Supplementary Material for CHROME: Clothed Human Reconstruction with Occlusion-Resilience and Multiview-Consistency from a Single Image

Arindam Dutta[1†] Meng Zheng[2] Zhongpai Gao[2] Benjamin Planche[2] Anwesa Choudhuri[2]
Terrence Chen[2] Amit K. Roy-Chowdhury[1] Ziyan Wu[2]
[1]University of California, Riverside [2]United Imaging Intelligence, Boston
{adutt020@, amitrc@ece.}ucr.edu, {first.last}@uii-ai.com

This supplementary material provides additional details and analyses of CHROME to complement the main paper. It begins with a discussion of implementation details, including training strategies, hyperparameters, and inference settings. Extended quantitative and qualitative results are presented to evaluate CHROME on scenarios such as stereo reconstruction and occlusion-resilient novel view synthesis and geometric reconstruction from single-view images. Also, we analyze the impact of pose estimation on multiview consistency and demonstrate how CHROME maintains robustness even with inaccurate pose inputs. We also provide comparative evaluations against existing large reconstruction models, showcasing the superior performance and generalizability of CHROME. Finally, an inference time analysis highlights its efficiency, making it suitable for real-world applications. Together, these results reinforce the robustness and versatility of CHROME across diverse datasets.

1. Implementation Details

For all our experiments, we use the PyTorch coding environment with all models being trained on 4× A40 GPUs.

We train $\mathcal{F}_{\mathcal{D}}$ for 100 epochs across all experiments. The training utilizes the AdamW optimizer with a cosine annealing learning rate schedule that peaks at 7×10^{-5} within 1000 warm-up steps, after which we use a constant learning rate of 5×10^{-6}. During inference, we employ classifier-free guidance with a guidance scale of 4 and 40 diffusion steps, utilizing an ancestral Euler sampling strategy.

For training $\mathcal{F}_{\mathcal{R}}$, we use a learning rate initialized at 4×10^{-4}, which decays following the Cosine Annealing strategy over 100 epochs. In Equation 4 (see main paper), λ_1 is set to 1.5 and λ_2 is set to 1.

† This work was done during Arindam Dutta's internship at United Imaging Intelligence, Boston, MA.

Table 1. Quantitative comparison for novel view texture reconstruction on regular THuman2.0 [1].

Algorithm	SMPL	PSNR ↑	SSIM ↑	LPIPS ↓
SiTH [2]	✓	17.12	0.843	0.155
GTA [3]	✓	18.05	-	-
SIFU [4]	✓	**22.10**	**.9230**	**.0790**
HSGD [5]	✗	17.37	.8950	.1300
PIFu [6]	✗	18.09	.9110	.1370
LGM [7]	✗	20.01	.8930	.1160
M123 [8]	✗	14.50	.8740	.1450
CHROME	✗	**20.80**	**.9114**	**.0878**

2. How Robust are Existing Large Reconstruction Models for Occlusion-Free Novel View Synthesis?

We showcase qualitative comparisons with three baseline algorithms: Zero123++ [9], ImageDream with LGM [7], and SV3D [10], utilizing their pretrained weights for the task of occlusion-free novel view reconstruction in Figure 2. The results reveal significant inconsistencies in these existing methods when applied to our task, underscoring the necessity of a specialized algorithm, CHROME.

3. Additional Quantitative Results

Clean THuman2.0: In Table 1, we evaluate the novel-view reconstruction capabilities of CHROME under standard occlusion-free conditions. The results show that CHROME outperforms all existing methods [3, 5–7, 11], except SIFU [4], for novel view synthesis across 16 views. However, it is important to note that SIFU relies on utilizing SMPL priors and 3D supervision, which may not be available in real-world scenarios. Furthermore, SIFU performs per-subject optimization, taking ≈ 6 minutes of computation time per image, while CHROME achieves

Figure 1. **Visual Results of CHROME on occluded THuman2.0 and CustomHumans:** Qualitative results of CHROME on occluded THuman2.0 and occluded CustomHumans.

