Matching While Perceiving: Enhance Image Feature Matching with Applicable Semantic Amalgamation

Shihua Zhang[1*], Zhenjie Zhu[1*], Zizhuo Li[1], Tao Lu[2], Jiayi Ma[1†]

[1]Electronic Information School, Wuhan University, Wuhan 430072,China
[2]School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205,China
suhzhang001@gmail.com, zejzhu@whu.edu.cn, zizhuo_li@whu.edu.cn, lutxyl@gmail.com, jyma2010@gmail.com

Abstract

Image feature matching is a cardinal problem in computer vision, aiming to establish accurate correspondences between two-view images. Existing methods are constrained by the performance of feature extractors and struggle to capture local information affected by sparse texture or occlusions. Recognizing that human eyes consider not only similar local geometric features but also high-level semantic information of scene objects when matching images, this paper introduces SemaGlue. This novel algorithm perceives and incorporates semantic information into the matching process. In contrast to recent approaches that leverage semantic consistency to narrow the scope of matching areas, SemaGlue achieves semantic amalgamation with the designed Semantic-Aware Fusion (SAF) Block by injecting abundant semantic features from the pre-trained segmentation model. Moreover, the Cross-Domain Alignment (CDA) Block is proposed to address domain alignment issues, bridging the gaps between semantic and geometric domains to ensure applicable semantic amalgamation. Extensive experiments demonstrate that SemaGlue outperforms state-of-the-art methods across various applications such as homography estimation, relative pose estimation, and visual localization.

Code — https://github.com/ZeJ-Zhu/SemaGlue

Introduction

Image matching, which aims to establish accurate correspondences between different images of the same scene, is a critical procedure for various complex vision applications (Ma et al. 2021), such as panoramic stitching (Brown and Lowe 2007), visual localization (Schönberger et al. 2018), 3D reconstruction (Koutsoudis et al. 2014), and neural rendering (Thies, Zollhöfer, and Nießner 2019). A well-developed and effective image matching pipeline begins with detecting and describing feature points in both images, where significant efforts have been devoted to practical detectors and descriptors (Lowe 2004; DeTone, Malisiewicz, and Rabinovich 2018). Subsequently, matching algorithms are conducted based on the visual descriptions to identify correspondences between two-view images. This paper focuses on determining accurate correspondences with existing feature points, better serving subsequent tasks.

A common method to establish point-to-point correspondences is the Nearest Neighbor (NN). Nevertheless, the generated matches are inevitably dominated by interference features due to the matcher's lack of inter-/intra-image message interactions. Thanks to the strength of deep learning, SuperGlue (Sarlin et al. 2020) as a precursor builds a graph neural network (GNN) (Wu et al. 2020) that aggregates information among feature points to enhance their descriptions. Though the following methods (Chen et al. 2021; Xue, Budvytis, and Cipolla 2023; Lindenberger, Sarlin, and Pollefeys 2023; Zhang and Ma 2024b) have elaborately designed more accurate and powerful matchers, subject to the myopic descriptors that focus on local and geometric features narrowly, they still behave unsatisfactorily, especially in the case of extreme situations, including sparse texture, illumination variations, and occlusions. In fact, besides the local and geometric features, humans that match two images with the naked eyes further consider the high-level semantic information about the scene and the objects, so that they rarely err even though in texture-less or wild scenes. Thus we lean toward asking the following questions: **Q1**: *Can we perceive the semantic information during matching to see more than only local or geometric features?* **Q2**: *How to mingle the semantic information applicably during the general matching process?*

The conjecture in **Q1** that additional semantic information facilitates the matcher in discriminating some ambiguous correspondences has been proved recently in MESA (Zhang and Zhao 2024) and OmniGlue (Jiang et al. 2024). They utilize semantic consistency, wherein established correspondences should possess identical semantic labels, narrowing down the exploration area for matching and enhancing the algorithm's robustness. However, as for **Q2**, both of them do not introduce the semantic messages into the matching process appropriately and effectively. Concretely, these methods that apply semantic priors to pre-select matching regions easily ignore true correspondences if the areas are not precise, and do not take full advantage of the abundant visual perception features. Besides, though using pre-trained foundation visual model (Kirillov et al. 2023; Oquab et al. 2024) could provide broad semantic knowledge, it also suffers from being computationally expensive and resource-

[*]These authors contributed equally.
[†]Corresponding author.

