SOFI-UGCL

Overview

This document outlines a method for recovering the full camera projection matrix

$$P = K[R \mid t]$$

from a single image, using vanishing-point geometry. The approach builds upon the outputs of SOFI (Multi-Scale Deformable Transformer for Camera Calibration), which provides estimates of the zenith vanishing point and the horizon line.

1 The Projection Matrix P

The camera projection matrix $P \in \mathbb{R}^{3\times 4}$ maps 3D world points $X = (X, Y, Z, 1)^T$ in homogeneous coordinates to 2D image points $x = (u, v, w)^T$ via:

$$x \sim PX \quad \Rightarrow \quad (u/w,\ v/w) = \text{image coordinates.}$$

We can write P in terms of its columns as:

$$P = \begin{bmatrix} p_1 & p_2 & p_3 & p_4 \end{bmatrix}$$

where (up to a scale factor for p_1, p_2, p_3):

- p_1 is the image of the world-X axis's vanishing point (vanishing point of lines parallel to world X-axis),
- p_2 is the image of the world-Y axis's vanishing point,
- p_3 is the image of the world-Z axis's vanishing point (i.e., vertical vanishing point),
- p_4 is the image of the world origin $(0, 0, 0)$.

More precisely, if $R = [r_1\ r_2\ r_3]$ where r_i are the world axis directions in camera coordinates, then $P = K[r_1\ r_2\ r_3\ t]$, implying $p_1 = Kr_1, p_2 = Kr_2, p_3 = Kr_3, p_4 = Kt$.

2 Decomposing $P = K[R \mid t]$

Intrinsic Matrix K

The camera's intrinsic matrix K is

$$K = \begin{pmatrix} f_x & s & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{pmatrix},$$

where f_x, f_y are the focal lengths in pixels, s is the skew (usually zero), and (u_0, v_0) is the principal point.

Forming f_x, f_y from Field-of-View

SOFI predicts a single horizontal field-of-view angle $\hat{f}$. For an image of width W pixels,

$$\tan\Bigl(\tfrac{\hat{f}}{2}\Bigr) = \frac{W/2}{f_x} \quad \Longrightarrow \quad f_x = \frac{W/2}{\tan\bigl(\frac{\hat{f}}{2}\bigr)}.$$

Assuming square pixels, we set

$$f_y = f_x.$$

Predicting the Principal Point

We append two small regression heads to SOFI's $\hat{q}_{\text{fov}}$ embedding:

$$\hat{u}_0 = w_u^\top \hat{q}_{\text{fov}} + b_u, \quad \hat{v}_0 = w_v^\top \hat{q}_{\text{fov}} + b_v.$$

Assembling $\hat{K}$

Putting it all together,

$$\hat{K} = \begin{pmatrix} \hat{f}_x & 0 & \hat{u}_0 \\ 0 & \hat{f}_y & \hat{v}_0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Extrinsic Parameters

The extrinsic parameters are:

- $R \in \mathbb{R}^{3\times 3}$: rotation matrix (world-to-camera),
- $t \in \mathbb{R}^3$: translation vector from world to camera frame,

so that

$$P = K\,[R \mid t], \quad \text{and} \quad p_4 = K\,t.$$

3 Obtaining R from SOFI Outputs

SOFI predicts:

- Zenith vanishing point $\hat{z} = (\hat{u}_z, \hat{v}_z, 1)^T$,
- Horizon line $\ell = (a, b, c)^T$.

Back-projecting to 3D

The world Z-axis direction (vertical) in the camera frame, r_3, is derived from the zenith vanishing point $\hat{z}$:

$$v_z = \hat{K}^{-1}\,\hat{z},$$
$$r_3 = \frac{v_z}{\|v_z\|}.$$

To obtain the full rotation matrix $R = [r_1\ r_2\ r_3]$, we need to determine r_1 and r_2. Since r_3 is the world's vertical direction, r_1 and r_2 must lie in the horizontal plane (i.e., be orthogonal to r_3) and be orthogonal to each other. The zenith vanishing point and horizon line alone do not uniquely determine the camera's "roll" around the vertical axis. To resolve this ambiguity, a common approach is to align the world X-axis (r_1) with the projection of a default camera direction (e.g., its own X-axis) onto the horizontal plane.

