# Landscape-Awareness for Geometric View Diffusion Model

Yan-Ting Chen*, Hao-Wei Chen*, Tsu-Ching Hsiao, and Chun-Yi Lee
Elsa Lab, National Taiwan University
r13922027@g.ntu.edu.tw, d13922023@csie.ntu.edu.tw,
joehsiao.x@gmail.com, cylee@csie.ntu.edu.tw
*Equal contribution

## Abstract

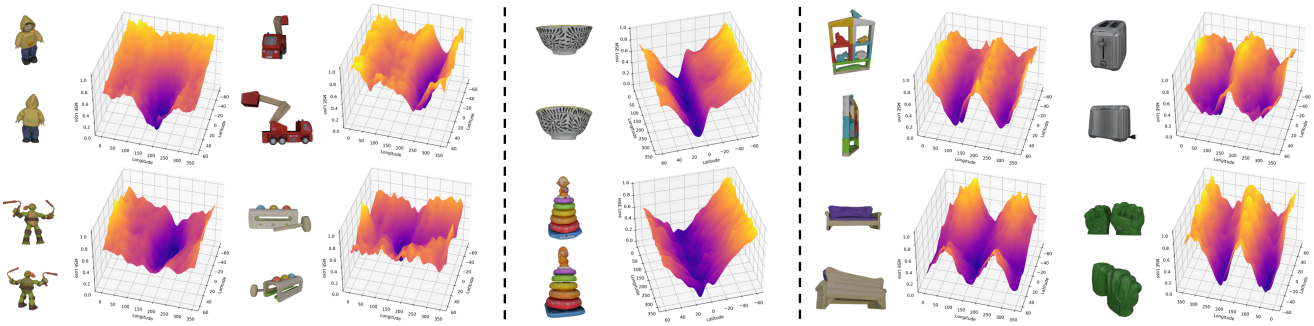
*Accurate camera viewpoint estimation under sparse-view conditions remains challenging, particularly in two-view scenarios. Recent approaches leverage diffusion models such as Zero123 to synthesize novel views conditioned on relative viewpoint, showing promising results when repurposed for viewpoint estimation via optimization with MSE loss. However, existing methods often suffer from non-convex loss landscape with numerous local minima, making them sensitive to initialization and reliant on naïve multi-start strategies. We analyze these optimization challenges and visualize failure cases, showing that geometric ambiguities, such as symmetry and self-similarity, can mislead gradient-based updates toward incorrect viewpoints. To address these limitations, we propose a score-based method that reshapes the optimization landscape to guide updates toward the ground-truth viewpoint, followed by a refinement stage using a viewpoint-conditioned diffusion model. Experiments show that our method improves convergence, reduces reliance on brute-force sampling, and achieves competitive accuracy with higher sample-efficiency.*

## 1. Introduction

Camera pose estimation constitutes a fundamental component in a wide range of applications, including robotics [62], structure from motion [45, 58], visual SLAM [12, 33], augmented reality and virtual reality [3, 37, 61], and 3D reconstruction [20, 32, 67]. Traditional methods typically establish pose estimation through feature correspondence, using either hand-crafted features [2, 24, 30] or learned features [8, 39] to extract and match keypoints across images [4, 44]. While feature-based approaches demonstrate strong performance in dense-view settings with sufficient overlap, they frequently fail in sparse-view scenarios where substantial viewpoint differences result in unreliable correspondences. To address these limitations, many recent works [26, 36, 47, 51, 61, 64] have directed their attention toward sparse-view scenarios. These methods diverge from the conventional pipeline and adopt data-driven techniques to learn geometric priors of objects. Nevertheless, contemporary approaches continue to face substantial challenges in highly sparse two-view settings. In such scenarios, large viewpoint differences lead to minimal feature overlap and introduce geometric ambiguities on occluded sides of objects, which compromise the reliability of pose estimation.

Since feature correspondence-based approaches often fail in sparse-view scenarios due to unreliable matches, recent methods leverage diffusion models [16, 49, 50], which can model complex image distributions while conditioning on modalities [63], such as text [21, 42, 65], mask [7], and pose [28, 46]. A representative example is Zero123 [28], which generates novel views from a reference image given a target relative pose parameterized in spherical coordinates. Its pose-conditioned generation has enabled applications such as novel view synthesis, image-to-3D generation [27–29, 59, 66], and pose estimation [6, 59, 66]. Building on this, methods like ID-Pose [6] and iFusion [59] reformulate pose estimation as an inverse problem by optimizing the pose via mean squared error (MSE) in the diffusion noise space. Given a reference and query image, they first compute the MSE in the noise space and backpropagate gradients with respect to the conditioned pose. Subsequently, they apply gradient descent to optimize the estimated pose such that the generated view closely aligns with the query image. These approaches demonstrate strong performance even under large viewpoint differences due to the robust generative capabilities of the underlying diffusion models.

Leveraging a pretrained pose-conditioned diffusion model for camera pose estimation offers two advantages. First, it exploits the generative capability of Zero123 to enable extreme two-view estimation. Second, unlike traditional energy-based methods [14, 17, 26, 34, 64], which rely on brute-force sampling to find the optimal pose, these approaches utilize the MSE loss computed in Zero123's noise space as an energy function, enabling direct gradient-based optimization of the pose. For instance, RelPose [64] sam-



(a) Single Minimum (b) Plateau (c) Multiple Local Minima

Figure 1. **3D MSE Loss Landscape.** Each object is associated with two views, serving as the reference image and the query image. The coordinates follow a spherical system, where the x- and y-axes denote latitude and longitude, and the z-axis indicates the normalized MSE magnitude. The generation procedure is detailed in Appendix D. (a) Some landscapes exhibit a single clear minimum, enabling gradient descent to reach the global optimum easily. (b) Some display a plateau along the longitudinal direction, reflecting continuous symmetry. (c) Others contain two distinct local minima. In the latter two cases, iFusion's optimization process is prone to getting stuck in a local minima, preventing convergence to the optimal solution. Additional visualizations of MSE loss landscapes are provided in Appendix D.

ples up to 50,000 candidate poses, which shows significant computational inefficiency. Approaches training via likelihood maximization often suffer from non-smooth energy landscapes, making gradient-based optimization unstable and requiring extensive sampling. In contrast, leveraging a pretrained pose-conditioned diffusion model as an energy function provides smoother gradients and supports end-to-end optimization. Despite this improvement, these methods still require multiple initializations to avoid convergence to incorrect viewpoints. This suggests that while the Zero123 noise-space MSE smooths the optimization landscape, we conjecture the presence of local minima on the landscape.

To examine the hypothesis that the optimization landscape contains local minima, Fig. 1 provides a visualization of this landscape. Specifically, the conditioned pose in Zero123 is varied while keeping the input image pair fixed and computing the corresponding loss defined in Eq. (2). The visualizations reveal distinct landscape characteristics across different objects. As shown in Fig. 1 (a), certain objects exhibit smooth surfaces with a single dominant basin, allowing gradient descent to consistently converge to the correct pose. In contrast, other examples in Figs. 1 (b) and (c) display multiple valleys and extended plateaus, indicating the presence of local minima that hinder convergence. To further analyze this phenomenon, Fig. 2 (d) presents the optimization trajectories of iFusion initialized from four distinct azimuth angles (i.e., $0°$, $90°$, $180°$, and $270°$), overlaid on the corresponding 2D MSE loss landscape. The visualization reveals how different starting points lead to different minima. Once a trajectory reaches a local minimum, it typically stagnates. These local minima often arise due to geometric ambiguities in the object. For instance, in Fig. 2 (a), the object exhibits symmetry between its front and back sides despite textural differences. This geometric characteristic gives rise to two deep valleys in the loss landscape, located 180 degrees apart in azimuth. Finally, Fig. 2 (c) illustrates the effect of initialization by presenting images generated from Zero123 with poses along the optimization trajectories. This highlights the substantial impact of initial conditions on the quality of the final outcome.

