Guess The Unseen: Dynamic 3D Scene Reconstruction from Partial 2D Glimpses Supplemental Material

Inhee Lee Byungjun Kim Hanbyul Joo
Seoul National University
{ininin0516, byungjun.kim, hbjoo}@snu.ac.kr
https://snuvclab.github.io/gtu/

A. Implementation Details

A.1. Baseline Implementation Details

HumanNeRF [17] does not support the simultaneous optimization of multiple people, so we optimize each person separately and merge them in the evaluation stage. Following the default HumanNeRF experiment settings, each person is optimized for 400k iterations using 4 NVIDIA RTX4090 GPUs which takes approximately 40 hours per person. For the ZJU-Mocap [11] dataset, we utilize the publicly available checkpoints shared by the authors.

***Shuai et al.* [15]** represents the scene as a composition of a background model and human model, both represented by a variant of NeRF [10, 11]. For the Panoptic dataset [5] and Hi4D dataset [19], we model the background using a time-conditioned NeRF defined on the surface of the cylinder fully covering the scene and the human model with NeuralBody [11]. We jointly optimize these models for 400k iterations using 2 NVIDIA RTX4090 GPUs which takes approximately 70 hours per scene. The remaining settings are the same as the original paper [15]. When we render the scene for evaluation, we discard the background and only render the human model.

InstantAvatar [4] reconstructs a single person from monocular video input. Hence, we optimize it on each person separately and merge them in the evaluation stage same as HumanNeRF [17]. We train the InstantAvatar for 50 epochs using a single RTX3090, following the default options used to optimize PeopleSnapShot [1] in the original paper.

A.2. Ours Implementation Details

Background pre-optimization. We first optimize background Gaussians $\mathcal{G}^{BG}$ with images that humans are masked out. The background Gaussians $\mathcal{G}^{BG}$ are initialized with point cloud obtained by SfM [14] or SLAM [16]. In the case of a fixed camera, we initialize Gaussians $\mathcal{G}^{BG}$ with a 3D sphere whose radius is 30m, together with background regularization loss to prevent it from occluding the people as follows:

$$\mathcal{L}_{reg}^{BG} = \lambda_{reg}^{BG} \sum_{i=0}^{N} ||\boldsymbol{\mu}_i^{BG} - 30||^2 \tag{1}$$

, where $\boldsymbol{\mu}_i^{BG}$ is the center of i-th background Gaussian. We scale the world's unit distance to be 1m before starting optimization. The background is optimized for 30k iterations following the default 3D-GS [6] experiment settings.

Human background joint optimization. After the pre-optimization of the background, we optimize human Gaussians $\mathcal{G}^h_{j\ j=1,\dots,N}$ and background Gaussians $\mathcal{G}^{BG}$ together. For the first $1.5k$ iteration of joint optimization, we fix the center of human Gaussians $\boldsymbol{\mu}_i$ on the initial points $\boldsymbol{x}_{i,init}$ and clamp the opacity o_i below 0.9 to avoid the body being transparent. We densify the human Gaussians in $[2000, 2500, 3000]$ iterations for detailed reconstruction and prune Gaussians which are exceptionally large or transparent every 500 iterations until the end of optimizations to reduce artifacts. The background Gaussians are densified only during pre-optimization stage and keep the same number of Gaussians in the joint optimization stage.

Optimization Details. We use Adam [7] optimizer with different learning rates for each component of 3D Gaussians. For the center of Gaussian $\boldsymbol{\mu}$, we set an initial learning rate as $1e^{-3}$ and decay it until $2e^{-6}$ during training. We use a fixed learning rate $2.5e^{-3}$ for color $\boldsymbol{c}$, $5e^{-2}$ for opacity $\boldsymbol{o}$, $5e^{-3}$ for scale $\boldsymbol{s}$, and $1e^{-3}$ for quaternion $\boldsymbol{q}$. We set the loss weight of SSIM loss $\lambda_{ssim} = 0.2$, MSE loss $\lambda_{rgb} = 0.8$, LPIPS loss $\lambda_{lpips} = 0.1$, and SDS loss $\lambda_{sds} = 1.0$. For hard surface regularization loss, we set the weight of loss λ_{hard} relative to reconstruction loss weight $\lambda_{recon} = 0.1 \times \lambda_{recon}$ to keep a balance of losses. We use a fixed reconstruction loss weight $\lambda_{recon} = 1.0$ before $1k$ iterations and then schedule the weight after $1k$ iterations to balance the reconstruction loss and SDS loss.

