Reconstructing Close Human Interaction with Appearance and Proxemics Reasoning

Buzhen Huang[1,2*] Chen Li[4,5] Chongyang Xu[3] Dongyue Lu[2] Jinnan Chen[2]
Yangang Wang[1] Gim Hee Lee[2]

[1]Southeast University [2]National University of Singapore [3]Sichuan University
[4]IHPC, Agency for Science, Technology and Research, Singapore
[5]CFAR, Agency for Science, Technology and Research, Singapore

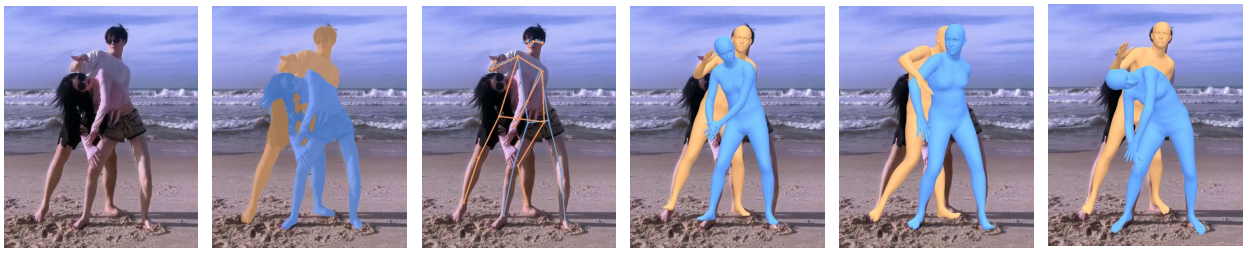

Figure 1. Due to the visual ambiguity, even state-of-the-art vision foundation models (*e.g.*, ViTPose++ [63] and temporal SAM [25, 37]) cannot clearly distinguish human semantics in close interactive cases. Consequently, human pose estimation methods based these basic human semantics tend to fail. In comparison, our dual-branch optimization framework that leverages human appearance, proxemics, and physics is capable of alleviating visual ambiguity to give better results.

Abstract

*Due to visual ambiguities and inter-person occlusions, existing human pose estimation methods cannot recover plausible close interactions from in-the-wild videos. Even state-of-the-art large foundation models (*e.g.*, SAM) cannot accurately distinguish human semantics in such challenging scenarios. In this work, we find that human appearance can provide a straightforward cue to address these obstacles. Based on this observation, we propose a dual-branch optimization framework to reconstruct accurate interactive motions with plausible body contacts constrained by human appearances, social proxemics, and physical laws. Specifically, we first train a diffusion model to learn the human proxemic behavior and pose prior knowledge. The trained network and two optimizable tensors are then incorporated into a dual-branch optimization framework to reconstruct human motions and appearances. Several constraints based on 3D Gaussians, 2D keypoints, and mesh penetrations are also designed to assist the optimization. With the proxemics prior and diverse constraints, our method is capable of estimating accurate interactions from in-the-wild videos captured in complex environments. We further build a dataset with pseudo ground-truth interaction annotations, which may promote future research on pose estimation and human behavior understanding. Experimental results on several benchmarks demonstrate that our method outperforms existing approaches. The code and data will be publicly available for research purpose.*

1. Introduction

Human interaction is an essential part of life and has many physical, mental and emotional benefits. Enabling machines to understand human interaction may promote a lot of downstream applications such as robotics, virtual reality, smart security, *etc*. As an effective tool for understanding human behaviour, 3D human pose and shape estimation has achieved profound progress in recent years. However, the existing methods are still not deployable for analysing close human interactions due to severe visual ambiguities and inter-person occlusions.

Specifically, single-person methods [9, 31, 49] only focus on pose accuracy and image-model alignment, while multi-person approaches [15, 18, 52, 66] tend to address penetration [18, 66] and spatial distribution reasonable-

[1]The work was done while Buzhen Huang is a visiting student at National University of Singapore.

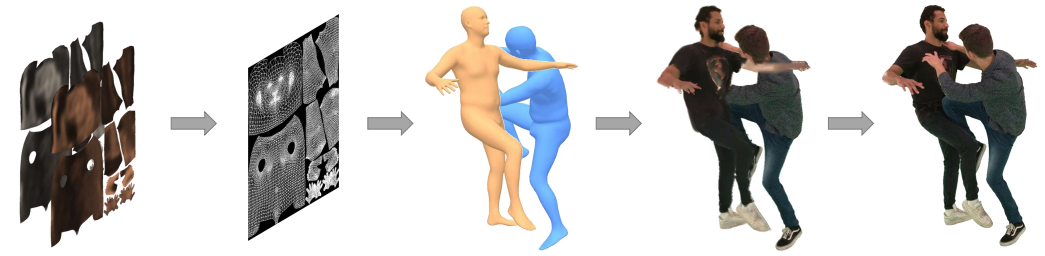

Figure 2. With predicted UV Gaussian maps, we can map the Gaussians to 3D space with a UV coordinate map and splat them to the image plane. We can then reason the depth ordinal relationship and image-model alignment with the rendered and original images. Since the Gaussians should also be consistent across non-occluded frames, the optimization adjusts poses to find an optimal solution in interactive frames, thereby producing accurate depth ordering and poses.

ness [15, 52]. They all ignore the important body contacts and proxemics in close interactions. Only a few recent works are designed for close human-human interactions [6, 16, 40, 56], which regularize the interactions with learned reaction priors [6, 16] or physical simulation [56]. However, the regression-based methods [6, 16, 56] rely on high-quality interaction data captured in indoor scenes, and thus show poor generalization ability on in-the-wild images. In contrast, BUDDI [40] fits human models to 2D keypoints via an optimization framework and can work in diverse environments. Nonetheless, even state-of-the-art large foundation models (*e.g.*, SAM [25] and ViTPose++ [63]) cannot clearly identify human semantics in complex interaction cases due to the visual ambiguity. Consequently, BUDDI still tend to fail.