We align the world X-axis (r_1) with the projection of the camera's X-axis onto the horizontal plane defined by r_3:

$$\begin{aligned}
&\text{Let } c_x = (1,0,0)^T \text{ be the camera X-axis direction in its own frame.} \\
&r_1' = c_x - (c_x^\top r_3) r_3 \quad \text{(Projection of } c_x \text{ onto plane orthogonal to } r_3\text{)} \\
&r_1 = \frac{r_1'}{\|r_1'\|}. \\
&r_2 = r_3 \times r_1 \quad \text{(Ensuring a right-handed orthonormal basis).}
\end{aligned}$$

Finally, assemble the rotation matrix:

$$R = \begin{bmatrix} r_1 & r_2 & r_3 \end{bmatrix}.$$

Training constraints for the rotation matrix:

$$L_{\text{ortho}} = \|R\,R^\top - I\|_F^2, \quad L_{\text{det}} = (\det R - 1)^2.$$

4 Recovering p_4: Image of the World Origin

The world XZ-plane ($Y = 0$) contains the world X-axis and Z-axis. Its image is the line $\ell_{Y=0}$. Since the vanishing points p_1 (for X-axis) and p_3 (for Z-axis) lie on this plane, its image line $\ell_{Y=0}$ must pass through them. Thus, $\ell_{Y=0} = p_1 \times p_3$. Similarly, the world YZ-plane ($X = 0$) contains the world Y-axis and Z-axis. Its image line $\ell_{X=0}$ must pass through their vanishing points p_2 and p_3. Thus, $\ell_{X=0} = p_2 \times p_3$.

The world origin $(0,0,0)$ lies on both the world XZ-plane and the world YZ-plane. Therefore, its image, p_4, must lie on the intersection of their image lines, $\ell_{Y=0}$ and $\ell_{X=0}$.

$$p_4 = \ell_{X=0} \times \ell_{Y=0}.$$

5 Recovering Translation t

Since $p_4 = K\,t$, we can recover the translation vector direction:

$$\tilde{t} = \hat{K}^{-1} p_4, \quad \hat{t} = \frac{\tilde{t}}{\|\tilde{t}\|}.$$

This recovers the translation direction up to an unknown scale, as is typical when estimating depth from a single image without additional information.

6 Final Projection Matrix

Thus, the full projection matrix can be constructed as:

$$P = \begin{bmatrix} \hat{K}\,r_1 & \hat{K}\,r_2 & \hat{K}\,r_3 & \hat{K}\,\hat{t} \end{bmatrix}.$$

7 5. SOFI Core Losses

For SOFI's three camera outputs and line queries, we use:

$$
\begin{aligned}
L_{\mathrm{zvp}} &= 1 \;-\; \frac{z^{\top}\hat{z}}{\|z\|\|\hat{z}\|}, \\
L_{\mathrm{hl}} &= \max\bigl(\|b_l - \hat{b}_l\|_1,\; \|b_r - \hat{b}_r\|_1\bigr), \\
L_{\mathrm{FoV}} &= \bigl|\, f - \hat{f}\,\bigr|, \\
L_{\mathrm{class}},\, L_{\mathrm{score}} &: \text{ Focal Loss on line class/confidence.}
\end{aligned}
$$

Then, the total SOFI loss is:

$$
L_{\mathrm{SOFI}} = 5L_{\mathrm{zvp}} + 5L_{\mathrm{hl}} + 5L_{\mathrm{FoV}} + L_{\mathrm{class}} + L_{\mathrm{score}}.
$$

8 6. UGCL-Style Geometric Constraints

6.1 Rotation Matrix Constraints

These losses ensure that the rotation matrix R is orthonormal.

$$
\begin{gathered}
L_{r_{12}} = (r_1^{\top} r_2)^2, \quad L_{r_{13}} = (r_1^{\top} r_3)^2, \quad L_{r_{23}} = (r_2^{\top} r_3)^2, \\
L_{R\mathrm{iso}} = \|R\,R^{\top} - I\|_F^2, \quad L_{\mathrm{det}} = (\det R - 1)^2.
\end{gathered}
$$

6.2 World-Origin Self-Consistency

This loss ensures that the derived image of the world origin p_4 is consistent with the estimated intrinsic matrix $\hat{K}$ and the recovered translation direction $\hat{t}$, effectively enforcing the relationship $p_4 \sim \hat{K}\,\hat{t}$. Note that this is an internal consistency check and does not require ground truth for translation.

$$
L_{\mathrm{wc}} = \bigl\|p_4 - \hat{K}\,\hat{t}\bigr\|_1.
$$

6.3 Horizon Line Consistency

The normal to the horizon plane in camera coordinates should be parallel to the zenith direction. Let $n_h = \hat{K}^{-\top}\ell$ be the normal to the horizon plane, and r_3 be the zenith direction.

$$
L_{\mathrm{horizon\_normal}} = \|n_h \times r_3\|^2.
$$

9 7. Final Unified Loss

$$
\boxed{
\begin{aligned}
L_{\mathrm{total}} = L_{\mathrm{SOFI}} \;+\; & \lambda_r\,(L_{r_{12}} + L_{r_{13}} + L_{r_{23}}) \;+\; \lambda_{\mathrm{iso}}\,L_{R\mathrm{iso}} \\
&+\; \lambda_{\mathrm{det}}\,L_{\mathrm{det}} \;+\; \lambda_{\mathrm{wc}}\,L_{\mathrm{wc}} \;+\; \lambda_{\mathrm{horizon\_normal}}\,L_{\mathrm{horizon\_normal}}.
\end{aligned}
}
$$

Each weight λ (e.g. 0.1) softly enforces its constraint while preserving SOFI's learned predictions.