UniSH: Unifying Scene and Human Reconstruction in a Feed-Forward Pass

Supplementary Material

5.1. Dataset Curation

Scalability and Data Requirements. A core advantage of UniSH is the ability to learn from unlabeled in-the-wild videos with minimal geometric assumptions. For the dataset curated in this work, we specifically target dance videos to capture high-quality, complex human motion where the subject remains the primary focus. However, this domain choice serves primarily as a proof-of-concept. Unlike prior work that necessitates laboratory settings or motion capture, our framework imposes no such restrictions. We only require monocular videos with a visible human subject. Consequently, our method can be effortlessly scaled up to significantly more diverse, general-purpose video sources available on the public internet.

Automated Filtering Pipeline. We curated a large-scale dataset from public platforms, employing a rigorous, multi-stage automated pipeline to ensure geometric and temporal consistency. First, to mitigate the impact of scene transitions and montages common in raw internet videos, we enforce strict temporal continuity. We employ the standard PySceneDetect library to identify shot boundaries based on content changes. Sequences exhibiting such discontinuities are automatically discarded, ensuring that the input to our temporal attention modules represents a continuous, unedited motion sequence.

Following temporal filtering, we employ the object detector adopted by [42] to isolate single-subject sequences. We retain only those clips where a unique person class is consistently detected, thereby eliminating crowd ambiguity. To guarantee sufficient resolution for surface refinement, we enforce a spatial prominence constraint, requiring the subject's average bounding box height to exceed 40% of the image height.

Finally, visibility is enforced by discarding sequences with bounding box truncation at image borders. Crucially, to ensure an unobstructed view, we filter out any sequence where the subject overlaps with other detected bounding boxes, effectively removing environmental occlusions. The resulting dataset comprises 1,354 unique sequences, totaling approximately 1.2 million frames.

Ethical Compliance. Our data collection protocol strictly aligns with the conference ethical guidelines. Acknowledging the impracticability of obtaining individual consent given the dataset scale, we restricted our acquisition exclusively to content publicly broadcast by original creators. We operate strictly within the scope of the fair use doctrine for academic research. To safeguard subject privacy, we enforce a policy against the retention of any personally identifiable metadata.

Furthermore, we uphold the right to be forgotten through a passive distribution mechanism. Rather than distributing raw video data, we release only Video IDs and corresponding timestamps. This ensures that any content removed by the creator from the hosting platform automatically becomes inaccessible within our dataset. This mechanism effectively preserves the creator's ultimate control over their content dissemination.

Bias and Limitations. We acknowledge potential biases in the data source. Online dance communities may skew towards specific demographics or body types. Users should be aware of these distribution shifts during deployment. However, our core contribution is the methodology for leveraging abundant in-the-wild videos. Since our framework requires no manual labels, this specific bias is not a limitation of the method itself. It can be readily mitigated by simply scaling the data collection to include more diverse sources.

6. Model Architecture Details

This section details the unified architecture of the UniSH framework. Our model integrates three specialized components. These modules collaborate to achieve joint metric-scale scene and human reconstruction in a single forward pass.

6.1. Scene Reconstruction Branch

This branch adapts the permutation-equivariant architecture of π^3 [59]. The goal is permutation-equivariant geometry estimation. A ViT-Large encoder [11] is adopted as the backbone. The core feature aggregation employs a Cross-Frame Transformer. This Transformer uses 36 layers. The hidden dimension is $D = 1024$. It is configured with 16 attention heads. Feature tokens alternate between spatial and global self-attention. The architecture omits frame index positional embeddings. Three parallel heads process the geometric features $\mathcal{F}_{geo}$. These heads predict the extrinsic camera poses E, the per-frame point map P, and the confidence map C.

6.2. Human Body Branch

We adopt the CameraHMR [33] framework for human pose and shape estimation, utilizing a ViTPose-Base backbone to

extract image features. The architecture employs a Transformer Decoder that processes the input human crop. To condition the reconstruction on scene-specific geometry, we inject bounding box and focal length information as done in [33]. Crucially, the focal length f is not estimated latently but is derived directly from the intrinsics predicted by our Scene Reconstruction Branch, explicitly coupling the two branches.

The SMPL Decoder outputs the per-frame pose parameters θ_i and the body shape parameters $\beta \in \mathbb{R}^{10}$. To ensure temporal consistency across the sequence, we compute the final body shape by averaging the per-frame β predictions over the entire video. In addition to the parametric outputs, the decoder extracts high-level human feature tokens $\mathcal{F}_{hmr}$, which serve as the query input for the subsequent AlignNet module.

