Supplementary for POISE: Pose Guided Human Silhouette Extraction under Occlusions

1. Generating Occluded Images

We employ Random Erase occlusion to introduce occlusions of varying severities K into the images. The severity of the occlusion directly affects the size of the occlusion patch, controlling its height and width. Specifically, we consider five different occlusion severities: 12, 16, 20, 24, and 28. To clarify, an occlusion severity of K implies that the height and width of the occlusion patch are randomly selected within the range of $2K$ to $2(K+8)$. For instance, for an occlusion severity of 12, the height and width of the occlusion patch are randomly chosen between 24 and 40 pixels. For the BRIAR dataset, we add COCO occlusions to the frames which helps in creating a more balanced training setup. Figure 1 shows a few examples of occluded images generated from the CASIA-B and the UP-S31 datasets.

Figure 1. **Examples of generated occluded images.** *Left:* Occluded CASIA-B image with Random Erase Occlusion (at K = 12). *Middle:* Occluded UP-S31 image with Random Erase Occlusion (at K = 12). *Right:* Occluded UP-S31 image with COCO Occlusion.

.

2. Implementation Details

The networks are trained for a total of 100 epochs and with a batch-size of 32 and results are reported at the 100^{th} epoch. The initial learning rate is set $1e-2$, which is decayed by a factor of 0.1 after the 5^{th} and 20^{th} epochs. For all experiments, λ_1 and λ_2 in the training objective were set to 0.1 and 1 respectively. λ_3 was also set to 1 for all experiments except human segmentation of UP-S31 dataset with COCO occlusions, where it was set to 0. The models are optimized with the Adam optimizer [5].

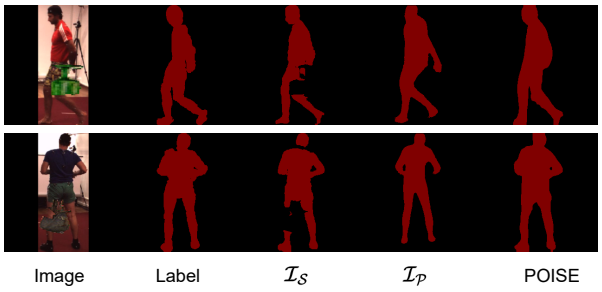

Figure 2. **Qualitative results on BRIAR.** Silhouettes extracted using POISE compared against $\mathcal{I}_{\mathcal{S}}$ and $\mathcal{I}_{\mathcal{P}}$ for COCO occlusions in the Humans3.6M dataset.

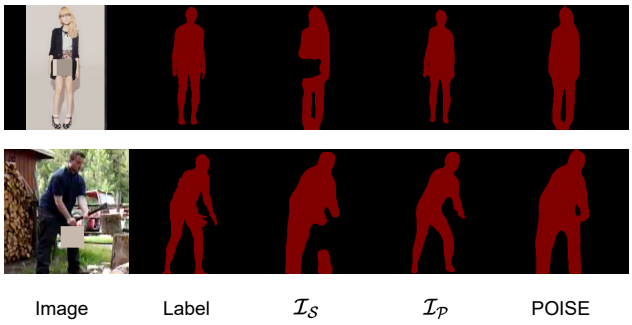

Figure 3. **Qualitative results on Random Erase occlusions.** Silhouettes extracted using POISE compared against $\mathcal{I}_{\mathcal{S}}$ and $\mathcal{I}_{\mathcal{P}}$ for Random Erase occlusions at severities of 12 and 20 on UP-S31.

3. Qualitative Results

In this section we provide additional qualitative results on Humans3.6M [4] , UP-s31 [7], CASIA-B [9] and BRIAR [2] datasets.

