SGAT: Learning Feature Matching with Singularity-enhanced Graph Attention Network

Yizhuo Zhang[1,2], **Kun Sun**[1,2*], **Chang Tang**[3], **Yuanyuan Liu**[1,2], **Xin Li**[4]
[1]School of Computer Science, China University of Geosciences(Wuhan), Wuhan 430074, China
[2]Key Laboratory of Geological Survey and Evaluation of Ministry of Education, China University of Geosciences(Wuhan), Wuhan 430074.
[3]School of Software Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
[4]Department of Computer Science, University at Albany, Albany 12222, USA
zyz@cug.edu.cn, sunkun@cug.edu.cn, tangchang@hust.edu.cn, liuyy@cug.edu.cn, xli48@albany.edu

Abstract

The task of image feature matching aims to establish correct correspondences between images from two different views. While approaches based on attention mechanisms have demonstrated remarkable advancements in image feature matching, they still encounter substantial limitations. Specifically, current graph attention network approaches face performance bottlenecks in complex scenarios, such as low-texture regions or occlusions. This limitation stems from the self-attention mechanism, which, when lacking effective guidance, can lead to divergent attention weights or incorrect focus on regions with low discriminability, resulting in matching failures in low-texture environments. Inspired by how humans focus on distinctive regions when performing cross-view matching, we enhance attention to singular points in images that are salient, unique and have high cross-view matching potential during information aggregation, thereby improving matching capability. To realize the aforementioned strategies, we develop a novel Singularity-enhanced Graph Attention Network (SGAT). SGAT leverages Co-potentiality and Multi-Scale Singularity as prior guidance, and designs a Singularity-aware Attention mechanism and a Co-potentiality Guided Attention mechanism, specifically enhancing the perception of singularity and matching potential during feature interaction. Experimental results on multiple datasets, including ScanNet1500, demonstrate that our method outperforms current state-of-the-art sparse matching methods. In particular, the improvement is most pronounced in complex scenarios such as low-texture environments, significantly enhancing the accuracy and robustness of image matching and its downstream tasks.

Code — https://github.com/tasi-z/SGAT

Introduction

Image matching is a crucial step in 3D vision tasks such as Structure from Motion (SfM)(He et al. 2024; Sun et al. 2023), Simultaneous Localization and Mapping (SLAM)(Mur-Artal, Montiel, and Tardós 2015; Mur-Artal and Tardós 2017; Liu et al. 2025b), visual localization(Cai et al. 2024a; Sarlin et al. 2021), and panorama stitching(Liao et al. 2023; Agarwal et al. 2009; Sun, Tao, and Qian 2019). Common image matching methods rely on sparse interest points, matching high-dimensional features that encode local visual appearances. However, stable matching remains challenging under complex conditions (such as viewpoint changes, illumination differences, and weak textures)(Fang et al. 2024; Sun and Tao 2019). Particularly in low-texture scenes with large viewpoint changes, the lack of distinctive features makes matching susceptible to noise and ambiguity.

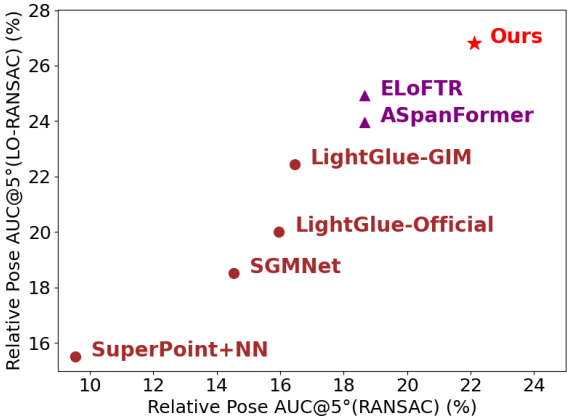

Figure 1: Camera Pose Estimation Accuracy Comparison. Our method SGAT (★), as a sparse matching approach, achieves superior accuracy compared to state-of-the-art sparse matchers (●) and semi-dense matchers (▲) on the textureless indoor scenes of ScanNet1500(Dai et al. 2017).

Recently, advanced methods have begun to explore attention mechanisms for interaction between keypoint features within and across views. The LightGlue algorithm(Lindenberger, Sarlin, and Pollefeys 2023) is a representative work that constructs densely connected graphs between image keypoints and aggregates feature information through attention graph neural networks, achieving significant improvements in matching capability. However, existing attention models lack effective range constraints when computing attention weights, making it difficult to aggregate useful information from effective regions in complex scenes and susceptible to interference from weakly textured or non-

*Corresponding author.

matching regions, leading to decreased matching accuracy.

Attention mechanisms in cross-view matching tasks enhance descriptor features by improving the perception of semantic context and structural consistency, thereby optimizing matching performance. Dense graph attention strategies without prior guidance, such as LightGlue, have limitations in complex matching scenarios. We observe that in failed matching cases, keypoint attention weights are often excessively distributed across regions with low specificity or low matching potential. This stems from two main issues: First, the feature over-smoothing problem, where irrelevant information from low-specificity regions contaminates feature representations, introducing redundant geometry-irrelevant information and reducing matching discriminative ability. Second, the feature interference problem, where attention mechanisms aggregate features across numerous non-overlapping regions, potentially obscuring keypoints' distinctive features with noise information, thus decreasing matching accuracy. These two issues jointly constitute the performance bottleneck of attention mechanisms without prior guidance: non-overlapping regions compromise feature representation purity, while low-specificity interference weakens feature space discriminability.

Regarding the problem of irrelevant feature interference, existing research has explored detector-free methods and identified that excessive attention to irrelevant regions negatively affects matching performance. Unnecessary regional feature interactions and matching searches may lead to incorrect matches. AdaMatcher(Huang et al. 2022) filters out most pixels by focusing on high-confidence candidate matching regions, thereby reducing the influence of irrelevant areas; MESA(Zhang and Zhao 2024) adopts precise local region matching strategies, effectively reducing matching redundancy. Both approaches have improved matching performance to some extent. On the other hand, in detector-based matching methods, OETR(Chen et al. 2022b) optimizes the matching process by regressing overlapping boxes, while OmniGlue(Jiang et al. 2024) attempts to reduce interference by pruning dense pairwise graph connections between keypoints in two images through evaluating foundation model feature similarities. However, these methods based on filtering strategies as hard constraints may limit robustness in complex scenarios.

