Mesh4D: 4D Mesh Reconstruction and Tracking from Monocular Video

Supplementary Material

In this **supplementary document**, we provide additional materials to supplement our main submission. In the **supplementary video**, we show more visual results using our method. The **code, models, benchmark splits, and evaluation framework** will be made publicly available for research purposes.

6. Implementation Details

6.1. Training details

Deformation VAE. The initial weights of our deformation VAE is loaded from HunYuan3D 2.1 shape VAE. All the last projection layers of the additional introduced modules are zero initialized, *i.e*. the skeleton injection layer and the spatio-temporal attention layer. The Deformation VAE is trained using AdamW with a learning rate of 1×10^{-5} and a batch size of 80. We set $M = 2048$ for the initial sampled aligned point cloud, and $N = 256$ for the number of point cloud after Farthest Point Sampling. The hidden dimension $c = 1024$ is set for the attention operation, and $c_0 = 64$ is set for the latent space. The weight of the KL divergence loss is set to $\lambda = 5{\times}10^{-5}$. Due to the limited computational resources, we only train our model with the frame number $T = 6$. However, thanks to our mesh representation, during animation, it is commonly to model only the key frames and do the shape interpolation between them, instead of training an interpolation model in L4GM [41] specialized for 3D-GS. In our supplementary video, we do the one frame shape interpolation between two key frames, resulting in a total 11 frames per sequence. During training, for each sample, we select 6 frames from the sequence, with the sampling stride randomly chosen from $[1, 2, 3, 4]$ to allow our model to adapt to input videos with various frame rates. Training is conducted on 4 NVIDIA H100 GPUs with a total training time of approximately 5 days.

Deformation diffusion. Our deformation diffusion model is initialized with the weights of HunYuan3D 2.1 [46] diffusion model and trained using AdamW with a learning rate of 1×10^{-5} and a batch size of 80. Similarly, we perform zero initialization to newly introduced modules, including canonical shape condition layer and temporal attention layer. The dimension of the latent feature from the HY3D ShapeVAE is 256×64. Training is conducted on 4 NVIDIA H100 GPUs with a total training time of approximately one week.

6.2. Inference details

During inference, given a monocular video of a moving object, we first segment the foreground moving object using a pre-trained model, and then resize to the same ratio as the training (90%) for condition. For the canonical shape reconstruction, we follow the same inference setting as HunYuan3D 2.1 [46], but using only 1 view (the first frame) as input. For the deformation diffusion model, we perform 50 steps of first-order Euler ordinary differential equation (ODE) to transform the sampled noise to the desire deformation latent, conditioned on the canonical shape latent and the image features from all input frames, Once we obtain the deformation latent, we reconstruct the per-vertex deformation field using the deformation decoder.

Method	Reconstruction			Tracking
	IoU ↑	P2S ↓	Chamfer ↓	ℓ_2-Corr ↓
w CFG	0.3949	0.0261	0.0243	0.0338
w/o CFG	**0.3973**	**0.0258**	**0.0238**	**0.0335**

Table 4. **Ablation study for classifier-free guidance (CFG).** The configuration without CFG obtains slightly better results.

Method	Reconstruction			Tracking
	IoU ↑	P2S ↓	Chamfer ↓	ℓ_2-Corr ↓
w/o pretrained	0.0819	0.0954	0.0854	0.2063
w pretrained	**0.3433**	**0.0327**	**0.0308**	**0.0601**

Table 5. **Quantitative evaluation** for using pretrained weights.

7. Additional Analysis

7.1. Ablation study for CFG

We perform classifier-free guidance (CFG) with the guidance weights of 5. As shown in Tab. 4, CFG does not help reconstruction quality. This observation aligns with other video-based diffusion reconstruction methods [18].

7.2. Ablation study for pretrained diffusion model

As claimed in Sec. 1, due to the limited size of existing 4D reconstruction datasets, we build our deformation diffusion model upon a pre-trained large-scale 3D generator, *i.e*. HunYuan3D 2.1 [46]. To validate this design choice, we conduct an ablation study. Specifically, we train two versions of our deformation diffusion model for the same number of iterations (3 days), one with HY3D pre-training weights and one without. The results are reported in Tab. 5. As can be seen, the model with HY3D pre-training weights outperforms the one without by a large margin in terms of all metrics, demonstrating that the pre-training weights obtained from large-scale 3D reconstruction dataset indeed benefit our deformation diffusion model.

Figure 7. **More visualization results.** The left column is two frames sampled from the input video, the others are the corresponding reconstruction results from 4 different views.

7.3. User study

We conducted a user study (3 in-the-wild and 3 synthetic sequences) with 24 people and let them rank the quality of reconstructed objects. The average ranking are 2.49/2.91/3.39/**1.20** for HY3D/L4GM/GVFD/**Ours** separately; the lower the better.

7.4. Abalation study for skeleton-free objects

The skeleton is a 'hint' that facilitates learning rigid parts, but not a prescription that only those exist. Our test dataset *does* contain skeleton-free objects like flags, flowers, in Tab. 6, we confirm that VAE trained using the skeleton still generalizes well to such objects.