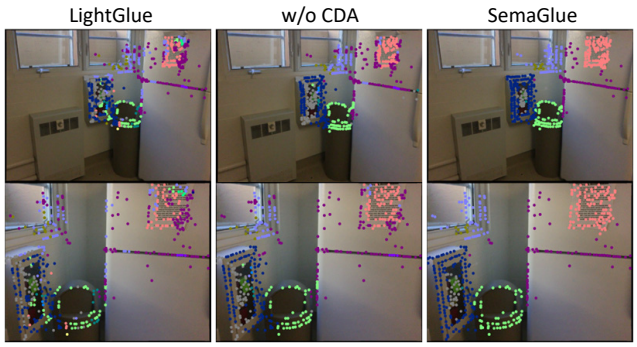

Figure 1: Visualizing descriptions of feature points. We reduce the dimensions of descriptions obtained from matchers to 3 and represent them in RGB. The left depicts the state-of-the-art method LightGlue, the middle directly fuses the semantic priors into a matching pipeline without bridging the domain gaps through CDA block (*i.e.*, w/o CDA), and the right is the full SemaGlue that amalgamates semantic priors applicably. SemaGlue draws feature points more consistently on the same class of objects or proximity semantics, which is expressed as similar colors. The descriptions obtained by w/o CDA are more cluttered in color similar to LightGlue with no semantic information injected.

intensive. Therefore, different from the above, we attempt to propose a new scheme to realize scene perception and semantic information injection by fully excavating semantic priors from a pre-trained segmentation model as well as enhancing local features without aggrandizing an unacceptable burden. As shown in Figure 1, after amalgamating the semantic information applicably, the representations of feature points located at the same semantic area within the two-view images are more similar to each other than before so mismatches are much harder to produce.

However, directly fusing semantic priors to enhance matching descriptions is unreasonable for ignoring the existing gaps between the semantic and geometric domains. Concretely, in the semantic domain, semantic features, which capture high-level contextual information, are often abstract and insensitive to geometric transformations. And in the geometric domain, local geometric features focus on spatial relationships and structures, making them sensitive to changes in scale, rotation, and perspective. These disparities in the nature of captured features lead to misalignment while fusing them directly, resulting in sub-optimal description representations for precise matching. As shown in Figure 1, the middle column that ignores the domain gaps does not amalgamate semantic information applicably so descriptions on the same semantic category are not consistent. Thus, we attempt to unify the features from these different domains hence inherently exhibiting a higher degree of alignment and then amalgamating the semantic priors applicably and effectively (Wang et al. 2024; Hu et al. 2021). Motivated by (Tzeng et al. 2017; Radford et al. 2021), we explore a co-domain of the semantic and geometric domains by seeking the relationships between their feature spaces, then align the semantic features from the pre-trained segmentation model with the patterns of local structures in the co-domain, hence bridging the domain gaps and yielding more applicable semantic priors for further amalgamation. The obtained amalgamation-applicable semantic priors preserve the semantic information together with local structural perception, thereby better serving the matching task.

Overall, to fully mine semantic priors to applicably enhance the geometric descriptions, we propose SemaGlue, the semantic amalgamated representation learning framework which consists of four essential parts: (a) **Semantic Extractor** that employs SegNext (Guo et al. 2022) as a lightweight model to extract semantic information; (b) **Cross-Domain Alignment (CDA) Block** that eliminates domain gaps between the semantic and geometric domains, searching for amalgamation-applicable semantic priors; (c) **Semantic-Aware Fusion (SAF) Block** that utilizes the priors as an instructional signal to enhance the semantic perceptual representations of local descriptions; (d) **Information Interaction Block** that further discriminates the feature representations to circumvent mismatches.