Our method is inspired by the human visual matching process. When humans match two views of a scene, they do not attend to all regions equally; instead, they instinctively anchor their attention on "landmarks" that are particularly salient, distinctive, and likely to be visible in the other view. Building on this intuition, we introduce two core concepts to guide the attention mechanism. The first is Singularity, which we define as an intrinsic property of a keypoint that renders it semantically salient and distinctive within the view. The second is Co-potentiality, which quantifies the likelihood that a keypoint in one image has a valid and discoverable correspondence in the other image. Unlike Singularity, which is an intra-image property, Co-potentiality is an inter-image property that accounts for factors such as overlapping field of view, occlusion, and pronounced appearance changes.

Based on this insight, our work proposes the Singularity-enhanced Graph Attention Network (SGAT). Its core idea is to first focus on prominent, unique, and easily matchable structures in the image, enhance attention to keypoints with high matching potential (co-potentiality) and global singularity during attention aggregation, and then perform unconstrained attention aggregation across the entire image after multiple layers of matching capability enhancement.

We propose and validate the effectiveness of using singularity and co-potentiality as prior knowledge to guide graph attention networks in feature matching. SGAT enhances attention interaction of intra-view features between keypoints with singularity and high matching potential by using keypoints' singularity and co-potentiality as priors, combined with Singularity-aware Attention (SA-Attention). Meanwhile, SGAT effectively suppresses the interference of "atypical" keypoints lacking matching potential in feature learning. To further enhance matching robustness, we use pre-perceived Co-potentiality Features as priors and adopt Co-potentiality Guided Attention (CoP-Attention) to implicitly enhance feature matching potential perception. This design enables the network to focus not only on keypoints' uniqueness but also on their matching potential during feature interaction, thereby more precisely guiding attention to keypoints with genuine matching value and significantly improving matching capability. As shown in Figure 1, in textureless indoor scenes, our method achieves better pose estimation accuracy compared to competitive sparse matching and semi-dense matching baselines.

Specifically, our approach includes the following key technical innovations:

- A Singularity-enhanced Graph Attention Network (SGAT) that achieves state-of-the-art performance in feature matching across complex scenes with extensive textureless areas.
- We propose a novel Co-potentiality Estimation Network and Multi-Scale Singularity Estimation module for precisely evaluating feature singularities at identified keypoints.
- Introduction of Singularity-aware Attention (SA-Attention) and Co-potentiality Guided Attention (CoP-Attention) mechanisms to enhance the robustness of feature matching.

Related Work

Descriptor-based Image Matching Methods

Traditional image matching methods (LoweDavid 2004) rely on hand-crafted features for keypoint detection, description, and matching. Deep neural networks have improved keypoint detection (Rosten and Drummond 2006; Savinov et al. 2017; Laguna et al. 2019) and description (Tian et al. 2019; Ebel et al. 2019), with end-to-end approaches (Dusmanu et al. 2019; DeTone, Malisiewicz, and Rabinovich 2018; Revaud et al. 2019; Luo et al. 2020; Tian et al. 2020) like DISK (Tyszkiewicz, Fua, and Trulls 2020b), ALIKED (Zhao et al. 2023), and DeDoDe (Edstedt et al. 2024) enhancing geometric invariance and 3D consistency.

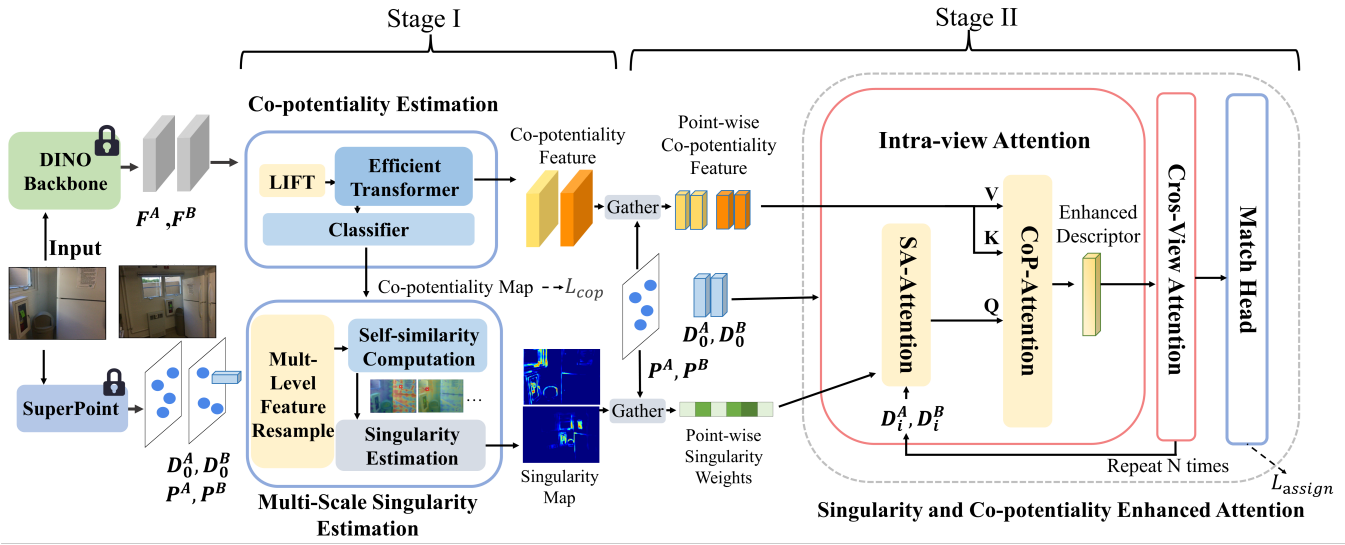

Figure 2: Pipeline Overview. **(1)** Given a pair of images, SuperPoint is used to detect keypoints and extract descriptors, while frozen DINOv2 Backbone extracts DINO features. **(2)** Co-potentiality Estimation is then performed, obtaining Co-potentiality Features through multi-layer Efficient Transformer on LIFT(Suri et al. 2024) upsampled features, and Co-potentiality Map through Classifier. **(3)** DINO features undergo Multi-Level Feature Resample and Self-similarity Computation, utilizing Co-potentiality Map to obtain Singularity Map. **(4)** Based on the keypoint positions, Point-wise Co-potentiality Features and Point-wise Singularity Weights are extracted from the Co-potentiality Feature and Singularity Map. **(5)** N alternating iterations of Intra-view Attention and Cross-view Attention are performed, with matching predicted at each layer, where Intra-view utilizes Singularity-aware Attention (SA-Attention) and Co-potentiality Guided Attention (CoP-Attention) for attention enhancement.