6.3. AlignNet

The AlignNet is our novel lightweight fusion network. Its primary function is aligning the human and scene predictions into a single metric-scale coordinate system.

The AlignNet is implemented as a two-layer transformer decoder with a default hidden dimension of $D_{hidden} = 512$. It operates with dedicated adapter layers. These layers map the input geometric features $\mathcal{F}_{geo}$ and human features $\mathcal{F}_{hmr}$ to the uniform hidden dimension D_{hidden}.

The query sequence Q is explicitly constructed. Q is formed by concatenating a learned scale token with the sequence of adapted Human Body features $\mathcal{F}_{hmr}$. The key K and value V sequences are derived from the consolidated Scene Geometry features $\mathcal{F}_{geo}$.

The prediction layers consist of a stack of two Cross-View Transformer Decoder layers. These layers utilize rotary position embeddings to encode the temporal sequence information.

The decoder output is used by two prediction heads. The dedicated Scale Head processes the Scale Token output. The predicted global scale s is obtained by applying the softplus activation function to the logits, ensuring positivity. The Translation Head predicts the camera translation t_i. The final translation vector t_i is constructed by scaling the raw (x, y) components by the final depth component z_{final}, then concatenating with z_{final}.

7. Implementation Details

This section outlines the specific numerical configurations and optimization routines used to train UniSH. We detail the general setup and the critical weighting parameters for each training stage.

7.1. General Optimization and Training Setup

The entire framework is trained using the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. We employ a cosine decay learning rate policy. The initial learning rate is 5×10^{-5} for all three training stages. Training runs for a total of 100 epochs per stage. We utilize a linear warm-up phase for the first 3 epochs, followed by the cosine decay schedule, decaying the learning rate to zero. We use a distributed training setup. The per-GPU batch size is configured to 2. We use gradient accumulation over 4 steps. This yields an effective total batch size of 64 with 8 GPUs. The weight decay parameter for regularization is fixed at 0.05. In each iteration, we sample a video clip spanning 5 seconds. This clip is temporally sampled to 30 frames at 6 FPS for training. The input image resolution is 518×294 pixels. Gradient Clipping is enabled with a maximum gradient norm of 1.0.

7.2. Weights of Losses

Stage 1: Human Surface Refinement This stage is designed to distill high-frequency geometric details from an expert depth model using unlabeled, in-the-wild videos. The training objective for this stage is given by:

$$\mathcal{L}_{stage1} = \frac{1}{N}\sum_{i=1}^{N}(\lambda_h \mathcal{L}_{h,i} + \lambda_{preg}||P_i - P_i^{orig}||_1)$$

The primary weight for the confidence-aware local human loss is $\lambda_h = 1$. The regularization weight penalizing deviation from the original pre-trained point map is $\lambda_{preg} = 0.1$. The local patch radius τ for computing $\mathcal{L}_{h,i}$ is set to 0.2.

Stage 2: Coarse-grained Alignment This stage establishes initial metric-scale consistency using synthetic data. The coarse alignment loss, $\mathcal{L}_{stage2}$, combines HMR supervision with a global scale loss:

$$\mathcal{L}_{stage2} = \frac{\lambda_{smpl}}{N}\sum_{i=1}^{N}\mathcal{L}_{smpl,i} + \lambda_{scale}||s - s_{opt}||_1$$

The weights are $\lambda_{smpl} = 1$ and $\lambda_{scale} = 0.1$. The component loss weights are set as $\lambda_v = 0.1$, $\lambda_{j3d} = 0.1$, $\lambda_{j2d} = 10$, $\lambda_{pose} = 0.1$, $\lambda_{shape} = 0.1$, and $\lambda_{trans} = 1$.

Stage 3: Fine-grained Alignment This final stage addresses the generalization gap by fine-tuning on unlabeled real-world data. The objective directly minimizes the geometric error between the reconstructed human point cloud and the predicted SMPL mesh. The total loss is:

$$\mathcal{L}_{stage3} = \frac{1}{N}\sum_{i=1}^{N}(\lambda_{align}\mathcal{L}_{align,i} + \lambda_{depth}\mathcal{L}_{dreg,i} + \lambda_{j2d}\mathcal{L}_{j2d,i})$$

We set the alignment weight $\lambda_{align} = 1$, the depth ordering regularization weight $\lambda_{depth} = 1$, and the $2D$ reprojection loss weight $\lambda_{j2d} = 10$.

