STRUCTURE FROM DIFFUSION: TAMING VIDEO DIFFUSION MODELS FOR CAMERA POSE ESTIMATION IN DYNAMIC VIDEOS

Anonymous authors
Paper under double-blind review

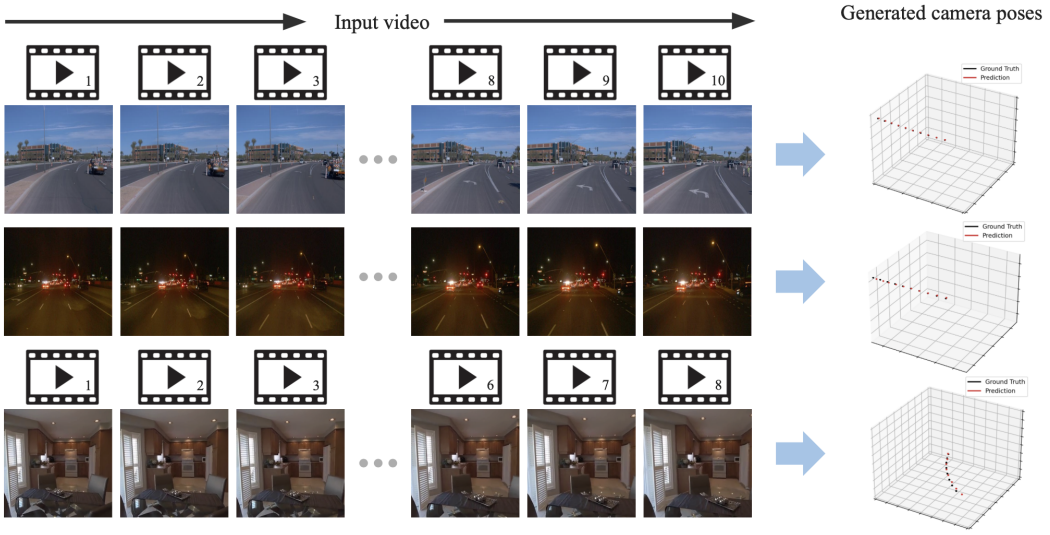


Figure 1: **SFD** is capable of generating camera poses from a video, even when the video contains dynamic objects and does not have strictly fixed frames. Our method produces accurate camera parameters and is fine-tuned on a pretrained video diffusion model, ensuring both simplicity and effectiveness in the pipeline.

ABSTRACT

Our research addresses the challenge of accurately predicting camera poses for in-the-wild dynamic videos—a task essential for applications in augmented reality, robotics, and visual perception systems. Unlike structured, lab-controlled environments, in-the-wild videos present diverse, complex scenes with significant variability in lighting, motion, and camera movement, making accurate pose estimation a persistent challenge. To tackle this, we propose a novel video diffusion model designed for camera pose prediction. Our model retargets a video generation model as a pose estimation tool by connecting a ray prediction model with a video encoder. Our model distills strong priors from pre-trained video generation models for camera motion and scene dynamics, leveraging the intrinsic temporal continuity of video features to ensure smooth and accurate pose estimation. We evaluate our approach on both dynamic and static datasets, demonstrating state-of-the-art performance. Compared to existing methods, our model achieves significant improvements in both accuracy and robustness, particularly in challenging real-world scenarios. Code will be open-sourced.

1 INTRODUCTION

Camera pose estimation in dynamic, in-the-wild videos is a crucial task for numerous applications, including augmented reality, robotics, and autonomous navigation. Traditional methods such as Structure-from-Motion (SfM) and Simultaneous Localization and Mapping (SLAM) rely on feature matching and geometric optimization. Despite achieving good performance on static scenes,

they struggle in challenging environments with dynamic objects, occlusions, and large viewpoint changes. Learning-based methods have emerged as a promising alternative, however, they often require extensive supervision and struggle with generalization to real-world scenarios.

In this paper, we propose a novel approach that leverages video diffusion models to estimate camera poses directly from dynamic video sequences. Diffusion models have demonstrated remarkable success in generative modeling for images, videos, and 3D scenes, effectively capturing complex data distributions. In particular, pre-trained video diffusion models exhibits strong scene generation capability, excelling at modeling dynamic contents with coherent motion. However, their potential for camera pose estimation has not been explored. Therefore, this paper aims to unleash the potential of video generation models—we exploit the inherent temporal coherence in videos to refine camera motion predictions, ensuring smooth and accurate pose estimation even in highly dynamic environments.

Our method, dubbed Structure from Diffusion (SFD), repurposes a video diffusion model that predicts camera poses from dynamic videos. Visual inputs are processed by a Stable Video Diffusion encoder—VQVAE—which encodes temporally coherent features that are crucial for video pose estimation. Then, we use a novel Plücker ray representation Plücker to represent camera poses, which naturally interfaces with modern deep learning frameworks. Unlike traditional pose parameterizations, Plücker rays explicitly encode geometric information and can be directly consumed by neural network architectures, significantly simplifying the pose representation pipeline. The diffusion model jointly integrates visual context extracted from a VQVAE encoder and geometric context from a dedicated ray encoder, enabling the model to exploit rich spatiotemporal correlations present in video sequences. We frame camera pose estimation as a ray-space denoising task. We introduce noise to Plücker rays and subsequently employ a diffusion-based model to predict the added noise from the integrated visual-ray features. This approach leverages the inherent spatial and temporal coherence priors available within video data, resulting in robust and accurate pose predictions.