Attention-based Sparse Matching Methods

SuperGlue (Sarlin et al. 2020) pioneered Transformers for matching, using attention graph neural networks for partial assignment. SGMNet (Chen et al. 2021), and MakeGNN (Li and Ma 2024) improved efficiency via sparse graphs and geometry-aware attention. LightGlue (Lindenberger, Sarlin, and Pollefeys 2023) adjusts network depth dynamically,OmniGlue (Jiang et al. 2024) and SemaGlue (Zhang et al. 2025b) leverage foundation models for cross-domain generalization. Meanwhile, methods such as ResMatch and AGNet(Deng et al. 2024; Chen et al. 2025; Li et al. 2024; Liu et al. 2025a; Zhang et al. 2025a) incorporate various prior knowledge to enhance matching performance. However, keypoint detection struggles in texture-poor regions. Unlike these, our model excels in weakly-textured scenes using singularity priors for attention enhancement.

Attention-based Semi-Dense Matching Methods

Detector-free methods like LoFTR (Sun et al. 2021) use Transformers for coarse-to-fine dense matching. ASpanFormer (Chen et al. 2022a) improves long-range dependencies via hierarchical attention. Efficient LoFTR (Wang et al. 2024) optimizes speed with aggregated attention. QuadTree (Tang et al. 2022) reduces complexity through token pyramids. MESA (Zhang and Zhao 2024) and PRISM (Cai et al. 2024b) reduce redundancy via segmentation and pruning. AdaMatcher (Huang et al. 2022) achieves adaptive assignment. These methods, though robust, often lack efficiency compared to sparse approaches.

Method

Overview

For a pair of images $\mathbf{A}$ and $\mathbf{B}$ with n and m keypoints (indexed by α and β), the core matching task is to find the optimal correspondence from the Cartesian product space $\alpha \times \beta$. To this end, we extract keypoint positions $\mathcal{P}_A, \mathcal{P}_B$ and descriptors $\mathbf{D}_0^A, \mathbf{D}_0^B$. These are supplemented by features $\mathbf{F}^A, \mathbf{F}^B$ from a frozen pre-trained DINOv2 model. The overall pipeline is shown in Figure 2.

Co-potentiality Estimation

To mitigate feature interference from non-matchable regions (e.g., occluded or out-of-view areas), we introduce a Co-potentiality estimation module. This module predicts the probability of each feature point successfully establishing a stable match, effectively guiding the attention mechanism.

To predict feature point co-potentiality, we first upsample the DINO features (1/14 scale) $\mathbf{F}^A$ and $\mathbf{F}^B$ of the image pair to (1/8 scale) through LIFT. Following the design of ELoFTR, this module employs efficient intra-view self-attention and inter-view cross-attention mechanisms.

The transformed feature maps are upsampled and fused with the original features to obtain co-potentiality enhanced features $\mathbf{F}_{CoP}^A$ and $\mathbf{F}_{CoP}^B$. To generate the Co-potentiality Map, a simple linear classification head is applied to each

patch region i of these enhanced features, yielding a Co-potentiality Score c_i. These scores collectively constitute the final Co-potentiality Map C, which is supervised by cross-entropy loss during training.

Multi-Scale Singularity Estimation

Global Singularity Computation The singularity of an image patch is determined by the number of highly similar regions within the image. When a patch exhibits few similar regions, its singularity becomes more pronounced, thereby providing a more stable and reliable basis for feature matching.To perform Global Singularity Computation (GSC), we compute the global singularity via Global Self-Similarity Computation (GSSC).For scale k, for patch region i to be predicted, we compute its local region singularity measure s_i^k, and finally obtain the singularity measure map S^k at scale k:

$$s_i^k = GSC(F^k, i) = 1 - \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}[\text{sim}(f_i^k, f_j^k) > \tau] \quad (1)$$

where $\text{sim}(f_i^k, f_j^k)$ represents the cosine similarity between pixel-level features at positions i and j in feature map F^k, N is the total number of pixels in the feature map F^k, τ is the similarity threshold, and $\mathbb{1}[\cdot]$ is the indicator function.

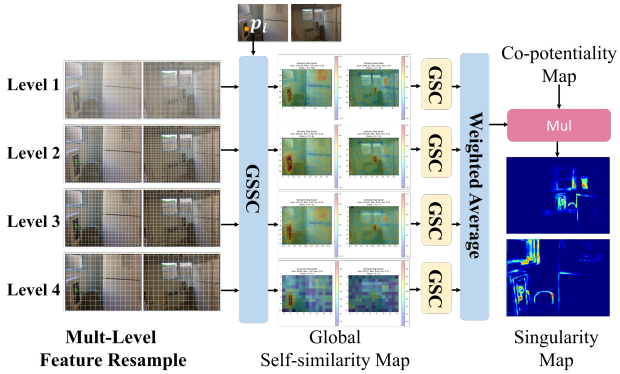

Figure 3: Multi-Scale Singularity Estimation Module.

Multi-Scale Singularity Computation Single-scale singularity measurement underestimates regional singularity when semantic elements contain repetitive regions with high self-similarity. We therefore propose a multi-scale measurement approach (shown in Figure 3) that fuses scores across different scales with the Co-potentiality Map C to compute the final singularity map S^{ms}.

$$S^{ms} = C \odot \sum_{k=1}^{4} w_k \cdot Upsample(S^k) \quad (2)$$

where w_k is the weight associated with scale k, C represents the Co-potentiality coefficient, and $Upsample$ denotes the upsampling operation.