In this work, we find that human appearances can provide straightforward cues to alleviate visual ambiguities and inter-person occlusions. As shown in Fig. 2, with the modeled human appearances and rendering techniques [23], we can directly leverage the original RGB image to infer the depth ordinal relationships and image-model alignment for occluded cases. Based on this observation, we design a novel dual-branch optimization framework constrained by appearance, proxemics and physics to reconstruct close human interactions. However, this straightforward idea requires simultaneously reconstructing human motions and appearances, which is a highly non-convex optimization aggravated by depth ambiguity.

To this end, we first propose a diffusion model to learn proxemic behaviors. In contrast to existing interaction priors [6, 16, 40], our model receives 2D observations and infers 3D interactions from both temporal and reactive information. After the training on interactive data with a mask strategy, the model can regularize the interaction from a noisy and partially observed input. Subsequently, the diffusion model with trained parameters is used as the motion branch in the optimization framework. During the optimization, the prior can produce a desired interaction by fine-tuning the network parameters, and this strategy is more robust to depth ambiguity and local minima. We then design an appearance branch with two optimizable tensors to constrain the reconstructed motions. Specifically, the tensors are decoded to Gaussian UV maps with a U-Net [47] backbone. The Gaussians are then mapped to 3D body surface with a UV coordinate map. By splatting the Gaussians, we can simultaneously optimize the motions and appearances. In addition, we also penalize mesh penetration and keypoints re-projection error to improve physical plausibility and pose accuracy.

Since our method can work well on different environments, we further collect 100 human-human interaction videos from Internet and build pseudo ground-truth interaction annotations. Experimental results show that the proposed dataset can improve the current regression-based method, and may promote future research on human interaction understanding. To summarize, the main contributions of this paper are as follows:

- We propose a dual-branch optimization framework constrained by appearance, proxemics and physics to reconstruct close human interactions, which can work well on in-the-wild videos.
- We demonstrate that human appearance can be an effective cue to alleviate visual ambiguities and inter-person occlusions in closely interactive scenarios.
- We build an in-the-wild dataset for close human interactions, which may promote future research on human interaction understanding.

2. Related Work

Human interaction reconstruction. Human pose and shape estimation has made tremendous progress in the past several years. However, most of works [9, 22, 31, 49] in this field focus on pose accuracy and image-model alignment for a single person, and ignore the important interactions between humans. Some works [4, 30, 46, 51] consider multiple humans in the same scene, but they just address the inter-person occlusions and do not reconstruct absolute positions for human interactions. Although some multi-person

methods can regress an approximate absolute translation for each human with projection geometry [3, 14, 66, 67], novel position representations [52, 68], or ordering-aware loss [15, 18, 24, 60], the coarse estimation is inadequate for delicate close human interactions. Reconstructing close human interaction is an open issue for decades due to the depth ambiguity, mesh penetration, and inter-person occlusion. Only a few recent works explicitly consider this problem by incorporating collision avoidance [7, 56], contact constraints [8, 50], or proxemics priors [6, 16, 40]. However, they all rely on detected 2D human semantics and are still confronted with visual ambiguities.

Human Gaussian splatting. 3D Gaussian Splatting [23] uses a set of 3D Gaussians to represent a scene and renders it by splatting and rasterizing the Gaussians, which has shown high efficiency and impressive performance on static objects. Recent works have introduced this technique to model dynamic 3D humans [39] and articulated objects [29]. Typically, a predicted human motion with SMPL representation [36] is used to initialize the Gaussians, and then the properties of each Gaussian is optimized by a rendering-and-compare strategy [29, 39, 45]. To model diverse cloth topologies, some methods adopt a more complex mesh template for the initialization [20, 32, 42] or directly use depth information as input [70]. Since these methods all rely on multi-view inputs, a few works [12, 13, 26, 61] further simplify the settings to use a monocular video to learn the human Gaussians. However, they still require the video to capture the complete observations of a human body. To address the occluded and invisible parts, OccGaussian [64] designs an occlusion feature query to reconstruct humans from partial observations. Lee *et al*. [28] also considered the similar obstacle in multi-person scenarios and used a 2D diffusion model to provide additional information. Nonetheless, they iteratively processed each human and still need accurate foreground masks. In contrast to these existing works, we simultaneously predict Gaussians for two characters, and use the reconstructed appearances to constrain the human motions.

Pseudo ground-truth generation. Historically, building 3D human annotations always relies on expensive marker-based [7, 17, 62] or multi-view [11, 33] systems in controlled environments. Although the captured poses are accurate, the model trained with this data shows poor generalization ability in in-the-wild scenarios due to its simple background and appearance. To close the domain gap, a few methods [1, 22, 43] have leveraged weak supervision for human reconstruction, but the results are still unsatisfactory due to the sparse constraint. Recently, pseudo ground-truth annotations [21, 27, 44] have enabled human pose and shape estimation to show impressive performance [9, 49]. With learned pose prior knowledge, some annotators [21, 34, 38] directly optimize the predictions by finetuning the network parameters. Additional constraints from camera perspective models [31], temporal dependencies [44], and crowd spatial distribution [15] are also incorporated to improve the annotation quality. Compared to common single- or multi-person data, close human-human interactions are more difficult to obtain. A very recent work, BUDDI [40], has designed a proxemics prior to generate interaction annotations. However, due to the visual ambiguity and the lack of temporal information, it is still struggle to produce high quality data in severely occluded cases.

3. Our Method

Fig. 3 shows an illustration of our framework. Given a monocular in-the-wild video with close interactions between two people, we propose a dual-branch optimization framework to reconstruct accurate body poses, natural proxemic relationships, and plausible physical contacts.

3.1. Human interaction representation

Motion representation. We adopt the SMPL model [36] with a 6D rotation representation [71] to describe the interaction, which consists of pose $\theta \in \mathbb{R}^{144}$, shape $\beta \in \mathbb{R}^{10}$, and translation $\tau \in \mathbb{R}^3$. For a video with N frames and 2 individuals, the reconstructed motions can be denoted as $\mathbf{x}^{1:N} = \left\{\mathbf{x}^{a,1:N}, \mathbf{x}^{b,1:N}\right\}$, where $\mathbf{x}^{a,1:N} = \left\{\theta^i, \beta^i, \tau^i\right\}_{i=1}^{N}$.