Figure 2 shows the efficacy of POISE over $\mathcal{I}_{\mathcal{S}}$ and $\mathcal{I}_{\mathcal{P}}$ on the Humans3.6M dataset under COCO occlusions. Similarly, figures 3 and 4 show the improvements obtained by POISE over $\mathcal{I}_{\mathcal{S}}$ and $\mathcal{I}_{\mathcal{P}}$ on UP-s31 dataset under Random erase and COCO occlusions respectively. Figure 5 shows the efficacy of POISE on the CASIA-B dataset. Figure 6 shows the efficacy of POISE in natural unconstrained settings in the BRIAR dataset, where despite atmospheric turbulence and heavy natural occlusions (due to inanimate objects and

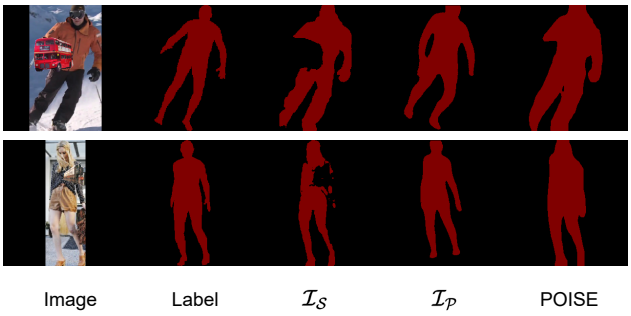

Figure 4. **Qualitative results on COCO occlusions.** Silhouettes extracted using POISE compared against $\mathcal{I}_S$ and $\mathcal{I}_P$ for COCO occlusions on UP-S31.

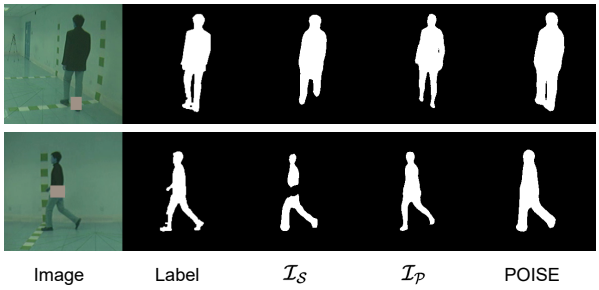

Figure 5. **Qualitative results on CASIA-B.** Silhouettes extracted using POISE compared against $\mathcal{I}_S$ and $\mathcal{I}_P$ for Random Erase occlusions at severities of 12 on CASIA-B dataset.

ground vegetation) it performs optimally.

These qualitative results show the inherent ability of POISE in learning robust feature representation which aids in obtaining complete human silhouettes under occlusions.

4. Experiments using SOTA instance segmentation models

In this section, we provide a study on recent state-of-the-art segmentation models which despite being trained on larger and more robust datasets fail to handle occlusions in practical real-world settings. We study two such recent models MaskDino [8] and Segment Anything (SAM) [6]. In Fig 7, we demonstrate the inadequacy of state-of-the-art (SOTA) methods in addressing occlusion due to their reliance on pixel-wise classification. When pixels are part of an occlusion in front of a human, these models fail to identify them, rendering them impractical. Moreover, our findings indicate that POISE enhances SOTA results, surpassing MaskDINO by approximately 5% (from $\approx$ 80 % to 85 %) in terms of segmentation accuracy on the Humans3.6M dataset.

5. POISE for Gait Recognition

In tables 3, 4 and 5 of the main draft, we observe that POISE outperforms both $\mathcal{I}_S$ and $\mathcal{I}_P$ by significant margins.

Figure 6. **Qualitative results on BRIAR.** Silhouettes extracted using POISE compared against $\mathcal{I}_S$ and $\mathcal{I}_P$ for natural occlusions in the BRIAR dataset.

Figure 7. Comparison of SOTA segmentation methods like MaskDINO and SAM for extracting human silhouette under occlusion. We can see that SOTA methods fail to capture body parts that are behind the occlusion.