To address the local minima issue caused by object-dependent geometric ambiguities, we introduce a score-based model that guides the optimization toward regions of high data likelihood. This motivates our two-stage optimization framework. In the first stage, a score-based model is trained to learn the score of the data distribution. This provides guidance that steers the optimization away from local minima. Having escaped poor local minima through the first stage, the second stage employs the pretrained diffusion model with an MSE loss to further refine the pose estimate. By employing these complementary stages, the framework significantly mitigates, or even eliminates, the need for multiple initialization points, which in turn improves the sample efficiency of the gradient-based solver. Moreover, since the primary goal involves reshaping the loss landscape, we further investigate the energy modeling approach. This approach trains the model to learn an energy function that represents the data distribution. After training, the score is obtained through differentiation of the energy, and optimization is performed via gradient descent. In the experiments, the score-based formulation demonstrates superior convergence behavior and competitive accuracy compared to state-of-the-art (SoTA) methods while requiring fewer samples and reducing inference time. The contributions of this work are summarized as follows:

- We introduce the *landscape perspective* for analyzing two-view pose estimation. In contrast to prior methods such as energy-based approaches and iFusion, we present the first systematic study examining how optimization landscapes essentially affect the pose estimation process.
- We propose a score-based method that learns to fundamentally reshape the optimization landscape and gradient field. This approach effectively mitigates local minima without resorting to dense multi-initialization strategies.
- We conduct comprehensive comparisons with SoTA approaches, achieving performance on par with existing ones while requiring fewer samples and faster inference.

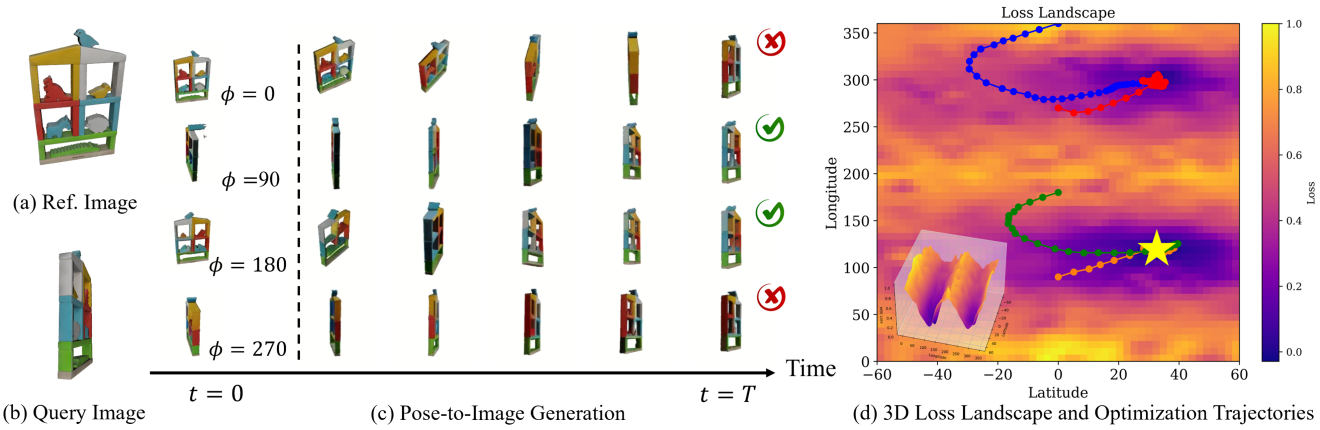



Figure 2. (a) & (b) Reference and query images of the object, captured from different camera poses. (c) Images generated by feeding the poses from optimization trajectory back into Zero123 [28]. Although all timestep-$T$ images appear visually similar to the query, only two accurately reproduce the object's correct appearance—white and blue in the front, yellow and red in the back—indicating correct pose alignment. (d) 2D MSE landscape with optimization trajectories initialized from four different starting poses at longitudes $0°$, $90°$, $180°$, and $270°$. Two of the trajectories converge to local minima. The bottom-left inset shows the 3D landscape for a clearer comparison.

## 2. Related Work

**Pose Estimation.** Traditional approaches estimate pose by detecting and matching keypoints across images, using either hand-crafted features [2, 13, 24, 30, 41] or learned descriptors [8, 11, 39, 53]. The matched correspondences are then used to recover relative poses. While these traditional methods [45] are effective in well-textured and dense-view scenarios, they often struggle under sparse-view or textureless conditions. To address these limitations, learning-based methods have emerged that bypass explicit feature matching. Some directly regress camera pose from input images [9, 22, 47, 51], while others adopt energy-based formulations [14, 17, 26, 34, 64] or leverage diffusion models [18, 55]. More recent trends predict dense ray or point maps, as in DUSt3R [57], MASt3R [25], and VGGT [56], providing stronger geometric constraints and thus more stable. Another emerging direction repurposes pretrained pose-conditioned diffusion models in reverse for pose estimation, which we discuss in the following paragraphs.

**Exploiting Pose-conditioned Diffusion Models on Pose Estimation.** Pose-conditioned diffusion models [5, 28, 43, 46] are generative models fine-tuned from pretrained diffusion models [40] to enable control over camera viewpoints, synthesizing images conditioned on a reference image and a given camera pose. Recent works invert these models for pose estimation. Methods such as ID-pose [6] and iFusion [66] estimate the relative pose between a reference image and a query image by inverting a pretrained pose-conditioned diffusion model. These approaches iteratively refine the pose by using Zero123 to predict the noise given an image pair and the current pose estimate, then comparing the predicted noise with the actual noise added to the query image. The pose is updated via gradient-based optimization to minimize this noise-space discrepancy.

## 3. Preliminary

This section reviews background concepts relevant to our work, including score-based and energy-based modeling techniques, and framework for leveraging pretrained pose-conditioned diffusion models in camera pose estimation.

To enable score-based learning over empirical data, it is essential to define a differentiable approximation of the underlying data distribution. Let $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$ denote a set of $i.i.d$ samples, where each $\mathbf{x}^{(i)} \in \mathbb{R}^d$ represents the $i^{\text{th}}$ observation drawn from an unknown distribution $p_{\text{data}}(\mathbf{x})$. A direct representation of this distribution can be expressed as a mixture of Dirac delta functions: $\frac{1}{N}\sum_{i=1}^{N}\delta(\|\mathbf{x}-\mathbf{x}^{(i)}\|)$, which exactly matches the observed samples but is inherently non-differentiable and thus unsuitable for gradient-based optimization. To enable a smooth and tractable approximation, Parzen density estimation [54] smooths each Dirac delta with an isotropic Gaussian kernel: $p_\sigma(\tilde{\mathbf{x}} \mid \mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}} \mid \mathbf{x}, \sigma^2\mathbf{I}_d)$, where $\sigma^2$ controls the degree of smoothing. This formulation yields a differentiable density estimate that forms the foundation for score-based learning.