We evaluate our method on multiple datasets, including both object-centric scenarios and challenging in-the-wild scenes. Our experiments show that SFD significantly outperforms existing approaches in terms of both accuracy and robustness. Our results suggest that video generation models provide strong priors for dynamic scene modeling and understanding. In summary, our contributions can be summarized as follows:

- We introduce a novel approach to camera pose estimation by utilizing video generation models. To our knowledge, this paper is the first to explore the potential of video generation models specifically for geometric estimation tasks. We show that a simple cross-attention layer can effectively produce spatiotemporal features from a video encoder and a geometric encoder.
- Our method exhibits strong empirical performance, with significant improvements over existing methods on multiple datasets. Our model improves over previous state-of-the-art methods by 20%. Compared to recent learning-based methods on dynamic scenes, our method achieves almost 10x performance improvement. These findings suggest that, beyond their capabilities in generating high-quality videos, video generation models inherently encode valuable priors beneficial for accurate geometric estimations.
- The proposed pipeline is general and straightforward, which can potentially tackle other similar geometric tasks with minimal modification. We will also open-source our code to facilitate reproducibility and future research in this field.

2 Related Work

Diffusion Models. Diffusion models have emerged as a powerful class of generative models, demonstrating remarkable success in image Ho et al. (2020); Song & Ermon (2019), video Ho et al. (2022); Singer et al. (2022); Blattmann et al. (2023); Rombach et al. (2022); Chen et al. (2024); Xing et al. (2024; 2025), and 3D generation Luo & Hu (2021); Melas-Kyriazi et al. (2023); Lan et al. (2025); Wang et al. (2023d). Denoising Diffusion Probabilistic Models (DDPMs) Ho et al. (2020) introduced a robust framework for iterative denoising, allowing high-quality sample generation by reversing a Gaussian noise corruption process. This was later improved by Denoising Diffusion Implicit Models (DDIMs) Song et al. (2020), which provided a more efficient non-Markovian for-

mulation, accelerating sampling while maintaining fidelity. Among notable advancements, Latent Diffusion Models (LDMs) Rombach et al. (2022) have improved computational efficiency by operating in a compressed latent space, while Stable Video Diffusion (SVD) Blattmann et al. (2023) and Sora Brooks et al. (2024) have extended diffusion-based generation to video synthesis with impressive temporal consistency. Additionally, conditional diffusion models allow control over generated outputs using auxiliary inputs such as text Saharia et al. (2022b), images Saharia et al. (2022a), and semantic maps Zhang et al. (2023), further enhancing their flexibility and applicability in various domains.

Structure-from-Motion and Simultaneous Localization and Mapping. Structure-from-Motion (SfM) and Simultaneous Localization and Mapping (SLAM) aim to recover camera poses and scene geometry from image sequences. Classical SfM methods Snavely et al. (2006); Harris et al. (1988); Tola et al. (2009) rely on detecting keypoints using handcrafted descriptors such as SIFT Lowe (1999); Low (2004) and SURF Bay et al. (2008), or learned feature extractors Yi et al. (2016), followed by feature matching and geometric verification using epipolar constraints Lucas & Kanade (1981). Camera poses are then estimated using five-point Li & Hartley (2006); Nistér (2004) or eight-point algorithms Hartley (1997), with RANSAC Fischler & Bolles (1981) and further refined via Bundle Adjustment (BA) Triggs et al. (2000). These techniques have been extended to large-scale datasets, achieving robust performance across thousands of images Furukawa et al. (2010); Sarlin et al. (2019). COLMAP Schonberger & Frahm (2016), a widely used open-source SfM and MVS framework, integrates these techniques into a robust end-to-end pipeline, making it a valuable tool for large-scale 3D reconstruction. In parallel, SLAM techniques can be categorized into indirect methods, which leverage feature correspondences Campos et al. (2021); Rosinol et al. (2020), and direct methods that optimize photometric errors Davison et al. (2007); Engel et al. (2017); Schops et al. (2019); Zubizarreta et al. (2020). While these advancements enhance robustness, both SfM and SLAM typically rely on sequential or overlapping image inputs, making them less suitable for sparse-view camera pose estimation.

Learning-Based Pose Estimation. Traditional geometric pose estimation methods struggle in sparse-view and wide-baseline settings, where few reliable correspondences can be established Choi et al. (2015). To overcome this, learning-based approaches directly predict camera motion without relying on explicit feature matching. These methods can be supervised using ground truth poses or trained unsupervisedly via photometric consistency Tang & Tan (2018); Ummenhofer et al. (2017); Zhou et al. (2017). Early works focused on object-centric or category-specific pose estimation Kehl et al. (2017); Ma et al. (2022); Wu et al. (2023; 2020), while more recent methods like RelPose Zhang et al. (2022) and RelPose++ Lin et al. (2023a) extend to category-agnostic pose estimation by refining rotation and translation separately. SparsePose Sinha et al. (2023) introduces an iterative refinement approach for pose estimation from sparse views, and methods like FORGE Jiang et al. (2024) leverage synthetic data to improve robustness. Recent diffusion-based approaches such as PoseDiffusion Wang et al. (2023b) generate camera poses through a denoising process, demonstrating promising results for wide-baseline pose estimation. Unlike methods that directly regress camera parameters, emerging techniques such as PF-LRM Wang et al. (2023c), DUSt3R Wang et al. (2024b) and MonST3R Zhang et al. (2024b) estimate sparse poses by leveraging pixel-aligned point clouds, followed by PnP optimization. These advancements push the boundaries of camera pose estimation in challenging sparse-view conditions, improving reliability and generalization.

