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

Shizhen Li
lyhlsz@stu.xjtu.edu.cn
Jingcheng Liu
JingchengLiu@stu.xjtu.edu.cn
Jianwu Fang
fangjianwu@mail.xjtu.edu.cn
DeZheng Gao
gaodezheng@stu.xjtu.edu.cn
Jianru Xue
jrxue@mail.xjtu.edu.cn

Xi'an Jiaotong University, China

Abstract

This paper introduces S³-Match, a common-view aligned image matching algorithm via self-supervised keypoint selection. The most common image matching methods depend on sparse interest points to minimize dependence on non-essential information and to effectively manage significant distortions, occlusions, or noise. Nonetheless, the repeatability of interest points and their reliable description often degrade in scenes with sparse textures or when there are changes in appearance due to varying viewpoints and lighting conditions.

To overcome these challenges, S³-Match employs a quality score that autonomously identifies feature points with high distinctiveness and stability during the training phase. Furthermore, it incorporates a cross-attention mechanism that aligns features within the common-view areas across images. This alignment provides consistent feature information across images, and focuses subsequent self-supervised keypoint extraction and feature description on these common-view regions. Experimental results demonstrate that S³-Match significantly outperforms SuperPoint in terms of keypoint selection consistency and uniformity. It also exhibits superior performance in pose estimation tasks and surpasses other advanced algorithms in computational efficiency. Additionally, we have validated a variant of S³-Match that does not rely on cross-image information, capable of meeting a broader range of application needs.

1 Introduction

Accurate and efficient image matching serves as a fundamental component in a wide range of computer vision applications[25, 26], including 3D scene reconstruction[17, 18], camera calibrations, and simultaneous localization and mapping(SLAM)[15], etc. In the domain of image matching, methods are typically classified into two categories: detector-based and detector-free. Compared to detector-free approaches, detector-based methods provide several significant advantages, such as computational efficiency with sparse features and scalable adaptability across various applications. An effective detector-based method should accurately identify salient points and provide descriptors that are invariant to changes in scale, viewpoint, illumination, and other environmental variables. Traditional methods[9, 10, 13, 29] seek to achieve robustness through analyzing local gradient orientations and magnitudes. However, these techniques often lack adaptability to new or complex environments. In response, the advancement of deep learning technologies has catalyzed the emergence of numerous data-driven approaches in recent years, such as MagicPoint[4], SuperPoint[5], LIFT[27], etc. Those learning-based methods show powerful generalization and strong performance on keypoint detection and local feature description[12, 19].

Despite the advancements in learning-based methods for keypoint detection and description, several challenges remain [11]. Some methods are restricted to train individual parts of the feature extraction pipeline [16], while others allow for end-to-end training but still require the use of outputs from traditional hand-crafted detectors to initiate the training [5, 27, 28]. For instance, SuperPoint [5] learns to detect corners by generating a large set of synthetic shapes with annotated corners as ground truth. However, this training pipeline is complex due to the necessity of labeled keypoints, and its multi-phase training process is challenging to tune. Additionally, fixed rules for feature point selection may fail under certain textures (e.g., corner detectors are unable to utilize linear textures), thus failing to effectively utilize all the textural information available in the images.

To address this issue, our approach initially attempts to generate interest points distributed across the entire image. Subsequently, we design an evaluation criterion to assess the quality of these interest points. Finally, we guide the network to preferentially select positions of higher-quality interest points based on this quality assessment.

Figure 1 presents the results of feature point selection using the S³-Match algorithm. The figure displays the effectiveness of feature point selection in images I_A and I_B. Utilizing camera pose and depth information available in the dataset, keypoints from I_B are projected onto I_A in red. When the same geometric locations in both images are recognized as keypoints, the overlay of red and green visually appears yellow. In a contrast, the right side of the figure showcases the repeatability of keypoints using the SuperPoint algorithm. Compared to SuperPoint, S³-Match achieves a higher detection rate of feature points, better uniformity in distribution, and significantly superior repeatability.

Additionally, many methods [5, 6, 22, 30] detect keypoints on downsampled feature maps to save computational resources, often at the expense of image detail. To overcome these limitations, we introduce a pyramid-like feature storage approach that detects keypoints at full resolutions while generating descriptors through the integration of multi-level and cross-image features, thereby enhancing both detail preservation and computational efficiency.