Subsequently, Point-wise Co-potentiality Features and Point-wise Singularity Weights are extracted at the keypoint level from S^{ms} and $\mathbf{F}_{CoP}$, respectively. Visualization results of the singularity score maps in different scenarios are presented in Figure 4.

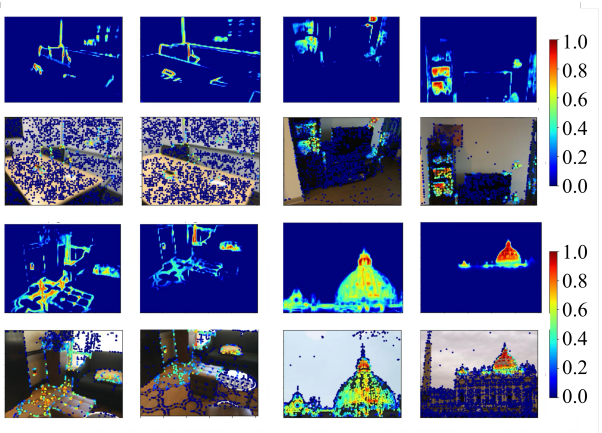

Figure 4: Visualization of Singularity Maps across four different scenarios. Red indicates higher confidence levels while blue indicates lower confidence.

Singularity and Co-potentiality Enhanced Attention

We sequentially apply three attention mechanisms to optimize pre-extracted keypoint descriptors: (1) Singularity-aware Attention (2) Co-potentiality Guided Attention and (3) Cross-View Attention.

Singularity-aware Attention To enhance attention effectiveness and focus on key matching points, we propose Singularity-aware Attention. This approach recognizes that keypoints have varying importance, with highly singular points deserving more weight during attention aggregation. We implement this by introducing a learnable bias matrix $\boldsymbol{\mathcal{A}}_l$ into the self-attention formula:

$$\mathbf{D}_l^{SA} = \text{softmax}\left(\frac{\mathbf{Q}_l\mathbf{K}_l^{\mathrm{T}}}{\sqrt{d_k}} + \boldsymbol{\mathcal{A}}_l\right)\mathbf{V}_l + \mathbf{D}_{l-1} \quad (3)$$

where $\mathbf{D}_l^{SA}$ is the layer l output features, and $\mathbf{Q}_l$, $\mathbf{K}_l$, $\mathbf{V}_l$ are query, key, and value matrices from input features $\mathbf{D}_{l-1}$. The bias matrix $\boldsymbol{\mathcal{A}}_l$ is constructed as follows: for any element a_l^{ij} in the attention score matrix, its bias value depends on the singularity score S_j^{ms} of the keypoint j :

$$a_l^{ij} = \alpha_l \cdot S_j^{ms} \quad (4)$$

Here, S_j^{ms} is the normalized singularity score of keypoint j, and α_l is a learnable scalar parameter. This bias value a^{ij} is broadcast to all rows in the j-th column of the matrix $\boldsymbol{\mathcal{A}}_l$. This design ensures that keys with higher singularity scores receive higher initial scores, leading to greater weights after the Softmax operation. Importantly, this is not a hard constraint but rather provides flexible guidance for the network to learn from.

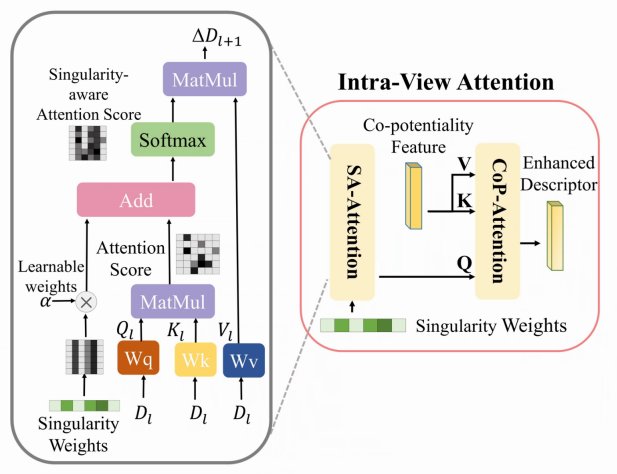

Figure 5: Singularity and Co-potentiality Enhanced Attention Module. SA-Attention utilizes the estimated Singularity Map for attention weighting to enhance singularity awareness, while CoP-Attention employs cross-attention with Co-potentiality Features to implicitly enhance feature matching potential awareness.

Co-potentiality Guided Attention Using pre-perceived Co-potentiality Features $\mathbf{F}^A_{CoP}$ and $\mathbf{F}^B_{CoP}$ as prior knowledge, we employ Co-potentiality Guided Attention (CoP-Attention) to implicitly enhance the feature matching potential awareness.

Specifically, we employ a Transformer-based enhancement module where the descriptor features refined by self-attention, $\mathbf{D}^{SA}_l$, serve as the Query. The co-potentiality features, $\mathbf{F}_{CoP}$, are utilized as both the Key and the Value. This process yields the co-potentiality guided attention output, denoted as $\mathbf{D}^{CoP}_l$.

Cross-View attention Subsequently, Cross-View attention is performed to achieve mutual updates between keypoints of the two images, resulting in $\mathbf{D}_l$. Next, a lightweight match head module is employed to estimate the matchability scores ${}^{\ell}\sigma^A_j$, ${}^{\ell}\sigma^B_j$, and the soft partial assignment matrix ${}^{\ell}\mathbf{P}ij$.

Supervision

The training process consists of two stages, with supervision losses designed separately for covisibility estimation and matching network.