Ray-Based Camera Parameterization. Ray-based camera models provide a flexible alternative to traditional pinhole models, particularly for complex lens systems such as fish-eye cameras Grossberg & Nayar (2001); Dunne et al. (2010). These models represent each pixel as a 3D ray, improving adaptability but often requiring complex calibration Schops et al. (2020). Recent works integrate ray-based parameterization with diffusion models for improved pose estimation. The Cameras as Rays approach treats cameras as ray bundles, leveraging denoising diffusion to sample plausible poses in sparse-view settings, achieving state-of-the-art performance Zhang et al. (2024a). These advances demonstrate the potential of ray-based diffusion models for accurate and scalable pose estimation, especially in wide-baseline and sparse-view scenarios.

3 METHOD

Given a set of unposed images, our goal is to recover camera poses and 3D scene geometry by leveraging video diffusion models. To achieve this, we fine-tune a Stable Video Diffusion (SVD) model Blattmann et al. (2023) to take a video as input and predict its corresponding camera poses. Our method, Structure from Diffusion (SFD), learns to infer camera motion by denoising Plücker ray representations, which encode the 6-degree-of-freedom (6-DoF) camera pose. This representation seamlessly integrates both position and orientation, allowing for robust and consistent pose estimation across diverse scenes. The encoded noisy ray embeddings, together with image features, are fed into a pre-trained SVD backbone to predict the responding ray-space noises. After denoising, the predicted Plücker rays can be converted into conventional global camera extrinsics and intrinsics, enabling direct camera pose recovery. Section 3.1 reviews the Plücker ray embedding and its relevance to our approach, while Section 3.2 details our model architecture and training pipeline.

3.1 PRELIMINARIES

Plücker Ray Representation. In 3D computer vision, cameras are parameterized using extrinsics, which describe the global transformation of an entire frame. Despite camera extrinsics provide a compact representation, they are not well compatible with modern deep learning architectures. Therefore, we opt for a Plücker ray representation, which is a dense grid parameterization of camera poses. The Plücker coordinate system provides a six-dimensional embedding of a ray that enables robust geometric computations, including efficient intersection tests and transformations. A Plücker ray is represented by a 6D vector per pixel/ray: $\mathbf{p} = \begin{bmatrix} \mathbf{d} \\ \mathbf{m} \end{bmatrix}$, where $\mathbf{d} \in \mathbb{R}^3$ is the direction vector of the ray and $\mathbf{m} = \mathbf{o} \times \mathbf{d} \in \mathbb{R}^3$ is the moment of the ray, encoding the relationship between the ray's direction and its position in space.

Camera Poses to Rays. To obtain a camera ray parameterization, our method starts with with intrinsic parameters K and extrinsic parameters (R, T). The extrinsic parameters define a camera-to-world transformation: $\mathbf{x}_{\text{world}} = R\mathbf{x}_{\text{cam}} + T$. For each pixel (u, v), we compute the corresponding 3D ray in world coordinates as follows. First, we convert the pixel coordinates to normalized camera coordinates using the camera intrinsics K: $\mathbf{x}_{\text{cam}} = K^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}$.

This provides the direction of the ray in the camera's coordinate system. Next, we transform this direction into world space using the camera extrinsics. The ray's origin is simply given by the camera translation: $\mathbf{o} = T$.

The direction of the ray in world coordinates is obtained by applying the camera rotation: $\mathbf{d} = R\mathbf{x}_{\text{cam}}$.

Finally, the Plücker coordinates for the ray are constructed as: $\mathbf{p} = \begin{bmatrix} \mathbf{d} \\ \mathbf{o} \times \mathbf{d} \end{bmatrix}$.

Rays to Camera Poses. Now that we have established how to covert camera poses into parameterized rays, we are equipped to discuss how to covert rays back to the camera extrinsics. Given a set of rays parameterized in Plücker coordinates $\mathbf{p} = (\mathbf{d}, \mathbf{m})$, we can infer camera parameters under certain constraints. The camera origin T can be recovered using: $\mathbf{o} = \frac{\mathbf{m} \times \mathbf{d}}{\|\mathbf{d}\|^2}$. The rotation is estimated by computing an optimal transformation from a canonical ray to the predicted direction $\mathbf{d}$.

If multiple rays are available, pose estimation techniques, such as Perspective-n-Point (PnP), can be used to estimate the optimal rotation R and translation T that best align the rays with their corresponding image coordinates.

3.2 STRUCTURE FROM DIFFUSION

Overview. While existing methods are adept at pose estimation for static scenes, they struggle with dynamic scenes due to challenging consistency issues. To tackle that, we opt to leverage video

diffusion model priors for consistency pose estimation. We repurpose a video generation model to a pose estimation model. In particular, we leverage a pretrained Stable Video Diffusion (SVD) model Blattmann et al. (2023) to enhance temporal consistency in camera pose generation from video. Our method, following SVD, operates via a two-stage process: first, a stochastic forward process injects Gaussian noise at a specific noise level into the rays. Then, a reverse process is applied to progressively remove this noise using a learnable denoiser D_θ, conditioning on image features.