Appearance representation. 3D Gaussian Splatting [23] is used to represent human appearance, which is parameterized by a set of 3D Gaussians. Conventionally, each Gaussian contains an offset relative to SMPL vertex $\mu \in \mathbb{R}^3$, color $c \in \mathbb{R}^3$, opacity $\sigma \in \mathbb{R}$, rotation $q \in \mathbb{R}^3$, and scale $s \in \mathbb{R}^3$. We can obtain the rendered appearance by splatting Gaussians to the image plane.

3.2. Proxemic prior

Directly optimizing SMPL parameters [2] for close human reconstruction encounters severe depth ambiguity and is highly sensitive to occlusions and local minima. To address these obstacles, we first train a diffusion model to learn pose and proxemics prior knowledge to assist the optimization.

Model architecture. We adopt a diffusion model to learn the prior, which iteratively predicts clean data from a pure noise conditioned on 2D observations. In addition to image features, we use 2D keypoints as an additional condition for the diffusion since the existing interaction datasets [33, 62] may not contain paired RGB images. When the RGB image is unavailable, we can still use these data to train the prior by setting image features to be zero. Specifically, the ground-truth two-person motions $\hat{\mathbf{x}}_0^{1:N}$ are first diffused towards a standard Gaussian distribution:

$$q(\mathbf{x}_t \mid \hat{\mathbf{x}}_0) = \sqrt{\hat{\alpha}_t}\hat{\mathbf{x}}_0 + \sqrt{1-\hat{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(0, \mathrm{I}), \quad (1)$$

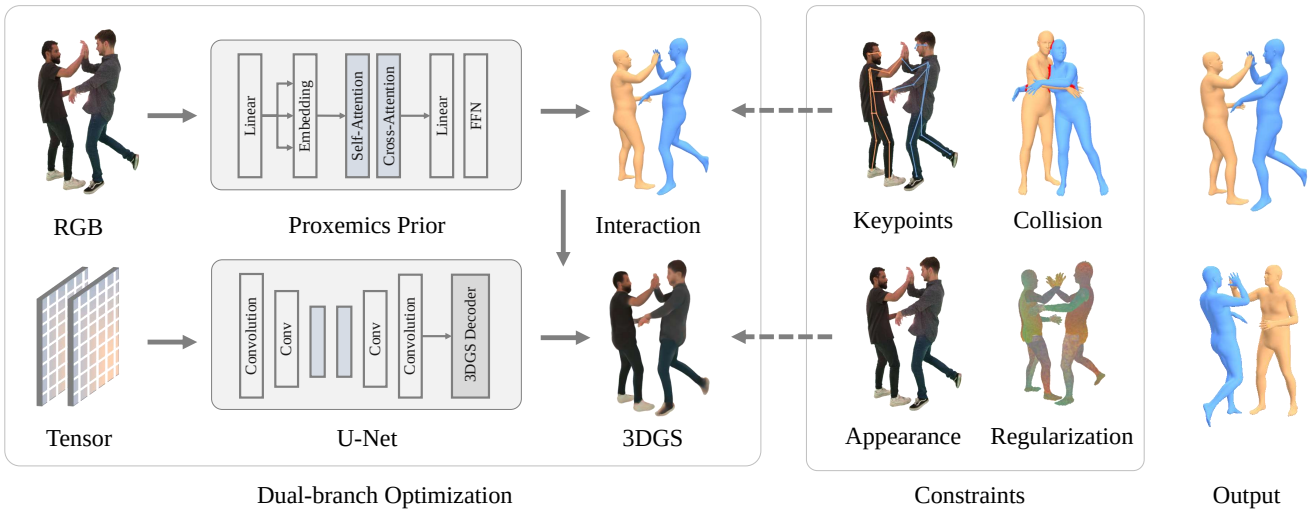

Figure 3. **Overview of our framework.** We propose a dual-branch optimization framework to reconstruct close human interactions from a monocular in-the-wild video. By optimizing the proxemics prior, U-Net backbone, and two optimizable tensors, the framework simultaneously predicts interactive motions and coarse appearances. With the constraints from 2D observations, physics, and prior knowledge, the framework can finally output 3D interactions with plausible body poses, natural proxemic relationships and accurate physical contacts.

where α_t and $\hat{\alpha}_t$ are constant hyper-parameters [41]. The noisy motions $\mathbf{x}_t^{1:N}$ are projected to high-dimensional vectors and then concatenated with image and keypoints features. We omit illustration of the full diffusion process in Fig. 3 for brevity. S transformer blocks are used to process the concatenated features. As shown in the top left of Fig. 3, the features of two individuals can share information with a cross-attention module [33] in each transformer block. Finally, the denoised motions are regressed with a feed-forward layer from the processed features. In each timestep, the diffusion model predicts the clean motions and then diffuses them to $\mathbf{x}_{t-1}^{1:N}$, which is defined as:

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\alpha(\mathbf{x}_t, c), \tilde{\gamma}_t \mathrm{I}), \tag{2}$$

where $\mu_\alpha(\mathbf{x}_t, c)$ is the estimated mean by the diffusion model under the condition of c. $\tilde{\gamma}_t$ is a hyper-parameter.

Mask strategy. To ensure that the prior is robust to occlusions, we adopt two mask strategies to learn temporal dependencies and proxemic behaviors, respectively. 1) We randomly mask the condition and input poses from a subset of frames, and then enforce the diffusion model to inpaint the missing information based on temporal relationships. 2) We may completely mask the inputs of one individual and compel the model to generate a reaction from the counterpart. With these two strategies, the prior can produce complete motions even when one individual is totally invisible.