**Score-based Modeling.** Given the smooth density approximation introduced above, score-based modeling [19, 54] aims to estimate the score of a data distribution. The score captures the direction that increases the data likelihood and thus provides a principled way to guide optimization toward regions of high probability. The score function is represented by a neural network $s_\theta(\mathbf{x})$ parameterized by $\theta$. The network is trained using the Denoising Score Matching (DSM) loss [54] $\mathcal{L}_{\text{DSM}}$, expressed as:

$$\mathcal{L}_{\text{DSM}}(\theta) = \tfrac{1}{2}\mathbb{E}_{\tilde{\mathbf{x}},\mathbf{x}}\left[\|s_\theta(\tilde{\mathbf{x}}) - \nabla_{\tilde{\mathbf{x}}}\log p_\sigma(\tilde{\mathbf{x}} \mid \mathbf{x})\|_2^2\right], \quad (1)$$

where $\nabla_{\tilde{\mathbf{x}}}\log p_\sigma(\tilde{\mathbf{x}} \mid \mathbf{x})$ is the denoising direction that can be computed analytically. Minimizing Eq. (1) trains $s_\theta$ to approximate the true score $\nabla_{\tilde{\mathbf{x}}}\log p_\sigma(\tilde{\mathbf{x}})$ of the smoothed

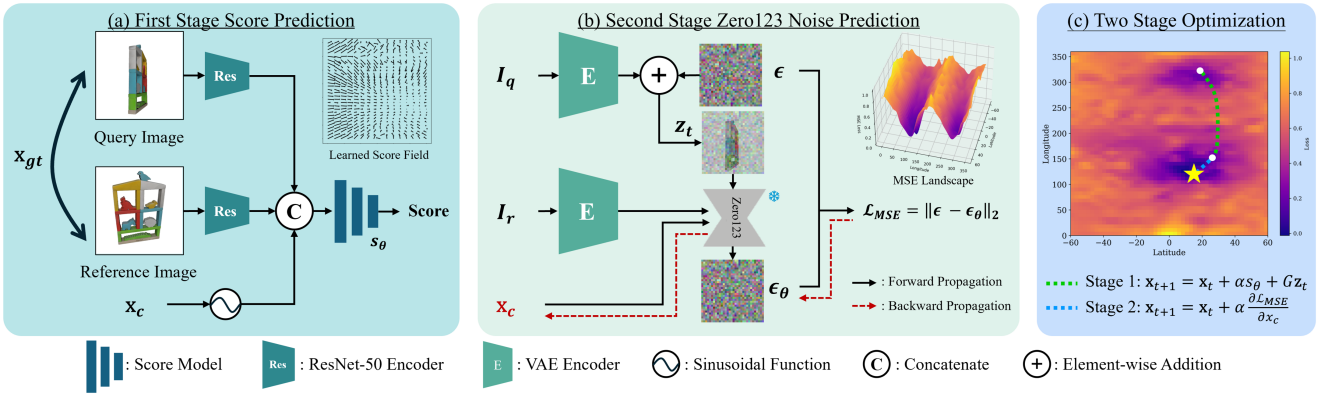



Figure 3. **Framework Overview.** (a) The first part shows our proposed score network, which uses a ResNet encoder to extract image features. The conditioned pose is encoded via a sinusoidal embedding, and these features are concatenated and fed into a MLP to predict score. The trained score function guides optimization trajectory toward the ground-truth pose, helping avoid local minima in the Zero123 MSE landscape. (b) The second stage uses Zero123 to refine via energy-based optimization. Given a pair of reference-query images and a conditioned pose, the frozen Zero123 model estimates the noise. The MSE between predicted and actual noise defines energy, which is minimized via gradient-based optimization to refine the pose. (c) The two stages are combined to form the overall optimization process.

data distribution [54]. Once trained, the score can be utilized in Langevin dynamics to iteratively sample from this distribution through: $\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_{t-1} + \frac{\alpha}{2} s_\theta(\tilde{\mathbf{x}}_{t-1}) + \sqrt{\alpha}\mathbf{z}_t$, where $\mathbf{z}_t \sim \mathcal{N}(0, I)$ represents Gaussian noise, $\alpha$ is the step size, and $t$ is iteration number. In our context, this learned score function later serves as a guide for optimization, steering updates toward more probable poses.

**Energy-Based Modeling.** Energy-based models (EBMs) provide an alternative but closely related perspective. They represent data distributions using an energy function $\mathcal{E}(\mathbf{x})$, where low-energy regions correspond to high-likelihood samples: $p(\mathbf{x}) = \frac{1}{Z}\exp(-\mathcal{E}(\mathbf{x}))$, where $Z$ denotes the normalization constant. The score is the gradient of the log-density, which can be expressed as $s(\mathbf{x}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}) = -\nabla_{\mathbf{x}}\mathcal{E}(\mathbf{x})$. This relationship shows that EBMs and score-based models define equivalent gradient fields over the data manifold, while EBMs model the energy function itself, score-based models directly approximate its gradient.

**Inverting Pose-conditioned Diffusion Model.** Zero123 generates novel views conditioned on a reference image and a relative camera pose. Let $I_r$ be the reference image and $I_q$ the query image. The query image is encoded by a VAE encoder: $\mathbf{z} = \mathrm{E}(I_q)$, and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ is added to produce the noisy latent $\mathbf{z}_t$. Zero123 is trained to predict this noise: $\mathcal{L}(I_q, (I_r, T)) = \mathbb{E}_{\mathbf{z},\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, (I_r, T))\|_2^2\right]$, where the condition includes the reference image $I_r$ and the relative camera pose $T$. Leveraging this formulation, methods such as ID-pose and iFusion invert Zero123 by treating the camera pose $T$ as an optimization variable while freezing the pretrained diffusion parameter $\theta$. The goal is to estimate the relative pose $\hat{T}_{r\to q}$ that minimize the diffusion denoising MSE loss:

$$\hat{T}_{r\to q} = \underset{T\in SE(3)}{\operatorname{argmin}}\, \mathcal{L}(I_q, (I_r, T)) + \mathcal{L}(I_r, (I_q, T^{-1})), \quad (2)$$

and gradients of the loss with respect to $T$ are used to iteratively update the pose via gradient-based optimization.

# 4. Methodology

## 4.1. Problem Formulation

Let the dataset consists of $N$ paired samples $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^N$, where each pair includes a relative camera pose $\mathbf{x}^{(i)}$ and an image pair $\mathbf{y}^{(i)} = (I_r^{(i)}, I_q^{(i)})$. Here, $I_r^{(i)}$ is the reference image and $I_q^{(i)}$ is the query image. We parameterize the camera pose in spherical coordinates $(\Theta, \Phi, \rho)$, following prior works such as iFusion and Zero123. The relative pose between the two images is defined as the difference between query and reference poses: $\mathbf{x} = (\Theta_q - \Theta_r, \Phi_q - \Phi_r, \rho_q - \rho_r)$. The objective is to estimate the transformation $\mathbf{x}$ from the provided image pair $\mathbf{y}$. It is assumed that each image pair $\mathbf{y}^{(i)}$ corresponds to a unique ground-truth relative pose $\mathbf{x}^{(i)}$. Formally, this implies that the conditional data distribution is deterministic and can be expressed as a Dirac delta function: $p(\mathbf{x} \mid \mathbf{y}^{(i)}) = \delta(\mathbf{x} - \mathbf{x}^{(i)})$.

## 4.2. The Proposed Framework

In this section, we introduce a two-stage optimization framework for estimating camera poses while mitigating the effect of local minima in the Zero123 energy landscape.