The contributions of S³-Match are summarized as follows:

- S³-Match introduces a self-supervised learning approach, eliminating the need for manually designed rules for feature point selection. This enables the algorithm to autonomously identify and select feature points that are highly distinctive and stable.
- It incorporates a global attention mechanism that aligns the features of corresponding parts in the common-view regions of two images, thereby facilitating the generation of

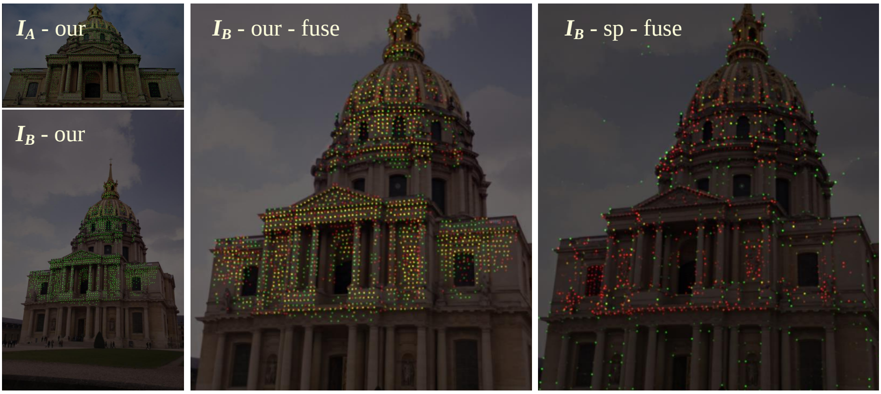

Figure 1: **Feature point detection results.** The figure displays the keypoint detection results by S³-Match in images A(I_A-our) and B(I_B-our), an integrated result achieved by projecting the keypoints from A onto B in red(I_B-our-fuse), and the comparative integration results from SuperPoint [5](I_B-sp-fuse).

descriptors and the selection of keypoints that focus on matchable areas.

- It features an efficient feature pyramid storage structure designed to achieve multi-scale, high-resolution dense descriptors, while also optimizing for computational efficiency.

2 Related Work

In the domain of learning-based image matching, there are detector-free models that directly produce semi-dense matches, and detector-based models, which generate keypoints along with their descriptions. This article primarily focuses on detector-based methods.

Keypoint detection and local feature matching have proven effective across a range of vision tasks. Traditional approaches typically detect keypoints before describing them [1, 24]. However, these ’detect-then-describe’ methods often suffer from reduced performance in keypoint detection and matching under conditions of significant changes in illumination, seasonal variations, and differing viewpoints. This decline is largely attributable to the keypoint detection process’s inherent emphasis on image details, which makes it sensitive to low-level information. Furthermore, training the detector and descriptor separately can result in information loss and inconsistencies between the keypoint detector and descriptor.

To overcome these challenges, the Joint Detection and Description approach has been proposed. This method unifies the tasks of keypoint detection and description, learning them concurrently within a single model. Such integration allows the model to leverage information from both tasks during optimization, thus enhancing its adaptability to specific vision tasks. A prominent example is SuperPoint [5], which adopts a self-supervised strategy to simultaneously determine keypoint locations at the pixel level and their descriptors.

In contrast to methods that rely on predefined corners to guide keypoint detection, approaches such as D2-Net [6], DISK [21], and ALIKED [30] enable models to autonomously discover keypoints in an end-to-end manner. Empirical evidence suggests that these methods offer superior matching performance, particularly under challenging conditions like strong variations in illumination and on weakly textured surfaces.

3 Methods

3.1 Algorithm Framework

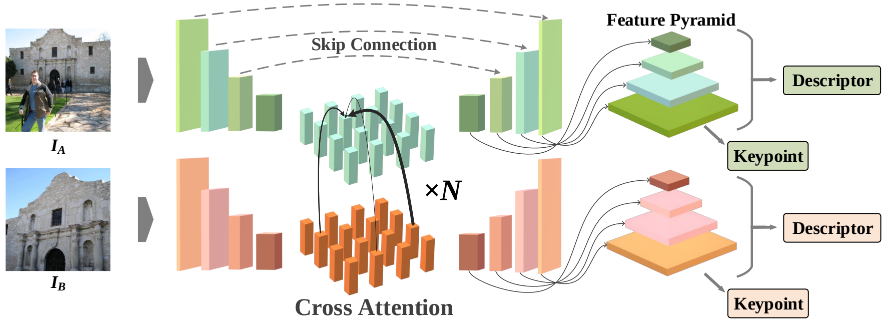

Figure 2: System Overview of S³-Match.

