Geometric Consistency Regularization for Monocular Depth Estimation via Surface Normal Loss

Mitigating Pseudo-Texture Artifacts in Diffusion-Based Models

Haruko386

October 17, 2025

Abstract

Diffusion-based models for monocular depth estimation, such as Marigold, have demonstrated exceptional performance in capturing detailed scene geometry. However, their iterative denoising process can sometimes introduce high-frequency, geometrically inconsistent details, manifesting as ”pseudo-textures” or noise on otherwise smooth surfaces. To address this, we introduce a surface normal loss term into the training objective. This loss acts as a geometric regularizer, penalizing inconsistencies in local surface orientation between the predicted and ground truth depth maps. By enforcing geometric consistency in the VAE's latent space, our approach effectively suppresses the generation of artifacts, resulting in smoother, more realistic, and geometrically coherent depth predictions without compromising structural detail.

1 Introduction

Monocular depth estimation is a fundamental task in computer vision. Recent advancements using diffusion models have shown remarkable success, leveraging their powerful generative capabilities to produce high-fidelity depth maps from single RGB images. The Marigold model, for instance, formulates depth estimation as an image-to-image translation problem within a latent diffusion framework, achieving state-of-the-art results.

Despite their strengths, a notable challenge in multi-step generative models is the emergence of artifacts. In the context of depth estimation, this often appears as *pseudo-textures*—fine-grained, noise-like patterns on surfaces that should be geometrically smooth (e.g., walls, floors). These artifacts, while visually subtle, represent significant geometric inaccuracies. They arise because the standard training objective, typically a pixel-wise loss (e.g., Mean Squared Error) in the latent space, does not explicitly enforce local geometric consistency.

To mitigate this issue, we propose integrating a **Surface Normal Loss** into the training objective of the Marigold model. This loss directly measures the consistency of local surface orientation, a crucial aspect of 3D geometry. By penalizing discrepancies between the surface normals derived from the predicted depth and those from the ground truth, we guide the model to learn smoother and more geometrically plausible surfaces, effectively suppressing the generation of pseudo-textures.

2 Methodology

Our approach introduces a regularization term to the standard diffusion loss. This term operates on the surface normals computed from the depth maps. To maintain computational efficiency, all calculations are performed within the VAE's compact latent space.

2.1 Surface Normal Estimation from Depth

A dense depth map can be interpreted as a 2.5D representation of a scene. From this representation, we can recover the 3D surface normal vector at each pixel. This process involves two main steps: back-projecting the depth map into a 3D point cloud, and then estimating the local surface orientation.

2.1.1 Back-Projection to 3D Point Cloud

Given a depth value $d(u, v)$ at pixel coordinate (u, v) and the camera intrinsic parameters—focal lengths (f_x, f_y) and principal point (c_x, c_y)—we can recover the corresponding 3D point $P(x, y, z)$ in the camera coordinate system using the following standard projection equations:

$$z = d(u, v) \tag{1}$$

$$x = \frac{(u - c_x) \cdot z}{f_x} \tag{2}$$

$$y = \frac{(v - c_y) \cdot z}{f_y} \tag{3}$$

Applying this transformation to all pixels yields a structured 3D point cloud.

2.1.2 Normal Vector Computation

The surface normal vector $\vec{n}$ at a point P is orthogonal to the surface's tangent plane at that point. We can approximate this normal by computing the cross product of two tangent vectors, which can be estimated from the partial derivatives of the 3D point positions with respect to the image grid coordinates, u and v.

$$\vec{t_u} = \frac{\partial P}{\partial u}, \quad \vec{t_v} = \frac{\partial P}{\partial v} \tag{4}$$

In our discrete grid, these partial derivatives are approximated using finite differences between adjacent points in the 3D point cloud. The normal vector $\vec{n}$ is then given by their cross product:

$$\vec{n} = \frac{\partial P}{\partial u} \times \frac{\partial P}{\partial v} \tag{5}$$

To ensure that our loss function only considers orientation, the resulting normal vector is normalized to a unit vector $\hat{n}$:

$$\hat{n} = \frac{\vec{n}}{\|\vec{n}\|} \tag{6}$$

2.2 Surface Normal Loss Function

The core of our proposed regularization is the Surface Normal Loss, $\mathcal{L}_{\text{normal}}$. It is formulated to minimize the angular distance between the predicted surface normals ($\hat{n}_{\text{pred}}$) and the ground truth normals ($\hat{n}_{\text{gt}}$). We use the cosine similarity, computed via the dot

product of the unit normal vectors, as our metric. The loss for a set of N valid pixels is defined as:

$$\mathcal{L}_{\text{normal}} = \frac{1}{N} \sum_{i=1}^{N} \left(1 - (\hat{n}_{\text{pred},i} \cdot \hat{n}_{\text{gt},i})\right) \tag{7}$$

This loss is bounded between $[0, 2]$. It reaches its minimum of 0 when the predicted and ground truth normals are perfectly aligned, and its maximum of 2 when they are oriented in opposite directions. Minimizing this loss encourages the model to generate surfaces with orientations that are locally consistent with the ground truth geometry.

2.3 Integration into the Training Framework

To avoid the high computational cost of decoding the latent representation to a full-resolution depth map at each training step, we perform the entire normal loss calculation in the VAE's latent space. The 4-channel latent representation from the UNet is first averaged across the channel dimension to produce a single-channel "pseudo-depth" map. This map, while not a true depth map, preserves the essential spatial and structural information required for robust normal estimation.

The final training objective combines the standard latent MSE loss from the diffusion model with our proposed surface normal loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \lambda \cdot \mathcal{L}_{\text{normal}} \tag{8}$$

where $\mathcal{L}_{\text{MSE}}$ is the mean squared error between the predicted and target noise (or sample) in the diffusion process, and λ is a hyperparameter that balances the influence of the geometric regularization term. This combined objective trains the model to be accurate in the diffusion target space while simultaneously adhering to local geometric constraints, thus leading to higher-quality depth predictions.