As shown in Figure 2, given an input video, SFD aims to generate the corresponding camera pose for each frame. Our model begins by encoding the input video using a pretrained VQVAE van den Oord et al. (2017) encoder, yielding a video latent representation z_{video}. Simultaneously, the pose is represented using Plücker rays, capturing both positional and directional information in a 6D format. Gaussian noise is then added to these rays, producing a noised latent representation z_{rays}. Note that the original VAE cannot directly encode the Plücker rays. Therefore, we deisgn a simple ray encoder that converts the 6-dimensional noised ray vector into ray embeddings. Next, the noised Plücker ray embeddings are concatenated with the video latent z_{video}, forming a joint latent representation. This concatenated latent is passed through the model, which directly denoises the rays. During inference, given a video clip, our model generates the camera pose corresponding to each frame.

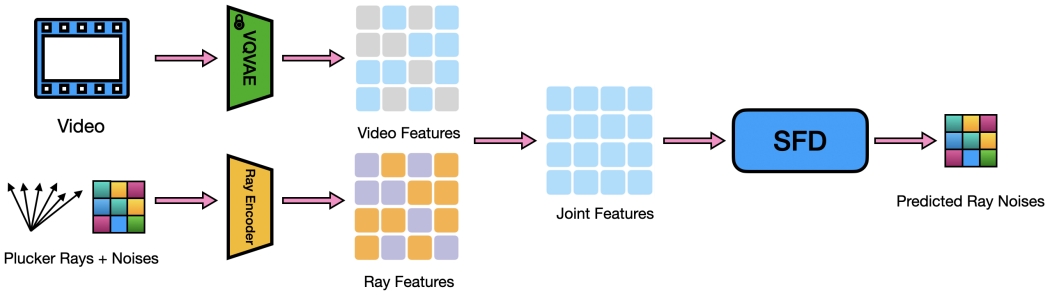


Figure 2: The overall architecture.

Training. We follow standard diffusion model training protocols to train our method. Our method takes an input F-frame RGB video sequence $\mathbf{x} \in \mathbb{R}^{F\times W\times H\times 3}$ along with its corresponding ray representation $\mathbf{r} \in \mathbb{R}^{F\times W\times H\times 6}$. At each training step, a noise level σ_t is randomly sampled for the entire sequence, following a normal distribution $\log \sigma_t \sim \mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2)$ Karras et al. (2022), where $P_{\text{mean}} = 0.7$ and $P_{\text{std}} = 1.6$.

Then, Gaussian noises, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, are applied directly to the ray representation $\mathbf{r}$, producing a noisy version $\mathbf{r}_t$ given by: $\mathbf{r}_t = \mathbf{r} + \sigma_t^2 \boldsymbol{\epsilon}$.

During the reverse process, the model learns a denoiser D_θ to estimate the clean ray representation $\hat{\mathbf{r}}_0$: $\hat{\mathbf{r}}_0 = D_\theta(\mathbf{r}_t; \sigma_t, \mathbf{z}^{(\mathbf{x})})$.

This process is optimized using the denoising score matching (DSM) objective Ho et al. (2020): $\mathcal{L} = \mathbb{E}_{\mathbf{r},\mathbf{z}^{(\mathbf{x})},\sigma_t}\left[\lambda(\sigma_t)\|\hat{\mathbf{r}}_0 - \mathbf{r}\|_2^2\right]$,

where the weighting function is defined as $\lambda(\sigma) = (1+\sigma^2)\sigma^{-2}$. Following EDM Hoogeboom et al. (2022), we represent the denoiser D_θ using the following formulation: $D_\theta(\mathbf{r}_t; \sigma_t, \mathbf{z}^{(\mathbf{x})}) = c_{\text{skip}}(\sigma_t)\mathbf{r}_t + c_{\text{out}}(\sigma_t)F_\theta(c_{\text{in}}(\sigma_t)\mathbf{r}_t, c_{\text{noise}}(\sigma_t), \mathbf{z}^{(\mathbf{x})})$,

where F_θ is a trainable U-Net, and the functions c_{skip}, c_{out}, c_{in}, and c_{noise} act as preconditioning mechanisms. In this formulation, the RGB video latent $\mathbf{z}^{(\mathbf{x})}$ serves as the conditioning input. It is incorporated into the model by concatenating it with the ray representation along the feature dimension.

Inference. At inference time, the ray sequence $\hat{\mathbf{r}}_0$ is reconstructed by initializing from a Gaussian noise sample $\hat{\mathbf{r}}_T \sim \mathcal{N}(\mathbf{0}, \sigma_T^2\mathbf{I})$. The reconstruction is performed iteratively through a denoising process, where the model is conditioned on the RGB video features and guided by the trained denoiser D_θ to produce $\tilde{\mathbf{r}}_0 = D_\theta(\mathbf{r}_t; \sigma_t, \mathbf{z}^{(\mathbf{x})})$, where $\hat{\mathbf{r}}_{t-1} = \frac{\hat{\mathbf{r}}_t - \tilde{\mathbf{r}}_0}{\sigma_t}(\sigma_{t-1} - \sigma_t) + \hat{\mathbf{r}}_t, \quad 0 < t \leq T$,

where $\sigma_0, ..., \sigma_T$ follow a predetermined variance schedule over T denoising steps.