Figure 2 illustrates the framework of the S³-Match algorithm, designed for dual-image matching in three-dimensional scenes, namely I^A and I^B. In this approach, a U-Net network encoder with shared parameters is first used to extract features from both images independently. Subsequently, a cross-attention mechanism is employed on the lower-resolution feature maps to align features within the common-view areas, promoting consistency in the corresponding regions within common-view areas.

Then, we further incorporate the decoder portion of the U-Net network, and we integrate the last feature maps at various levels of the decoder to construct a feature pyramid.

In the S³-Match, the extraction of keypoints occurs at the lowest level of the feature pyramid, which corresponds to the terminal layers of the U-Net network. The generation of descriptors is based on sampling at the corresponding locations on the entire feature pyramid using bilinear interpolation at the keypoint coordinates and concatenating these sampled results into a multi-scale descriptor. This strategy avoids generating dense descriptors with many channels on high-resolution feature maps, thereby reducing information redundancy and wastage of computational resources.

3.2 Common-View Aligned Bottleneck Features

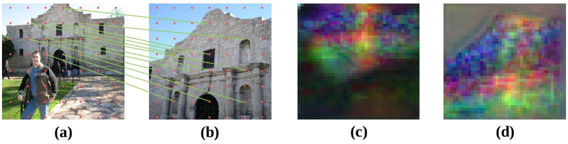

Figure 3: Global Consistency in Feature Backgrounds: (a, b) Illustrate the matching relationships between the original image and its specific regions. (c, d) Show the feature background images, where the channel values of the feature map are mapped to RGB channels and averaged.

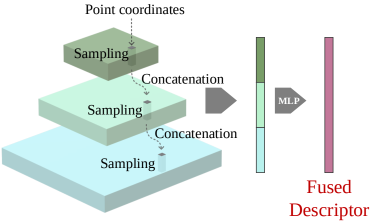

Figure 4: Multi-Scale Descriptor Sampling Module

S³-Match performs a flattening operation on the deepest bottleneck feature maps of the

U-Net, represented as F^A and F^B, to execute the cross-attention mechanism. We utilized the linear transformer architecture from LoFTR, which is detailed in the appendix. Notably, we omit the self-attention mechanism, replacing it instead with a convolution operation to facilitate intra-image information exchange after each cross-attention iteration.

S³-Match employs a bidirectional cross-attention mechanism applied four times to provide a common-view aligned consistent feature maps for descriptor generation and keypoint selection across two images, as illustrated in Figure 3. The figure clearly demonstrates that the spatial features corresponding to the two images are highly similar, and these features are focused on the matchable areas within the common-view regions. This focus facilitates the exclusion of interest point selection in non-common-view areas. We utilize ground truth of region correspondences to guide the generation of this features, with specific details on the supervision function available in the appendix.

3.3 Multi-Scale Fusion Descriptors

To efficiently extract and integrate feature information, we construct a feature pyramid that captures the final feature map from each scale of the U-Net decoder. We also introduce a descriptor sampling module that extracts descriptors through bilinear interpolation at feature point coordinates across these multi-scale feature maps, concatenating them along the feature dimension to form a multi-scale descriptor. These descriptors are then processed by a Multilayer Perceptron (MLP) for further feature integration. The entire process of obtaining descriptors from point coordinates is represented as $D = \mathcal{S}(P)$, as shown in Figure 4.

Assuming the fused descriptors for the feature points are $D_A = \mathcal{S}(P_A)$ and $D_B = \mathcal{S}(P_B)$, we computes the confidence matrix for matching relationships between feature points, denoted as $\mathcal{Q}$, as follows:

$$\mathcal{C}(i,j) = \frac{1}{\tau} \cdot \langle D_A(i), D_B(j) \rangle, \tag{1}$$

$$\mathcal{Q}(i,j) = \text{softmax}_1 \left(\mathcal{C}(i,\cdot)\right)_j \cdot \text{softmax}_1 \left(\mathcal{C}(\cdot,j)\right)_i . \tag{2}$$

In this calculation, $\mathcal{C}(i,j)$ quantifies the feature similarity between the ith point in I^A and the jth point in I^B, adjusted by a temperature parameter τ to modulate the sensitivity of the similarity metric. We introduce an improved softmax function that does not require the sum of matching probabilities to equal one, specifically designed for points without correspondences:

$$\text{softmax}_1(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j} + 1}. \tag{3}$$

Based on the confidence matrix $\mathcal{Q}$, we identifies matching pairs whose confidence exceeds the threshold θ_q and that also satisfy the nearest-neighbor matching criteria. As a result, we obtain N sets of matching pairs: $\{p_A^i, p_B^i\}$, where $i = 1, 2, ..., N$. For detailed information on descriptor supervision, please refer to the appendix.