**Framework Overview.** The proposed framework, as illustrated in Fig. 3, comprises two stages: *the score-based optimization stage* and *the energy-based refinement stage*. In the first stage, our framework employs a score network $s_\theta(I_r, I_q, \tilde{\mathbf{x}})$ that predicts the update direction of the pose for a reference image $I_r$, a query image $I_q$, and a noised pose $\tilde{\mathbf{x}}$. The underlying principle is that the score model learns to approximate the gradient of the log-probability density of plausible poses conditioned on the image pair. Through iterative updates along this learned gradient, the pose is encouraged to move toward high-probability regions of the pose space, which effectively escapes local minima. More specifically, the score network adopts a lightweight

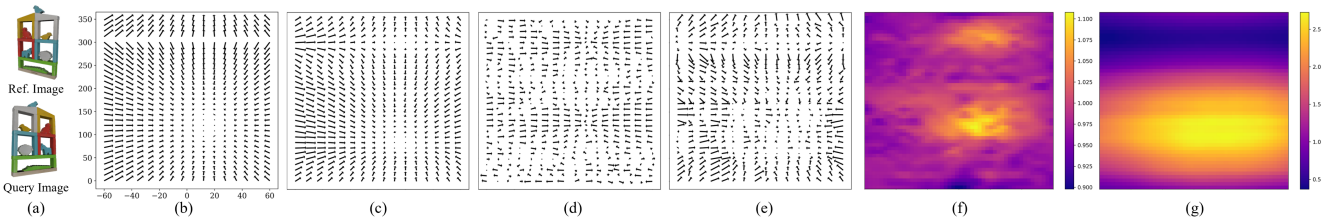



Figure 4. **Toy Example.** (a) Reference and query images; (b) Oracle score field; (c) Score field from our score-based model; (d) Score field from Zero123 MSE; (e) Score field from our energy-based model; (f) Probability landscape from Zero123 MSE loss; (g) Probability landscape from our energy-based model. The landscapes in (f) and (g) represent the probability distribution and are plotted as $\exp(-\mathcal{E}_\theta(\mathbf{x}))$.

design: image features are extracted by a ResNet-50 [15] backbone, while the conditional pose is encoded via sinusoidal embeddings. The concatenated features are processed by a three-layer MLP to predict the score function.

Once the pose has been guided toward a geometrically consistent region, the second stage employs the pretrained Zero123 model to refine it through energy-based optimization. In particular, Gaussian noise is injected into the latent representation of the query image to obtain a noisy latent $\mathbf{z}_t$. The Zero123 model then estimates the noise conditioned on the reference image and the current pose estimate. The MSE between the predicted and original noise serves as the energy function, where the gradient of this energy with respect to the pose provides a refinement direction. This energy-driven gradient descent further aligns the pose with the cross-view consistency encoded in the Zero123 model.

This two-stage pipeline forms a coherent optimization strategy: the score model first provides global guidance to avoid suboptimal local minima, while the diffusion-based energy model subsequently delivers fine-grained local corrections. Both stages rely on gradient-based updates but differ in the manner through which the gradient is obtained. The first stage derives it explicitly from the learned score, while the second obtains it implicitly from the energy loss.

**Training Objective.** Our training objective for the score model $s_\theta$ follows the denoising score matching (DSM) principle, extended to the conditional setting:

$$\mathcal{L}(\theta) = \tfrac{1}{2}\mathbb{E}_{\mathbf{x},\mathbf{y}}\mathbb{E}_{\tilde{\mathbf{x}}\sim \mathrm{U}} \left\| s_\theta(\tilde{\mathbf{x}}, \mathbf{y}) - \nabla_{\tilde{\mathbf{x}}} \log p_\sigma(\tilde{\mathbf{x}} \mid \mathbf{x}, \mathbf{y}) \right\|_2^2 . \quad (3)$$

Since our score model operates in the low-dimensional pose space, we adopt a simplified objective compared to NCSN [49]. Specifically, $\tilde{\mathbf{x}}$ is sampled from a uniform distribution U, and the noise scale is fixed at $\sigma = 1$, allowing the model to disregard noise-level conditioning. This uniform sampling enables the model to learn a score function that captures the global gradient structure over the entire pose space, rather than focusing only on a local neighborhood around $\mathbf{x}$. Despite this simplification, we show in Section 4.4 that the optimal solution remains theoretically equivalent to that obtained with a Gaussian kernel. Implementation details and hyperparameters are provided in the Appendix B. We further compare our score-based formulation with an energy-based alternative in Appendix A.

### 4.3. Two Stage Optimization Process

After training, the learned score model enables guidance of arbitrary initial poses toward higher-density regions of the pose distribution. Once the optimization escapes poor local minima, the Zero123 MSE loss further refines the estimated pose. Since precisely determining when a local minima has been escaped is challenging in practice, we adopt a fixed iteration threshold, after which the optimization proceeds using the MSE gradient. Consequently, the overall process consists of two distinct stages. In the first stage, the pose is updated using the learned score model to guide the initial optimization toward regions of higher data likelihood:

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_{t-1} + \alpha s_\theta(\tilde{\mathbf{x}}_{t-1}, \mathbf{y}) + G\mathbf{z}_t, \quad \mathbf{z}_t \sim \mathcal{N}(0, \mathbf{I}_3), \quad (4)$$

where $G = \mathrm{diag}(\gamma_1, \gamma_2, \gamma_3)$ controls the noise scale for each coordinate. This update resembles Langevin dynamics, where the learned score provides a drift toward high-likelihood regions and the Gaussian noise encourages exploration of the pose space. This dynamics ensures that the norm of the expected pose error decays exponentially.

$$\|\mathbb{E}[\tilde{\mathbf{x}}_t - \mathbf{x}_{\mathrm{gt}}]\| = M(1-\alpha)^t, \quad \mathrm{Var}[\tilde{\mathbf{x}}_t] \approx \frac{G^2}{2\alpha}, \quad (5)$$

where $M$ is a constant depending on the initial distance. A detailed proof is provided in the Appendix A. In the second stage, the pretrained Zero123 model is employed as an energy function, and gradient-based methods is performed to solve the optimization problem defined in Eq. (2).

**Joint Reasoning across Multiple Views.** A straightforward extension of a two-view method to the multi-view setting is to process each image pair independently; however, this ignores multi-view consistency. To address this, we formulate a unified objective that performs energy-based optimization in a high-dimensional pose space, as shown in Eq. (6). Enforcing global consistency allows reliable relations to correct erroneous ones, improving robustness.

$$\hat{\mathcal{T}} = \underset{\{T_1,\ldots,T_n\}\subset SE(3)}{\arg\min} \sum_{i=1}^{N} \sum_{j\neq i} \mathcal{L}\big(I^{(j)}, (I^{(i)}, T_i^{-1}T_j)\big), \quad (6)$$

where $\{T_1, \ldots, T_n\}$ denotes the set of absolute camera poses for all views. Despite its advantages, optimizing Eq. (6) is challenging. The solution space grows exponentially with the number of views, making multi-start strategies computationally prohibitive and prone to local minima.