Table 1: Evaluation on Static Scenes

	Co3D		Realestate10K	
	Rot ↓	CC ↓	Rot ↓	CC ↓
Colmap Schonberger & Frahm (2016)	33.6	0.934	45.0	0.0708
Dust3r Wang et al. (2024a)	40.7	0.287	2.14	0.0503
Monst3r Zhang et al. (2024b)	39.7	0.287	2.17	0.0505
Relpose++ Lin et al. (2023b)	22.0	0.1346	-	-
PoseDiffsion w/o GGS Wang et al. (2023a)	31.6	0.1842	1.90	0.2814
PoseDiffsion Wang et al. (2023a)	29.1	0.1650	1.38	0.2580
RayDiffsion Zhang et al. (2024a)	13.3	0.0979	0.645	0.0049
Ours	**10.6**	**0.096**	**0.557**	**0.0041**

Table 2: Evaluation on Dynamic Scenes

	Waymo		Argoverse	
	Rot ↓	CC ↓	Rot ↓	CC ↓
Colmap Schonberger & Frahm (2016)	Fail	Fail	Fail	Fail
Dust3r	12.9	0.971	4.31	6.01
Monst3r Zhang et al. (2024b)	4.15	0.2517	1.26	3.5039
Ours	**0.30**	**0.094**	**0.38**	**0.092**

4 EXPERIMENTS

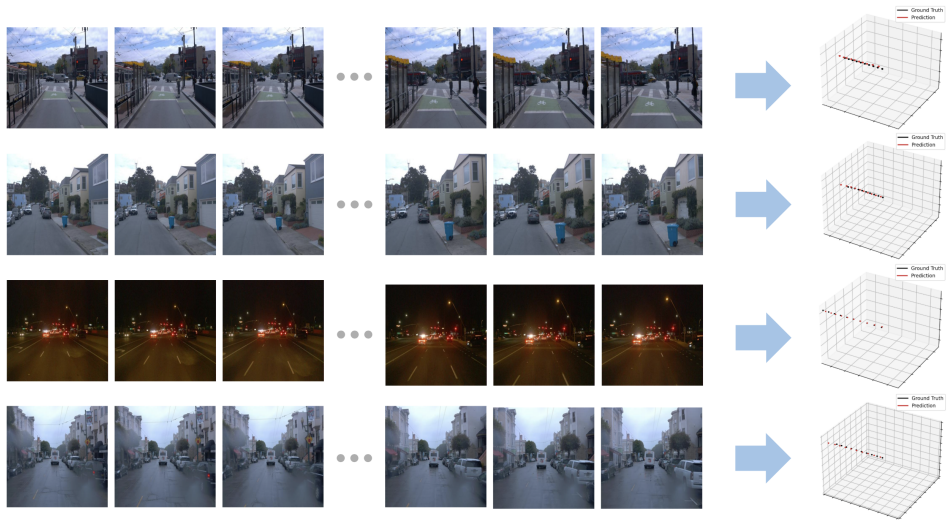


Figure 3: Qualitative results for Waymo.

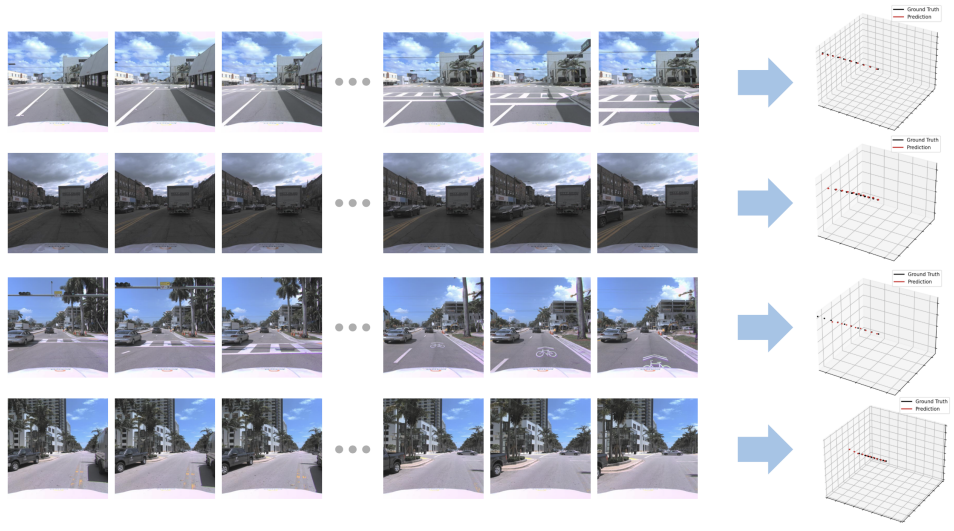


Figure 4: Qualitative results for Argoverse.

4.1 EXPERIMENTAL SETUP

Datasets. We quantitatively evaluate our model on 4 datasets, covering a diverse range of scenarios, including object-centric and scene-level settings, as well as dynamic and static environments.