3.4 Adaptive Feature Point Selection

S³-Match generates probability maps M_A and M_B consistent with the original image resolution directly within the final layer of the U-Net network, utilizing a 1×1 convolution and a

Sigmoid activation function. Non-maximum suppression is employed to select a fixed number of keypoints from the probability maps, yielding two sets of keypoints P_A and P_B. Subsequently, based on the aforementioned matching rules, we identify N sets of corresponding matching pairs: $\{p_A^i, p_B^i\}$, where $i = 1, 2, ..., N$.

To tackle the challenges of repeatability in feature point detection across varying viewpoints and the sparsity of feature points, the S³-Match algorithm introduces an adaptive interest point exploration mechanism. This mechanism dynamically adjusts the strategy for feature point extraction.

When evaluating the reliability of matching pairs p_A^i, p_B^i generated by current deep learning networks, two main criteria are considered: minimal error between selected keypoints to ensure their stability across images, and a high confidence level in the matches to affirm the distinctiveness and uniqueness of the feature points. Low confidence can result in unstable results and mismatches, adversely affecting match quality.

We have designed criteria for assessing the advantages and disadvantages of feature point selection. The matching quality score, denoted as $\mathcal{R}$, is defined as follows:

$$\mathcal{R}_i = \exp(-\beta \cdot \varepsilon_{p_A^i, p_B^i}) \cdot \mathcal{Q}_{p_A^i, p_B^i}, \quad i = 1, 2, \ldots, N, \tag{4}$$

where ε denotes the matching error between two candidate-matched points. When the dataset provides accurate camera poses and depth maps, we can compute the exact corresponding position of p_A^i in image B as $\hat{p_A^i}$, with ε being the Euclidean distance $\|p_A^i - \hat{p_A^i}\|$. If the depth maps in the dataset lack sufficient accuracy, ε can be estimated by computing the epipolar error based on camera poses. β is a constant coefficient used to scale the error to a suitable magnitude, while $\mathcal{Q}_{p_A^i, p_B^i}$ represents the confidence probability between the matching pairs, as defined in Equation 2, reflecting the distinguishability of the two feature points.

During the training process of deep learning networks, improvements in network performance typically lead to gradual increases in $\mathcal{R}$. To ensure that the distribution of $\mathcal{R}$ remains stable despite network improvements, we normalize $\mathcal{R}$. Initially, we calculate the mean and standard deviation of all the matching quality scores $\mathcal{R}_i$:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} R_i, \quad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(R_i - \mu)^2}. \tag{5}$$

Then, each $\mathcal{R}_i$ is normalized to control its distribution range:

$$\mathcal{R}_i = \max(\min(0.5 + \frac{\mathcal{R}_i - \mu}{2\sigma}, 1), 0), \quad i = 1, 2, \ldots, N. \tag{6}$$

To adaptively adjust the rules for interest point selection, we utilize a self-supervised loss function $\mathcal{L}_{det}$, which is composed of the following components:

$$\begin{aligned} \mathcal{L}_{det} =& \frac{5}{N}\sum_{i=1}^{N}\{(M_A^{p_A^i} - R_i)^2 + (M_B^{p_B^i} - R_i)^2\} + \\ & \frac{1}{L}\sum_{j=1}^{L}\{(M_A^{\neg p_A^j})^2 + (M_B^{\neg p_B^j})^2\} + \frac{1}{HW}\sum_{k=1}^{HW}(M_A^k + M_B^k). \end{aligned} \tag{7}$$

In the first component, $M_A^{p_A^i}$ and $M_B^{p_B^i}$ denote the keypoint probability values at the locations of candidate matching points p_A^i and p_B^i within the probability maps M_A and M_B, respectively. The second component, $\neg p_A^j$ and $\neg p_B^j$, pertains to points identified as keypoints by the network that fail to form matching relationships across images. The final component is designed to ensure that the probability values of the non-keypoint regions in the probability maps are minimized.

