ViT-Based COLMAP

CheWei Hsu [1] Jui-Hung Chen [1]

[1]Universitiy of Southern California

76

1. Motivations

Traditional 3D reconstruction systems like COLMAP rely on classic computer vision algorithms such as SIFT, which often fail in low-texture or repetitive environments. This project explores the use of Vision Transformer (ViT)-based models as a replacement for SIFT in the Structure-from-Motion (SfM) pipeline.

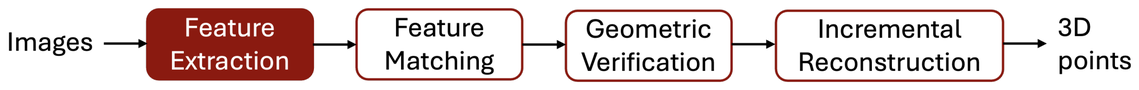


Figure 1. SfM Pipeline

2. Methods

Ground Truth Generation

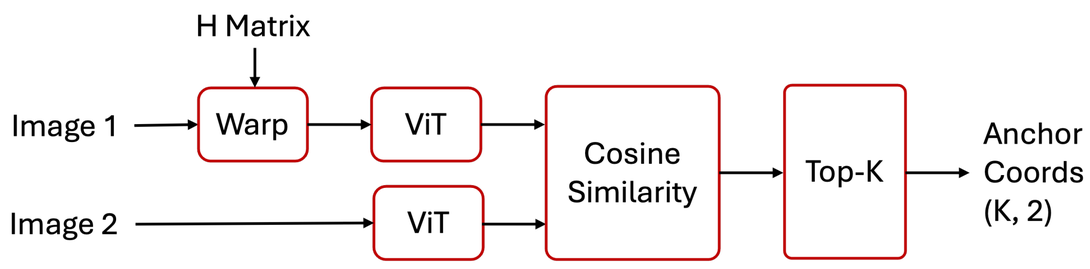


Figure 2. Flow to Generate Positives

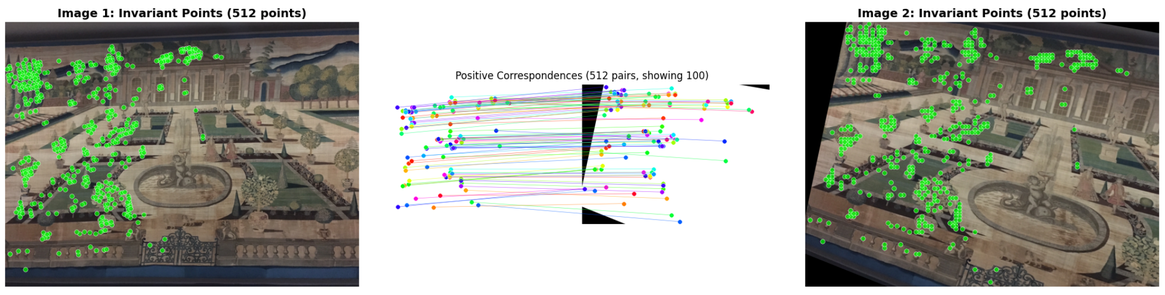
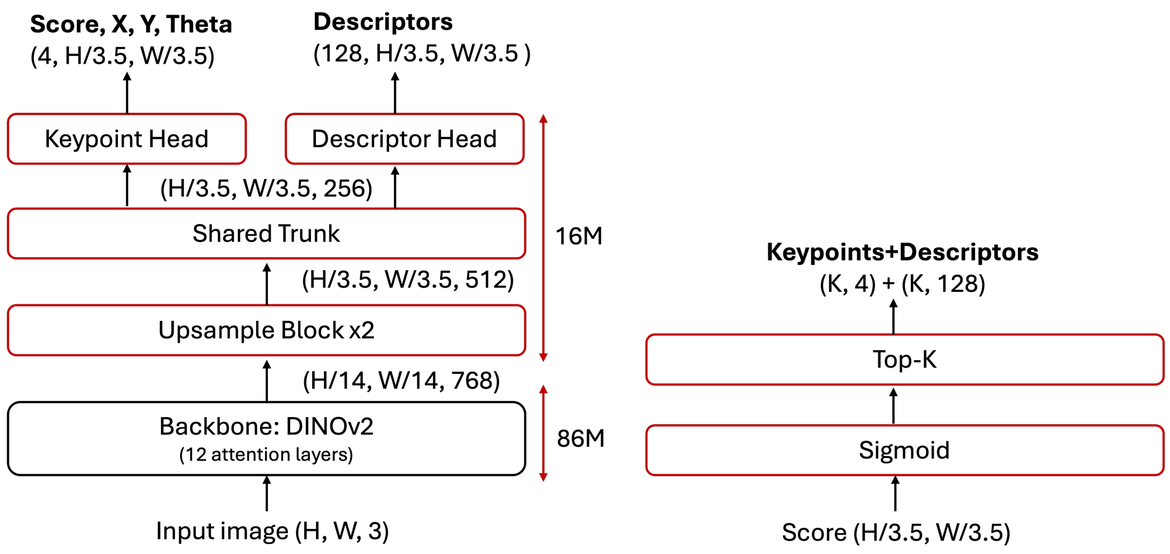