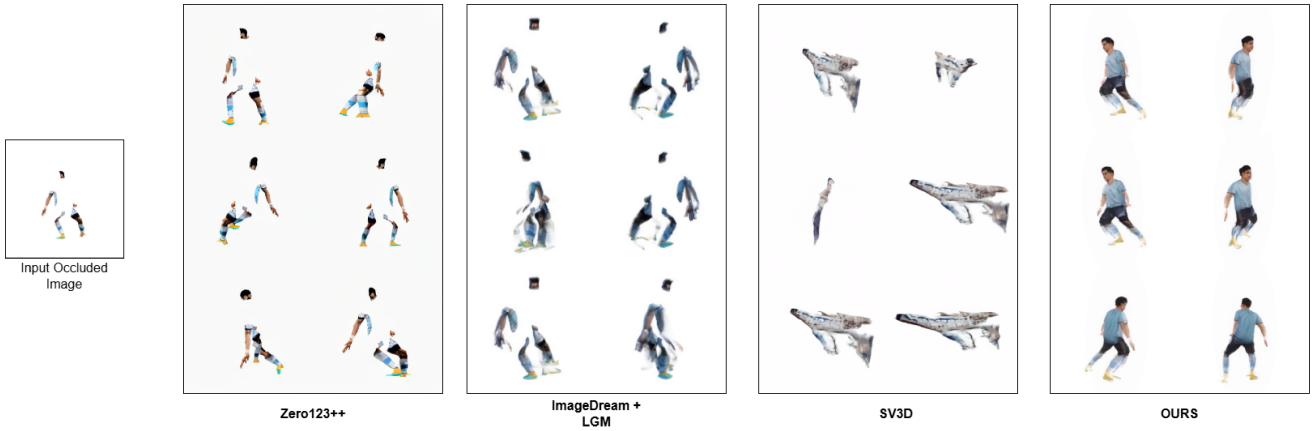

Figure 2. Comparative Qualitative Analysis of Existing Large Reconstruction Models for Zero-Shot Novel View Texture Reconstruction from Occluded Single View Images.

comparable results in ≈ 15 seconds.

Table 2. Quantitative comparison for zero-shot novel view texture reconstruction on Occluded CAPE.

Algorithm	SMPL	PSNR ↑	SSIM ↑	LPIPS ↓
PIFu [6]	✗	14.77	.8779	.1353
GTA [3]	✓	13.90	.8955	.1274
SIFU [4]	✓	13.93	.8939	.1273
SiTH [2]	✓	13.28	.8782	.1527
CHROME	✗	**18.54**	**.9130**	**.0850**

Occluded CAPE: In Table 2, we present quantitative results against baseline algorithms for novel view synthesis on occluded CAPE wherein CHROME successfully outperforms all baseline algorithms.

Methods	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**
SD-XL+SIFU [4]	16.27	.8649	.1525
CHROME	**20.54**	**.9098**	**.0893**

Table 3. Novel View Synthesis (NVS) using Inpainting for Deocclusion on Occluded THuman2.0.

Occl.	**SIFU**	**CHROME**
25%	15.39/.877/.110	19.51/.909/.090
50%	14.62/.882/.115	19.27/.907/.092
75%	14.04/.880/.123	19.06/.904/.094

Table 4. Sensitivity to Occlusion Sizes.

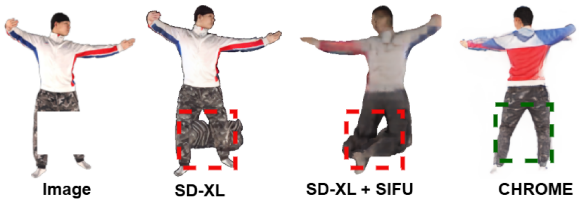

Figure 3. SD-XL + SIFU vs CHROME (zoom in on limbs).

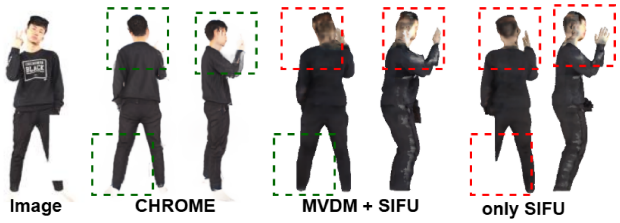

Figure 4. MVDM + SIFU vs CHROME.

Inpainting before Reconstruction: While a potential solution to our problem could be using a foundational inpainting model such as Stable Diffusion Inpainting, we observe that this model is hardly able to retain the texture or anatomy of the human leading to far from accurate 3D reconstruction, as quantitatively shown in Table 3 and qualitatively in Figure 3.