SDS loss details. We use a publicly available SD1.5 [13] and OpenPose ControlNet [20] checkpoint for the SDS loss. Similar to other methods using SDS [12], we use a high CFG

scale of 50 to generate detailed texture on unseen parts. We sample the noise time step τ of SDS loss from $\mathcal{U}[0.5, 0.98]$ for the first $2k$ iterations and then smoothly anneal it into $\mathcal{U}[0.02, 0.3]$ over following $2k$ iterations similar to the prior work [21]. We also schedule the weight of reconstruction loss λ_{recon} with a maximum time step τ_{max} on each iteration to balance the reconstruction loss and SDS loss as follows:

$$\lambda_{recon} = 10^6 \times \tau_{max}^2. \tag{2}$$

We apply SDS loss from $1k$ iteration of the joint optimization. For every single iteration of reconstruction loss, we apply SDS loss on all humans who appeared in the scene.

We sample random unseen cameras for SDS loss from the surface of a sphere with a radius of 2.2, centered on the human pelvis. The azimuth φ and elevation ϑ of cameras are drawn from $\varphi \sim \mathcal{U}[-\pi, \pi]$ and $\vartheta \sim \mathcal{U}[-0.3\pi, 0.3\pi]$. Additionally, we choose a view-augmented prompt [*side, front, back*] based on the sampled azimuth φ and SMPL global rotation. For the initial $3k$ iterations of optimization with SDS loss, we mainly render the full body of posed human Gaussians $\mathcal{G}_j^h(\boldsymbol{\theta}_{j,t})$ and canonical human Gaussians $\mathcal{G}_j^h(\boldsymbol{\theta}_c)$ for SDS loss. In the subsequent iterations, we also randomly sample from zoomed-in views of the head, upper body, and lower body together with the full body of the posed human, and the full body of the canonical with a uniform probability of 0.2. This two-stage random camera sampling facilitates the detailed reconstruction of unseen parts and head.

B. Dataset Preprocessing

B.1. Panoptic Dataset [5]

We trim the last round of the ultimatum 160422 sequence, extracting 540 multi-view images of 6 individuals by subsampling every 4 frames. Among the 31 HD cameras in the Panoptic Dome, we specifically choose cameras 0, 3, 5, 8, 22, 24, and 25 for evaluation, while camera 16 serves as the input. To simulate a challenging scenario, we intentionally pick the input view camera that excludes the entrance of the Panoptic Dome [5] where individuals enter one by one.

To acquire the SMPL parameters $\boldsymbol{\theta}_{t,j}$ and $\boldsymbol{\beta}_j$ of individuals, we optimize them by minimizing the distance between 3D SMPL joints and provided pseudo ground truth COCO 3D joints. Our optimization process incorporates pose prior, angle shape regularization, and 3D joint error, as outlined in [2]. We leverage SMPL joints and SAM [8] to obtain each individual's mask in the input frames. Initially, we arrange individuals based on their depth which is calculated as the distance between the pelvis of SMPL and the camera center. Starting with the individual closest to the camera, we obtain a mask by querying the projected SMPL joints which is not occluded into SAM [8]. We assume the joints is occluded if it's projected on the masks of nearer people.

(a) CFG 20 (b) CFG 100

Figure 1. **Ablation study for the classifier-free guidance scale.** We cropped out the black blurry artifacts near the feet due to lack of space. We can check that a low CFG scale (a) generates a smooth monotonic texture in unseen parts while a high CFG scale (b) synthesizes and enhances wrinkles of clothing on both seen and unseen parts (lower row), but also introduces more artifacts. (upper row)

B.2. In-the-wild Videos

In handling in-the-wild videos, we categorize them into two scenarios: static camera and moving camera. For the camera moving cases, we employ DROID-SLAM [16] to estimate the initial camera pose and Goel et al. [3] to track people with regressing SMPL parameters. Subsequently, we refine the estimated parameters by minimizing the reprojection error between estimated 2D body joints [18]. In cases with a static camera, we skip the camera pose estimation step.