In summary, this work makes the following contributions:

- We propose SemaGlue, which perceives semantic information from a pre-trained segmentation model and integrates it into the image feature matching process.
- We propose a Cross-Domain Alignment Block to align the feature representations in semantic and geometric domains thereby eliminating the domain gaps and yielding the amalgamation-applicable semantic priors. Then the semantic prior further guides the information propagation during the matching process with the assistance of a designed Semantic-Aware Fusion Block.
- Experiments are conducted across extensive tasks, and the reported state-of-the-art results demonstrate the superiority of SemaGlue. We also provide further analysis and ablation studies on the interpretability of our model.

Related Work

Classic Image Feature Matching

Traditional image feature matching can be divided into two processes. The former focuses on obtaining interpretable local descriptions and feature points, like SIFT (Lowe 2004), ORB (Rublee et al. 2011), and Convolutional Neural Network (CNN)-based approaches (DeTone, Malisiewicz, and Rabinovich 2018; Zhao et al. 2023). While the latter seeks to remove false correspondences in the coarse matching set, including traditional methods (Ma et al. 2014, 2019) and learning-based ones (Yi et al. 2018; Li, Zhang, and Ma 2023; Zhang and Ma 2024a). However, unlike the two-phased pipeline, we directly determine correspondences from feature points and descriptions in an end-to-end manner.

End-to-End Image Feature Matching

SuperGlue (Sarlin et al. 2020) first accomplishes end-to-end image feature matching relying on GNN's powerful capabilities. Then, researchers attempt to further enhance the network, *e.g.*, KeyGNN (Jiang et al. 2023) focuses