Training loss. We use the following loss functions to train the prior:

$$\mathcal{L} = \mathcal{L}_{\text{reproj}} + \mathcal{L}_{\text{smpl}} + \mathcal{L}_{\text{joint}} + \mathcal{L}_{\text{vel}} + \mathcal{L}_{\text{int}}. \tag{3}$$

The reprojection loss is given by:

$$\mathcal{L}_{\text{reproj}} = \|\Pi\left(J_{3D} + \tau\right) - \hat{J}_{2D}\|_2^2, \tag{4}$$

where $\hat{J}_{2D}$ is ground-truth 2D pose. It should be noted that the reprojection loss is important for regressing the absolute translations. Previous single-person methods [9, 49] always use a weak-perspective camera and may result in unreasonable spatial distribution in interaction reconstruction. We thus follow CLIFF [31] to use a perspective camera $\Pi(\cdot)$ with a common diagonal Field-of-View 55° to project the 3D joints to 2D image. The remaining terms include the supervisions from the SMPL parameters: $\mathcal{L}_{\text{smpl}} = \|[\beta, \theta] - [\hat{\beta}, \hat{\theta}]\|_2^2$, 3D joint positions: $\mathcal{L}_{\text{joint}} = \|J_{3D} - \hat{J_{3D}}\|_2^2$, and velocities: $\mathcal{L}_{\text{vel}} = \|\dot{J}_{3D} - \hat{\dot{J}}_{3D}\|_2^2$ on each individual. To enforce more plausible interaction, we also penalize the relative distance between two characters:

$$\mathcal{L}_{\text{int}} = \||J_{3D}^a - J_{3D}^b| - |\hat{J}_{3D}^a - \hat{J}_{3D}^b|\|_2^2. \tag{5}$$

Once the training is completed, the diffusion model can predict close interactions from RGB images and 2D keypoints. However, due to the visual ambiguity and occlusion, the results may not be consistent with image observations. To mitigate this issue, we use the predicted motions and trained network parameters as initial values, and further refine the motions with a dual-branch optimization.

3.3. Dual-branch optimization

Due to the visual ambiguity, the current regression models [25, 63] cannot clearly distinguish human semantics in closely interactive cases, and thus feed-forward human reconstruction tends to fail. We therefore design a dual-branch optimization to leverage human appearance, proxemics, and physics to address these problems.

Motion branch. We utilize the trained diffusion model from the previous section to construct the motion branch.

Initially, motions $\mathbf{x}_0^{1:N}$ are regressed from image and keypoints features. As these motions may exhibit image-model misalignment and lack reliability, we then diffuse them to $\mathbf{x}_1^{1:N}$. Unlike traditional reverse diffusion processes that adjust motions under additional guidance [16, 49], we finetune the network parameters π_m using several loss functions to update $\mathbf{x'}_0^{1:N}$. This approach enhances the controllability of reconstructed motions and ensures their consistency with observations. Moreover, the pretrained network parameters can provide pose and proxemic prior knowledge to alleviate the depth ambiguity and occlusion.

Appearance branch. Previous optimization-based methods iteratively fit the model to 2D measurements like 2D keypoints [2, 43], silhouette [58], or part segmentation [27]. However, even state-of-the-art large foundation models [25, 63] struggle to produce accurate human semantics for close human interactions due to visual ambiguity. We find that RGB images can provide reliable dense correspondences and can serve as a constraint with a reconstructed human appearance. Consequently, we design the appearance branch to predict 3D Gaussians for appearance modeling as shown in the bottom left of Fig. 3. The Gaussian properties are encoded in a UV map [19] since it is difficult to directly optimize tens of thousands of independent Gaussians. We use a U-Net backbone to regress the map from an optimizable tensor which works as a latent code for the human appearance. The Gaussian UV map has 14 channels containing offset μ, color c, opacity σ, rotation q, scale s, and identity d. The convolution layers build the dependencies for Gaussians on the UV map, and each Gaussian can be mapped to 3D space with the UV coordinate map. By optimizing the input tensors and U-Net, we can reconstruct the human appearances. With the poses from the motion branch, we can also render the appearances to 2D images via Gaussian splatting.

Objective function. We formulate several objective functions to constrain the output of the dual-branch framework. We first penalize the appearance loss:

$$\mathcal{L}_{\text{app}} = \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{ssim}} + \mathcal{L}_{\text{lpips}}, \tag{6}$$

where $\mathcal{L}_{\text{rgb}}$, $\mathcal{L}_{\text{ssim}}$, and $\mathcal{L}_{\text{lpips}}$ are the L1, SSIM [59], and LPIPS [69] loss between rendered and original images. We also calculate the re-projection loss Eq. (4) with 2D keypoints to prevent large pose deviations. Since we can access the region of the Gaussians rendered on the image with identity d, we only use the keypoints that fall within the rendered region of each individual, which differs from the common formulation and helps alleviate the impact of some incorrect detections. To prevent inter-person penetrations in close human reconstruction, we adopt a differentialable 3D distance fields [55] to reflect the mesh collision:

$$\mathcal{L}_{\text{pen}} = \sum_{(f_a, f_b) \in \mathcal{C}} \left\{ \sum_{v_a \in f_a} \| -\Psi_{f_b}(v_a)\, n_a \|_2^2 + \sum_{v_b \in f_b} \| -\Psi_{f_a}(v_b)\, n_b \|_2^2 \right\} \tag{7}$$

where f_a, f_b are two colliding triangles in the detected colliding triangles $\mathcal{C}$. v and n are vertex position and normal, respectively, and $\Psi(\cdot)$ is the distance field. We also use a smoothness term to enforce smooth motions:

$$\mathcal{L}_{\text{smooth}} = \sum_{i=1}^{N-1} \| J_{3D}^{i+1} - J_{3D}^{i} \|_2^2. \tag{8}$$

We further regularize the predicted motion and appearance parameters by:

$$\mathcal{L}_{\text{reg}} = \|\theta - \theta'\|_2^2 + \|\tau - \tau'\|_2^2 + \|\beta - \beta'\|_2^2 + \mathcal{L}_{\text{offset}} + \mathcal{L}_{\text{scale}}, \tag{9}$$

where θ', β', and τ' are the initial predictions from the diffusion model. $\mathcal{L}_{\text{offset}} = \|\mu\|_2^2$ and $\mathcal{L}_{\text{scale}} = \|s\|_2^2$ calculate the L2-norm of the predicted offsets and scales, respectively.