MVDM w/ Existing Algorithms: We evaluate SIFU on deoccluded images generated by our MVDM ($\mathcal{F}_\mathcal{D}$) and show its qualitative performance in Fig. 4).

Sensitivity to Occlusion levels: To assess sensitivity to changing occlusion sizes, we evaluated performance at three levels of occlusion (Table 4; where we find that CHROME maintains consistent reconstruction quality even

as occlusion severity increases.

4. CHROME for Stereo Reconstruction

As detailed in the main paper (Section 3), our method, CHROME, can be seamlessly adapted to stereo reconstruction scenarios, demonstrating its versatility. In Table 5, we provide comprehensive quantitative results that demonstrate the effectiveness of CHROME in stereo reconstruction, *i.e.*, when using two input views. These results underscore the flexibility and robustness of CHROME in handling stereo data and achieving high-quality occlusion-resilient reconstructions. Note that, CHROME can be trivially extended to handle as many views as the user would like and is upper bounded only by hardware constraints.

Table 5. Quantitative comparison of Novel View Texture Reconstruction given stereo inputs on occluded THuman2.0, where the angle represents the separation between the two views relative to the first frame, which is front-facing to the camera.

Stereo Angle	PSNR ↑	SSIM ↑	LPIPS ↓
45°	24.32	.9280	.0542
90°	24.70	.9310	.0521
135°	24.78	.9313	.0511

Table 6. Analysis of the Inference Time for Baseline Algorithms with respect to CHROME on a NVIDIA A40 GPU

Algorithm	Inference Time (Seconds) ↓
PIFu	33
GTA	57
SIFU	330
SiTH	≈ 300
CHROME	**15**

5. Analyzing Inference Time

In Table 6, we present a comparative analysis of the inference time of CHROME versus baseline algorithms. The results demonstrate that CHROME achieves superior inference time performance, making it more suitable for real-time applications compared to existing algorithms.

6. Additional Qualitative Results

We provide more qualitative results on the occluded THuman2.0 and occluded CustomHumans as an extension of the main paper in Figure 1. We provide a qualitative analysis of $\mathcal{F}_D$ for reliable occlusion-free novel view synthesis in Figure 5. We also provide a visualization of normal maps against baselines in Figure 6. Furthermore, we provide

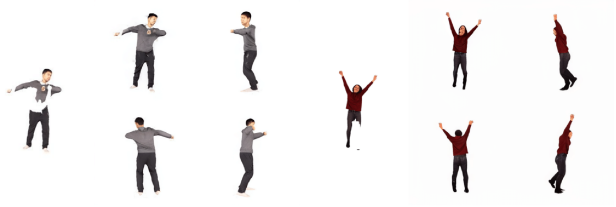

Figure 5. **Qualitative analysis of Pose Conditioned MVDM:** Qualitative analysis of our ($\mathcal{F}_D$) reveals that our pose conditioned MVDM generates reliable occlusion-free images which can later be utilized for 3D reconstruction.

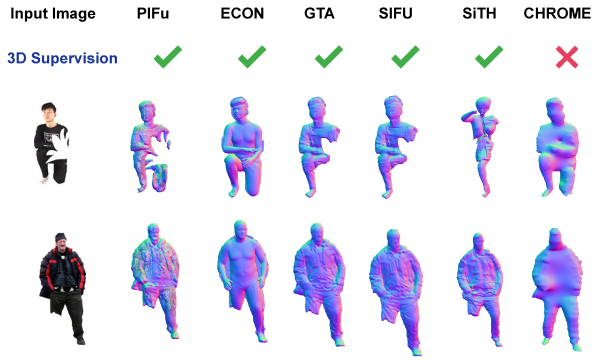

Figure 6. **Qualitative analysis of geometric reconstruction via normal maps:** Qualitative comparisons of CHROME against state-of-the-art methods for geometric reconstruction via normal consistency. Note that CHROME does not require 3D mesh supervision during training whereas all baselines necessitate the same.