8. Impact of Human Surface Refinement

Our surface refinement strategy is designed to overcome the inherent trade-off between geometric fidelity and multi-view consistency. The baseline π^3 model [59], shown on the left, provides strong cross-frame consistency for the scene and coarse human structure. However, the resulting human point cloud often exhibits poor geometric fidelity. It fails to accurately conform to the detailed shape of the human body.

Conversely, a specialized monocular depth estimator, such as the expert MoGe-2 model [55] (center), can predict exceptionally high-fidelity human surface details on a per-frame basis. This approach, however, fundamentally suffers from poor temporal consistency and global scene alignment.

Our method successfully integrates these dual requirements (Fig. 6). By employing confidence-aware distillation from the MoGe-2 expert, UniSH effectively injects high-frequency geometric information into the robust π^3 framework. The result is a system that maintains the global coherence and cross-frame consistency inherited from the π^3 backbone, while simultaneously achieving significantly higher fidelity and detail in the reconstructed human surface. This qualitative synthesis validates the necessity and efficacy of our specialized two-stage human surface refinement approach.

9. Impact of Fine-grained Alignment

Our framework employs a coarse-to-fine strategy to align the reconstructed human mesh with the scene. We first emphasize the necessity of this curriculum. The initial coarse alignment stage is a strict prerequisite for convergence. Attempting to optimize the geometric fine-tuning objectives directly from random initialization results in training collapse.

Once initialized by the coarse stage, the fine-grained alignment handles the critical sim-to-real adaptation. We visually ablate the components of this stage in Figure 7. The second column shows the performance without the geometric alignment loss ($\mathcal{L}_{align}$). It fails to generalize to in-the-wild inputs due to the domain gap. This results in incorrect global scale predictions and erroneous SMPL translations that drift from the visual evidence.

The third column demonstrates the model trained without the depth-ordering regularization ($\mathcal{L}_{dreg}$). The alignment loss successfully pulls the SMPL mesh near the human point cloud. However, the lack of physical depth constraints causes the SMPL body to be placed incorrectly. It often floats in front of the reconstructed surface. This violates the physical principle that the camera-visible surface should occlude the internal body volume. Our full model successfully integrates both objectives. It achieves accurate metric scale and translation while maintaining correct depth ordering. The SMPL mesh is placed coherently to align with the reconstructed human surface.

10. Additional Evaluation on Sloper4D

To further demonstrate the generalization and robustness of UniSH, we additionally evaluate our method on the Sloper4D dataset [10], which contains ground-truth annotations for both human motion and scene reconstruction.

As shown in Table 4 (a), our method achieves competitive global motion performance against specialized HMR methods. Furthermore, to evaluate our predicted metric scene scale, we directly compare our raw scale predictions with the metric ground truth. We construct a baseline that applies ZoeDepth's metric scale to the scale-agnostic π^3 predictions. As shown in Table 4 (b), UniSH outperforms the baselines, validating our scale alignment strategy.

Table 4. **Quantitative evaluation on the Sloper4D dataset.** (a) Comparison of global human motion estimation accuracy. (b) Comparison of metric scene reconstruction. UniSH demonstrates competitive motion tracking accuracy and superior absolute scale prediction compared to strong baselines.

(a) Global Motion (Sloper4D)

Method	WA-MPJPE↓	W-MPJPE↓	RTE↓
WHAM	297.7	1272.3	10.5
TRAM	215.1	1285.3	3.0
UniSH (Ours)	**254.8**	**1277.5**	**5.3**

(b) Metric Recon. (Sloper4D)

Method	AbsRel↓	$\delta_{1.25}$ ↑
π^3 + ZoeDepth	1.68	0.22
CUT3R	0.35	0.44
UniSH (Ours)	**0.28**	**0.53**

11. Inference Cost Analysis

A key advantage of our feed-forward architecture is its efficiency compared to optimization-based methods. On a single NVIDIA RTX 4090 GPU, UniSH achieves an inference speed of approximately **20 FPS**.

Crucially, our proposed feature fusion module, AlignNet, incurs a negligible computational overhead. As illustrated in Figure 8, AlignNet accounts for only ∼0.5% of the total inference time. The majority of the computational cost is dominated by the Scene Reconstruction branch (∼97.5%) and the Human Body branch (∼2%). This demonstrates that our human-scene alignment strategy is both highly effective and computationally lightweight.