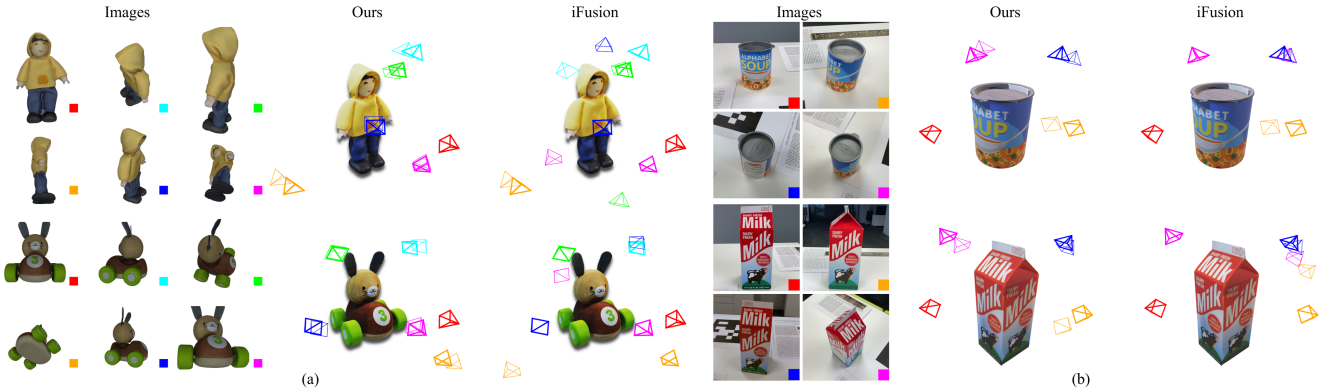



Figure 5. **Qualitative Results.** Visualization of predicted camera poses (thin) compared to ground truth poses (bold). For each object, we randomly select two initial viewpoints and estimate the relative poses of all target views from a reference image, shown in red. (a) Results on GSO objects: our method consistently converges to the correct pose, while iFusion often gets stuck in local minima, leading to incorrect predictions. (b) Results on HOPEv2 real-world objects: our method accurately recovers the correct poses despite strong geometric symmetries, as the distinct textures allow effective disambiguation, whereas iFusion frequently converges to incorrect local optima.

To address this, we extend our two-stage framework to the multi-view scenario. Given $N$ images, the first stage uses the learned score function to infer pairwise relative poses $\mathcal{T} = \{T_{i\to j}\}_{i\neq j}$. We then perform a global optimization to obtain a consistent set of absolute poses $\overline{\mathcal{T}} = \{\overline{T}_i\}_{i=1}^N$, where $\overline{T}_{i\to j} = \overline{T}_i^{-1}\overline{T}_j$. This reparameterization removes redundancy and enforces global consistency across all views. The resulting estimate $\overline{\mathcal{T}}$ provides a strong initialization for the subsequent refinement. Then we apply a Zero123-based energy optimization, using Eq. (6), to refine the poses and obtain the final transformation set $\hat{\mathcal{T}}$.

### 4.4. Theoretical Justification

We provide a theoretical analysis of the objective in Eq. (3).

**Proposition 1.** Consider the objective

$$\mathcal{L}(\theta) = \tfrac{1}{2}\mathbb{E}_{\mathbf{y},\mathbf{x},\tilde{\mathbf{x}}}\,\|s_\theta(\tilde{\mathbf{x}},\mathbf{y}) - \nabla_{\tilde{\mathbf{x}}}\log p_\sigma(\tilde{\mathbf{x}}\mid\mathbf{x})\|_2^2, \quad (7)$$

where the expectation is taken over $p(\mathbf{y}), p(\mathbf{x}\mid\mathbf{y})$, and $p(\tilde{\mathbf{x}}\mid\mathbf{x})$. If $p_\sigma(\tilde{\mathbf{x}}\mid\mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}};\mathbf{x},\sigma^2 I)$, then the optimal solution $s^\star(\tilde{\mathbf{x}},\mathbf{y})$ for fixed $(\tilde{\mathbf{x}},\mathbf{y})$ is

$$s^\star(\tilde{\mathbf{x}},\mathbf{y}) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}\mid\mathbf{y},\tilde{\mathbf{x}})}[\nabla_{\tilde{\mathbf{x}}}\log p_\sigma(\tilde{\mathbf{x}}\mid\mathbf{x})]. \quad (8)$$

Equivalently, in integral form,

$$s^\star(\tilde{\mathbf{x}},\mathbf{y}) = \frac{\int(\mathbf{x}-\tilde{\mathbf{x}})\,p(\tilde{\mathbf{x}}\mid\mathbf{x})\,p(\mathbf{x}\mid\mathbf{y})\,d\mathbf{x}}{\sigma^2\int p(\tilde{\mathbf{x}}\mid\mathbf{x})\,p(\mathbf{x}\mid\mathbf{y})\,d\mathbf{x}}. \quad (9)$$

To simplify the sampling strategy, consider the case where $\tilde{\mathbf{x}}$ is sampled from a uniform distribution U. This motivates the following lemma.

**Lemma 1.** Given $\mathbf{y}$ and $\tilde{\mathbf{x}}\sim\mathrm{U}$, where U denotes a uniform distribution independent of $\mathbf{x}$, the optimal solution of the loss function $\mathcal{L}(\theta)$ in Eq. (7) is

$$\begin{aligned} s_U^\star(\tilde{\mathbf{x}},\mathbf{y}) &= \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}\mid\mathbf{y})}[\nabla_{\tilde{\mathbf{x}}}\log p_\sigma(\tilde{\mathbf{x}}\mid\mathbf{x})] \\ &= \frac{1}{\sigma^2}\int(\mathbf{x}-\tilde{\mathbf{x}})p(\mathbf{x}\mid\mathbf{y})\,d\mathbf{x}. \end{aligned} \quad (10)$$

**Lemma 2.** In general, $s^\star(\tilde{\mathbf{x}},\mathbf{y}) \neq s_{\mathrm{U}}^\star(\tilde{\mathbf{x}},\mathbf{y})$, except in the special case where $p(\mathbf{x}\mid\mathbf{y})$ collapses to a Dirac delta distribution, i.e., $p(\mathbf{x}\mid\mathbf{y}^{(i)}) = \delta(\mathbf{x}-\mathbf{x}^{(i)})$.

Lemma 2 shows that, under the assumption of a unique solution for each conditional image pair, the simplified objective shares the same optimum as Eq. (7). Details proofs of Proposition 1 as well as Lemma 1 and 2 are provided in Appendix A for completeness.

### 4.5. Toy Example

We present a toy example trained on a single object from the GSO [10] dataset. This example demonstrates that both score-based and energy-based modeling can learn the underlying data distribution. As illustrated in Fig. 4, we visualize the score field for the score-based method and the probability density $\exp(-\mathcal{E}_\theta(\mathbf{x}))$ for the energy-based method. For the latter, we also display the induced score field obtained via gradient computation. Fig. 4 (f) depicts the probability density computed from the Zero123 MSE landscape, which exhibits two modes due to the object's geometry, and Fig. 4 (d) visualizes the corresponding score field. In contrast, Fig. 4 (c) presents the score field from the score-based modeling, which aligns well with the oracle score field in Fig. 4 (b). Fig. 4 (g) reveals the probability density learned by the energy-based model and demonstrates a smooth and coherent landscape. However, its corresponding score field in Fig. 4 (e) appears noisy and inferior to that of the score-based method. We conjecture that this limitation arises from the indirect prediction process, which requires gradient computation on the energy function. A more comprehensive comparison of these two modeling approaches is provided in Section 5.4.