Stage I: Co-potentiality Estimation Loss In the first stage, we employ the Binary Cross-Entropy (BCE) loss function to train the Co-potentiality estimation module. This module models the Co-potentiality of each grid region as a binary classification problem, aiming to predict the matching probability between grid points across two images. Based on the above definition, the Co-potentiality Estimation Loss $\mathcal{L}_{\text{cop}}$ can be expressed as:

$$\mathcal{L}_{\text{cop}} = \frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{W}\text{BCE}(l_{ij}, M_{ij}) \tag{5}$$

where: $l_{ij} \in \mathbb{R}$ represents the logit value output by the model at position (i,j), $M_{ij} \in \{0,1\}$ is the binary label at the corresponding position in the mask image, H, W are the height and width of the feature map respectively, $N = H \times W$ is the total number of grid points, and BCE denotes the binary cross-entropy loss.

Stage II: Matching Network Loss In the second stage, the matching network learns to generate an assignment matrix $\mathbf{P}$ that describes image matching relationships. Similar to Stage I, the ground truth matching set $\mathcal{M}$ is determined through two-view geometry, excluding outlier sets $\overline{\mathcal{A}}$ and $\overline{\mathcal{B}}$. The total loss function for the matching network is formulated as:

$$\mathcal{L}_{\text{assign}} = -\sum_{\ell}\Bigg[\frac{1}{|\mathcal{M}|}\sum_{(i,j)\in\mathcal{M}}\log({}^{\ell}\mathbf{P}_{ij}) + \frac{1}{2|\overline{\mathcal{A}}|}\sum_{i\in\overline{\mathcal{A}}}\log(1-{}^{\ell}\sigma^A_i) + \frac{1}{2|\overline{\mathcal{B}}|}\sum_{j\in\overline{\mathcal{B}}}\log(1-{}^{\ell}\sigma^B_j)\Bigg] \tag{6}$$

where the first term represents the matching loss that encourages high probabilities for correct correspondences, and the second and third terms represent the outlier losses for images A and B respectively, which promote proper identification of outlier features.

Experiments

In this section, we comprehensively evaluate the proposed SGAT method. We first introduce the implementation details, then compare with existing methods on three main tasks: indoor relative pose estimation, outdoor relative pose estimation, and homography estimation. Finally, we analyze the effectiveness of each component through ablation studies.

Implementation Details

We train on the MegaDepth dataset, which comprises over 1 million images from 196 landmarks with SfM reconstructions. For co-potential estimation, we extract patch features from DINOv2 ViT-B/14 (patch size of 14, feature dimension of 768). The matching network consists of 9 attention layers (each with 4 heads and 256 dimensions): the initial 4 layers incorporate Singularity Maps in SA-Attention, while the remaining 5 use unconstrained self-attention.

The training process is divided into two stages: (1) training the Covisibility Estimation module from scratch using the AdamW optimizer (learning rate of 4×10^{-3}, batch size of 4, for 30 epochs); (2) fine-tuning initialized with GIM-LightGlue weights, extracting 2048 SuperPoint keypoints (learning rate of 1×10^{-5}, batch size of 16, for 10 epochs).

Regarding network hyperparameter configuration, in the Global Singularity Computation module, the similarity threshold τ is set to 0.7. For multi-scale feature fusion, the weights w_i are uniformly set to $\frac{1}{4}$. The attention weighting coefficient α_1 is adaptively learned through end-to-end training. Detailed hyperparameter ablation studies are provided in the appendix.

Method	ScanNet			ScanNet++			MegaDepth1500			Times(ms)
	@5°	@10°	@20°	@5°	@10°	@20°	@5°	@10°	@20°	
SP+NN	9.54	21.68	**37.17**	3.4	**8.57**	**16.3**	31.7	46.8	60.1	**18.1**
DISK+NN	8.62	19.81	33.01	2.96	7.17	13.2	36.7	52.9	65.9	43.2
ALIKED+NN	8.93	20.95	35.45	3.46	8.26	15.09	41.9	58.4	71.7	28.6
DeDoDe+NN	**10.22**	**22.07**	35.37	**3.78**	8.56	15.08	**49.4**	**65.5**	**77.7**	85.8
LoFTR	16.73	33.63	49.66	6.72	14.4	23.5	52.8	69.2	81.2	117.9
ASpanFormer	**18.66**	**36.97**	54.20	**8.57**	**17.66**	**27.9**	55.3	71.5	83.1	171.5
ELoFTR	**18.66**	36.84	**54.38**	7.78	16.33	25.97	**56.4**	**72.2**	**83.5**	**65.2**
SuperGlue	17.20	35.13	53.67	7.59	16.83	27.92	48.5	66.2	79.3	49.9
LightGlue-O	15.96	33.49	51.26	6.52	15.07	25.63	49.9	66.8	79.9	39.0
LightGlue-G	16.45	34.62	52.54	7.63	16.71	27.58	50.3	67.9	**80.9**	**38.9**
SGMNet	14.53	31.02	48.62	6.44	14.52	24.66	43.2	61.6	75.6	45.1
OmniGlue	13.52	28.1	44.22	5.2	11.68	20.04	44.31	61.75	75.56	123.5
SGAT(Ours)	**22.84**	**43.29**	**61.45**	**9.82**	**20.79**	**33.25**	**50.7**	**68.2**	80.7	50.1

Table 1: RANSAC relative pose estimation results across ScanNet1500, ScanNet++, and MegaDepth1500 datasets. We evaluate all methods using models trained on the MegaDepth dataset. The AUC of pose errors under different thresholds is shown, where all input images are resized to 640×480. Second-best results are underlined, and best results are shown in bold. LightGlue-O and LightGlue-G denote the official pre-trained LightGlue model and the LightGlue model trained by GIM, respectively.