3.5 Fine-tuning of Matching Relationships

We propose a descriptor-based fine-tuning mechanism to enhance the precision of the matching process. Initially, descriptors are extracted from the multi-scale feature pyramid:

$$d_A^i = \mathcal{S}(p_A^i), \quad d_B^i = \mathcal{S}(p_B^i). \tag{8}$$

These descriptors provide detailed information on the differences between feature points, allowing for precise adjustments to their positions:

$$\Delta p_A^i = \mathrm{MLP}[d_A^i, d_B^i]. \tag{9}$$

The network predicts the position offset Δp_A^i, resulting in the adjusted matching pair $\{p_A^i + \Delta p_A^i, p_B^i\}$. The Mean Absolute Error (MAE) loss function is used to supervise this fine-tuning process:

$$\mathcal{L}_{bias} = \frac{1}{N} \sum_{i=1}^{N} \| p_A^i + \Delta p_A^i - \hat{p_B^i} \|, \tag{10}$$

where $\hat{p_B^i}$ represents the true corresponding point of p_B^i in I^A.

4 Experiments

4.1 Implementation Details

The S³-Match model was trained on the MegaDepth [7] and ScanNet [3] datasets. For the MegaDepth dataset, the model used the Adam optimizer with an initial learning rate of 4×10^{-3} and a batch size of 4. It was trained on four RTX 4090 GPUs for two days until convergence, processing images at a resolution of 1280×1280. Additionally, a strategy of halving the learning rate every four epochs was employed to optimize training efficiency. For the ScanNet dataset, the training configuration was similar, but the initial learning rate was set to 1×10^{-3}, and the images were processed at a resolution of 640×480.

Due to limited training resources, the paper did not perform detailed adjustments to the weights of each loss component. The final network loss function integrated contributions from various parts as follows:

$$\mathcal{L} = \mathcal{L}_{align} + \mathcal{L}_{desc} + \mathcal{L}_{det} + \mathcal{L}_{bias}. \tag{11}$$

The overall loss function combines supervision for common-view alignment, descriptor match, self-supervised keypoint selection, and matching refinement. These elements work together to significantly improve the network's ability to accurately detect and match keypoints.

4.2 Outdoor Pose Estimation

Pose estimation involves determining the orientation and position of an object or camera in three-dimensional space. In outdoor environments, pose estimation generally requires establishing matching relationships between two or more images. These matches are then used to compute the fundamental matrix using the Random Sample Consensus (RANSAC) algorithm, which helps mitigate the impact of erroneous matches and reconstructs changes in camera perspective.

We evaluate the effectiveness of S³-Match for outdoor pose estimation using the MegaDepth dataset [7], which comprises data reconstructed from 196 different locations via COLMAP SfM/MVS. We adopted the evaluation protocol proposed by DISK [21], assessing model performance by calculating the Area Under Curve (AUC) across different pose error thresholds (5°, 10°, 20°). As depicted in Table 1, S³-Match exhibits superior performance in handling challenging outdoor scenes compared to competing techniques.

Table 1: Performance Comparison of Pose Estimation on MegaDepth[7] and ScanNet[3] Datasets

Method	**MegaDepth (AUC)(%)**			**ScanNet (AUC)(%)**		
	@5°	@10°	@20°	@5°	@10°	@20°
D2-Net [6] + NN	-	-	-	5.3	15.0	28.0
SP [5] + NN	31.7	46.8	60.1	9.4	21.5	36.4
SP + SuperGlue [14]	42.2	61.2	76.0	16.2	33.8	51.8
DRC-Net [8]	27.0	42.9	58.3	7.7	17.9	30.5
PDC-Net+ [20]	51.5	67.2	78.5	20.3	39.4	57.1
MatchFormer [23]	53.3	69.7	82.0	24.3	43.9	61.4
ASpanFormer [2]	55.3	71.5	83.1	**25.6**	**46.0**	**63.3**
LoFTR [19]	52.8	69.2	81.2	22.1	40.8	57.6
S³-Match	**56.5**	**72.7**	**84.1**	22.4	41.2	57.4

To underscore the computational efficiency of S³-Match under identical hardware conditions, we conduct a time evaluation of image processing using various advanced methods, all under uniform environmental and hardware setups. As shown in Figure 5, S³-Match processes an image with a resolution of 1280 pixels in merely 69 milliseconds. This total processing time includes 16 milliseconds for the encoding phase of U-Net, 24 milliseconds for the decoding phase, 17 milliseconds for the attention mechanism, and 12 milliseconds for additional components.