Table 1. **Evaluation results on the synthetic dataset.** Results on the GSO and OO3D datasets show that our two-stage optimization framework improves success rate and recall across thresholds. **Red** indicates our best result, and **blue** denotes the second best result.

| Dataset | Method | @5 | | @15 | | @30 | | @5 | | @15 | | @30 | | Rot. ↓ | Trans. ↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R ↑ | R(R) ↑ | R ↑ | R(R) ↑ | R ↑ | R(R) ↑ | SR ↑ | SR(R) ↑ | SR ↑ | SR(R) ↑ | SR ↑ | SR(R) ↑ | | |
| GSO | **DUSt3R** [25] | 0.530 | 0.534 | 0.903 | **0.923** | **0.957** | **0.986** | - | - | - | - | - | - | 4.63 | 0.053 |
| | **VGGT** [56] | **0.752** | **0.800** | 0.866 | **0.945** | 0.869 | **0.960** | - | - | - | - | - | - | **2.14** | **0.050** |
| | **ID-Pose** [6] | 0.223 | 0.247 | 0.541 | 0.624 | 0.607 | 0.723 | 0.039 | 0.049 | 0.118 | 0.152 | 0.146 | 0.201 | 10.29 | 0.134 |
| | **iFusion** [59] | **0.700** | **0.704** | **0.904** | 0.916 | 0.918 | 0.938 | **0.275** | **0.278** | **0.365** | **0.374** | **0.382** | **0.398** | **3.07** | **0.035** |
| | **Ours** | 0.645 | 0.650 | **0.907** | 0.921 | **0.927** | 0.945 | **0.585** | **0.591** | **0.811** | **0.840** | **0.836** | **0.878** | 3.63 | 0.051 |
| OO3D | **DUSt3R** | 0.395 | 0.404 | 0.791 | 0.832 | **0.889** | **0.969** | - | - | - | - | - | - | 6.49 | **0.054** |
| | **VGGT** | 0.477 | **0.514** | 0.739 | 0.830 | 0.757 | 0.875 | - | - | - | - | - | - | **4.81** | 0.085 |
| | **ID-Pose** [6] | 0.134 | 0.146 | 0.454 | 0.531 | 0.546 | 0.694 | 0.022 | 0.028 | 0.084 | 0.111 | 0.113 | 0.167 | 13.79 | 0.147 |
| | **iFusion** | **0.516** | **0.523** | **0.841** | **0.866** | 0.882 | 0.930 | **0.189** | **0.192** | **0.306** | **0.316** | **0.332** | **0.359** | **4.76** | **0.047** |
| | **Ours** | **0.479** | 0.489 | **0.838** | **0.875** | **0.905** | **0.970** | **0.447** | **0.464** | **0.780** | **0.842** | **0.848** | **0.949** | 5.15 | 0.055 |

# 5. Experimental Results

## 5.1. Experimental Setups

**Dataset.** Following the setup in iFusion [59], we adopt the GoogleScannedObject (GSO) [10] and OmniObject3D (OO3D) [60] 3D model datasets for our experiments. Note that additional dataset details are provided in Appendix B. For real-world evaluation, we use the HOPEv2 [52] dataset from the BOP Challenge [35], which contains 28 grocery objects captured in 50 scenes across diverse household and office environments under varying lighting conditions.

**Evaluation Metrics.** We evaluate our method using the following metrics: Recall (R), Success Rate (SR), Rotation Error (Rot.), and Translation Error (Trans.). Recall measures the proportion of final predictions that satisfy predefined thresholds. Specifically, in gradient-based optimizer, we typically sample $N$ initial poses and select the lowest loss among the $N$ trials. Recall considers this best prediction. In contrast, Success Rate evaluates all $N$ predictions and reports the percentage that meet the thresholds, thereby reflecting the method's robustness to varying initializations. For both Success Rate and Recall, we adopt rotation thresholds of $5°$, $15°$, $30°$, and a translation distance threshold of $0.2$. To isolate rotational performance, we also report Rotation Recall and Rotation Success Rate, denoted as R(R) and SR(R), which assess accuracy solely based on rotation thresholds. Finally, we compute Rotation Error and Translation Error as the median across all evaluated samples to reduce sensitivity to outliers caused by convergence failures.

**Evaluations.** We evaluate our framework on camera pose estimation tasks and compare it against ID-Pose [6], iFusion [59], DUSt3R [57], and VGGT [56]. We consider two initialization strategies. The first strategy employs nine uniformly distributed poses with latitudes of $-30°$, $0°$, and $30°$, and longitudes from $0°$ to $360°$ in $120°$ increments. The second strategy varies the number of randomly sampled initial poses to evaluate the robustness through recall.

## 5.2. Quantitative Results

Table 1 presents the evaluation results on the GSO and OO3D datasets. Our method achieves substantial improvements in success rate, which demonstrates that the proposed score-based framework effectively mitigates the sample inefficiency inherent in directly applied gradient-based optimization with the Zero123 MSE loss. This improvement is particularly significant as it addresses a fundamental limitation of existing diffusion-based pose estimation methods. Moreover, our method achieves comparable performance to other SoTA methods across recall, rotation error, and translation error metrics. Table 2 further reports the evaluation results on real-world data from the HOPEv2 dataset. The results demonstrate robust performance under various challenging real-world conditions, including lighting variation, diverse object textures, material effects, and natural image blur. Furthermore, Table 3 presents the performance of our framework on unseen objects. The second-stage refinement leverages the strong generative capability of Zero123 to effectively guide the optimization trajectory toward the correct pose. As a result, our approach achieves performance on par with SoTA methods, even though the first-stage score model is trained on a more limited dataset compared to VGGT. This outcome highlights the effectiveness of our two-stage framework in leveraging pretrained generative priors for robust pose estimation on novel objects.

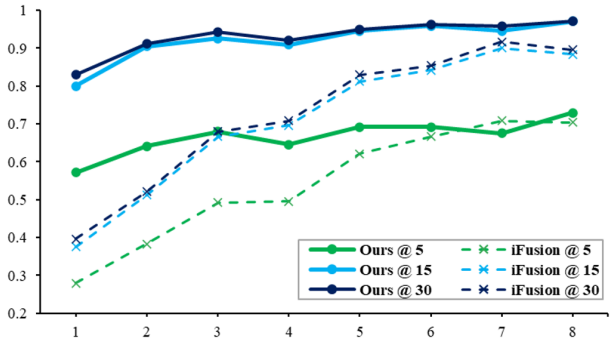



Figure 6. **Evaluation under varying numbers of samples.** The figure compares our framework and iFusion in recall for various numbers of random initial poses. The x-axis represents the number of random initial poses, and the y-axis represents the recall.