RealEstate10K Zhou et al. (2018) consists of videos primarily sourced from YouTube, containing both indoor and outdoor scenes, and the per-frame poses are computed using structure-from-motion

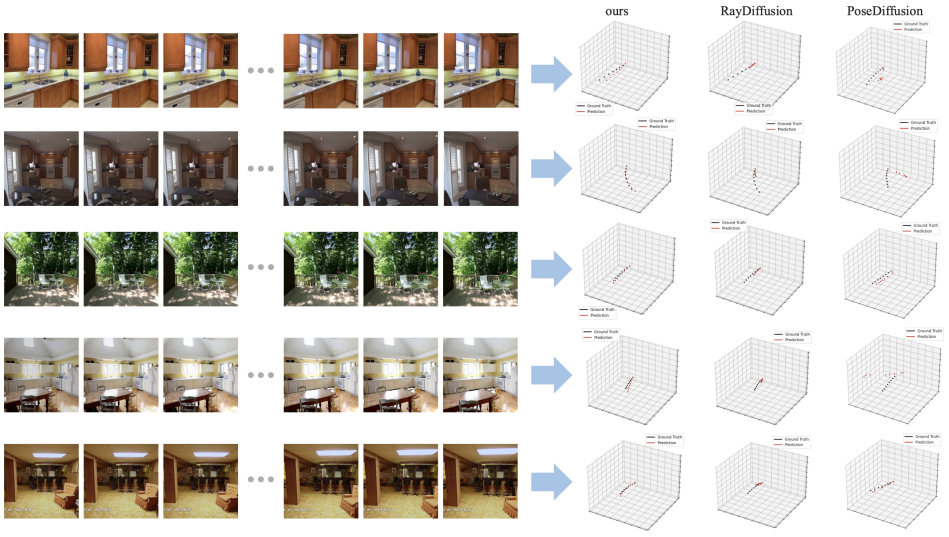


Figure 5: Qualitative results for Realestate10K.

(Colmap Schonberger & Frahm (2016)). In our experiments, we use totally 63,214 training videos, and videos that contain more than 200 frames for evaluation, resulting 1,600 test scenes. During inference, we randomly select 500 scenes for evaluation. We extract images at the given frame rate, center-crop them, and resize these extracted images to 256×256. Other datasets also use the same image resolution.

CO3Dv2 Reizenstein et al. (2021) is an object-centric dataset that consists of 51 categories of objects. Each frame is annotated with poses estimated using COLMAP Schonberger & Frahm (2016). In line with Raydiffusion Zhang et al. (2024a), we use 41 categories for training while the remaining 10 for testing.

Waymo Sun et al. (2020) contains 1,000 sequences of driving logs: 798 sequences for training and 202 for validation. Each sequence consists of a 20-second video recorded at 10 FPS, and each frame includes three cameras. For our experiments, we sample 10 frames per sequence by selecting evenly spaced frames to ensure a diverse representation of motion while maintaining temporal consistency.

Argoverse2 Wilson et al. (2023) contains 1,000 driving scenes, split into 700 for training, 150 for validation, and 150 for testing. Similar to Waymo, we sample 10 evenly spaced frames from each sequence.

Baselines. We evaluate our method against three approaches: Structure from Motion, SLAM-based methods, and Diffusion model-based methods. For static scenes, we compare with RelPose++ Lin et al. (2023b), PoseDiffusion Wang et al. (2023a), Dustr Wang et al. (2024a) and RayDiffusion Zhang et al. (2024a). For dynamic scenes, we compare primarily with a recent state-of-the-art method Monst3r Zhang et al. (2024b). And all datasets use COLMAP as a baseline of Structure from Motion.

4.2 Implementation Details

In our study, we optimized the training process by building on top of the SVD implementation, an enhanced version of the original SVD. For data processing, we utilize VQVAE to pre-extract features from images, which are resized and center-cropped to 256×256 pixels. Post VQ-VAE encoding, these images are encoded into features of $32 \times 32 \times 4$.

During training, we randomly select a sequence of images as the conditioning input. Given a video clip of shape $(b, f, 3, h, w)$, we encode it using pretrained VQVAE, obtaining a video latent representation of shape $(b, f, 4, 32, 32)$. Simultaneously, we compute the ray map for the same video clip, resulting in a tensor of shape $(b, f, 6, h, w)$. Noise is first added to the ray map. The noised ray, after being encoded by an additional encoder, is then concatenated with the video latent representation. This combined input is fed into the SVD model, which is a Spatiotemporal U-Net, to predict

the noise. We leverage the SVD model that was pretrained on 14-frames and fine-tune all layers to optimize performance.

Regarding how we encode camera poses, we treat the first frame (frame 0) of the video as the identity pose. Each subsequent frame's camera pose is represented relatively to the first frame, ensuring that the first camera is positioned at the origin (0, 0, 0) with an identity rotation and zero translation. In the visualization, the central camera position corresponds to (0, 0, 0), which represents frame 0.

Metrics. We evaluate the discrepancy between predicted and ground-truth camera poses. RayDiffusion Zhang et al. (2024a) reports the proportion of camera centers within 0.1 scene scale units for translation and the proportion of relative rotations within 15° for rotation. However, we find such thresholded proportions to be coarse and potentially misleading. Instead, we report the relative rotation error and translation error directly, providing a more precise and continuous evaluation. For the Waymo and Argoverse2 datasets, we use 10 frames per sequence; for Real10K and CO3D, we follow prior work and use 8 frames. Since consecutive frames in CO3D and Real10K exhibit limited motion, we sample clips using a 5-frame interval to ensure meaningful pose variation.

For static datasets, we compare our method with traditional SfM approaches, SLAM-based camera pose estimation methods, and diffusion model based methods. As shown in Table 1, our approach achieves the lowest rotation and translation errors among all methods.

