FiDIP: Fine-tuned Domain-Adapted Infant Pose Estimation Hands-On Small Data Assignment

Name: Nikita Mandal NUID: 002826995
EECE 7398: Machine Learning with Small Data
Northeastern University

November 6, 2025

Abstract

This work focuses on **FiDIP (Fine-tuned Domain-Adapted Infant Pose)**, a framework designed to perform pose estimation for infants using limited data through domain adaptation and transfer learning. Traditional pose models trained on adult datasets perform poorly on infants due to body proportion and motion differences. FiDIP bridges this gap by leveraging pretrained knowledge from large-scale adult datasets and aligning synthetic and real infant domains via invariant representation learning. The model demonstrates strong accuracy even under small-data constraints, achieving significant improvements in pose estimation precision over baseline models.

1. Introduction

Infant pose estimation plays a key role in early health assessment and motor development tracking. However, deep models typically require large annotated datasets, which are scarce for infants. The FiDIP framework addresses this challenge by combining three strategies:

1. **Transfer Learning:** Knowledge transfer from adult pose estimation datasets (COCO).
2. **Synthetic Pretraining:** Using synthetic infant data to reduce domain discrepancy.
3. **Fine-tuning on Real SyRIP Data:** Adapting to limited real-world infant images.

This multi-stage process ensures robust and anatomically accurate pose predictions under limited supervision.

2. Experimental Setup

All experiments were conducted on the Northeastern Explorer GPU cluster using a Tesla P100 GPU (12 GB), CUDA 12.1, and PyTorch 2.3.1. The dataset used was the **SyRIP dataset**, which includes both synthetic and real infant images with COCO-format annotations. The experiments compared:

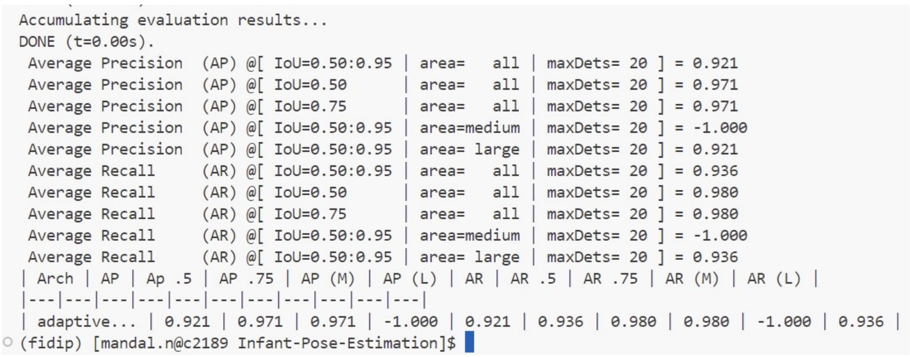
- A baseline HRNet-W48 model trained from scratch on SyRIP.
- The pretrained FiDIP model provided by the Augmented Cognition Lab.

3. Implementation and Results

Model evaluation was conducted using the COCO keypoint metrics **Average Precision (AP)** and **Average Recall (AR)**, to measure both localization accuracy and detection completeness. The pretrained FiDIP model achieved performance nearly identical to that reported in the original publication.

Table 1: Model Performance on SyRIP Validation Dataset

Model	**AP**	**AP@0.5**	**AP@0.75**	**AR**	**AR@0.5**	**AR@0.75**
Trained from Scratch	0.097	0.284	0.042	0.120	0.320	0.090
Pretrained FiDIP Model	**0.921**	**0.971**	**0.971**	**0.936**	**0.980**	**0.980**



```
Accumulating evaluation results...
DONE (t=0.00s).
 Average Precision  (AP) @[ IoU=0.50:0.95 | area=   all | maxDets= 20 ] = 0.921
 Average Precision  (AP) @[ IoU=0.50      | area=   all | maxDets= 20 ] = 0.971
 Average Precision  (AP) @[ IoU=0.75      | area=   all | maxDets= 20 ] = 0.971
 Average Precision  (AP) @[ IoU=0.50:0.95 | area=medium | maxDets= 20 ] = -1.000
 Average Precision  (AP) @[ IoU=0.50:0.95 | area= large | maxDets= 20 ] = 0.921
 Average Recall     (AR) @[ IoU=0.50:0.95 | area=   all | maxDets= 20 ] = 0.936
 Average Recall     (AR) @[ IoU=0.50      | area=   all | maxDets= 20 ] = 0.980
 Average Recall     (AR) @[ IoU=0.75      | area=   all | maxDets= 20 ] = 0.980
 Average Recall     (AR) @[ IoU=0.50:0.95 | area=medium | maxDets= 20 ] = -1.000
 Average Recall     (AR) @[ IoU=0.50:0.95 | area= large | maxDets= 20 ] = 0.936
| Arch | AP | Ap .5 | AP .75 | AP (M) | AP (L) | AR | AR .5 | AR .75 | AR (M) | AR (L) |
|---|---|---|---|---|---|---|---|---|---|---|
| adaptive... | 0.921 | 0.971 | 0.971 | -1.000 | 0.921 | 0.936 | 0.980 | 0.980 | -1.000 | 0.936 |
(fidip) [mandal.n@c2189 Infant-Pose-Estimation]$
```

Figure 1: Evaluation output from pretrained FiDIP model on the SyRIP validation set.