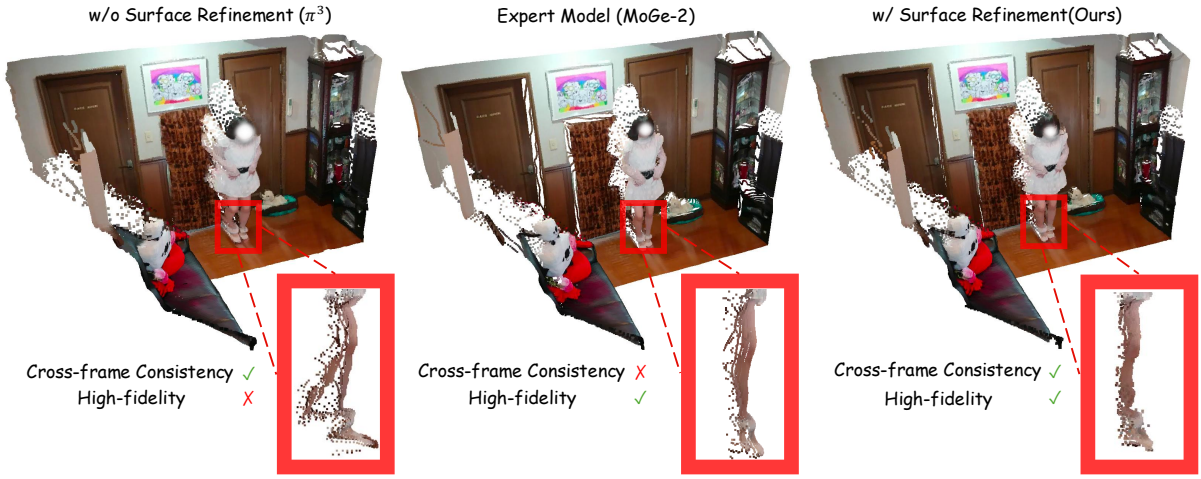

Figure 6. **Qualitative Impact of Human Surface Refinement.** Comparison illustrating the effectiveness of our specialized surface refinement strategy. The π^3 baseline (Left) provides good cross-frame consistency but the reconstructed human geometry is coarse and does not conform well to the body shape. The Expert Monocular Model (MoGe-2) (Center) achieves high fidelity but lacks multi-view consistency. Our full method (Right) successfully distills the high-frequency surface details into the multi-view reconstruction framework, achieving both high fidelity and strong cross-frame consistency.

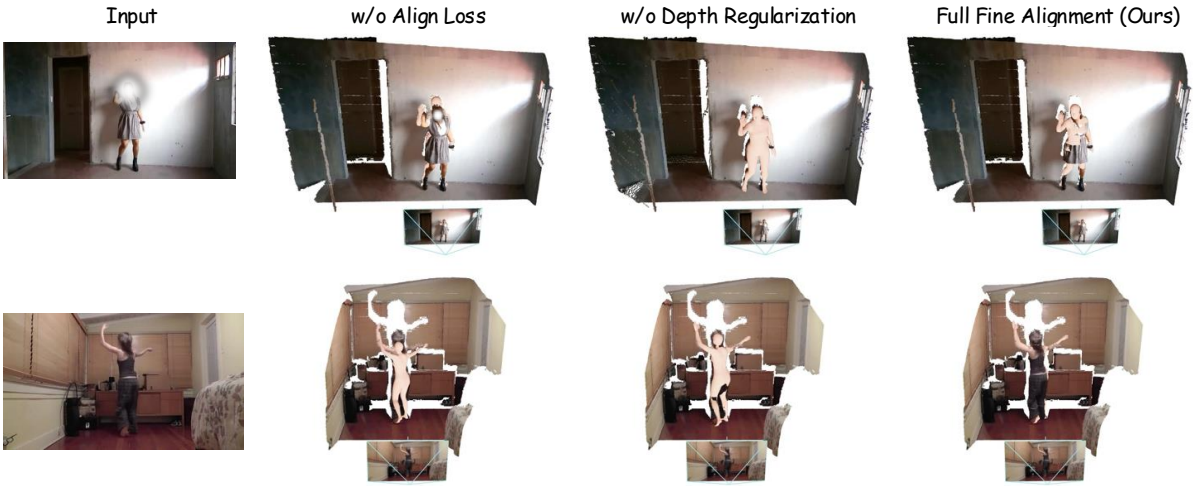

Figure 7. **Visual ablation of the Fine-grained Alignment stage.** We validate the necessity of our unsupervised geometric losses on in-the-wild data. **w/o Align Loss:** Without explicit geometric alignment ($\mathcal{L}_{align}$), the model fails to predict the correct scene scale and SMPL placement due to the domain gap. **w/o Depth Regularization:** Removing the depth constraint ($\mathcal{L}_{dreg}$) results in physical inconsistencies, where the SMPL mesh is not correctly positioned to align with the visible human point cloud. **Full Fine Alignment (Ours):** Our complete method achieves accurate global alignment and correct depth ordering. **Note:** We omit the visualization for "w/o Coarse Alignment" as removing the initial coarse stage leads to training non-convergence.

