针对性适配技术**，从而兼顾泛化能力与实际部署需求。

近年方法总结（2019–2025）

为了更清晰地总结上述内容，下表对若干近期在特定人群姿态估计领域的代表性方法进行了比较：

方法（年份）	架构类型	目标人群/应用场景	适应技术	关键贡献与结果
Sapiens (Meta, 2024)	ViT（0.3–2B 参数）	通用（基础模型）	在 3 亿多样人体图像上预训练；微调轻量级预测头	用于四类任务（姿态、分割、深度、法向）的**基础模型**，支持 17、133 和 **308 个关键点**；微调后在 Humans-5K 上实现姿态估计 SOTA 表现，并具备强泛化能力。
ViTPose++ (2023)	ViT（纯 Transformer）	通用（多任务：人类、动物）	知识因子分解（任务无关与特定层）；知识 token 蒸馏	单模型可适配异构关键点体系（人体、全身、动物姿态），在 COCO、WholeBody、AP-10K 上同时达到 SOTA，不降低速度；展示了将 **10 亿参数模型知识迁移到小模型** 的能力。
AggPose (IJCAI 2022)	ViT（多尺度 Transformer）	婴儿（临床姿态评估）	在 COCO 成人数据上预训练；在新婴儿数据集上完全微调	首个**大规模婴儿姿态数据集**（2 万图像）+ 纯 Transformer 架构；在婴儿数据上超过混合模型 HRFormer，达成 **95.0 AP**；在成人 COCO 数据上也有略微提升。
FiDIP (FG 2021)	CNN（HRNet/ResNet）	婴儿（小样本学习）	不变特征学习 + 合成婴儿数据（SyRIP）	首个针对婴儿领域的**领域自适应方法**；通过对抗学习对齐成人与婴儿特征分布，在多个 CNN 主干上均显著提升 AP（相较直接微调提升 1–3 点）。

方法（年份）	架构类型	目标人群/应用场景	适应技术	关键贡献与结果
SHIFT (CVPRW 2025)	CNN（HRNet + 自定义模块）	婴儿（无标注图像）	无监督自适应：mean-teacher + 姿态先验 + 轮廓一致性正则项	首个面向婴儿姿态的**无监督领域自适应（UDA）方法**，提出**婴儿姿态流形先验**与分割一致性机制，在不使用真实标签的前提下**超越部分监督模型**，跨数据集性能提升达**16%**。
Zhou 等人 (WACV 2024)	CNN（优化式 2D → 3D）	婴儿（3D 姿态估计）	扩散模型生成婴儿风格 3D 姿态 + 基于生成先验的优化式回归	利用“**婴儿化生成器**”将成人 3D 姿态风格转化为婴儿，结合 2D–3D 优化重建策略，即使缺乏真实 3D 标注，也能在 SyRIP 数据集上取得 SOTA（MPJPE 43.6mm）；展示了**生成式数据增强策略** 的潜力。
WheelPose (2024)（待发表）	ViT（ViTPose++ 派生）	轮椅使用者（坐姿姿态）	渲染轮椅合成图像 + 关键点系统扩展（包括轮椅接触点等）	研究针对轮椅人群扩展骨架定义，结合合成图像训练；改进坐姿姿态估计准确率，使用 ViTPose++ 的任务特定模块实现迁移与适配（具体结果待正式发表）。

注：上述表格涵盖了从早期基于 CNN 的领域迁移方法（如 FiDIP、SHIFT）到近年 Transformer 架构基础模型（如 Sapiens、ViTPose++）的演进。Transformer 模型原生支持跨任务迁移与关键点系统扩展，而 CNN 方法则开创了领域不变表示、合成数据引导等关键策略。
两者并非对立，**基础模型的通用性与任务特定策略相结合**，才构成了当前最先进的适应方法体系。

结论

将人体姿态估计器适配至非常规人群是一项正在快速发展的研究方向，融合了**计算机视觉与生物力学/临床需求**。近年的文献共识是：虽然大规模预训练模型提供了良好的起点，但要在婴儿、老年人等特定人群上达到最佳精度，仍需引入**专用技术手段**。领域自适应方法——包括少量标注数据上的微调、使用合成数据、或引入解剖学先验——都能在这些人群中显著提升性能。

跨年龄群体的泛化能力可通过在多样化数据上训练、并采用架构灵活的模型（能够适配不同身体尺寸与关键点定义）来增强。值得注意的是，随着**Transformer 架构基础模型**（如 ViTPose++、Sapiens 等）的发展，知识迁移变得更加容易：这类模型因其结构的灵活性与强大的表达能力，能更好地支持多任务、多人群的姿态建模。

同时，针对特定挑战（如极端姿态或遮挡问题）的定制解决方案依然不可或缺，如：

- 增加合成数据；
- 引入姿态先验；
- 使用遮挡一致性机制；
- 添加关键点扩展模块。

这些策略结合基础模型能力，共同推动了对以往“超出适用范围”的人群的姿态识别性能。

简而言之，要将人体姿态估计器微调至婴儿、老年人或其他**代表性不足的人群**，最有效的路径包括：

1. 以强大的预训练模型为基础（如已在多样人体图像上训练的模型）；
2. 引入适当的领域适应方法（有监督或无监督），以应对外观与姿态分布的偏移；
3. 根据需要调整输出关键点结构或解码器，适应新的骨架定义；
4. 注入先验知识或合成样本，引导模型更准确地学习新领域特征。

遵循上述原则，近年来的研究已在婴幼儿等传统模型“力不能及”的人群中实现了显著性能提升，这也使我们更接近实现真正**包容性的人体姿态估计系统**——适用于任何年龄、体型与身体状态。