Supplementary for FastPose-ViT: A Vision Transformer for Real-Time Spacecraft Pose Estimation

Pierre Ancey[1] Andrew Price[1] Saqib Javed[1] Mathieu Salzmann[1,2]

[1]EPFL [2]Swiss Data Science Center
Lausanne, Switzerland

{firstname.lastname}@epfl.ch

0.1. Derivation for Spatial Augmentation

In the main paper, we state that for an in-plane rotation matrix M and camera intrinsic matrix K, the term $K^{-1}MK$ simplifies to a pure 3D rotation matrix $R_z(\theta)$. Here we provide the proof.

The intrinsic matrix K and its inverse K^{-1} are given by:

$$K = \begin{pmatrix} f_x & 0 & W/2 \\ 0 & f_y & H/2 \\ 0 & 0 & 1 \end{pmatrix}, K^{-1} = \begin{pmatrix} \frac{1}{f_x} & 0 & -\frac{W}{2f_x} \\ 0 & \frac{1}{f_y} & -\frac{H}{2f_y} \\ 0 & 0 & 1 \end{pmatrix} \tag{1}$$

The in-plane rotation matrix $M(\theta)$ which rotates an image by angle θ around its center $(W/2, H/2)$ is:

$$M(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta & \frac{W}{2}(1-\cos\theta) + \frac{H}{2}\sin\theta \\ \sin\theta & \cos\theta & \frac{H}{2}(1-\cos\theta) - \frac{W}{2}\sin\theta \\ 0 & 0 & 1 \end{pmatrix} \tag{2}$$

By performing the matrix multiplication for $K^{-1}M(\theta)$ and then $K^{-1}M(\theta)K$, and assuming $f_x = f_y$ (a common case), the expression simplifies significantly. The final result of the multiplication $K^{-1}M(\theta)K$ is:

$$K^{-1}M(\theta)K = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} = R_z(\theta) \tag{3}$$

This is the standard rotation matrix for a counter-clockwise rotation of angle θ about the z-axis. It is an orthogonal matrix with a determinant of 1, confirming it is a valid rotation matrix in $SO(3)$.

0.2. ViT Backbone Performance Trade-Off

In Table 1 we summarize the performance of various ViT backbones. As expected, larger models with higher resolution and smaller patch sizes generally perform better but require more floating point operations. We focus our analysis on the ViT-B-224/16 and ViT-B-384/16 due to their favorable trade-off between performance and latency.

Table 1. Performance of various ViT backbones on SPEED.

ViT Backend	GFLOPS	Params (M)	E_T [m] ↓	E_R [deg] ↓
ViT-B-224/16	17.6	86.6	0.0231	0.4469
ViT-B-384/16	55.5	86.9	0.0221	0.4183
ViT-B-224/32	4.4	88.2	0.0425	0.7907
ViT-L-224/16	61.55	304.3	0.0218	0.4068
ViT-L-512/16	361.99	305.2	**0.0182**	**0.3380**
ViT-L-224/32	15.38	306.5	0.0437	0.8079
ViT-H-224/14	167.29	632.0	0.0211	0.3896
ViT-H-518/14	1016.7	633.5	0.0188	0.3438