Indoor Pose Estimation

ΔT	Method	RANSAC AUC			
		@5°	@10°	@20°	mAA
1	SGAT(Ours)	0.1479	**0.3180**	**0.5112**	**0.2102**
	LightGlue-G	0.1354	0.2937	0.4793	0.1948
	LightGlue-O	0.1217	0.2728	0.4534	0.1806
	ELoFTR	**0.1514**	0.3091	0.4895	0.2054
2	SGAT(Ours)	**0.1624**	**0.3331**	**0.5063**	**0.2157**
	LightGlue-G	0.1307	0.2868	0.4502	0.1856
	LightGlue-O	0.1257	0.2730	0.4416	0.1792
	ELoFTR	0.1550	0.3001	0.4540	0.1979
3	SGAT(Ours)	**0.1562**	**0.3129**	**0.4755**	**0.2041**
	LightGlue-G	0.1185	0.2631	0.4161	0.1692
	LightGlue-O	0.1084	0.2425	0.3922	0.1579
	ELoFTR	0.1363	0.2738	0.4101	0.1775
4	SGAT(Ours)	**0.1530**	**0.3037**	**0.4535**	**0.1969**
	LightGlue-G	0.1016	0.2368	0.3757	0.1508
	LightGlue-O	0.1008	0.2190	0.3517	0.1427
	ELoFTR	0.1235	0.2470	0.3677	0.1588

Table 2: Relative pose estimation results on the ScanNet-IS dataset. The table shows AUC values for pose error at different thresholds on Interval-Sampled ScanNet (larger intervals indicate less overlap and more challenging matching). All input images are resized to 640×480 resolution. Second-best results are underlined, while best results are shown in bold.

We evaluate our method on ScanNet1500(Dai et al. 2017), ScanNet-IS(Interval-Sampled from the ScanNet test set) , and ScanNet++(Yeshwanth et al. 2023). Comparisons are made against descriptor-based methods (SuperPoint (DeTone, Malisiewicz, and Rabinovich 2018), DISK (Tyszkiewicz, Fua, and Trulls 2020a), ALIKED (Zhao et al. 2023), DeDoDe (Edstedt et al. 2024)), semi-dense matchers (LoFTR (Sun et al. 2021), ELoFTR (Wang et al. 2024), ASpanFormer (Chen et al. 2022a)), and sparse matchers (SuperGlue (Sarlin et al. 2020), SGMNet (Chen et al. 2021), OmniGlue (Jiang et al. 2024), LightGlue (Lindenberger, Sarlin, and Pollefeys 2023; Shen et al. 2024).

Experiment	MMA		DLT		RANSAC	
	@1px	@3px	@1px	@3px	@1px	@3px
ALIKED+NN	**47.2**	79.5	**6.1**	**19.69**	21.64	51.39
DISK+NN	37.9	**81.6**	1.74	5.2	**34.34**	**55.19**
SP+NN	23.9	65.4	2.92	7.62	32.59	49.6
ELoFTR	66.6	90.1	29.85	54.36	**38.84**	**61.65**
ASpanFormer	**72.0**	**94.3**	**39.67**	**65.99**	28.04	57.2
SGMNet	26.5	80.7	29.45	62.86	32.73	55.05
SuperGlue	28.4	86.2	32.56	65.84	33.77	56.93
LightGlue-O	29.6	88.4	36.13	68.09	34.71	**57.78**
SGAT(Ours)	**29.9**	**90.2**	**37.02**	**68.77**	**34.72**	57.21

Table 3: Homography estimation results on the HPatches dataset. The Mean Matching Accuracy (%) at different thresholds and the AUC (%) of corner reprojection error for homography estimation using both non-robust estimator (DLT) and robust estimator (RANSAC) are reported.

SGAT consistently outperforms competing methods across these datasets. As shown in Table 1, on ScanNet1500, it surpasses ELoFTR at all thresholds. On ScanNet++, it achieves a higher mAA compared to ASpanFormer. Computational times for different method categories are also reported to highlight efficiency.

On ScanNet-IS (Table 2), the advantages become more pronounced with larger sampling intervals, demonstrating robustness to minimal overlap.

For descriptor-based methods, we report the total time for descriptor extraction and nearest neighbor matching. For sparse and semi-dense matching methods, we report only the

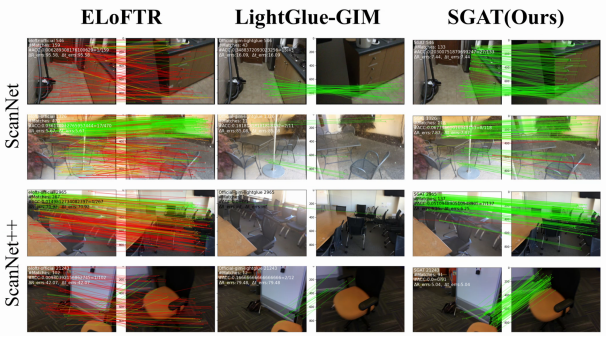

Figure 6: Qualitative Results. Our method is compared with the sparse matching pipeline SuperPoint (DeTone, Malisiewicz, and Rabinovich 2018)+LightGlue-GIM (Shen et al. 2024) and the semi-dense matcher ELoFTR(Wang et al. 2024). In image pairs containing textureless regions and large viewpoint changes, our method can perform matching robustly. Red indicates an epipolar error exceeding 5×10^{-4} (in normalized image coordinates).

inference time of the matching network.

Outdoor Pose Estimation

We evaluate our method on the MegaDepth dataset using 1,500 image pairs from the "Sacre Coeur" and "St. Peter's Square" scenes. As shown in Table 1, SGAT achieves superior performance among sparse matching methods.

Homography Estimation

We evaluate homography estimation on the HPatches(Balntas et al. 2017) dataset, which contains planar scene image sequences captured under different viewpoints and lighting conditions. As shown in Table 3, among sparse matching methods, SGAT achieves the best results on both MMA and DLT metrics.

Ablation Studies

Method	Pose Estimation AUC		
	@5°	@10°	@20°
1) LightGlue-G (baseline)	16.45	34.62	52.54
2) CoP Feature → frozen DINOv2	14.79	30.44	47.32
3) Remove CoP-Attn	21.95	41.65	59.87
4) SA-Attn → Vanilla	21.44	41.58	60.14
5) SW: Add → Mult	19.91	38.75	57.16
6) Singularity → Single-scale	22.6	42.55	60.88
7) DINOv2 → ResNet-50	21.38	41.62	60.24
Full Model	**22.84**	**43.29**	**61.45**

Table 4: Ablation Studies.

In this section, we conduct a series of ablation experiments to analyze the effectiveness of key components in our proposed Singularity-enhanced Graph Attention Network (SGAT). All ablation experiments are performed on the ScanNet1500 dataset using pose estimation AUC as the evaluation metric, with results shown in Table 4.