Figure 3. Extracted Positive Keypoints

Model Architecture



Training Losses

$$L_{\text{total}} = \lambda_{\text{det}} L_{\text{det}} + \lambda_{\text{ori}} L_{\text{ori}} + \lambda_{\text{des}} L_{\text{des}}$$

- **Detector Loss**: $L_{\text{det}} = \textbf{BCE}(\text{Score}_{\text{pred}}, \text{Score}_{\text{gt}})$
  - Score_{gt}: Gaussian heatmap computed from top-K invariant points
- **Orientation Loss**: $L_{\text{ori}} = \textbf{MSE}(\text{atan}(\sin(\Delta_k), \cos(\Delta_k))$
  - $\Delta_k = \theta_{k2} - (\theta_{k1} + \theta_{\text{rot}_k})$
- **Descriptor Loss**: $L_{\text{des}} = L_{\text{pos}} + L_{\text{neg}}$
  - $L_{\text{pos}} = 1 - \langle \text{anchor, pos} \rangle$
  - $L_{\text{neg}} = \text{ReLU}(m + d(\text{anchor, pos}) - \min_i d(\text{anchor, neg}_i))$

Negative Sampling Strategies

(a) In-Image Negatives (b) Cross-Image Negatives (c) Hard Negatives

3. Experiments

ViT Model Selection: DINOv2 or BEiT?