For dynamic datasets, since diffusion model-based methods lack publicly available code and fail to produce results in this dynamic setting, we compare our approach against traditional SfM and a recent learning-based method Monstr. Table 2 suggests that Colmap fails all testing cases, highlighting the difficulty of dynamic scene pose estimation. Our method significantly outperforms Monst3r, demonstrating superior performance in dynamic scenes as well.

Quantitive Results. Table 1 and Table 2 presents the quantitative comparison of our method with the baseline approaches. Our method significantly outperforms all baselines, suggesting that video diffusion models provide strong geometric priors for pose estimation.

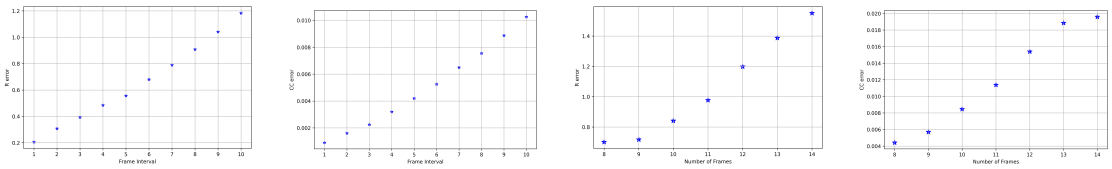
Visualization. Our qualitative results are presented in Figure 3 and Figure 4 for autonomous driving scenarios. The datasets feature highly dynamic scenes with elements such as moving pedestrians and vehicles, as well as environmental variations like changing weather conditions and day-night transitions. Despite these complexities, our model consistently maintains its robustness.

Figure 5 provides a visualization comparing our method with other diffusion model-based approaches on the Realestate10K dataset. To ensure an equitable comparison, we fine-tuned these alternative models on the dataset and evaluated them using an 8-frame video clip, identical to our own testing conditions. The visualization reveals that competing models exhibit difficulty in effectively capturing the intrinsic coherence of video data, resulting in large translation errors. In contrast, our approach produces camera poses that are demonstrably smoother and more consistent.

5 Ablation

Frame Interval Varies. As discussed in the Metrics section, datasets such as CO3D and Real10K exhibit relatively minor inter-frame motion. Consequently, our evaluation protocol for video clips utilizes a 5-frame interval. To investigate the impact of this parameter, Figure 6 (a-b) presents an ablation study illustrating performance on the Real10K dataset across varying frame intervals. It is observed that as the frame interval increases from 1 to 10, the diminishing spatio-temporal overlap between frames correspondingly elevates the difficulty of camera pose prediction, thereby leading to a discernible decline in performance. The selection of a 5-frame interval for evaluation purposes is thus predicated on achieving an optimal balance, ensuring that the captured motion is sufficiently pronounced for robust analysis without becoming excessively large, which could introduce confounding factors.

Video Clip Length. For the Co3D and RealEstate10K datasets, results are reported using 8-frame sequences to maintain methodological consistency and ensure a fair comparison with established baseline methods. It is pertinent to note, however, that our proposed model is not constrained to a fixed input of 8 frames; its architecture is inherently designed to accommodate sequences of varying lengths. To investigate the influence of frame count on camera pose estimation accuracy, a series of experiments were conducted under multiple frame number settings. Due to computational resource



(a) (b) (c) (d)

Figure 6: In the RealEstate10K dataset, we show how the rotation error and the translation error increases as the frame interval increases (a-b) and as the number of frames increases (c-d).

limitations, this analysis was performed on frame sequences ranging from 8 to 14 frames. The study was executed on the RealEstate10K dataset, and the empirical results, presented in Figure 6 (c-d), indicate a marginal decrease in performance as the number of frames increases within this tested range.

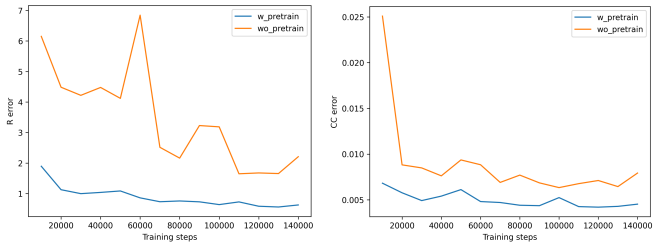


Figure 7: In the RealEstate10K dataset, we ablate whether incorporating priors from video generation models facilitates more robust pose estimation.

Pretrained weights. Finally, a comparative analysis was conducted between model variants initialized with pretrained weights from SVD and those trained from random initialization (i.e., without pretrained weights). The empirical results, illustrated in Figure 7, demonstrate that the model leveraging pretrained weights consistently outperforms its counterpart. This finding substantiates the hypothesis that incorporating priors from video generation models facilitates more robust pose estimation.

6 Conclusion

Limitations. Our model requires the input to be a video, as it is built on top of a pretrained video diffusion model. Datasets like CO3D are more akin to multi-view, object-centric image collections rather than true videos. Additionally, for autonomous driving datasets, we use frames captured at different timestamps from the same camera. However, if frames are taken from different cameras and concatenated, they do not form a real video. While our model can still work in such cases, the error tends to be higher compared to video-based data.