Table 2. **Evaluation results on the HOPEv2 dataset.** Our method maintains strong performance under real-world conditions, as reflected by recall and success rate. **Red** indicates the best result, and **blue** the second best.

| Dataset | Method | @5 | | @15 | | @30 | | @5 | | @15 | | @30 | | Rot. ↓ | Trans. ↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R ↑ | R(R) ↑ | R ↑ | R(R) ↑ | R ↑ | R(R) ↑ | SR ↑ | SR(R) ↑ | SR ↑ | SR(R) ↑ | SR ↑ | SR(R) ↑ | | |
| HOPEv2 [52] | VGGT | **0.208** | **0.279** | **0.571** | **0.798** | **0.631** | **0.893** | - | - | - | - | - | - | **8.10** | **0.132** |
| | iFusion | 0.095 | 0.107 | 0.411 | 0.506 | 0.494 | 0.619 | **0.040** | **0.048** | **0.169** | **0.222** | **0.206** | **0.291** | 14.78 | 0.151 |
| | Ours | **0.214** | **0.214** | **0.679** | **0.702** | **0.851** | **0.887** | **0.164** | **0.165** | **0.534** | **0.546** | **0.786** | **0.837** | **8.96** | **0.059** |

Table 3. **Evaluation on unseen objects**. We sample 10 additional objects from the GSO dataset that were not included in training.

| Method | @5 | | @15 | | @30 | | Rot. ↓ | Trans. ↓ |
|---|---|---|---|---|---|---|---|---|
| | R ↑ | R(R) ↑ | R ↑ | R(R) ↑ | R ↑ | R(R) ↑ | | |
| DUSt3R | 0.147 | 0.170 | 0.523 | 0.636 | 0.602 | 0.818 | 9.96 | 0.114 |
| VGGT | **0.579** | **0.647** | 0.807 | **0.954** | 0.818 | **0.965** | **3.31** | 0.088 |
| iFusion | **0.625** | **0.625** | **0.875** | **0.910** | **0.921** | **0.954** | **3.65** | **0.044** |
| Ours | 0.489 | 0.500 | **0.818** | 0.841 | **0.864** | 0.898 | 5.07 | **0.072** |

Table 4. **Multi-view joint reasoning.** We evaluate our framework for multi-view estimation, reporting recall at thresholds of 15° and 30°. To analyze the contributions of each stage, we include ablation results: without Stage 1 (score-based initialization), without Stage 2 (Zero123 refinement), and using both Stage 1 and Stage 2.

| | # of Images | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| R@15 | w/o Stage 1 | 0.200 | 0.103 | 0.065 | 0.075 | 0.074 | 0.071 | 0.025 |
| | w/o Stage 2 | 0.280 | 0.230 | 0.292 | 0.294 | 0.295 | 0.299 | 0.289 |
| | Stage 1 & 2 | **0.540** | **0.513** | **0.568** | **0.589** | **0.616** | **0.662** | **0.643** |
| R@30 | w/o Stage 1 | 0.230 | 0.137 | 0.112 | 0.141 | 0.119 | 0.114 | 0.050 |
| | w/o Stage 2 | 0.450 | 0.443 | 0.507 | 0.528 | 0.544 | 0.550 | 0.521 |
| | Stage 1 & 2 | **0.580** | **0.583** | **0.662** | **0.690** | **0.727** | **0.767** | **0.786** |

Table 5. **Ablation on different modeling approaches.** We compares score-based modeling and energy-based modeling on the GSO dataset with 10 objects.

| Method | R@5 | R@15 | R@30 | SR@5 | SR@15 | SR@30 | Rot. | Trans. |
|---|---|---|---|---|---|---|---|---|
| Score | **0.700** | **0.963** | **0.963** | **0.631** | **0.900** | **0.914** | **3.12** | 0.051 |
| Energy | 0.563 | 0.850 | 0.888 | 0.456 | 0.726 | 0.778 | 4.25 | **0.044** |

### 5.3. Qualitative Results

Fig. 5 (a) presents qualitative comparisons on the synthetic GSO dataset. Our method consistently converges to correct poses, whereas iFusion often fails due to entrapment in local minima. Fig. 5 (b) illustrates results on the real-world HOPEv2 dataset, where geometric symmetry introduces ambiguity. In these challenging cases, iFusion frequently converges to incorrect symmetric views that satisfy local optimality but fail to capture the true pose. In contrast, our method resolves these ambiguities, validating that our learned score function effectively guides the optimization away from suboptimal local minima and achieves robust convergence even with limited pose initialization.

### 5.4. Ablation Study

**Evaluation under Various Numbers of Initializations.** We evaluate the robustness of our method under different numbers of initializations and compare it with iFusion, as shown in Fig. 6. Under a 30° rotation threshold, our method achieves similar recall with only two initial poses, whereas iFusion requires eight. This empirical result clearly demonstrates the sample efficiency and robustness of our method, particularly in low-sample regimes.

**Ablation on Two Stage Design.** To validate our two-stage framework, we evaluate the contribution of each stage on the multi-view pose estimation task. For each configuration, the number of input views varies from two to eight. As shown in Table 4, removing Stage 1 leads to a significant performance degradation. This occurs since the solution space grows exponentially with the number of views, rendering optimization highly sensitive to initialization. This confirms that random sampling is ineffective and that the score model is essential for guiding optimization toward promising regions in the multi-view estimation task. Moreover, the incorporation of Stage 2 refinement consistently improves recall, substantiating the effectiveness of using the pretrained Zero123 as an energy-based refinement module.

**Comparison of Score-Based & Energy-Based Methods.** To assess the effectiveness of our score-based design, we compare it with the energy-based formulation introduced in Appendix A. Both formulations aim to guide the pose update toward the correct pose. The results in Table 5 show that the score-based approach outperforms the energy-based approach. We attribute this to the fact that the score-based method predicts the score directly, whereas the energy-based model learns an energy function whose gradient is an indirect approximation of the score for pose updates.

## 6. Conclusion

In this work, we introduced a novel perspective on two-view camera pose estimation by analyzing it through the landscape perspective. We provided clear visualizations of the MSE landscape of the Zero123 model, and highlighted the local minima issues encountered by iFusion. Unlike previous methods that rely heavily on dense sampling, our approach leverages a learned score model to reshape the optimization dynamics, effectively guiding pose estimates toward the ground-truth and mitigating the impact of poor local minima. By incorporating this insight into a two-stage optimization scheme, our method achieves performance on par with with state-of-the-art methods while requiring fewer samples and significantly reducing inference time. These results demonstrate that a landscape-aware formulation not only enhances robustness to initialization but also paves the way for more sample-efficient and gradient-driven approaches to pose estimation.

## 7. Acknowledgement

The authors gratefully acknowledge the support from the National Science and Technology Council (NSTC) in Taiwan under grant numbers NSTC 114-2221-E-002-069-MY3, NSTC 113-2221-E-002-212-MY3, and NSTC 114-2218-E-A49-026, as well as the support from the Academia Sinica Scholar Award (ASSA) under grant number AS-ASSA-115-02, NTU Artificial Intelligence Center of Research Excellence, and Taiwan Centers of Excellence in Artificial Intelligence. This research was also supported by the NVIDIA Academic Grant Program. The authors would also like to express their appreciation for the donation of the GPUs from NVIDIA Corporation and NVIDIA AI Technology Center (NVAITC) used in this work. Furthermore, the authors extend their gratitude to the National Center for High-Performance Computing (NCHC) for providing computational and storage resources. The authors also thank the NVIDIA Taipei-1 supercomputer for providing essential computing resources.