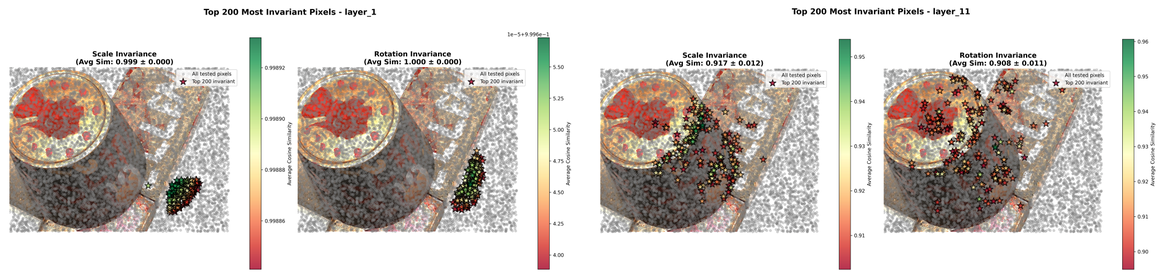


Figure 4. BEiT Invariant Points

Figure 5. DINOv2 Invariant Points

ViT Extractor Selection

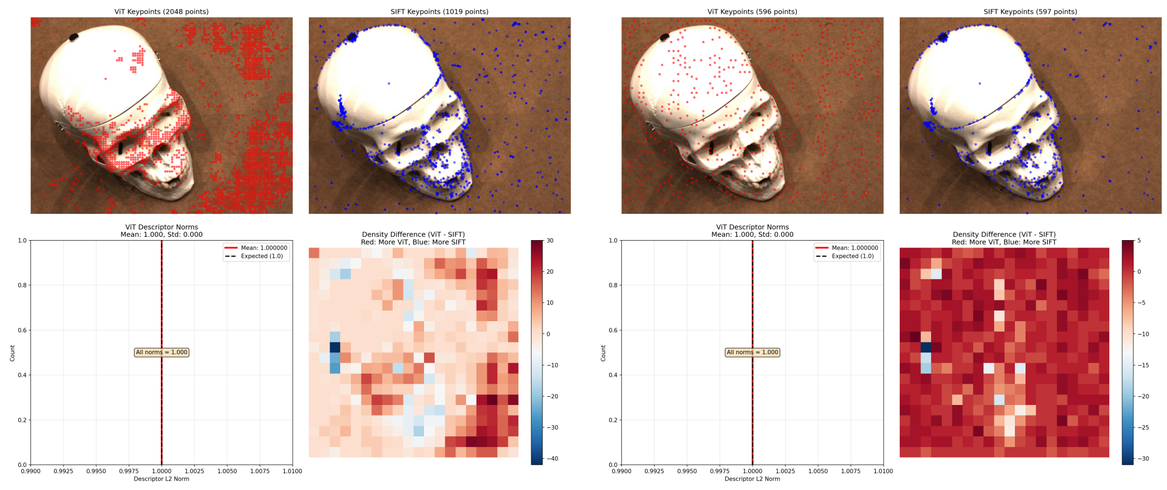
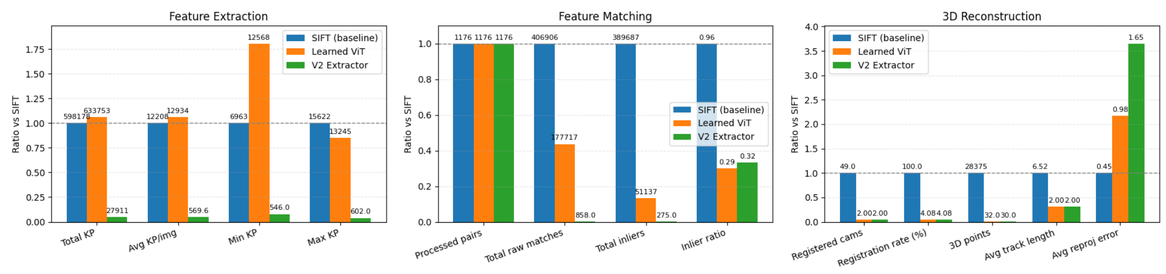


Figure 6. V1 baseline extractor

Figure 7. V2 extractor

4. Results & Discussions

SIFT, V2 Extractor, and Learned Extractor

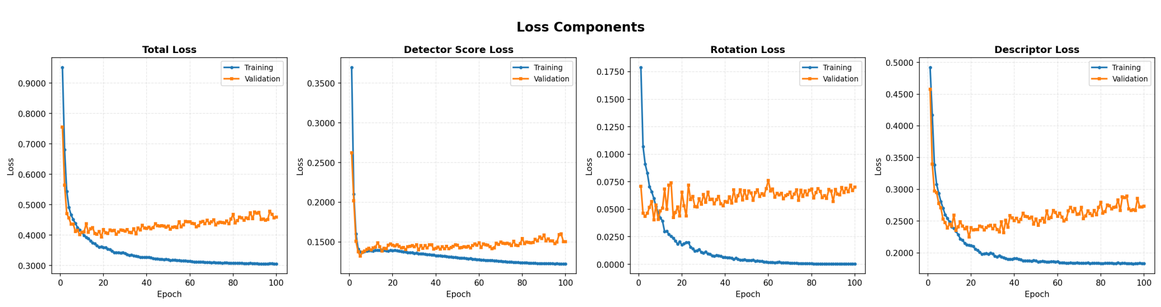


Matched Points

Figure 8. SIFT

Figure 9. Learned Extractor

Overfitting



5. Future Work

Data Augmentation: **1) Synthetic data**: Applying random homographies to create training pairs with controlled transformations. **2) Threshold-based positive selection**: Selecting points based on similarity threshold instead of a fixed number of top-K points.

Loss Function Improvement: Modify the loss function to optimize for geometrically consistent, repeatable features that produce high-quality inlier matches passing RANSAC, which is required for COLMAP's 3D reconstruction initialization.