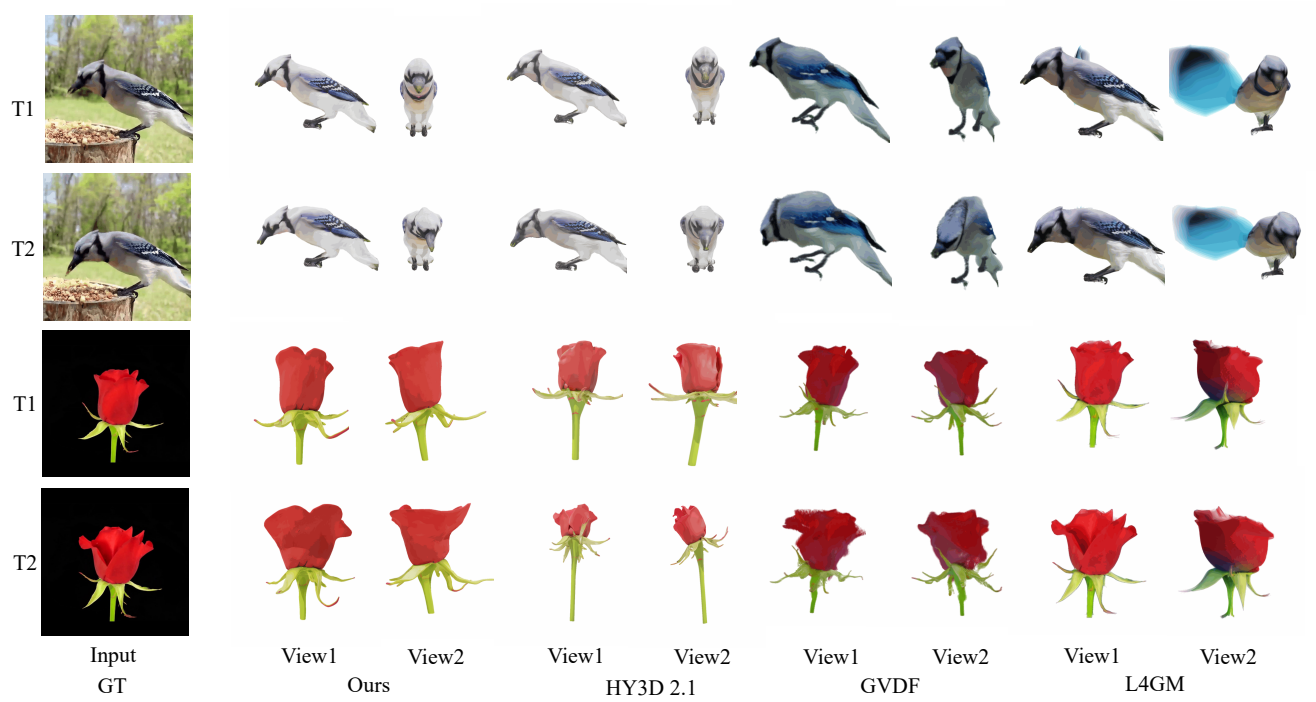

Figure 8. **In-the-wild results** from Consistent4D dataset. Our method can generalize well on in-the-wild video sequences even for skeleton-free objects, such as flowers.

Method	IoU ↑	P2S ↓	Chamfer ↓	ℓ_2-Corr ↓
VAE (trained w/o skeleton)	0.737	0.029	0.018	0.014
VAE (trained with skeleton)	**0.766**	**0.015**	**0.017**	**0.012**

Table 6. **VAE results** on objects **without** skeleton information.

7.5. Relation with mesh motion generation works

Both DriveAnyMesh [42] and AnimateAnyMesh [61] generate motion for given 3D mesh and text prompt, whereas we reconstruct 4D mesh from monocular videos. **Motion Encoding:** DriveAnyMesh encodes motion in a [*pair*]-wise manner, while we encode it in a [*sequence*]-level manner by using spatio-temporal attention. AnimateAnyMesh encodes *fixed* length trajectory into [*channel*] dimension, while we consider it as additional temporal dimension which naturally supports *flexible* length trajectory. **Architecture:** We fine-tuned our VAE+Diffusion from HY3D to utilize the 3D priors, which is verified to be useful in Tab. 5. Furthermore, none of them considers using the skeleton as an additional supervision to enhance VAE's ability to represent motion. Note that this skeleton is only used during training to help learn a better latent space, and it is not required during inference.

8. Visualization

8.1. More visualization results.

As shown in Fig. 7, our method can generalize well on various objects and motions.

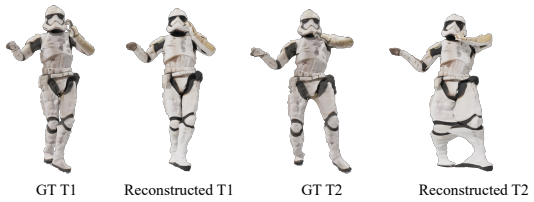

Figure 9. **Failure case of Mesh4D.**

8.2. visualization results for in-the-wild videos

we show qualitative results in Fig. 8 for in-the-wild sequences from the Consistent4D dataset. Our method can generalize well on in-the-wild video sequences even for skeleton-free objects, such as flowers.

9. Limitations

Although our method performs well and generalizes to a wide range of objects and animations, it can fail when the topology changes a lot during animation, or the method fails to reconstruct correct topology or shape for canonical mesh. As shown in Fig. 9, when the 3D reconstruction model fails to predict separate legs in the first frame, even if our model predicts the deformation field for the following frames, the topology of the subsequent mesh remains unchanged, leading to the incorrect 4D reconstruction. However, this prob-

lem can be easily solved by choosing a different frame to reconstruct the canonical shape and perform a backward and forward deformation field reconstruction. However, how to choose a good reference frame for the canonical mesh is orthogonal to our contribution and beyond the scoop of this paper.