Joint optimization for motion and appearance. Given an in-the-wild video with two-person close interactions, we first track each human with AutoTrackAnything [37] to produce bounding-boxes and masks. VitPose [63] is also used to detect 2D keypoints for each human. To be noted that other methods require precise segmentation or keypoints of each individual for reconstruction. In contrast, our framework simultaneously splats two interactive humans and requires only the segmentation of all individuals as a whole to mask the background, which avoids the individual semantic parsing in close interactions. Subsequently, we predict the initial motions with the trained proxemics prior. After the initialization, the overall optimization objective is defined as:

$$\arg\min_{\pi_m, \pi_a} \mathcal{L} = \mathcal{L}_{\text{app}} + \mathcal{L}_{\text{reproj}} + \mathcal{L}_{\text{pen}} + \mathcal{L}_{\text{smooth}} + \mathcal{L}_{\text{reg}}. \tag{10}$$

The optimization variables are π_m and π_a, where π_a represents the parameters of the optimizable tensors and U-Net in the appearance branch. We adopt the Adam optimizer with learning rates of 0.00002 and 0.003 for the motion and appearance branches, respectively. The optimization typically takes $\sim 3-5$ minutes for a video with 128 frames.

4. WildCHI Dataset

Despite the prosperous development of 3D human datasets in recent years, data for close human interaction remains scarce due to complex body contacts, extreme inter-person

Dataset	Motions	Frames	Scene	3D Pose Format	Scheme	Temporal	RGB image
CHI3D [7]	373	63K	indoor	SMPLX	MoCap	✓	✓
Hi4D [65]	100	11K	indoor	SMPL	mRGB	✓	✓
ExPI [10]	115	30K	indoor	Skeleton	mRGB	✓	✓
InterHuman [33]	6,022	1.7M	indoor	SMPL	mRGB	✓	✗
Inter-X [62]	11,388	8.1M	indoor	SMPLX	MoCap	✓	✗
Flickr Fits [40]	–	11K	outdoor	SMPLX	Pseudo	✗	✓
WildCHI	100	20K	outdoor	SMPL	Pseudo	✓	✓

Table 1. **Comparisons of existing human-human interaction datasets.**

occlusions, and severe visual ambiguities. Although some recent works have introduced several large-scale human-human interaction datasets [33, 62], they lack RGB images and cannot be used for the human interaction reconstruction task. Furthermore, other datasets are captured in controlled environments, leading to a domain gap from in-the-wild images. The only outdoor dataset available is Flickr Fits [40], which annotates images with close interactions under a contact constraint. However, no in-the-wild dataset exists to learn temporal dependencies for close human interactions, which are crucial for addressing occlusions and depth ambiguities. To this end, we collect 100 videos with diverse environments and subjects from TikTok [54], and build pseudo ground-truth with the proposed method. We also manually filter out incorrect estimations. Tab. 1 demonstrates the strengths of our in-the-wild close human interaction (**WildCHI**) dataset, which has a similar amount of motion as commonly used indoor datasets (Hi4D and ExPI). We also train CloseInt [16], a regression-based close human interaction method, on the proposed dataset. The experimental results in Tab. 2 show that WildCHI can improve the performance of CloseInt [16] in both indoor and outdoor scenarios. According to the terms of service of TikTok [53], we will release our dataset for research purposes. More animatable samples of the proposed dataset can be found in the Supplementary Material.

5. Experiments

5.1. Datasets

Inter-X [62] and **InterHuman [33]** are large-scale human-human interaction datasets. Inter-X covers 40 daily interaction categories with 89 distinct subjects having different social relationships. InterHuman also contains diverse two-person interactions. Due to the lack of color images, we use these datasets to train the proxemixs prior only. **Hi4D [65]** is an accurate multi-view dataset capturing closely interacting humans. It contains 20 unique pairs of participants with varying body shapes and clothing styles performing diverse interaction motion sequences. We follow [16] to use 5 pairs (23, 27, 28, 32, 37) as testset. The remaining sequences are used for training. **CHI3D [7]** captures 3 pairs of people in close interaction scenarios with a Vicon MoCap system and 4 additional RGB cameras. We use the standard splits of this dataset. **3DPW [57]** also contains several sequences with two-person interactions. We use these sequences for evaluation only.

5.2. Metrics

We report Mean Per Joint Position Error (MPJPE) and MPJPE after Procrustes Analysis (PA-MPJPE) on close human interaction datasets. The joint PA-MPJPE [40] is also used, which applies Procrustes Analysis on the pair. Additionally, we utilize the Mean Per Vertex Position Error (MPVPE) to measure mesh quality. Following [16], we use an interaction error to assess the quality of reconstructed interactive behaviors. Moreover, we incorporate the average penetration depth (A-PD) [5] to evaluate body contact and penetration depth.

5.3. Comparison to State-of-the-Art Methods

We compare the proposed dual-branch optimization framework with some state-of-the-art baseline methods on different datasets to demonstrate our superiority. We first evaluate Human4D [9] and CLIFF [31] on Hi4D [65] and 3DPW [57]. These two methods are designed for single-person scenarios, which directly regress SMPL parameters from a single image. As shown in Tab. 2, Human4D [9] achieves high joint accuracy with a ViT backbone but struggles to produce accurate spatial distributions due to the using of weak-perspective camera. On the other hand, CLIFF [31] estimates each human in original camera coordinates but still encounters a high interaction loss due to depth ambiguity. We also compare with BEV [52] and GroupRec [15], which explicitly consider multi-person scenarios by introducing constraints for crowds. While these methods predict humans with more reasonable distributions, they tend to ignore close interactions.