C. Effect of Classifier-Free Guidance Scale

To explore the impact of changing the classifier-free guidance (CFG) scale, we conduct an ablation study using Hi4D [19] `pair00-dance` sequence. As illustrated in the lower row of Fig. 1, a high CFG scale synthesizes detailed unseen parts such as cloth wrinkles and uniform numbers, while a low CFG scale produces a smooth, monotonic texture without any wrinkles. Notably, a high CFG scale introduces more artifacts such as green stains which are amplified by the light reflected from the floor shown in the upper row of Fig. 1. This study shows the importance of selecting a

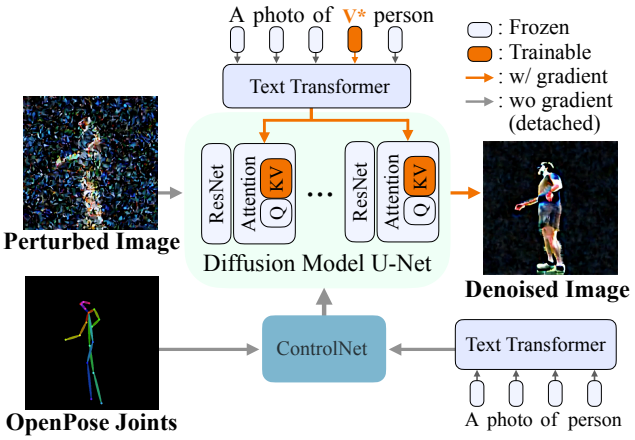

Figure 2. **Overall Pipeline of Textual Inversion in our method** The orange part is what we optimize during textual inversion. **V*** indicates textual inversion token <person-j> which is training target. As shown here, we use CustomDiffusion [9] together with ControlNet [20] to obtain individuals' inversion token <person-j> and fine-tuned diffusion model ϕ_j.

proper CFG scale to reconstruct a detailed human avatar with minimal artifacts.

D. Details of Textual Inversion

To obtain an individual's text-token <person-j> and specified fine-tuned diffusion, we run CustomDiffusion on each individual's observations with modifications as shown in Fig. 2. We use OpenPose ControlNet [20] during Textual Inversion to avoid possible overfitting on observed body pose and camera pose. To obtain an individual's text-token <person-j> and specified fine-tuned diffusion, we first randomly perturb the observed image and then estimate the added noise of the perturbed image. By minimizing the MSE loss between the added noise and the estimated noise, we optimize the text-token and fine-tune the diffusion model. As we use the latent diffusion model [13] here, the training objective is as follows:

$$\mathcal{L}_{textual} = \text{MSE}(\epsilon_\phi(\boldsymbol{z}_\tau; \boldsymbol{y}, \tau) - \epsilon) \tag{3}$$

, where $\boldsymbol{z}_\tau$ is a perturbed latent corresponding to perturbed image in Fig. 2 and ϵ is the added noise. During optimization, we randomly sample τ from $\tau \sim \mathcal{U}[0, 1]$.

We optimize textual token and fine-tune diffusion using Adam [7] optimizer with learning rate $5e^{-6}$ and batch size 4 for 1000 iterations. To mitigate the situation where the text token learns the background, we mask out the background and randomly fill it with white or black color. We do not use prior preservation loss here to overfit the text token on observed images. The text-token <person-j> is queried only in Diffusion U-Net and not queried in the ControlNet module as shown in Fig. 2.

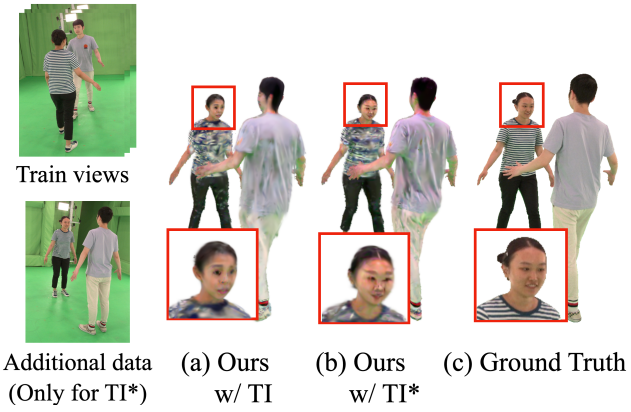

Figure 3. **Ablation study of adding additional data during Textual Inversion.** TI* means the textual inversion used in SDS loss is trained with a single additional image of the frontal view. Both (a) and (b) are optimized with train views and the only difference is in the Textual Inversion.