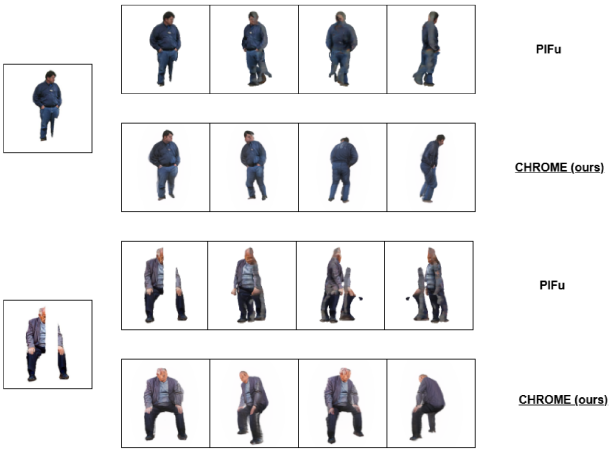

Figure 7. **Qualitative analysis of CHROME on AHP:** Qualitative comparisons of CHROME with state-of-the-art method PIFu [6] on the naturally occluded AHP dataset in zero-shot settings. Clearly, the predictions from PIFu are not occlusion-resilient whereas CHROME effectively handles occlusions, producing multiview consistent reconstructions.

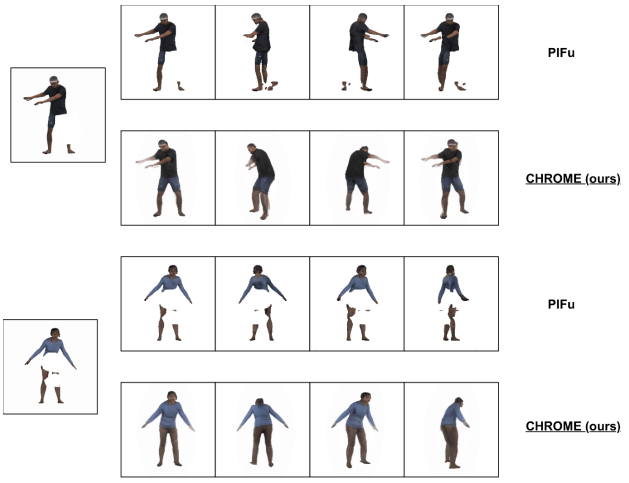

Figure 8. **Qualitative analysis of CHROME on artificially occluded CAPE:** Qualitative comparisons of CHROME with state-of-the-art method PIFu [6] on artificially occluded CAPE in zero-shot settings. Clearly, the predictions from PIFu are not occlusion-resilient whereas CHROME effectively handles occlusions, producing multiview consistent reconstructions.

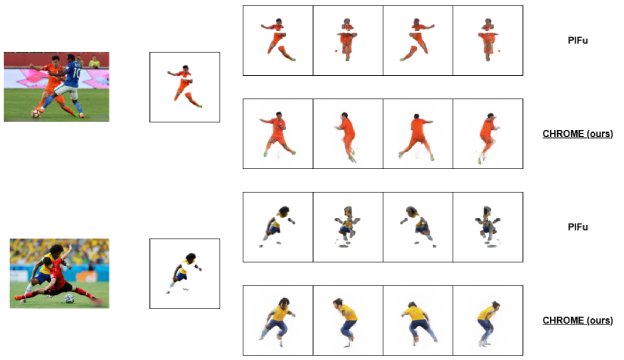

Figure 9. **Qualitative analysis of CHROME on OCHuman:** Qualitative comparisons of CHROME with state-of-the-art method PIFu [6] on the naturally occluded OCHuman dataset in zero-shot settings. Clearly, the predictions from PIFu are not occlusion-resilient whereas CHROME effectively handles occlusions, producing multiview consistent reconstructions.

qualitative results on the AHP [12], artificially occluded CAPE [13], OCHuman [14] and MultiHuman [15] datasets in Figures 7, 8, 9 and 10. AHP, OCHuman, and MultiHuman feature instances of natural occlusions, where existing state-of-the-art (SOTA) algorithms, such as PIFu [6], tend to perform poorly. In contrast, CHROME shows superior performance by providing high-quality reconstructions that are robust to occlusions. We show qualitative results only against PIFu as we find it to be the best performing baseline algorithm in terms of quantitative performance.