The main reason behind such results is that POISE is able to have a complete human silhouette under occlusions - which preserves both body-shape specific information and walking pattern. While $\mathcal{I}_{\mathcal{S}}$ retains body-shape, it might fail to preserve the walking pattern for persistent lower-body occlusions in a video. Similarly, utilization of Pose2sil (g) in isolation does not yield optimal outcomes as it lacks the inherent human body shape, despite preserving the walking pattern.

When the BRIAR dataset [2] is juxtaposed to CASIA-B dataset [9], we observe that, several videos suffer from heavy occlusions from inanimate objects (such as traffic cones) and ground vegetation, wherein the entire lower body of the human is hidden behind the occlusion, as shown in figure 6. In such situations, the walking pattern information is absent in $\mathcal{I}_{\mathcal{S}}$, which is reasonably preserved in $\mathcal{I}_{\mathcal{P}}$. This explains the superior performance of $\mathcal{I}_{\mathcal{P}}$ over $\mathcal{I}_{\mathcal{S}}$ on the BRIAR dataset.

For the CASIA-B dataset [9], small artificial occlusions only affect a limited region of the image as compared to the large occlusions in the BRIAR dataset that conceal almost the entire lower half of the human body. Due to these small occlusions, $\mathcal{I}_{\mathcal{S}}$ is fairly close to the complete human silhouette seen in figure 5. This explains why performance of $\mathcal{I}_{\mathcal{S}}$ is superior to $\mathcal{I}_{\mathcal{P}}$ for CASIA-B dataset.

It is important to highlight that our fusion of $\mathcal{I}_{\mathcal{S}}$ and $\mathcal{I}_{\mathcal{P}}$ in POISE , effectively retains both body-shape specific information and the distinctive walking style. This greatly contributes to the superior performance of POISE in comparison to $\mathcal{I}_{\mathcal{S}}$ and $\mathcal{I}_{\mathcal{P}}$ across both datasets. This outcome underscores the robustness and generalization ability of POISE .

5.1. Effects of enhanced silhouettes

In an ideal scenario, an improved silhouette should directly lead to enhanced gait performance using the same feature extractor weights. To test this hypothesis, we utilize the BRIAR dataset with $\mathcal{I}_{\mathcal{S}}$, $\mathcal{I}_{\mathcal{P}}$ and POISE silhouettes, with the model [3] pre-trained on CASIA-B [9]. Table 1 shows that simply using POISE silhouettes without any further training of the gait recognition model can lead to much higher gait recognition performance as compared against $\mathcal{I}_{\mathcal{S}}$ and $\mathcal{I}_{\mathcal{P}}$, thus underscoring the effectiveness of POISE .

Gait Algo.	Method	Acc@Top1	Acc@Top2	Acc@Top3
GaitBase [3]	$\mathcal{I}_{\mathcal{S}}$	5.95	13.00	18.93
	$\mathcal{I}_{\mathcal{P}}$	15.02	21.61	24.90
	POISE	**15.84**	**21.81**	**26.34**
Gaitset [1]	$\mathcal{I}_{\mathcal{S}}$	8.56	14.77	18.56
	$\mathcal{I}_{\mathcal{P}}$	13.15	20.57	26.13
	POISE	**13.58**	**22.43**	**27.16**

Table 1. Results for Gait recognition on BRIAR dataset with fixed feature extractor.

References

[1] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8126–8133, 2019. 3

[2] David Cornett, Joel Brogan, Nell Barber, Deniz Aykac, Seth Baird, Nicholas Burchfield, Carl Dukes, Andrew Duncan, Regina Ferrell, Jim Goddard, et al. Expanding accurate person recognition to new altitudes and ranges: The briar dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 593–602, 2023. 1, 3

[3] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9707–9716, June 2023. 3

[4] 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

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

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 2

[7] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[8] Feng Li, Hao Zhang, Huaizhe xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation, 2022. 2

[9] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th international conference on pattern recognition (ICPR'06)*, volume 4, pages 441–444. IEEE, 2006. 1, 3