E. Enhancing Identity with Additional Images

By employing additional image sources for the target identity, if they are known in advance, we can enhance the identity of the person with sparse observations. Specifically, training the Textual Inversion (TI) with an extra face image of the target person, assuming this information is available beforehand, enables our method to produce results that more closely resemble the target human, even in scenarios with an extreme lack of frontal train views. We further show such scenario in Fig. 3 (b), where training the TI with just a single additional frontal image substantially improves the resemblance of the outputs, compared to Fig. 3 (a). This demonstrates the unique advantage of using textual inversion for reconstruction, a method that is difficult to leverage using only reconstruction loss.

Table 1. Table of notations.

Symbol	Description
	Index
i	Gaussian index, $i \in \{1, \ldots, N\}$ in 3D Gaussian attributes
j	Human index, in human Gaussians $\mathcal{G}_j^h$ and SMPL parameters $\theta_{j,t}, \beta_j$
t	Time index, $t \in \{1, \ldots, T\}$ in SMPL pose parameters, input images
k	Joint index, $k \in \{1, \ldots, N_{joint}\}$ in LBS skinning
	Learnable Attributes of 3D Gaussians
$\boldsymbol{\mu}_i \in \mathbb{R}^3$	Center of i-th Gaussian
$\boldsymbol{q}_i \in SO(3)$	Covariance Matrix's Quaternion Component of i-th Gaussian
$\boldsymbol{s}_i \in \mathbb{R}^3$	Covariance Matrix's Scale Component of i-th Gaussian
$\boldsymbol{c}_i \in \mathbb{R}^3$	Color of i-th Gaussian
$o_i \in \mathbb{R}$	Opacity of i-th Gaussian
G_i	i-th Gaussian consists of $\{\boldsymbol{\mu}_i, \boldsymbol{q}_i, \boldsymbol{s}_i, \boldsymbol{c}_i, o_i\}$
	Parameters of Diffusion Model
ϕ/ϕ_j	Diffusion model / Diffusion model fine-tuned on j-th person
τ	noise time-step of diffusion model $\tau \in [0, 1]$
$\boldsymbol{z}_0$	Encoded latent of the queried RGB images on diffusion model
$\boldsymbol{z}_\tau$	Perturbed latent with noise time-step $\tau \in [0, 1]$
ϵ	Noise added to the latent
ϵ_ϕ	Noise estimated by diffusion model ϕ
	Parameters of Human Deformation
$\boldsymbol{\theta}_{j,t} \in \mathbb{R}^{72}$	SMPL pose parameter of j-th Human in time $t \in \{1, \ldots, T\}$
$\boldsymbol{\beta}_j \in \mathbb{R}^{10}$	SMPL shape parameter of j-th Human
$\boldsymbol{\theta}_c \in \mathbb{R}^{72}$	Canonical pose parameter shared for all humans
	Rendered and Observed Images
R_t/I_t	Rendered / Observed RGB image in time $t \in \{1, \ldots, T\}$
R_v^h	Rendered RGB image of a human with camera v

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, 2018. 1

[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*, 2016. 2

[3] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 2

[4] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *CVPR*, 2023. 1

[5] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *TPAMI*, 2017. 1, 2

[6] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *SIGGRAPH*, 2023. 1

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1, 3

[8] 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*, 2023. 2

[9] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 3

[10] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1

[11] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 1

[12] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 1

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 3

[14] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1

[15] Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. Novel view synthesis of human interactions from sparse multi-view videos. In *SIGGRAPH*, 2022. 1

[16] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. In *NeurIPS*, 2021. 1, 2

[17] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, 2022. 1

[18] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022. 2

[19] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *CVPR*, 2023. 1, 2

[20] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 1, 3

[21] Joseph Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d with advanced diffusion guidance. In *ICLR*, 2024. 2