# Supplementary Material for Visibility guided Self-Supervised Occlusion-Resilient Human Pose Estimation

Arindam Dutta[1] Sarosij Bose[1] Rohit Kundu[1] Calvin-Khang Ta[1,2,†] Saketh Bachu[1,3,†]
Konstantinos Karydis[1] Amit K. Roy-Chowdhury[1]

[1]University of California, Riverside, CA [2]Dolby Laboratories, USA [3]EvenUp, San Francisco, USA

{adutt020, sbose007, rkund006, cta003, sbach008, karydis, amitrc}@ucr.edu

## 1. Implementation Details

For our pose estimation model, we utilize the ResNet101 [1] pretrained on the ImageNet dataset in combination with the Simple Baseline decoder [2] as the feature extractor $\mathcal{P}$, following [3, 4]. Augmentations such as rotation, translation, and shear as specified in [4] were used. The hyperparameter $\tau$ in Eqn. 3 and Eqn. 5, with set to 0.5, based on UniFrame [4]. For the loss function in Eqn. 11, we used $\lambda_a = 1 \times 10^{-5}$ and $\lambda_v = 1$ in all experiments. Consistent with [5], the human pose prior model $\mathcal{G}$ is implemented as a 7-layer MLP network, comprising two encoding layers and five decoding layers. All models were trained with a batch size of 32 and 500 iterations per epoch for a total of 70 epochs with Adam optimizer [6] with an initial learning rate of 1e-04, decaying by a factor of 0.1 after the $45^{th}$ and $60^{th}$ epochs. PyTorch was used as the coding framework, and all experiments were performed on a single 24 GB NVIDIA RTX 3090.

## 2. Datasets

We utilize the **Surreal** dataset [7] as the source dataset for all of our experiments due to its extensive collection of more than 6 million synthetic images that depict artificial humans in an indoor environment. Our primary target datasets are **3DOH50K** [8], which includes images of humans in various occluded scenarios within an indoor setting and **BOW** with 4000 images, equally split between images from [9] and [10], covering both indoor and outdoor scenes. Following [11–13], we also create synthetically occluded datasets using **Humans3.6m (H36M)** [14] and **Leeds Sports Pose (LSP)** [9], by adding occlusion objects from the Pascal VOC dataset [15]. This allows us to have occlusion-free ground-truth poses for quantitative evaluation. We refer to these artificially occluded datasets as **Ocl-H36M** and **Ocl-LSP**, respectively. The adaptation and evaluation splits for both **Ocl-H36M** and **Ocl-LSP** are identical to those used for **H36M** and **LSP** in existing works [3, 4, 16].

## 3. Baselines and Metrics

We evaluate **VisOR** against recent state-of-the-art algorithms for human pose estimation [3, 4, 16]. We also report the performance of the *Source-only* model, which serves as a lower bound. This refers to the model's performance on the unlabeled target data when trained exclusively on labeled source data. We report PCK@0.05 for twelve joints: left and right shoulders (Sld.), elbows (Elb.), wrists, hips, knees, and ankles, as well as their average. PCK@0.05 measures the percentage of correct keypoint predictions within 5% of the image size.

## 4. Additional Qualitative Results

Additional qualitative results on the 3DOH50K and **BOW** benchmarks are presented in Figures 1 and 2. Figure 3 shows qualitative results on Ocl-H36M and Ocl-LSP datasets. Figure 4 shows performance of **VisOR** in occlusion free settings on LSP dataset.

## 5. Additional Quantitative Results

**Recent Baselines and Datasets:** Table 1 reports quantitative results comparing **VisOR** to ViTPose [17], where **VisOR** achieves superior performance for the SURREAL $\rightarrow$ Ocl-LSP, thereby showcasing the effectiveness of **VisOR** for pose estimation under occlusions.