BUDDI [40] is a recent work designed specifically for close interaction reconstruction using an optimization-based framework. It fits two SMPL models to detected 2D keypoints and can handle in-the-wild images. However, the current state-of-the-art pose detectors [63] still struggle to produce reliable keypoints for close interactive cases due to visual ambiguity. Although it significantly improves the pose accuracy, the model training relies on a lot of high-

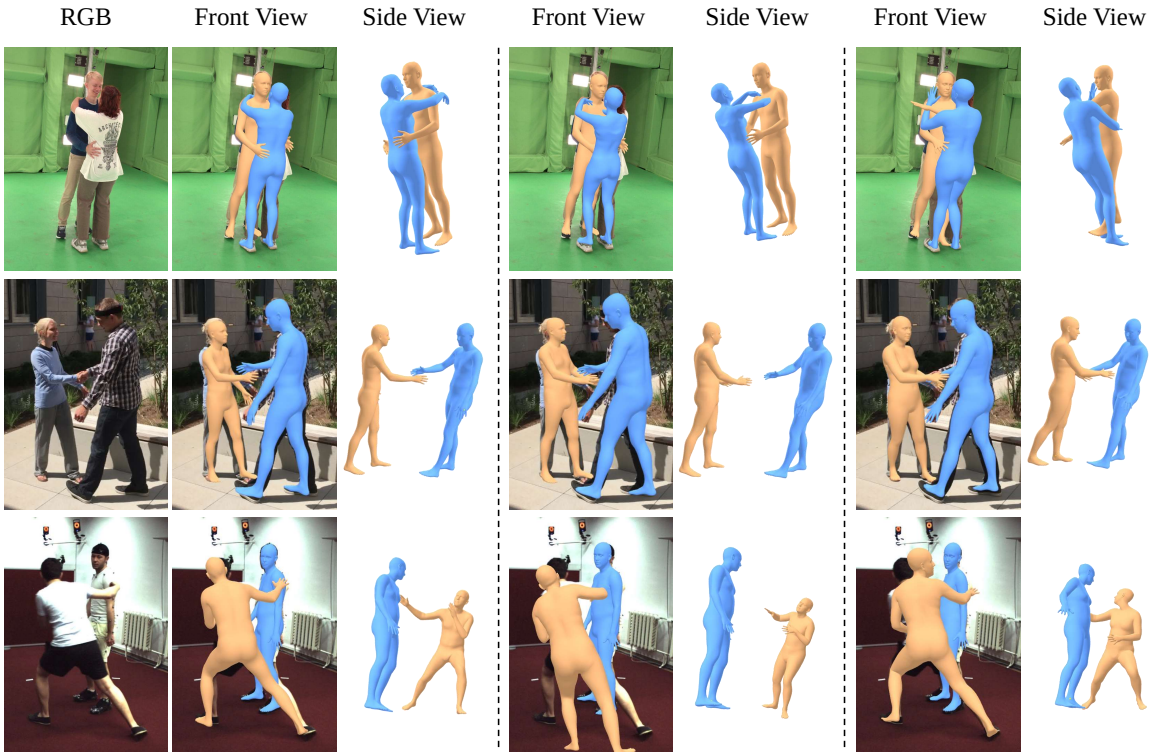

Figure 4. Qualitative comparison with BUDDI [40] and CloseInt [16]. Our method is more robust to visual ambiguity.

Method	Hi4D				3DPW			
	MPJPE	PA-MPJPE	MPVPE	Inter	MPJPE	PA-MPJPE	MPVPE	Inter
Human4D [9]	72.1	52.4	88.6	–	72.9	49.1	81.8	–
CLLIF [31]	91.3	53.6	109.6	141.5	–	–	–	–
BEV [52]	91.8	52.2	101.2	131.0	78.3	48.5	82.3	136.4
GroupRec [15]	82.4	51.6	88.6	98.8	73.3	48.7	81.2	110.6
BUDDI [40]	96.8	70.6	116.0	102.6	83.6	53.6	93.8	113.1
CloseInt [16]	63.1	47.5	76.4	81.4	70.6	51.4	80.6	100.3
CloseInt [16] w/ WildCHI	61.4	45.1	75.4	80.5	66.4	48.3	77.4	**95.9**
Ours	**59.1**	**44.3**	**72.0**	**80.2**	**64.5**	**45.6**	**75.2**	96.4

Table 2. **Comparisons on Hi4D and 3DPW.** Our method can achieve state-of-the-art performance in both indoor and outdoor scenarios. "–" means the results are not available.

quality interaction data. Since these data can only be obtained in a controlled environment, the trained model shows poor generalization ability in outdoor scenarios. Our approach differs from the above works as we simultaneously reconstruct human motions and appearances, directly utilizing the RGB image as a constraint. This strategy alleviates visual ambiguity by comparing rendered and original images. As depicted in Fig. 4, our method estimates interactions with more accurate body poses, depth ordinal relationships, and model-image alignment.

We also conduct experiments on outdoor images. On 3DPW dataset, we follow CloseInt [16] to use all interactive sequences as a benchmark for the evaluation. Tab. 2 shows that our method can achieve state-of-the-art in terms of most metrics, which demonstrates the superiority of our method in in-the-wild scenarios. As shown in Fig. 4, our method can work well under diverse environments. In addition, the close interaction data produced by our method can also significantly improve the current regression-based method (*e.g.*, CloseInt). We also show some qualitative images and videos in Supplementary Material, which also demonstrate the effectiveness of our method.

5.4. Ablation Study

Human appearance. We investigate the importance of the proposed appearance constraint by removing the appearance branch, and supervise the optimization by only motion-level loss functions. As shown in Fig. 5, although

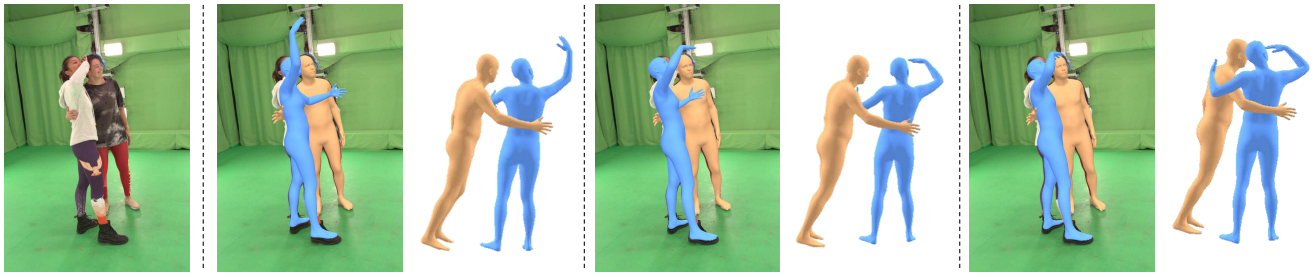

Figure 5. Ablation study. The initial prediction is severely affected by visual ambiguity and cannot reconstruct accurate interaction. With the proposed optimization, the body pose can be improved with the additional constraints. In addition, we find that appearance constraint is important for the depth ordinal relationships.