12. Ablation on Depth Teacher Model

In our Surface Refinement stage, we utilize a pre-trained expert depth model to distill high-frequency geometric details. During the development of our method, we ablated the choice of the teacher model, specifically comparing the single-frame **MoGe-2** against the video-based **MegaSAM**.

As shown in Table 5, while using MegaSAM improves upon the baseline (without refinement), it lacks the fine-grained precision of image-based MoGe-2. Video-based baselines tend to over-smooth details, which is detrimental to our alignment process. The results confirm that

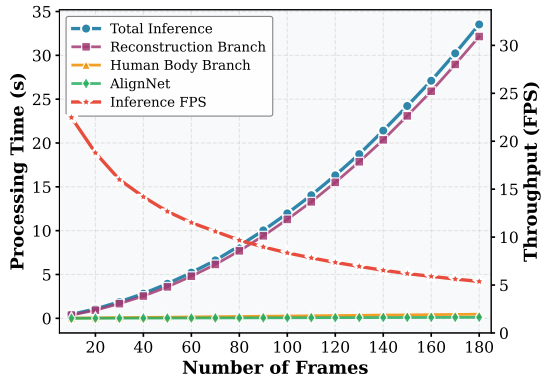

Figure 8. **Inference time breakdown of UniSH on an RTX 4090 GPU.** The proposed AlignNet introduces negligible computational overhead (∼0.5%), while the majority of the processing time is allocated to the Scene Reconstruction (∼97.5%) and Human Body (∼2%) branches. The overall system operates efficiently at approximately 20 FPS.

the single-frame MoGe-2 offers superior geometric fidelity, providing a more precise geometric anchor that is essential for accurate SMPL alignment.

Table 5. **Ablation study on the depth teacher model for Surface Refinement.** We compare the single-frame MoGe-2 against the video-based MegaSAM. MoGe-2 provides superior geometric fidelity, which yields a more precise anchor for human-scene alignment and ultimately leads to lower HMR errors.

	EMDB-2			Bonn	
Method	**WA↓**	**W↓**	**RTE↓**	**Abs↓**	$\delta_{1.25}$↑
w/o Refine	130.2	308.7	6.4	0.049	0.975
MegaSAM	125.6	301.5	6.1	0.043	0.976
MoGe-2 (Ours)	**118.5**	**270.1**	**5.8**	**0.035**	**0.980**

13. More Visualization Results

This section presents additional qualitative results demonstrating the robustness of UniSH across diverse, challenging scenes and complex human motions. In Fig. 9, the temporal sequence is explicitly encoded by the color gradient, ranging from light blue (earlier frames) to dark blue (later frames), which is consistently applied to the reconstructed cameras and the SMPL meshes.

The upper sequence showcases reconstruction of an extreme, highly articulated human pose (rock climbing). Our framework accurately aligns the SMPL body model to the reconstructed scene geometry (the climbing wall), confirming stability even under non-standard articulation. The lower sequence illustrates coherent, long-term tracking of human motion within a large, in-the-wild urban environment. The color-coded trajectory clearly shows the consistent movement of both the predicted SMPL meshes and their corresponding camera poses over time. The framework successfully reconstructs the complex irregular wall structure and maintains stable metric-scale alignment of the human trajectory over multiple frames. These visualizations confirm the generalization capability and metric stability of the joint scene and human reconstruction provided by UniSH.

14. 4D Visualizations

We provide a supplementary video file to further illustrate our method. This video focuses on two primary aspects of our performance. First, we demonstrate the temporal consistency of the UniSH framework. We show the dynamic evolution of the reconstructed scene and human mesh on continuous video sequences. This highlights the stability of our method in maintaining geometric coherence over time.

Second, we include dynamic visualizations of our ablation study. We compare our full model against the ablated variants discussed in Section 9. This video comparison provides a clearer perspective on how our alignment strategies function in dynamic scenarios compared to static image figures. Please refer to the attached file `demo_video.mp4`.