Table 1. ViTPose vs **VisOR** on Ocl-LSP.

| **Algorithm** | Sld. | Elb. | Wrist | Hip | Knee | Ankle | **Avg.** |
|---|---|---|---|---|---|---|---|
| VitPose | 49.1 | 66.7 | 59.2 | 77.8 | 66.5 | 67.4 | 64.5 |
| **VisOR** | 59.5 | 75.7 | 67.8 | 74.7 | 65.2 | 70.2 | 68.9 |

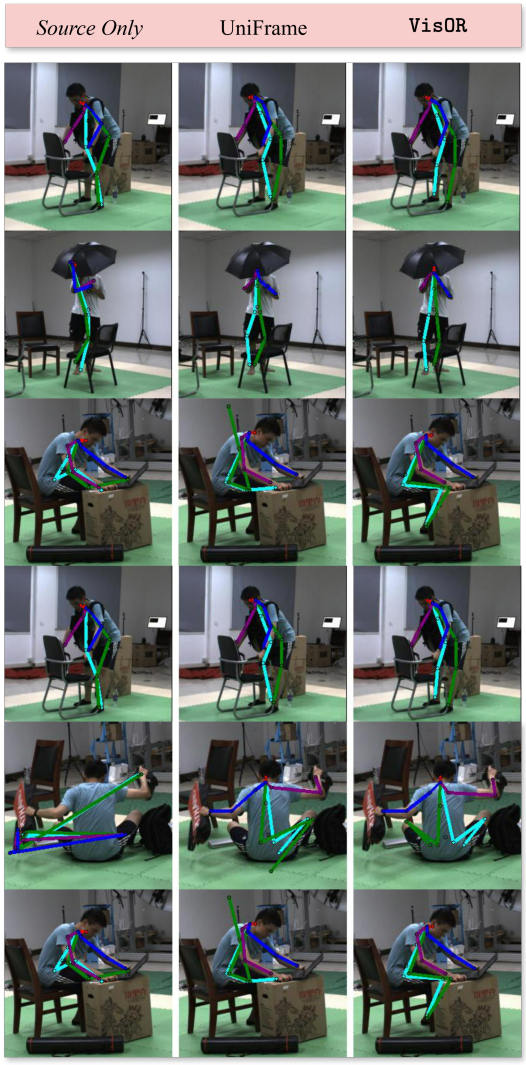


Figure 1. **Qualitative Results for Surreal→ 3DOH50K**

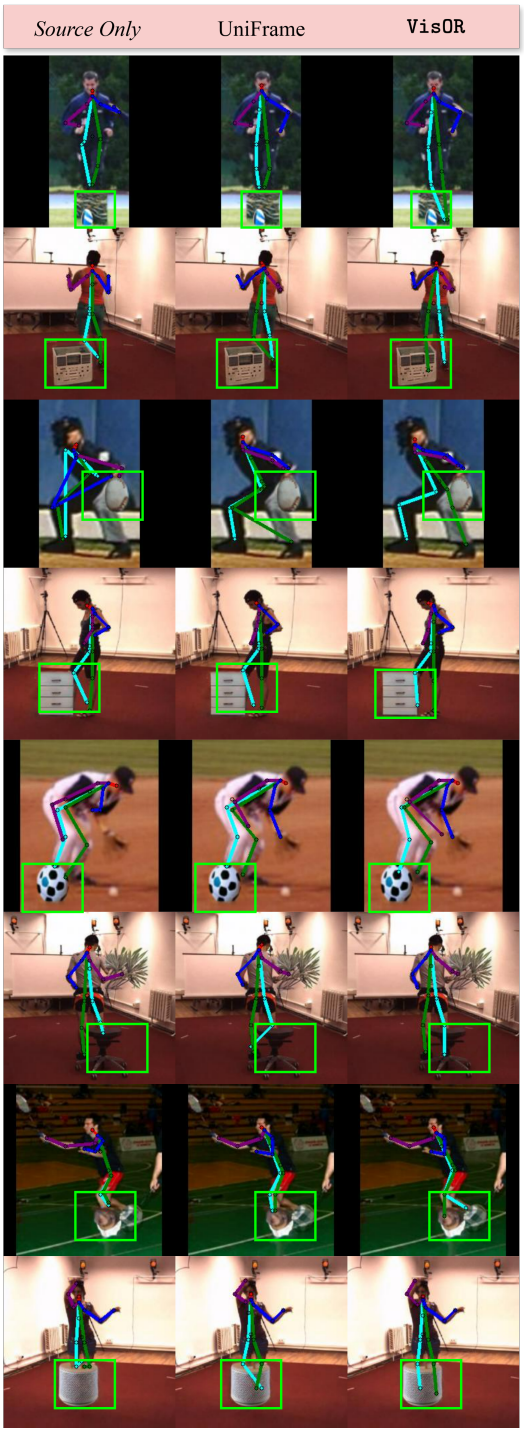


Figure 2. **Qualitative Results for Surreal→ BOW**

**Different Occlusion Types:** We compare VisOR and UniFrame [4] at five increasing severity levels, where the size of the occlusions increases with severity, as shown in 2. At severity level 1, occlusions are approximately 48x48 pixels in 256x256 images, while at severity level 5, they increase to around 96x96 pixels. Our results indicate that as the size of the occlusion increases, the performance of VisOR declines, but the degradation in the performance of UniFrame is significantly more pronounced. This shows that VisOR remains more robust than UniFrame even under larger occlusions. Consequently, it can be concluded that VisOR is substantially more reliable than state-of-the-art algorithms when dealing with occlusions of varying types, shapes, and sizes.

We also evaluated VisOR with simple, non-contextual occlusions by overlaying black patches of varying sizes (min: 32x32, max: 96x96) for SURREAL → LSP scenario, clearly VisOR significantly outperforms UniFrame, as shown in Table 3.