the framework without the appearance constraint can still produce accurate body poses, the depth ordinal relationship is incorrect. We find that human appearance is effective to prevent this problem, even with coarse texture. We simultaneously splat the two-person Gaussians onto the 2D image during human appearance reconstruction to reflect occlusion relationships in the results. By comparing the rendered image with the original RGB input, we can reason the depth ordinal relationship and enforce better interactions. Direct use of RGB images as a constraint also promotes more accurate body poses and model-image alignment since it does not introduce noises compared to 2D keypoints and masks. Additionally, physical constraints always limit the solution space of the optimization [48]. When the physical constraint is not applied, the appearance and proxemics loss promote the optimization to find more accurate 3D joints without considering the mesh penetration.

Proxemics prior. The primary limitations of optimization-based human reconstruction are the sensitivity to depth ambiguity and local minima. To alleviate the impact of these two obstacles, we propose a proxemics prior learned from extensive interaction data to assist the optimization. We formulate this prior as a diffusion model, allowing it to denoise noisy motions and generate clean data. During the optimization, we finetune the pretrained network parameters, and enforce the network to output accurate motions under various supervisions. In Tab. 3, we compare our approach with a strategy that directly optimizes SMPL parameters without the proxemics prior. By leveraging learned network parameters containing pose and interaction prior knowledge, optimization with the prior proves to be more accurate and efficient.

6. Limitation and Future Work

Limitation. Although the proposed framework can produce 3D close human interactions from a monocular in-the-wild video, there are still some limitations. First, our method cannot reconstruct high-quality complete human textures when the light condition is changed or the human is partially observed. Although a coarse texture is sufficient for constraining the underlying body motion, the quality of reconstructed appearances can still be improved by incorporating light embedding [35] or large vision foundation models [28]. Second, the current design can only capture two-person close interactions. Without sufficient data, we cannot train the proxemics prior for interactive behaviours with more than 2 people. As a result, building an interaction dataset for crowd is also a promising direction for future human behaviour understanding related tasks. In addition, the input video should contain some frames with little or no contact for constraining appearances.

Method	MPJPE	PA-MPJPE	MPVPE	Interaction	A-PD
Initial Prediction	65.05	48.54	78.35	86.20	1.16
Ours w/o Appearance	60.68	45.86	73.52	81.01	0.83
Ours w/o Proxemics	61.52	47.13	74.84	87.13	0.85
Ours w/o Physics	57.01	42.67	69.57	78.50	1.30
Ours	59.06	44.29	71.99	80.18	0.81

Table 3. **Ablations on Hi4D.** "Initial Prediction" denotes the results directly predicted by the pretrained proxemics prior without the optimization. "w/o Appearance", "w/o Proxemics" and "w/o Physics" represent the frameworks without appearance constraint, proxemics prior, and physical constraint, respectively.

7. Conclusion

We propose a novel dual-branch optimization framework to reconstruct two-person close interactions from a monocular in-the-wild video. To alleviate the depth ambiguity and insufficient visual information, we first introduce a proxemics prior based on diffusion model to assist the optimization. We then build a appearance branch with 3D Gaussian splatting to address the notorious visual ambiguity. Compared to previous works that rely on keypoints, masks, and pure RGB information, our method is more robust to diverse environments and can produce more accurate results. Based on the propose framework, we further build an in-the-wild close interaction dataset to promote related research.

Acknowledgement. This research is supported by China Scholarship Council under Grant Number 202306090192.

References

[1] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 3

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, pages 561–578, 2016. 3, 5

[3] Junuk Cha, Muhammad Saqlain, GeonU Kim, Mingyu Shin, and Seungryul Baek. Multi-person 3d pose and shape estimation via inverse kinematics and refinement. In *ECCV*, pages 660–677, 2022. 3

[4] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *CVPR*, pages 1475–1484, 2022. 2

[5] Rong et al. Monocular 3d reconstruction of interacting hands via collision-aware factorized refinements. In *3DV*, 2021. 6

[6] Qi Fang, Yinghui Fan, Yanjun Li, Junting Dong, Dingwei Wu, Weidong Zhang, and Kang Chen. Capturing closely interacted two-person motions with reaction priors. In *CVPR*, 2024. 2, 3

[7] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *CVPR*, pages 7214–7223, 2020. 3, 6

[8] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Reconstructing three-dimensional models of interacting humans. *arXiv preprint arXiv:2308.01854*, 2023. 3

[9] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *ICCV*, pages 14783–14794, 2023. 1, 2, 3, 4, 6, 7

[10] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *CVPR*, pages 13053–13064, 2022. 6

[11] TomasSimon HanbyulJoo, HaoLiu XulongLi, LinGui Lei-Tan, and TimothyGodisart SeanBanerjee. Panoptic studio: A massively multiview system for social interaction capture. *TPAMI*, 41(1), 2019. 3

[12] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *CVPR*, 2024. 3

[13] Shoukang Hu and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *CVPR*, 2024. 3

[14] Buzhen Huang, Tianshu Zhang, and Yangang Wang. Pose2uv: Single-shot multiperson mesh recovery with deep uv prior. *TIP*, 31:4679–4692, 2022. 3

[15] Buzhen Huang, Jingyi Ju, Zhihao Li, and Yangang Wang. Reconstructing groups of people with hypergraph relational reasoning. In *ICCV*, pages 14873–14883, 2023. 1, 2, 3, 6, 7