4.3 Indoor Pose Estimation

ScanNet is a widely used dataset for indoor scene understanding and 3D reconstruction, featuring a diverse array of indoor environments. It includes image pairs with wide baselines and large textureless areas, presenting significant challenges for image analysis.

To ensure fairness in the evaluation process, this study adhered to the training and testing protocols of LoFTR [19]. Despite the well-documented challenges that detector-based methods encounter in textureless scenes, the test results unequivocally show that S³-Match outperforms all detector-based methods in terms of pose accuracy. The performance is documented in Table 1.

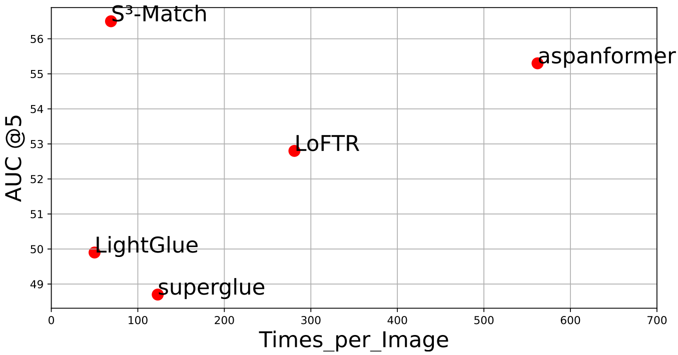

Figure 5: Performance and Efficiency Comparison between S³-Match and Other Open-Source Methods

4.4 Qualitative Analysis of Keypoint Selection

Figure 6 illustrates the local feature point selection results for S³-Match and SuperPoint within the dataset, as well as the integrated outcomes. Compared to SuperPoint, S³-Match more effectively exploits the subtle textural information in images, yielding keypoints that are uniformly distributed, with high detection rates, excellent repeatability, and enhanced stability.

4.5 Ablation Study

To rigorously evaluate the effectiveness of the innovative modules within the S³-Match framework, we evaluated various network variants. Initially, the fine-tuning mechanism for matching relationships was removed. Subsequently, the cross-image consistent background informed by cross-attention was eliminated, along with the associated second constraint of $\mathcal{L}_{det}$. This left the generation of keypoints and descriptors dependent solely on individual image information, resembling a more traditional approach to feature detection and descriptor extraction.

We contrast these features with the SuperPoint algorithm, which relies on synthetically generated datasets, to demonstrate the superiority of S³-Match's self-supervised keypoint selection. The results of the ablation study are detailed in Table 2.

Table 2: Ablation Study Results on Pose Estimation in MegaDepth

Method	AUC%			Runtimes per image
	@5°	@10°	@20°	at 1280 × 1280 resolution
Full S³-Match	56.5	72.7	84.1	69ms
No Fine-tuning	54.1	70.9	82.8	66ms
No Fine-tuning & Attention	42.8	60.1	75.0	43ms
SuperPoint+NN	31.7	46.8	60.1	38ms

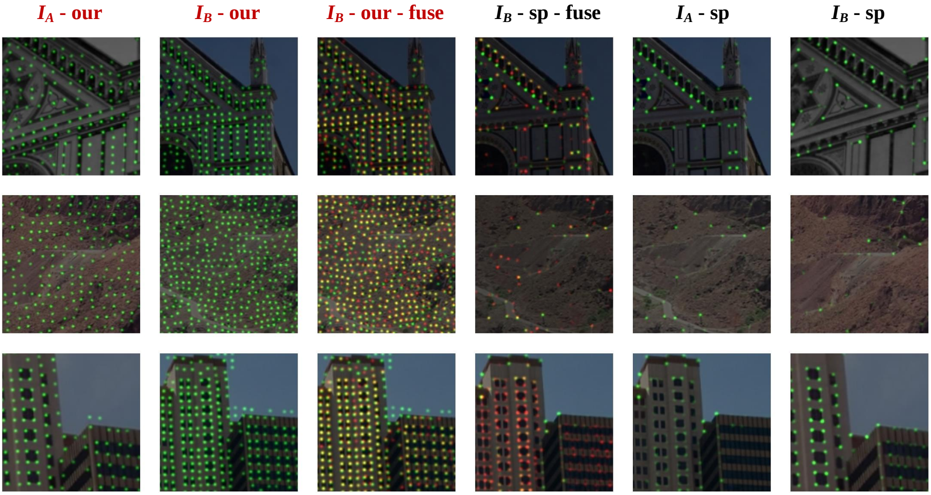

Figure 6: **Qualitative Analysis Results on Keypoint Selection in MegaDepth.** Keypoints in I_A are projected onto I_B in red. When the keypoints from the two images are consistent, the overlay of red and green appears as yellow.