**Results on Occlusion Free Target Data:** In Table 4, our evaluation of the SURREAL → LSP benchmark—where

| Algorithm | Sev-1 | Sev-2 | Sev-3 | Sev-4 | Sev-5 |
|---|---|---|---|---|---|
| UniFrame [4] | 70.6 | 69.0 | 64.3 | 64.0 | 56.9 |
| VisOR | **73.4** | **71.5** | **68.9** | **68.7** | **64.9** |

Table 2. PCK@0.05 across different severities for **Surreal → Ocl-LSP**. Best results in **bold**.

Table 3. SURREAL → LSP w/ black patch occlusions (cut-out).

| **Algorithm** | Sld. | Elb. | Wrist | Hip | Knee | Ankle | **Avg.** |
|---|---|---|---|---|---|---|---|
| UniFrame | 44.0 | 59.2 | 55.3 | 74.4 | 68.2 | 63.0 | 60.7 |
| VisOR | 51.5 | 67.9 | 58.7 | 80.0 | 73.4 | 67.4 | 66.5 |

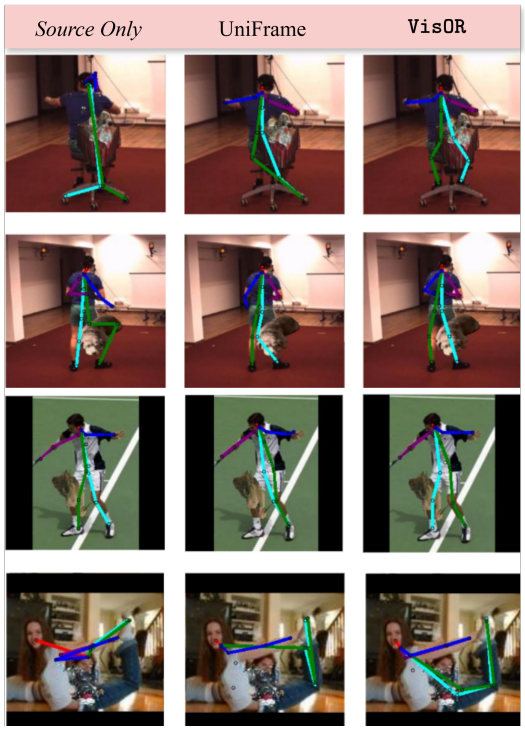


Figure 3. **Qualitative Results for Surreal→ Ocl-H36M and Surreal→ Ocl-LSP.**

the target data set contains minimal occlusions—shows that VisORachieves performance comparable to the state-of-the-art UniFrame[4], while outperforming RegDA [3] by approximately 6%. This shows that VisOR not only excels in occlusion-heavy scenarios, but also retains competitive accuracy in clean, unoccluded settings. Although EPIC[16] achieves slightly higher overall accuracy in this particular benchmark, it is based on adversarial learning, which is known to suffer from training instability and poor generalization under occlusion. In particular, EPIC performs significantly worse when evaluated in occlusion-rich environments, making it less suitable for real-world deployment where occlusions are prevalent.

