From Projection to Perception: A Mathematical Exploration of Shadow-based Neural Reconstruction

A research report submitted to the Scientific Committee of the Hang Lung Mathematics Award

Team Number
2596873

Team Members
Wong Yuk To, Hung Kwong Lam
Cheung Tsz Lung, Chan Ngo Tin, Zhou Lam Ho

Teacher
Mr. Chan Ping Ho

School
Po Leung Kuk Celine Ho Yam Tong College

Date
July 5, 2025

Abstract

This paper explores ShadowNeuS [LWX23], a neural network that reconstructs 3D geometry from single-view camera images using shadow and light cues. Unlike traditional 3D reconstruction methods relying on multi-view cameras or sensors, ShadowNeuS leverages a neural signed distance field (SDF) for accurate 3D geometry reconstruction. Analysis of the training process reveals deep connections to projective geometry, spatial reasoning in $\mathbb{R}^3$, and the network's perception of three-dimensional space.

Contents

1 Background

1.1 What is 3D Reconstruction from Images?

The goal of 3D reconstruction is to recover the structure of a 3D scene using only 2D images. Consider a 3D scene represented by a set of points $\mathbf{P} = (P_x, P_y, P_z) \in \mathbb{R}^3$. Each image taken of the 3D scene contains a set of pixel points $\mathbf{p} = (p_x, p_y) \in \mathbb{R}^2$. The process of capturing a 3D point in a 2D image I_n can be modeled as a projection function π_n

$$\pi_n : \mathbb{R}^3 \to \mathbb{R}^2, \quad (P_x, P_y, P_z) \mapsto (p_x, p_y) \tag{1}$$

This function represents how a camera maps a 3D point to a 2D pixel in the n-th image. To reconstruct the 3D scene, we need to solve the inverse problem π_n^{-1}.

$$\pi_n^{-1}(\mathbf{p}) = \left\{ \mathbf{P} \in \mathbb{R}^3 \mid \pi_n(\mathbf{P}) = \mathbf{p} \right\} \tag{2}$$

However, this inverse problem is typically **ill-posed**, as multiple 3D points may project to the same 2D pixel, leading to ambiguity. We will detail in Section 1.4.

1.2 Information Encoded in 2D Images

A 2D image I_n can provide multiple types of information, such as color and texture.

The information available from an image includes:

- **Pixel coordinates**: $\mathbf{p} = (p_x, p_y) \in \mathbb{R}^2$, represents the location of each pixel in the image
- **Color values**: $C_n(\mathbf{p}) = [r, g, b] \in [0,1]^3$, represents the RGB value of each pixel in the image
- **Image gradient**:

$$\nabla I_n(\mathbf{p}) = (\nabla r(\mathbf{p}), \nabla g(\mathbf{p}), \nabla b(\mathbf{p})) = \left(\underbrace{\left[\frac{\partial r}{\partial p_x}, \frac{\partial r}{\partial p_y} \right]^\top}_{\text{red channel}}, \underbrace{\left[\frac{\partial g}{\partial p_x}, \frac{\partial g}{\partial p_y} \right]^\top}_{\text{green channel}}, \underbrace{\left[\frac{\partial b}{\partial p_x}, \frac{\partial b}{\partial p_y} \right]^\top}_{\text{blue channel}} \right) \in \mathbb{R}^6 \tag{3}$$

  It captures local changes in intensity, indicating edges or texture information in the image

- **Learned features**: $\phi(I_n)(\mathbf{p}) \in \mathbb{R}^d$, represents high-dimensional features extracted from the image using methods like convolutional neural networks (CNNs) or other feature extractors

These data result from projecting 3D structures through a camera. For example, the color $C_n(\mathbf{p})$ may correspond to the visible surface of a 3D object, while $\nabla I_n(\mathbf{p})$ may hint the edge of the shape of that 3D object, etc.

1.3 The Forward Projection: From 3D World to 2D Image

We formalize the perspective projection process that projects a 3D point $\mathbf{P} = (P_x, P_y, P_z)$ to a pixel point $\mathbf{p} = (p_x, p_y)$ using the camera parameters.