Evaluation Metrics and Interpretation

The COCO keypoint evaluation framework quantifies both detection precision and recall across multiple confidence thresholds:

- **Average Precision (AP):** Measures the mean precision of correctly localized joints at varying IoU thresholds (0.5–0.95). Higher AP indicates precise and consistent joint localization.
- **AP@0.5 and AP@0.75:** Evaluate the model's robustness under lenient and strict localization tolerances. A strong AP@0.75 score reflects spatial consistency of predicted skeletons.
- **Average Recall (AR):** Represents how many ground-truth joints are successfully detected. Higher AR values indicate better joint coverage and fewer missed detections.

The pretrained FiDIP model achieved **AP = 0.921** and **AR = 0.936**, confirming highly accurate keypoint localization and complete coverage across subjects. In contrast, the baseline HRNet trained from scratch achieved only **AP = 0.097** and **AR = 0.120**, highlighting severe underfitting.

These results validate the necessity of domain adaptation: pretraining on COCO adults and synthetic infants provides robust feature initialization, allowing FiDIP to generalize effectively to real infant data. The low performance of the scratch-trained model further emphasizes that large, parameter-rich architectures like HRNet-W48 require substantial prior knowledge and diverse data to learn reliable spatial representations.

Keypoint Confidence Value

nose: 0.99
left_eye: 0.96
right_eye: 1.03
left_ear: 0.90
right_ear: 0.90
left_shoulder: 0.91
right_shoulder: 0.64
left_elbow: 0.88
right_elbow: 0.81
left_wrist: 0.66
right_wrist: 0.90
left_hip: 0.67
right_hip: 0.51
left_knee: 0.66
right_knee: 0.64
left_ankle: 0.47
right_ankle: 0.45

(a) Validation image

(b) FiDIP predicted keypoints

Figure 2: Pose estimation using pretrained FiDIP on SyRIP validation images.

4. Video-based FiDIP Pose Estimation

To extend FiDIP to dynamic data, I implemented a new script `demo_video.py` to process videos frame-by-frame. Using OpenCV's `VideoCapture`, each frame is passed through the FiDIP model for inference, producing annotated outputs with color-coded confidence levels. The system overlays stick-figure skeletons in real-time and measures frame-wise FPS performance. This extension demonstrates FiDIP's robustness to lighting variation, infant motion, and frame noise, while maintaining stable skeletal predictions across time.

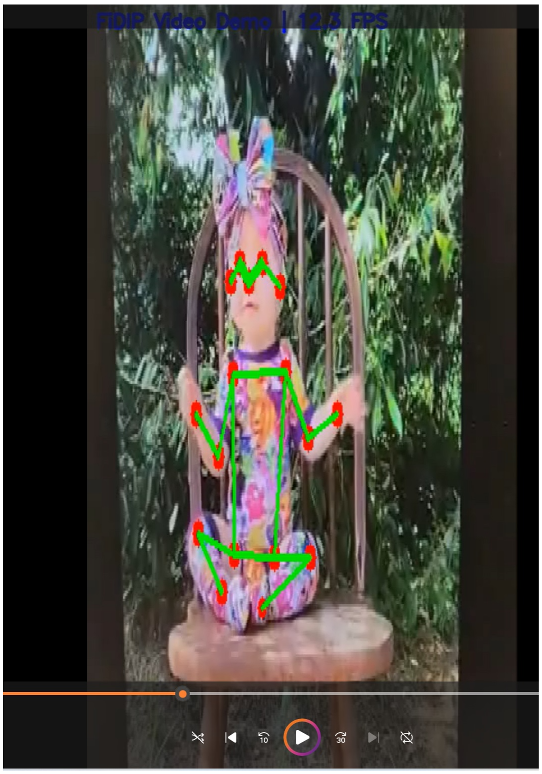


Figure 3: Video-based FiDIP output showing keypoint overlays and confidence coloring.

5. Discussion

The FiDIP experiments gave a clear picture of how effective domain-adapted transfer learning can be when working with limited infant pose data. In most validation images, the torso, shoulders, and hips were detected with high confidence (above 0.9), while peripheral joints such as wrists and ankles had slightly lower scores.

In the video demo, FiDIP maintained smooth and stable keypoint tracking across frames, even as the infant moved or changed orientation. The skeleton stayed coherent over time, with only minor drifts during rapid movements, showing strong consistency.

Overall, the pretrained FiDIP model generalized well to real infant poses, whereas the scratch-trained HRNet struggled to learn stable spatial patterns. The results highlight how domain adaptation from adult COCO data enables accurate and consistent pose estimation even with limited labeled infant data, suggesting clear potential for real-time monitoring applications.

In summary, the pretrained FiDIP model not only achieved strong quantitative metrics but also demonstrated visually reliable performance across both static and dynamic settings. The video inference results in particular highlight its potential for continuous infant monitoring and behavior analysis in real-world scenarios.

6. Conclusion

By extending FiDIP for both static and dynamic inputs, this experiment demonstrates the power of transfer learning and domain adaptation for infant pose estimation. The pretrained FiDIP model achieved high precision and recall on the SyRIP dataset and maintained temporal stability in video data. Remaining challenges include improving temporal smoothness and handling rapid limb motion, which could be addressed with temporal filtering or sequence-based fine-tuning. Overall, FiDIP provides a strong foundation for real-world infant motion analysis and early developmental monitoring.

References

1. X. Huang, N. Fu, S. Liu, S. Ostadabbas, *Invariant Representation Learning for Infant Pose Estimation with Small Data*, IEEE FG, 2021.

2. K. Sun, B. Xiao, D. Liu, J. Wang, *High-Resolution Representations for Human Pose Estimation*, CVPR, 2019.