S³-Match: Common-View Aligned Image Matching via Self-Supervised Keypoint Selection

Shizhen Li [1], Jingcheng Liu [1], Jianwu Fang [1], DeZheng Gao [1], Jianru Xue [1]

[1]Xi'an Jiaotong University, China

Introduction

- Common image matching methods rely on sparse interest points to manage distortions, occlusions, and noise. However, they often struggle with repeatability and reliable description in scenes with sparse textures, varying viewpoints, or lighting.
- S³-Match overcomes these challenges by autonomously identifying distinctive and stable feature points using a quality score during training.

Contributions

- **Self-supervised Learning**: Automatically identifies and selects distinctive, stable keypoints without the need for manually designed rules.
- **Global Attention Mechanism**: Aligns features in common-view areas, enhancing keypoint matching and descriptor generation.
- **Efficient Feature Pyramid**: Produces high-resolution, multi-scale descriptors while optimizing computational efficiency.

Methods

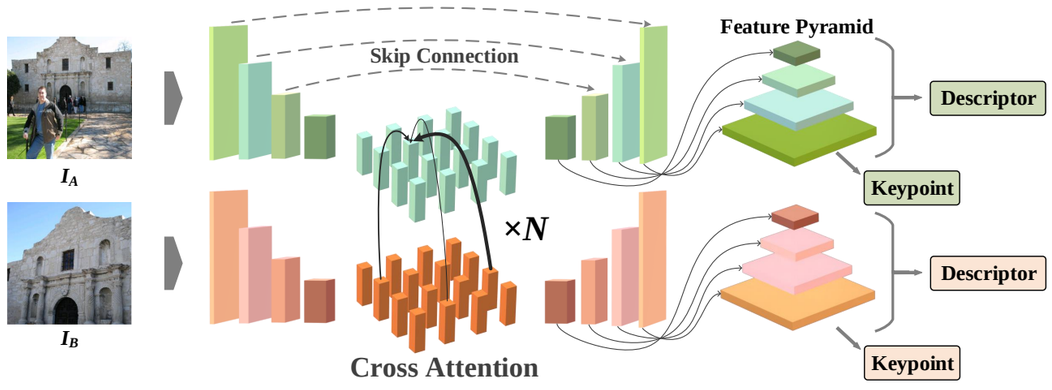

Figure 1. System Overview of S³-Match.

The framework of the S³-Match algorithm, designed for dual-image matching in three-dimensional scenes. In this approach, a U-Net encoder network is first used to extract features from both images independently. Subsequently, a cross-attention mechanism is applied on lower-resolution feature maps to align features within the common-view areas.The extraction of keypoints occurs at the lowest level of the feature pyramid, The generation of descriptors is based on sampling on the entire feature pyramid at the keypoint coordinates and concatenating these sampled results into a multi-scale descriptor.

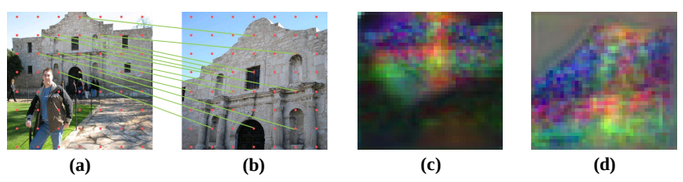

Figure 2. Global Consistency in Feature Backgrounds: (a, b) illustrate the matching relationships between the original image and its specific regions; (c, d) show the feature maps.

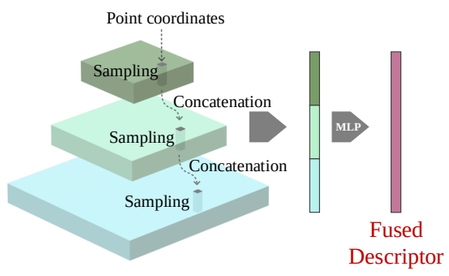

Figure 3. Multi-Scale Descriptor Sampling Module

S³-Match utilizes a bidirectional cross-attention mechanism, to provide consistent, common-view aligned feature maps for descriptor generation and keypoint selection across the two images.

Adaptive Feature Point Selection

The S³-Match algorithm introduces an adaptive interest point exploration mechanism. We have designed criteria for assessing the advantages and disadvantages of feature point selection. The matching quality score, denoted as R, is defined as follows:

$$R_i = \exp(-\beta \cdot \epsilon_{p_A^i, p_B^i}) \cdot Q_{p_A^i, p_B^i}, \quad i = 1, 2, \ldots, N, \tag{1}$$

where ϵ denotes the matching error between two candidate-matched points. β is a constant coefficient used to scale the error to a suitable magnitude, while $Q_{p_A^i, p_B^i}$ represents the confidence probability between the matching pairs.