Camera parameters:

- **Extrinsic** (world-to-camera transformation)
  - **Camera center**: $\mathbf{C} = (C_x, C_y, C_z) \in \mathbb{R}^3$, represents the position of the camera in the world coordinate.
  - **Rotation matrix**: $R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \in SO(3)$, represents a 3x3 matrix that rotates the world to align with the orientation of the camera.
  - **Translation vector**: $\mathbf{t} = -R\mathbf{C} \in \mathbb{R}^3$, represents the translation that aligns the camera center with the world origin.
  - **Homogeneous transformation matrix**: $T = \begin{bmatrix} R & \mathbf{t} \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{4\times 4}$
- **Intrinsic** (projection to image plane)
  - **Intrinsic matrix**:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \in \mathbb{R}^{3\times 3} \tag{4}$$

    where f_x, f_y are the focal lengths in pixels and c_x, c_y is the principal point (the pixel coordinates where the camera's lens is optically centered)

Forward Projection Pipeline:

We use homogeneous coordinates $\mathbf{P}_{\text{hom}}$ where an extra variable is added to handle scaling. The process involves:

1. **World to camera coordinate**: Transform $\mathbf{P} = (P_x, P_y, P_z)$ to camera coordinates

$$\mathbf{P} \to T \begin{bmatrix} \mathbf{P} \\ 1 \end{bmatrix} = \begin{bmatrix} R\mathbf{P} + \mathbf{t} \\ 1 \end{bmatrix} \tag{5}$$

2. **Perspective projection**:

$$\mathbf{P}_{\text{hom}} = K[R|\mathbf{t}] \begin{bmatrix} \mathbf{P} \\ 1 \end{bmatrix} = K(R\mathbf{P} + \mathbf{t}) = \begin{bmatrix} p'_x \\ p'_y \\ z' \end{bmatrix} \tag{6}$$

3. **Normalization**: Convert to 2D pixel coordinates by the scaling factor z'

$$\begin{bmatrix} p_x \\ p_y \end{bmatrix} = \frac{1}{z'} \begin{bmatrix} p'_x \\ p'_y \end{bmatrix}, \quad z' \neq 0 \tag{7}$$

Result:

$$\mathbf{P}_{\text{hom}} = \begin{bmatrix} p'_x \\ p'_y \\ z' \end{bmatrix} = K[R|\mathbf{t}] \begin{bmatrix} \mathbf{P} \\ 1 \end{bmatrix}, \quad \begin{bmatrix} p_x \\ p_y \end{bmatrix} = \frac{1}{z'} \begin{bmatrix} p'_x \\ p'_y \end{bmatrix}, \quad z' \neq 0 \tag{8}$$

1.4 The Inverse Problem: From 2D Image to 3D World

We attempt to invert the forward projection and recover the 3D point $\mathbf{P} = (P_x, P_y, P_z)$ from its 2D image projection $\mathbf{p} = (p_x, p_y)$. From Section 1.3, the forward projection is given by (refer to equation (8))

$$\mathbf{P}_{\text{hom}} = \begin{bmatrix} p'_x \\ p'_y \\ 1 \end{bmatrix} z' = K(R\mathbf{P} + \mathbf{t}) \tag{9}$$

Since $p'_x = z' p_x$ and $p'_y = z' p_y$ (refer to equation (7)), the homogeneous image coordinates are

$$\begin{bmatrix} p'_x \\ p'_y \\ z' \end{bmatrix} = z' \begin{bmatrix} p_x \\ p_y \\ 1 \end{bmatrix} \tag{10}$$

To recover $\mathbf{P}$, we invert the projection (refer to equation (9)):

$$R\mathbf{P} + \mathbf{t} = K^{-1} z' \begin{bmatrix} p_x \\ p_y \\ 1 \end{bmatrix} \tag{11}$$

$$\mathbf{P} = R^{-1} \left(z' K^{-1} \begin{bmatrix} p_x \\ p_y \\ 1 \end{bmatrix} - \mathbf{t} \right) \tag{12}$$

Since $\mathbf{t} = -R\mathbf{C}$, we have $-R^{-1}\mathbf{t} = -R^{-1}(-R\mathbf{C}) = \mathbf{C}$. We obtain:

$$\mathbf{P}(z') = \mathbf{C} + z' \cdot \left(R^{-1} K^{-1} \begin{bmatrix} p_x \\ p_y \\ 1 \end{bmatrix} \right) \tag{13}$$

This can be reformulated as a camera ray (refer to equation (13)):

$$\mathbf{P}(\lambda) = \mathbf{C} + \lambda \cdot \mathbf{d}, \quad \lambda > 0, \quad \mathbf{d} = \left(R^{-1} K^{-1} \begin{bmatrix} p_x \\ p_y \\ 1 \end{bmatrix} \right) \tag{14}$$

Note that $\mathbf{d}$ is the viewing direction ray in 3D space starting from the camera center $\mathbf{C}$.
The problem is **ill-posed** because the depth λ is unknown, meaning p defines a ray of possible 3D points rather than an unique P. To make the problem well-posed, additional constraints are needed, such as stereo vision or depth sensors, which provide depth information or multiple viewpoints to determine an unique λ.

References

[LWX23] Jingwang Ling, Zhibo Wang, Feng Xu. *ShadowNeuS: Neural SDF Reconstruction by Shadow Ray Supervision*. arXiv: 2211.14086, 2023.