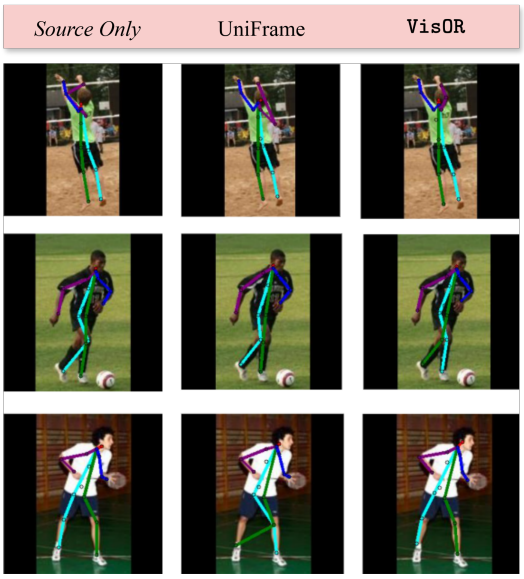


Figure 4. **Qualitative Results for Surreal→ LSP**

| **Algorithm** | **Sld.** | **Elb.** | **Wrist** | **Hip** | **Knee** | **Ankle** | **Avg.** |
|---|---|---|---|---|---|---|---|
| *Source only* | 51.5 | 65.0 | 62.9 | 68.0 | 68.7 | 67.4 | 63.9 |
| EPIC [16] | 72.1 | 86.4 | 85.2 | 87.7 | 87.0 | 86.2 | 84.8 |
| RegDA [3] | 62.7 | 76.7 | 71.1 | 81.0 | 80.3 | 75.3 | 74.6 |
| UniFrame [4] | 69.2 | 84.9 | 83.3 | 85.5 | 84.7 | 84.3 | 82.0 |
| VisOR | 68.1 | 82.9 | 80.5 | 86.9 | 85.6 | 83.5 | 81.2 |

Table 4. PCK@0.05 for **SURREAL → LSP**.

## References


[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[2] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 1

[3] Junguang Jiang, Yifei Ji, Ximei Wang, Yufeng Liu, Jianmin Wang, and Mingsheng Long. Regressive domain adaptation for unsupervised keypoint detection. In *CVPR*, 2021. 1, 3

[4] Donghyun Kim, Kaihong Wang, Kate Saenko, Margrit Betke, and Stan Sclaroff. A unified framework for domain adaptive pose estimation. In *ECCV*, 2022. 1, 2, 3

[5] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *ECCV*, 2022. 1

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[7] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2017. 1

[8] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single

color image. In *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2020. 1

[9] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, volume 2, page 5. Aberystwyth, UK, 2010. 1

[10] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2013. 1

[11] Qiang Zhou, Shiyin Wang, Yitong Wang, Zilong Huang, and Xinggang Wang. Human de-occlusion: Invisible perception and recovery for humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3691–3701, 2021. 1

[12] Arindam Dutta, Rohit Lal, Dripta S Raychaudhuri, Calvin-Khang Ta, and Amit K Roy-Chowdhury. Poise: Pose guided human silhouette extraction under occlusions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6153–6163, 2024.

[13] Rajat Modi, Vibhav Vineet, and Yogesh Rawat. On occlusions in video action detection: Benchmark datasets and training recipes. *Advances in Neural Information Processing Systems*, 36:57306–57335, 2023. 1

[14] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 1

[15] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 1

[16] Qucheng Peng, Ce Zheng, Zhengming Ding, Pu Wang, and Chen Chen. Exploiting aggregation and segregation of representations for domain adaptive human pose estimation. *arXiv preprint arXiv:2412.20538*, 2024. 1, 3

[17] Yufei Xu, Jing Zhang, Qiming Zhang, Wei Liu, Nick Barnes, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *arXiv*, abs/2204.12484, 2022. 1