5 Conclusion

We introduces a common-view aligned image matching algorithm with self-supervised keypoint selection, named S³-Match. S³-Match utilizes self-supervised learning to autonomously identify and select feature points that are highly distinctive and stable. Moreover, the algorithm aligns features within common-view areas, ensures consistent feature information across images, and directs subsequent self-supervised keypoint extraction and feature description efforts towards these common-view regions.

The approach described in this paper significantly outperforms the SuperPoint algorithm in terms of consistency, repeatability, and density of keypoint selection. Compared to other advanced algorithms, our method achieves nearly the lowest computational overhead and exhibits the best performance in pose estimation tasks.

6 Acknowledgments

This work is supported by the NSFC under Grants 62036008 and 62273057, and Telcent-CSIG project 2021122900001.

References

[1] Matthew Brown, Gang Hua, and Simon Winder. Discriminative learning of local image descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):43–57, 2010.

[2] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David Mckinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching

with adaptive span transformer. In *European Conference on Computer Vision*, pages 20–36. Springer, 2022.

[3] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[4] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Toward geometric deep slam. *arXiv preprint arXiv:1707.07410*, 2017.

[5] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[6] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint detection and description of local features. *arXiv preprint arXiv:1905.03561*, 2019.

[7] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018.

[8] Jinjiang Liu and Xueliang Zhang. Drc-net: Densely connected recurrent convolutional neural network for speech dereverberation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 166–170. IEEE, 2022.

[9] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

[10] Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, 2003.

[11] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. Lf-net: Learning local features from images. *Advances in neural information processing systems*, 31, 2018.

[12] Jerome Revaud, Philippe Weinzaepfel, César De Souza, Noe Pion, Gabriela Csurka, Yohann Cabon, and Martin Humenberger. R2d2: repeatable and reliable detector and descriptor. *arXiv preprint arXiv:1906.06195*, 2019.

[13] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011.

[14] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[15] Torsten Sattler, Akihiko Torii, Josef Sivic, Marc Pollefeys, Hajime Taira, Masatoshi Okutomi, and Tomas Pajdla. Are large-scale 3d models really necessary for accurate visual localization? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1646, 2017.

[16] Nikolay Savinov, Akihito Seki, Lubor Ladicky, Torsten Sattler, and Marc Pollefeys. Quad-networks: unsupervised learning to rank for interest point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1822–1830, 2017.

[17] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[18] Johannes L Schönberger, Marc Pollefeys, Andreas Geiger, and Torsten Sattler. Semantic visual localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6896–6906, 2018.

[19] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021.

[20] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[21] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020.

[22] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. Learning feature descriptors using camera pose supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 757–774. Springer, 2020.

[23] Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Matchformer: Interleaving attention in transformers for feature matching. In *Proceedings of the Asian Conference on Computer Vision*, pages 2746–2762, 2022.

[24] Xing Wei, Yue Zhang, Yihong Gong, and Nanning Zheng. Kernelized subspace pooling for deep local descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1875, 2018.

[25] Rongtao Xu, Changwei Wang, Bin Fan, Yuyang Zhang, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. Domaindesc: Learning local descriptors with domain adaptation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2505–2509. IEEE, 2022.

[26] Shibiao Xu, Shunpeng Chen, Rongtao Xu, Changwei Wang, Peng Lu, and Li Guo. Local feature matching using deep learning: A survey. *arXiv preprint arXiv:2401.17592*, 2024.

[27] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 467–483. Springer, 2016.

[28] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2666–2674, 2018.

[29] Hu Zhang, Zhaohui Tang, Yongfang Xie, and Weihua Gui. Rpi-surf: A feature descriptor for bubble velocity measurement in froth flotation with relative position information. *IEEE Transactions on Instrumentation and Measurement*, 70:1–14, 2021.

[30] Xiaoming Zhao, Xingming Wu, Weihai Chen, Peter CY Chen, Qingsong Xu, and Zhengguo Li. Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Transactions on Instrumentation and Measurement*, 2023.