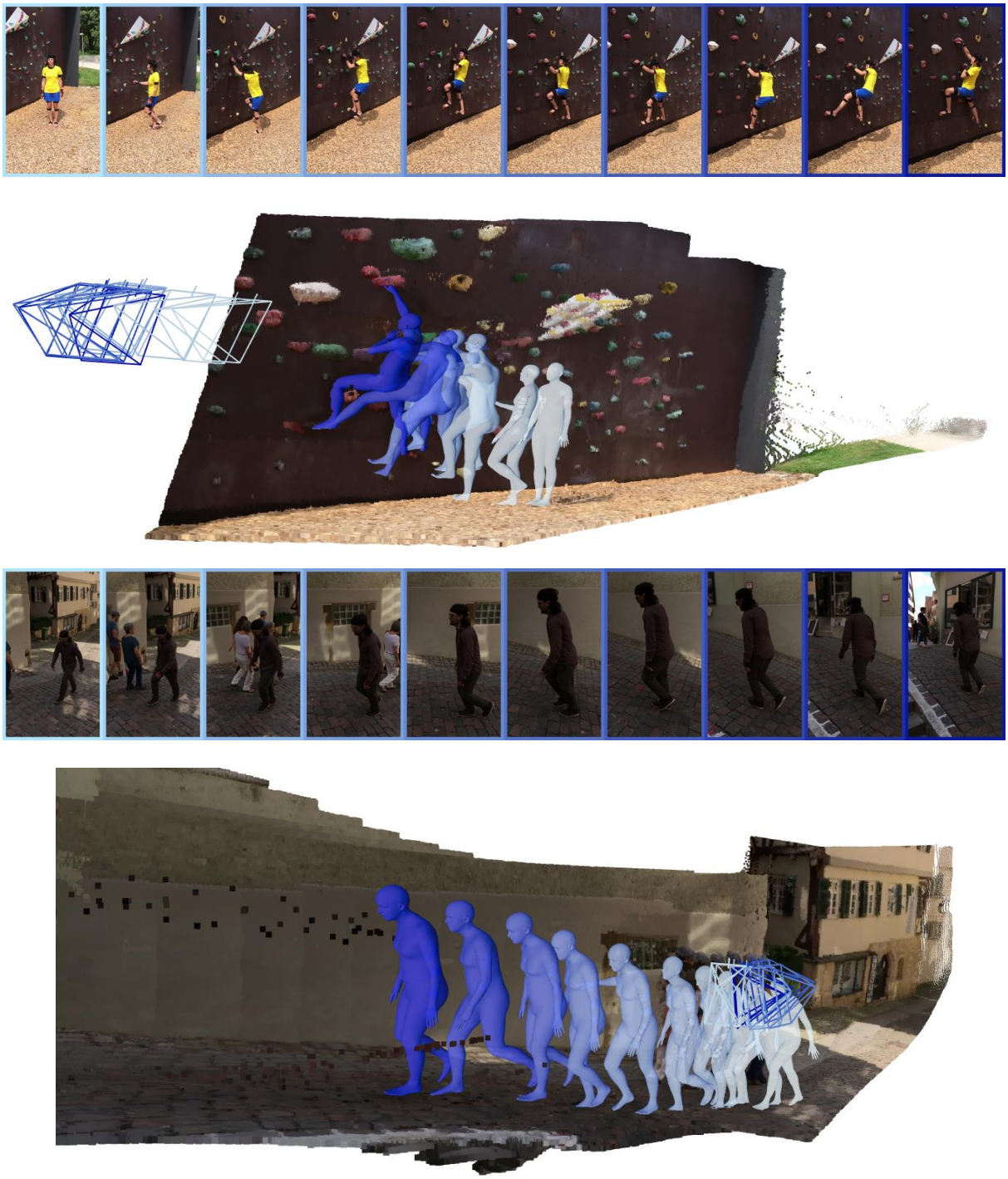

Figure 9. **Qualitative Visualization of Joint Scene and Human Reconstruction.** The examples demonstrate the robustness and metric consistency of our framework. The color gradient (light blue to dark blue) consistently encodes the temporal sequence across both the reconstructed camera poses and the SMPL meshes. The upper example illustrates robustness in reconstructing highly articulated poses (rock climbing) and accurately aligning the SMPL mesh with the scene geometry. The lower example demonstrates coherent, long-term tracking of human motion in a complex urban environment, verifying the metric stability and generalization of our joint reconstruction framework.