[16] Buzhen Huang, Chen Li, Chongyang Xu, Liang Pan, Yangang Wang, and Gim Hee Lee. Closely interactive human reconstruction with proxemics and physics-guided adaption. In *CVPR*, 2024. 2, 3, 5, 6, 7

[17] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2014. 3

[18] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, pages 5579–5588, 2020. 1, 3

[19] Yujiao Jiang, Qingmin Liao, Xiaoyu Li, Li Ma, Qi Zhang, Chaopeng Zhang, Zongqing Lu, and Ying Shan. Uv gaussians: Joint learning of mesh deformation and gaussian textures for human avatar modeling. *arXiv preprint arXiv:2403.11589*, 2024. 5

[20] Yuheng Jiang, Zhehao Shen, Penghao Wang, Zhuo Su, Yu Hong, Yingliang Zhang, Jingyi Yu, and Lan Xu. Hifi4g: High-fidelity human performance rendering via compact gaussian splatting. In *CVPR*, 2024. 3

[21] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *3DV*, pages 42–52, 2021. 3

[22] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 2, 3

[23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *TOG*, 42(4):1–14, 2023. 2, 3

[24] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *CVPR*, pages 1715–1725, 2022. 3

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 1, 2, 4, 5

[26] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. In *CVPR*, 2024. 3

[27] 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 *CVPR*, pages 6050–6059, 2017. 3, 5

[28] Inhee Lee, Byungjun Kim, and Hanbyul Joo. Guess the unseen: Dynamic 3d scene reconstruction from partial 2d glimpses. In *CVPR*, 2024. 3, 8

[29] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *CVPR*, 2024. 3

[30] Haoyuan Li, Haoye Dong, Hanchao Jia, Dong Huang, Michael C Kampffmeyer, Liang Lin, and Xiaodan Liang.

Coordinate transformer: Achieving single-stage multi-person mesh recovery from videos. In *ICCV*, pages 8744–8753, 2023. 2

[31] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, pages 590–606, 2022. 1, 2, 3, 4, 6, 7

[32] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *CVPR*, 2024. 3

[33] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *arXiv preprint arXiv:2304.05684*, 2023. 3, 4, 6

[34] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *CVPR*, pages 21159–21168, 2023. 3

[35] Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, et al. Vastgaussian: Vast 3d gaussians for large scene reconstruction. *arXiv preprint arXiv:2402.17427*, 2024. 8

[36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, 34(6):1–16, 2015. 3

[37] Roman Lyskov. Autotrackanything, 2024. 1, 5

[38] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Neuralannot: Neural annotator for 3d human mesh training sets. In *CVPRW*, pages 2298–2306, 2022. 3

[39] Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *CVPR*, 2024. 3

[40] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. In *CVPR*, 2024. 2, 3, 6, 7

[41] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 4

[42] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *CVPR*, 2024. 3

[43] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 3, 5

[44] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. In *CVPR*, pages 1485–1495, 2022. 3

[45] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *CVPR*, 2024. 3

[46] Zhongwei Qiu, Qiansheng Yang, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Chang Xu, Dongmei Fu, and Jingdong Wang. Psvt: End-to-end multi-person 3d pose and shape estimation with progressive video transformers. In *CVPR*, pages 21254–21263, 2023. 2

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 2

[48] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics (ToG)*, 39(6):1–16, 2020. 8

[49] Anastasis Stathopoulos, Ligong Han, and Dimitris Metaxas. Score-guided diffusion for 3d human recovery. In *CVPR*, 2024. 1, 2, 3, 4, 5

[50] Sanjay Subramanian, Evonne Ng, Lea Müller, Dan Klein, Shiry Ginosar, and Trevor Darrell. Pose priors from language models. *arXiv preprint arXiv:2405.03689*, 2024. 3

[51] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, pages 11179–11188, 2021. 2

[52] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, pages 13243–13252, 2022. 1, 2, 3, 6, 7

[53] TikTok. Tiktok terms of service. https://www.tiktok.com/legal/page/row/terms-of-service/en, 2021. Accessed: 2024-11-01. 6

[54] TikTok. Tiktok. https://www.tiktok.com/, 2021. Accessed: 2024-11-01. 6

[55] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 118(2):172–193, 2016. 5

[56] Nicolas Ugrinovic, Boxiao Pan, Georgios Pavlakos, Despoina Paschalidou, Bokui Shen, Jordi Sanchez-Riera, Francesc Moreno-Noguer, and Leonidas Guibas. Multiphys: Multi-person physics-aware 3d motion estimation. In *CVPR*, 2024. 2, 3

[57] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 6

[58] Yangang Wang, Yebin Liu, Xin Tong, Qionghai Dai, and Ping Tan. Outdoor markerless motion capture with sparse handheld video cameras. *TVCG*, 24(5):1856–1866, 2017. 5

[59] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, pages 600–612, 2004. 5

[60] Hao Wen, Jing Huang, Huili Cui, Haozhe Lin, Yu-Kun Lai, Lu Fang, and Kun Li. Crowd3d: Towards hundreds of people reconstruction from a single image. In *CVPR*, pages 8937–8946, 2023. 3

[61] Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alexander G Schwing, and Shenlong Wang. Gomavatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. In *CVPR*, 2024. 3

[62] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, Yunhui Liu, Wenjun Zeng, and Xiaokang Yang. Inter-x: Towards versatile human-human interaction analysis. *arXiv preprint arXiv:2312.16051*, 2023. 3, 6

[63] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 2, 4, 5, 6

[64] Jingrui Ye, Zongkai Zhang, Yujiao Jiang, Qingmin Liao, Wenming Yang, and Zongqing Lu. Occgaussian: 3d gaussian splatting for occluded human rendering. *arXiv preprint arXiv:2404.08449*, 2024. 3

[65] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *CVPR*, pages 17016–17027, 2023. 6

[66] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *CVPR*, pages 2148–2157, 2018. 1, 3

[67] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. *NeurIPS*, 31, 2018. 3

[68] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body meshes as points. In *CVPR*, pages 546–556, 2021. 3

[69] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 5

[70] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *CVPR*, 2024. 3

[71] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, pages 5745–5753, 2019. 3