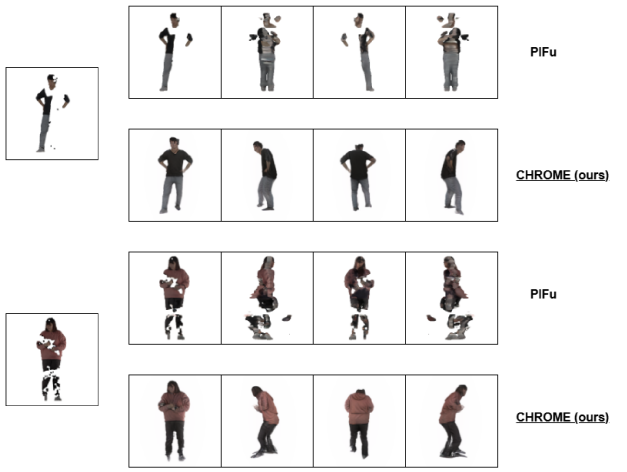

Figure 10. **Qualitative analysis of CHROME on MultiHuman:** Qualitative comparisons of CHROME with state-of-the-art method PIFu [6] on the naturally occluded MultiHuman dataset in zero-shot settings. Clearly, the predictions from PIFu are not occlusion-resilient whereas CHROME effectively handles occlusions, producing multiview consistent reconstructions.

Figure 11. Analyzing the Impact of Pose Estimation on Multiview Reconstruction: Observe the differences between the pose estimates derived from the input occluded image and those obtained from the synthesized multiview images. The conditioning of $\mathcal{F}_D$ on the input occluded image ensures that the synthesized images preserve information originating from the input occluded image.

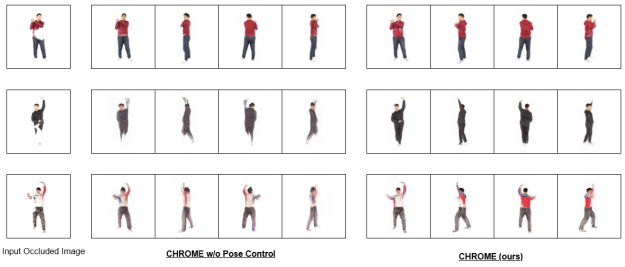

Figure 12. Additional Qualitative Results highlighting the importance of using Pose estimates as explicit control guidance.

7. Analyzing the Effect of Pose Estimator on Multiview Reconstruction

In Section 3 of the main paper, we highlight that even when pose estimation is inaccurate, the MVDM model ($\mathcal{F}_D$) en-

sures that the generated multiview images remain consistent with the input image. Specifically, the model preserves the visible regions of the input image and reconstructs only the occluded parts based on the provided pose information and the occluded input image itself. This is qualitatively illustrated in Figure 11, where the 2D projections of the estimated 3D pose are noticeably incorrect and implausible (particularly for the legs). Despite this, $\mathcal{F}_D$ successfully generates multiview reconstructions that align with the occluded input image and pose conditioning, producing plausible multiview outputs. Additional qualitative results for novel view synthesis using the $\mathcal{F}_\mathcal{D}$ trained without incorporating pose information (discussed in Ablation Study, Section 4 of the main paper) is presented in Figure 12.

Limitations and Weaknesses

Limitations: It should be noted that our solution may suffer from the domain gap between training and inference poses. Prior-based augmentations could be considered in future work, to improve generalizability.

Societal Impacts: While CHROME may be used for unwanted re-identification and surveillance, we believe that our method can positive impact our community by lowering technical barriers for broader participation, *e.g.*, in VR/AR applications and other creative processes.

References

[1] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, June 2021. 1

[2] I Ho, Jie Song, Otmar Hilliges, et al. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 538–549, 2024. 1, 3

[3] Zechuan Zhang, Li Sun, Zongxin Yang, Ling Chen, and Yi Yang. Global-correlated 3d-decoupling transformer for clothed avatar reconstruction. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3

[4] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9936–9947, 2024. 1, 3

[5] Badour AlBahar, Shunsuke Saito, Hung-Yu Tseng, Changil Kim, Johannes Kopf, and Jia-Bin Huang. Single-image 3d human digitization with shape-guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 1

[6] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2304–2314, 2019. 1, 3, 4, 5

[7] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 1

[8] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 1

[9] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 1

[10] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2025. 1

[11] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 1

[12] 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. 5

[13] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3d people in generative clothing. In *Computer Vision and Pattern Recognition (CVPR)*, June 2020. 5

[14] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 889–898, 2019. 5

[15] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *IEEE Conference on Computer Vision (ICCV 2021)*, 2021. 5