Co-potentiality Feature's Impact and CoP-Attention Necessity: To verify conditional features' importance in CoP-Attention, we replaced them with frozen DINOv2 features (Table 4, Row 2), showing performance degradation compared to the baseline and complete model. This indicates pre-perceived Co-potentiality Features are crucial for attention guidance. We also validated CoP-Attention by removing it while keeping SA-Attention (Row 3), which improved over baseline but remained below the complete model, showing both mechanisms' importance.

SA-Attention Effectiveness and Feature Processing: Replacing SA-Attention with standard Vanilla Self-Attention while keeping CoP-Attention (Row 4) decreased performance similarly to removing CoP-Attention. We studied singularity weighting (SW) by changing from addition to multiplication (Row 5), finding significant degradation. Testing single-scale versus multi-scale singularity measurement (Row 6) demonstrated multi-scale's advantages in comprehensive feature capture.Substituting the frozen DINOv2 with an ImageNet-pretrained ResNet-50 for prior computation (Row 7) still yielded substantial improvements compared to the baseline, indicating that the performance gains are not solely dependent on DINOv2's characteristics but rather stem from the inherent design of SGAT.

These ablation experiments validate each SGAT network module's effectiveness, showing SA-Attention and CoP-Attention's collaborative enhancement of image feature matching in complex scenes.

Conclusions

In this paper, we propose the Singularity-enhanced Graph Attention Network (SGAT) to address the performance bottleneck of existing attention-based image feature matching methods in challenging scenarios such as low-texture environments. SGAT enhances the perception of high-matching-potential regions through multi-scale singularity measurements as prior knowledge, combined with Singularity-Aware Attention (SA-Attention) and Co-potentiality Guided Attention (CoP-Attention) mechanisms. Experimental results on multiple benchmark datasets demonstrate that this method significantly outperforms existing state-of-the-art approaches, exhibiting superior matching performance and robustness particularly in complex scenarios with weak textures.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.62176242), Hubei Provincial Natural Science Foundation of China (No.2025AFB832) and Open Research Fund of State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University (No.23E03), also in part by the Opening Fund of Key Laboratory of Geological Survey and Evaluation of Ministry of Education (No.GLAB2024ZR09) and the Fundamental Research Funds for the Central Universities.

References

Agarwal, S.; Furukawa, Y.; Snavely, N.; Simon, I.; Curless, B.; Seitz, S. M.; and Szeliski, R. 2009. Building Rome in a day. *ICCV*.

Balntas, V.; Lenc, K.; Vedaldi, A.; and Mikolajczyk, K. 2017. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *CVPR*, 5173–5182.

Cai, X.; Wang, Y.; Huang, Z.; Shao, Y.; and Li, D. 2024a. VOLoc: Visual Place Recognition by Querying Compressed Lidar Map. *arXiv preprint arXiv:2402.15961*.

Cai, X.; Wang, Y.; Luo, L.; Wang, M.; Li, D.; Xu, J.; Gu, W.; and Ai, R. 2024b. PRISM: PRogressive dependency maximization for Scale-invariant image Matching. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5250–5259.

Chen, G.; Fu, T.; Chen, H.; Teng, W.; Xiao, H.; and Zhao, Y. 2025. RDD: Robust Feature Detector and Descriptor using Deformable Transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6394–6403.

Chen, H.; Luo, Z.; Zhang, J.; Zhou, L.; Bai, X.; Hu, Z.; Tai, C.-L.; and Quan, L. 2021. Learning to match features with seeded graph matching network. In *ICCV*, 6301–6310.

Chen, H.; Luo, Z.; Zhou, L.; Tian, Y.; Zhen, M.; Fang, T.; Mckinnon, D.; Tsin, Y.; and Quan, L. 2022a. Aspanformer: Detector-free image matching with adaptive span transformer. In *ECCV*, 20–36. Springer.

Chen, Y.; Huang, D.; Xu, S.; Liu, J.; and Liu, Y. 2022b. Guide local feature matching by overlap estimation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 365–373.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 5828–5839.

Deng, Y.; Zhang, K.; Zhang, S.; Li, Y.; and Ma, J. 2024. Resmatch: Residual attention learning for feature matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1501–1509.

DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2018. SuperPoint: Self-Supervised Interest Point Detection and Description. *CVPRW*.

Dusmanu, M.; Rocco, I.; Pajdla, T.; Pollefeys, M.; Sivic, J.; Torii, A.; and Sattler, T. 2019. D2-net: A trainable cnn for joint description and detection of local features. In *CVPR*, 8092–8101.

Ebel, P.; Mishchuk, A.; Yi, K. M.; Fua, P.; and Trulls, E. 2019. Beyond cartesian representations for local descriptors. In *ICCV*, 253–262.

Edstedt, J.; Bökman, G.; Wadenbäck, M.; and Felsberg, M. 2024. DeDoDe: Detect, don't describe—Describe, don't detect for local feature matching. In *2024 International Conference on 3D Vision (3DV)*, 148–157. IEEE.

Fang, C.; Sun, K.; Li, X.; Li, K.; and Tao, W. 2024. ODNet: Orthogonal descriptor network for multiview image keypoint matching. *Inf. Fusion*, 105: 102206.

He, X.; Sun, J.; Wang, Y.; Peng, S.; Huang, Q.; Bao, H.; and Zhou, X. 2024. Detector-free structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21594–21603.

Huang, D.; Chen, Y.; Xu, S.; Liu, Y.; Wu, W.-Q.; Ding, Y.; Wang, C.; and Tang, F. 2022. Adaptive Assignment for Geometry Aware Local Feature Matching. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5425–5434.

Jiang, H.; Karpur, A.; Cao, B.; Huang, Q.; and Araujo, A. 2024. Omniglue: Generalizable feature matching with foundation model guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19865–19875.

Laguna, A. B.; Riba, E.; Ponsa, D.; and Mikolajczyk, K. 2019. Key.Net: Keypoint Detection by Handcrafted and Learned CNN Filters. *ICCV*, 5835–5843.