Summary. In this paper, we introduced Structure from Diffusion (SFD), a novel framework leveraging video diffusion models for camera pose estimation in dynamic video scenes. By utilizing Plücker ray representations, our approach ensures stable and geometrically consistent pose predictions. Through extensive experiments, we demonstrated that SFD outperforms existing methods on both static and dynamic datasets, achieving higher accuracy and robustness in real-world scenarios.

Our findings highlight the potential of diffusion models for video-based camera pose estimation, offering a promising alternative to traditional learning-based and geometric methods. Future work will explore improvements in computational efficiency, the integration of additional geometric constraints, and the extension of our approach to more complex, multi-camera settings.

References

Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. 2024. *URL https://openai. com/research/video-generation-models-as-world-simulators*, 3, 2024.

Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024.

Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. Robust reconstruction of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5556–5565, 2015.

Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1052–1067, 2007.

Aubrey K Dunne, John Mallon, and Paul F Whelan. Efficient generic calibration method for general cameras with single centre of projection. *Computer Vision and Image Understanding*, 114(2): 220–233, 2010.

Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. Towards internet-scale multi-view stereo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 1434–1441. IEEE, 2010.

Michael D Grossberg and Shree K Nayar. A general imaging model and a method for finding its parameters. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pp. 108–115. IEEE, 2001.

Chris Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pp. 10–5244. Citeseer, 1988.

Richard I Hartley. In defense of the eight-point algorithm. *IEEE Transactions on pattern analysis and machine intelligence*, 19(6):580–593, 1997.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8867–8887. PMLR, 2022. URL `https://proceedings.mlr.press/v162/hoogeboom22a.html`.

Hanwen Jiang, Zhenyu Jiang, Kristen Grauman, and Yuke Zhu. Few-view object reconstruction with unknown categories and camera poses. In *2024 International Conference on 3D Vision (3DV)*, pp. 31–41. IEEE, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the IEEE international conference on computer vision*, pp. 1521–1529, 2017.

Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy. Ln3diff: Scalable latent neural fields diffusion for speedy 3d generation. In *European Conference on Computer Vision*, pp. 112–130. Springer, 2025.

Hongdong Li and Richard Hartley. Five-point motion estimation made easy. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 1, pp. 630–633. IEEE, 2006.

Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. *arXiv preprint arXiv:2305.04926*, 2023a.

Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. *arXiv preprint arXiv:2305.04926*, 2023b.

David G Low. Distinctive image features from scale-invariant keypoints. *Journal of Computer Vision*, 60(2):91–110, 2004.

David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pp. 1150–1157. Ieee, 1999.

Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, volume 2, pp. 674–679, 1981.

Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2837–2845, 2021.

Wei-Chiu Ma, Anqi Joyce Yang, Shenlong Wang, Raquel Urtasun, and Antonio Torralba. Virtual correspondence: Humans as a cue for extreme-view geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15924–15934, 2022.

Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12923–12932, 2023.

David Nistér. An efficient solution to the five-point relative pose problem. *IEEE transactions on pattern analysis and machine intelligence*, 26(6):756–770, 2004.

Julius Plücker. Analytisch-geometrische entwicklungen. URL `https://api.semanticscholar.org/CorpusID:122419536`.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotný. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 10881–10891. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01072. URL `https://doi.org/10.1109/ICCV48922.2021.01072`.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1689–1696. IEEE, 2020.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022a.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022b.

Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12716–12725, 2019.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

Thomas Schops, Torsten Sattler, and Marc Pollefeys. Bad slam: Bundle adjusted direct rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 134–144, 2019.

Thomas Schops, Viktor Larsson, Marc Pollefeys, and Torsten Sattler. Why having 10,000 parameters in your camera model is better than twelve. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2535–2544, 2020.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Samarth Sinha, Jason Y Zhang, Andrea Tagliasacchi, Igor Gilitschenski, and David B Lindell. Sparsepose: Sparse-view camera pose regression and refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21349–21359, 2023.

Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pp. 835–846. 2006.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020.

Chengzhou Tang and Ping Tan. Ba-net: Dense bundle adjustment network. *arXiv preprint arXiv:1806.04807*, 2018.

Engin Tola, Vincent Lepetit, and Pascal Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):815–830, 2009.

Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings*, pp. 298–372. Springer, 2000.

Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5038–5047, 2017.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6306–6315, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html.

Jianyuan Wang, Christian Rupprecht, and David Novotný. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 9739–9749. IEEE, 2023a. doi: 10.1109/ICCV51070.2023.00896. URL https://doi.org/10.1109/ICCV51070.2023.00896.

Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9773–9783, 2023b.

Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023c.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024a.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024b.

Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4563–4573, 2023d.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023.

Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1–10, 2020.

Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8792–8802, 2023.

Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Tooncrafter: Generative cartoon interpolation. *ACM Transactions on Graphics (TOG)*, 43 (6):1–11, 2024.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pp. 399–417. Springer, 2025.

Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pp. 467–483. Springer, 2016.

Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *European Conference on Computer Vision*, pp. 592–611. Springer, 2022.

Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. *arXiv preprint arXiv:2402.14817*, 2024a.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arxiv:2410.03825*, 2024b.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1851–1858, 2017.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4):65, 2018. doi: 10.1145/3197517.3201323. URL https://doi.org/10.1145/3197517.3201323.

Jon Zubizarreta, Iker Aguinaga, and Jose Maria Martinez Montiel. Direct sparse mapping. *IEEE Transactions on Robotics*, 36(4):1363–1370, 2020.

A Appendix

You may include other additional sections here.