During the training process of deep learning networks, improvements in network performance typically lead to gradual increases in R. To ensure that the distribution of R remains stable despite network improvements, we normalize R:

$$R_i = \max(\min(0.5 + \frac{R_i - \mu}{2\sigma}, 1), 0), \quad i = 1, 2, \ldots, N. \tag{2}$$

We utilize a self-supervised loss function L_{det}, which is composed of the following components:

$$\begin{aligned} L_{det} = & \frac{5}{N} \sum_{i=1}^{N} \{(M_A^{p_A^i} - R_i)^2 + (M_B^{p_B^i} - R_i)^2\} + \\ & \frac{1}{L} \sum_{j=1}^{L} \{(M_A^{\neg p_A^j})^2 + (M_B^{\neg p_B^j})^2\} + \frac{1}{HW} \sum_{k=1}^{HW} (M_A^k + M_B^k). \end{aligned} \tag{3}$$

Experiments

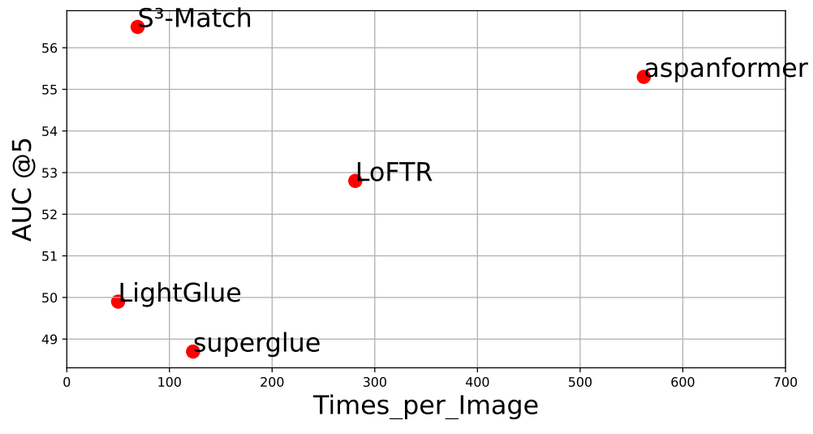

Figure 4. Performance and Efficiency Comparison between S³-Match and Other Open-Source Methods

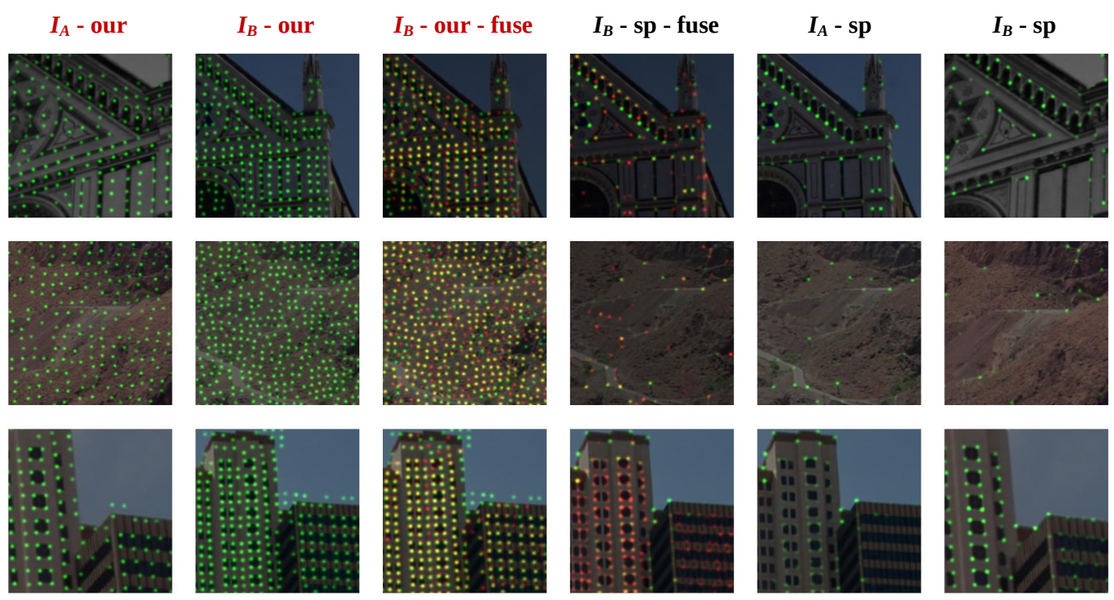

Figure 5. **Feature Point Detection Results.** The figure shows keypoint detection results by S³-Match in images A (I_A-our) and B (I_B-our). The integrated result is obtained by projecting keypoints from A onto B in red (I_B-our-fuse), compared with results from SuperPoint (I_B-sp-fuse).