Li, Z.; Fang, F.; Wang, T.; and Zhang, G. 2024. Homography Estimation With Adaptive Query Transformer and Gated Interaction Module. *IEEE Transactions on Circuits and Systems for Video Technology*.

Li, Z.; and Ma, J. 2024. Learning feature matching via matchable keypoint-assisted graph neural network. *IEEE Transactions on Image Processing*.

Liao, K.; Nie, L.; Lin, C.; Zheng, Z.; and Zhao, Y. 2023. RecRecNet: Rectangling rectified wide-angle images by thin-plate spline model and DoF-based curriculum learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10800–10809.

Lindenberger, P.; Sarlin, P.-E.; and Pollefeys, M. 2023. LightGlue: Local Feature Matching at Light Speed. In *ICCV*.

Liu, X.; Wang, C.; Shi, G.; Zhang, X.; Miao, Q.; and Fan, M. 2025a. SGAD: Semantic and Geometric-aware Descriptor for Local Feature Matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 27095–27104.

Liu, Y.; Sun, K.; Tang, C.; Qian, Y.; and Li, X. 2025b. TPDepth: Leveraging Text Prompts with ControlNet to Boost Diffusion-based Depth Estimation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, 4290–4299. New York, NY, USA: Association for Computing Machinery. ISBN 9798400720352.

LoweDavid, G. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*.

Luo, Z.; Zhou, L.; Bai, X.; Chen, H.; Zhang, J.; Yao, Y.; Li, S.; Fang, T.; and Quan, L. 2020. Aslfeat: Learning local features of accurate shape and localization. In *CVPR*, 6589–6598.

Mur-Artal, R.; Montiel, J. M. M.; and Tardós, J. D. 2015. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *TR*, 31: 1147–1163.

Mur-Artal, R.; and Tardós, J. D. 2017. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5): 1255–1262.

Revaud, J.; de Souza, C. R.; Humenberger, M.; and Weinzaepfel, P. 2019. R2D2: Reliable and Repeatable Detector and Descriptor. In *NeurIPS*.

Rosten, E.; and Drummond, T. 2006. Machine learning for high-speed corner detection. In *ECCV*, 430–443. Springer.

Sarlin, P.; Unagar, A.; Larsson, M.; Germain, H.; Toft, C.; Larsson, V.; Pollefeys, M.; Lepetit, V.; Hammarstrand, L.; Kahl, F.; and Sattler, T. 2021. Back to the Feature: Learning Robust Camera Localization From Pixels To Pose. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 3247–3257. Computer Vision Foundation / IEEE.

Sarlin, P.-E.; DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2020. SuperGlue: Learning Feature Matching with Graph Neural Networks. In *CVPR*.

Savinov, N.; Seki, A.; Ladicky, L.; Sattler, T.; and Pollefeys, M. 2017. Quad-networks: unsupervised learning to rank for interest point detection. In *CVPR*, 1822–1830.

Shen, X.; Cai, Z.; Yin, W.; Müller, M.; Li, Z.; Wang, K.; Chen, X.; and Wang, C. 2024. Gim: Learning generalizable image matcher from internet videos. *arXiv preprint arXiv:2402.11095*.

Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; and Zhou, X. 2021. LoFTR: Detector-free local feature matching with transformers. In *CVPR*, 8922–8931.

Sun, K.; and Tao, W. 2019. A center-driven image set partition algorithm for efficient structure from motion. *Inf. Sci.*, 479: 101–115.

Sun, K.; Tao, W.; and Qian, Y. 2019. Guide to match: Multi-layer feature matching with a hybrid gaussian mixture model. *IEEE Transactions on Multimedia*, 22(9): 2246–2261.

Sun, K.; Yu, J.; Tao, W.; Li, X.; Tang, C.; and Qian, Y. 2023. A unified feature-spatial cycle consistency fusion framework for robust image matching. *Inf. Fusion*, 97: 101810.

Suri, S.; Walmer, M.; Gupta, K.; and Shrivastava, A. 2024. Lift: A surprisingly simple lightweight feature transform for dense vit descriptors. In *European Conference on Computer Vision*, 110–128. Springer.

Tang, S.; Zhang, J.; Zhu, S.; and Tan, P. 2022. QuadTree Attention for Vision Transformers. *ICLR*.

Tian, Y.; Balntas, V.; Ng, T.; Barroso-Laguna, A.; Demiris, Y.; and Mikolajczyk, K. 2020. D2d: Keypoint extraction with describe to detect approach. In *ACCV*.

Tian, Y.; Yu, X.; Fan, B.; Wu, F.; Heijnen, H.; and Balntas, V. 2019. Sosnet: Second order similarity regularization for local descriptor learning. In *CVPR*, 11016–11025.

Tyszkiewicz, M.; Fua, P.; and Trulls, E. 2020a. DISK: Learning local features with policy gradient. *NeurIPS*.

Tyszkiewicz, M. J.; Fua, P.; and Trulls, E. 2020b. DISK: Learning local features with policy gradient. *arXiv: Computer Vision and Pattern Recognition*.

Wang, Y.; He, X.; Peng, S.; Tan, D.; and Zhou, X. 2024. Efficient LoFTR: Semi-dense local feature matching with sparse-like speed. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21666–21675.

Yeshwanth, C.; Liu, Y.-C.; Nießner, M.; and Dai, A. 2023. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12–22.

Zhang, J.; Xia, Z.; Dong, M.; Shen, S.; Yue, L.; and Zheng, X. 2025a. CoMatcher: Multi-View Collaborative Feature Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21970–21980.

Zhang, S.; Zhu, Z.; Li, Z.; Lu, T.; and Ma, J. 2025b. Matching while perceiving: Enhance image feature matching with applicable semantic amalgamation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10094–10102.

Zhang, Y.; and Zhao, X. 2024. MESA: Matching Everything by Segmenting Anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20217–20226.

Zhao, X.; Wu, X.; Chen, W.; Chen, P. C.; Xu, Q.; and Li, Z. 2023. Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Transactions on Instrumentation and Measurement*, 72: 1–16.