## References

[1] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016. 16

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: speeded up robust features. In *Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part I*, pages 404–417. Springer, 2006. 1, 3

[3] Hayet Belghit, Abdelkader Bellarbi, Nadia Zenati, and Samir Otmane. Vision-based pose estimation for augmented reality: a comparison study. *arXiv preprint arXiv:1806.09316*, 2018. 1

[4] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXII*, pages 20–36. Springer, 2022. 1

[5] Yabo Chen, Jiemin Fang, Yuyang Huang, Taoran Yi, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Cascadezero123: One image to highly consistent 3d with self-prompted nearby views. In *European Conference on Computer Vision*, pages 311–330. Springer, 2024. 3, 19

[6] Weihao Cheng, Yan-Pei Cao, and Ying Shan. Id-pose: Sparse-view camera pose estimation by inverting diffusion models. *arXiv preprint arXiv:2306.17140*, 2023. 1, 3, 7, 18

[7] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 1

[8] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 1, 3

[9] Siyan Dong, Shuzhe Wang, Shaohui Liu, Lulu Cai, Qingnan Fan, Juho Kannala, and Yanchao Yang. Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 16739–16752. Computer Vision Foundation / IEEE, 2025. 3

[10] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas Barlow McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation, ICRA 2022, Philadelphia, PA, USA, May 23-27, 2022*, pages 2553–2560. IEEE, 2022. 6, 7, 20

[11] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint detection and description of local features. *arXiv preprint arXiv:1905.03561*, 2019. 3

[12] Jakob Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: large-scale direct monocular SLAM. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II*, pages 834–849. Springer, 2014. 1

[13] Christopher G. Harris and Mike Stephens. A combined corner and edge detector. In *Proceedings of the Alvey Vision Conference, AVC 1988, Manchester, UK, September, 1988*, pages 1–6. Alvey Vision Club, 1988. 3

[14] Rasmus Laurvig Haugaard, Frederik Hagelskjær, and Thorbjørn Mosekjær Iversen. Spyropose: SE(3) pyramids for object pose distribution estimation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pages 2074–2083. IEEE, 2023. 1, 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 5, 16

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising

diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1

[17] Timon Höfer, Benjamin Kiefer, Martin Messmer, and Andreas Zell. Hyperposepdf hypernetworks predicting the probability distribution on SO(3). In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 2368–2378. IEEE, 2023. 1, 3

[18] Tsu-Ching Hsiao, Hao-Wei Chen, Hsuan-Kung Yang, and Chun-Yi Lee. Confronting ambiguity in 6d object pose estimation via score-based diffusion on se (3). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 352–362, 2024. 3

[19] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, 2005. 3

[20] 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)*, pages 31–41. IEEE, 2024. 1

[21] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6007–6017, 2023. 1

[22] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2938–2946. IEEE Computer Society, 2015. 3

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 16

[24] Axel Barroso Laguna and Krystian Mikolajczyk. Key.net: Keypoint detection by handcrafted and learned CNN filters revisited. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):698–711, 2023. 1, 3

[25] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024. 3, 7

[26] Amy Lin, Jason Y. Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. In *International Conference on 3D Vision, 3DV 2024, Davos, Switzerland, March 18-21, 2024*, pages 106–115. IEEE, 2024. 1, 3

[27] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36:22226–22246, 2023. 1

[28] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 9264–9275. IEEE, 2023. 1, 3

[29] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 1

[30] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004. 1, 3

[31] Matthew Matl. Pyrender. https://github.com/mmatl/pyrender, 2019. 15

[32] 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 *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, pages 405–421. Springer, 2020. 1

[33] Raul Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Trans. Robotics*, 31(5): 1147–1163, 2015. 1

[34] Kieran Murphy, Carlos Esteves, Varun Jampani, Srikumar Ramalingam, and Ameesh Makadia. Implicit-pdf: Non-parametric representation of probability distributions on the rotation manifold. *arXiv preprint arXiv:2106.05965*, 2021. 1, 3

[35] Van Nguyen Nguyen, Stephen Tyree, Andrew Guo, Mederic Fourmy, Anas Gouda, Taeyeop Lee, Sungphill Moon, Hyeontae Son, Lukas Ranftl, Jonathan Tremblay, Eric Brachmann, Bertram Drost, Vincent Lepetit, Carsten Rother, Stan Birchfield, Jiri Matas, Yann Labbé, Martin Sundermeyer, and Tomas Hodan. BOP challenge 2024 on model-based and model-free 6d object pose estimation. *CoRR*, abs/2504.02812, 2025. 7

[36] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *IEEE/CVF*

*Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5470–5480. IEEE, 2022. 1

[37] Paschalis Panteleris, Damien Michel, and Antonis A. Argyros. Toward augmented reality in museums: Evaluation of design choices for 3d object pose estimation. *Frontiers Virtual Real.*, 2:649784, 2021. 1

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035, 2019. 16

[39] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 32, 2019. 1, 3

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. 3

[41] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part I*, pages 430–443. Springer, 2006. 3

[42] 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, 2022. 1

[43] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single image. *arXiv preprint arXiv:2310.17994*, 2023. 3

[44] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 4937–4946. Computer Vision Foundation / IEEE, 2020. 1

[45] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4104–4113. IEEE Computer Society, 2016. 1, 3

[46] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 1, 3, 19

[47] Samarth Sinha, Jason Y. Zhang, Andrea Tagliasacchi, Igor Gilitschenski, and David B. Lindell. Sparsepose: Sparse-view camera pose regression and refinement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 21349–21359. IEEE, 2023. 1, 3

[48] Joan Solà, Jérémie Deray, and Dinesh Atchuthan. A micro lie theory for state estimation in robotics. *CoRR*, abs/1812.01537, 2018. 16

[49] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11895–11907, 2019. 1, 5

[50] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1

[51] Hao Tang, Weiyao Wang, Pierre Gleize, and Matt Feiszli. Aden: Adaptive density representations for sparse-view camera pose estimation. In *European Conference on Computer Vision*, pages 111–128. Springer, 2024. 1, 3

[52] Stephen Tyree, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Jeffrey Smith, and Stan Birchfield. 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2022, Kyoto, Japan, October 23-27, 2022*, pages 13081–13088. IEEE, 2022. 7, 8

[53] Michal J. Tyszkiewicz, Pascal Fua, and Eduard Trulls. DISK: learning local features with policy gradient. In *Advances in Neural Information Processing Systems*

*33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3

[54] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Comput.*, 23(7): 1661–1674, 2011. 3, 4

[55] 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*, pages 9773–9783, 2023. 3

[56] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotný. VGGT: visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 5294–5306. Computer Vision Foundation / IEEE, 2025. 3, 7, 16

[57] 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*, pages 20697–20709, 2024. 3, 7, 16

[58] Xingkui Wei, Yinda Zhang, Zhuwen Li, Yanwei Fu, and Xiangyang Xue. Deepsfm: Structure from motion via deep bundle adjustment. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, pages 230–247. Springer, 2020. 1

[59] Chin-Hsuan Wu, Yen-Chun Chen, Bolivar Solarte, Lu Yuan, and Min Sun. ifusion: Inverting diffusion for pose-free reconstruction from sparse views. *arXiv preprint arXiv:2312.17250*, 2023. 1, 7, 15, 16, 18

[60] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 803–814. IEEE, 2023. 7

[61] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4578–4587. Computer Vision Foundation / IEEE, 2021. 1

[62] Cunjun Yu, Zhongang Cai, Hung Pham, and Quang-Cuong Pham. Siamese convolutional neural network for sub-millimeter-accurate camera pose estimation and visual servoing. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2019, Macau, SAR, China, November 3-8, 2019*, pages 935–941. IEEE, 2019. 1

[63] Zheyuan Zhan, Defang Chen, Jian-Ping Mei, Zhenghe Zhao, Jiawei Chen, Chun Chen, Siwei Lyu, and Can Wang. Conditional image synthesis with diffusion models: A survey. *arXiv preprint arXiv:2409.19365*, 2024. 1

[64] Jason Y. Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXI*, pages 592–611. Springer, 2022. 1, 3

[65] 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*, pages 3836–3847, 2023. 1

[66] Qitao Zhao and Shubham Tulsiani. Sparse-view pose estimation and reconstruction via analysis by generative synthesis. *Advances in Neural Information Processing Systems*, 37:111899–111922, 2024. 1, 3

[67] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 12588–12597. IEEE, 2023. 1