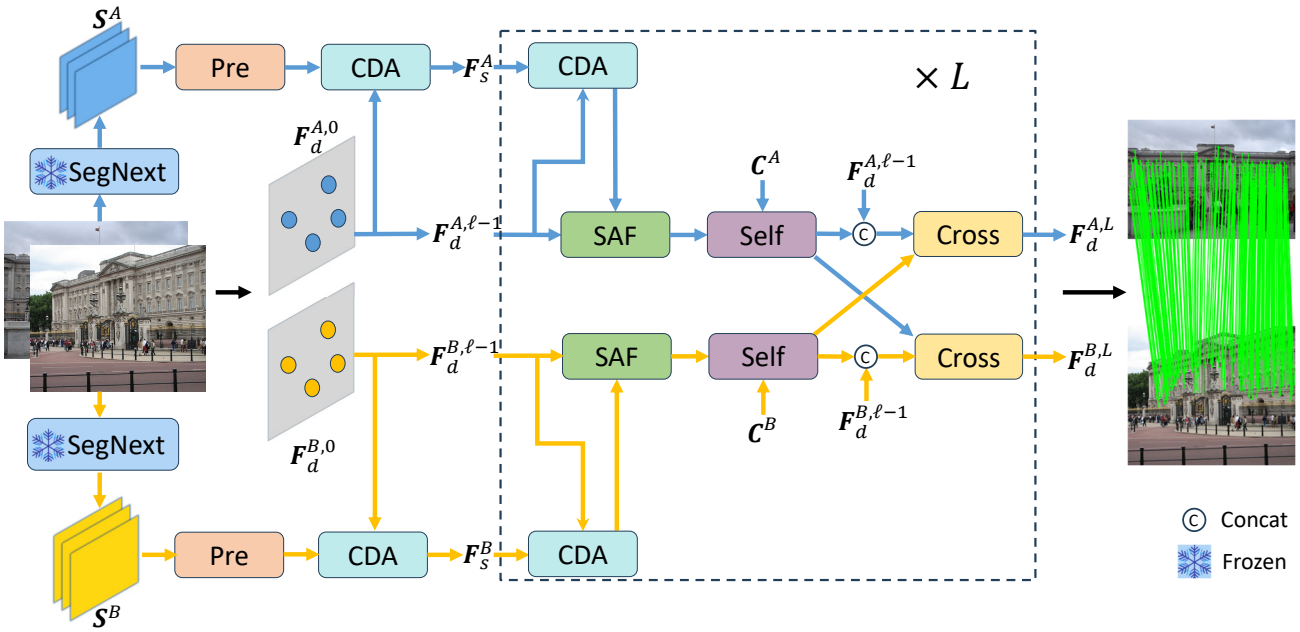

Figure 2: The framework of SemaGlue. We use a frozen SegNext to capture semantic information S^A, S^B, and transform them into semantic priors F_s^A, F_s^B through Pre-Processing (Pre) and an additional Cross-Domain Alignment (CDA) Block. Subsequently, at each layer, the semantic priors are first updated with the CDA block to eliminate the domain ambiguities. Then the Semantic-Aware Fusion (SAF) Block is employed to augment the semantic perceptual presentation of geometric features, followed by the Information Interaction Block (Self & Cross) to further circumvent mismatches. Finally, after passing through the stacked L layers, accurate correspondences are established by performing dual-softmax on the output features $F_d^{A,L}, F_d^{B,L}$.

more on structure-important feature points, ParaFormer (Lu et al. 2023) designs parallel attention and U-type GNN, IMP (Xue, Budvytis, and Cipolla 2023) emphasizes the geometric information with the attitude consistency loss, LightGlue (Lindenberger, Sarlin, and Pollefeys 2023) makes inference faster on image pairs that are intuitively easy to match. While such a direction deserves further exploration, these matching methods that rely on the quality of feature points sometimes falter in challenging scenarios like sparse texture and occlusions. Compared to the local features, high-level semantic information is more robust to appearance changes and more consistent with the matching tendency of human eyes. Therefore, in this paper, we joint semantic-geometric cues to guide end-to-end image feature matching.

Semantic Related Matching

On account of the semantic information is helpful to many tasks, leveraging robust semantic image representations is a promising avenue toward image matching. For area matching, SGAM (Zhang, Zhao, and Qian 2023) provides an intuitive way to predict true area matches with semantic labels, and MESA (Zhang and Zhao 2024) searches area matching based on SAM (Kirillov et al. 2023) segmentation by building a novel multi-relational graph structure. For sparse matching, OmniGlue (Jiang et al. 2024) incorporates DINOv2 (Oquab et al. 2024) and location coordinates to produce semantic features and only aggregates information from the DINOv2-pruned potential matching set. However, these approaches do not effectively integrate semantic information. Hence we propose a novel approach to genuinely perceive and amalgamate semantic information.

Methodology

To capture semantic priors and implement amalgamation effectively and appropriately, we first excavate semantic insights from a frozen Semantic Extractor together with Pre-Processing and an additional CDA block. Then in each layer, we attempt to perceive the scenes more than local geometric features with the CDA block and SAF block, and finally enhance the semantic information-assisted features with Information Interaction Block to yield the correspondence results. The framework of SemaGlue is shown in Figure 2. We will commence with an overall elucidation of the problem formulation, succeeded by a meticulous presentation of the implementation specifics for each block.

Problem Formulation

Given an image $I \in \{A, B\}$, local feature coordinates $\boldsymbol{C}^I = \{\boldsymbol{c}_i^I\}, \boldsymbol{c}_i^I \in \mathbb{R}^2$ and descriptions $\boldsymbol{D}^I = \{\boldsymbol{d}_i^I\}, \boldsymbol{d}_i^I \in \mathbb{R}^{C_d}$ can be obtained by an off-the-shelf feature extractor, where i means the i-th feature point in image I. For learnable image feature matching methods, the correct correspondences $\mathcal{M}$ can be established as:

$$\boldsymbol{F}_d^I = \mathcal{T}(\boldsymbol{C}^I, \boldsymbol{D}^I), \tag{1}$$

$$\boldsymbol{P} = \mathrm{Softmax}\left(\boldsymbol{F}_d^A (\boldsymbol{F}_d^B)^T\right) \odot \mathrm{Softmax}\left(\boldsymbol{F}_d^B (\boldsymbol{F}_d^A)^T\right)^T, \tag{2}$$

$$\mathcal{M} = \text{Index}\left(\boldsymbol{P} \geq \delta\right), \tag{3}$$

where $\boldsymbol{F}_d^I \in \mathbb{R}^{N^I \times C}$ is the position-embed local features, $\mathcal{T}(\cdot,\cdot)$ is a multilayer perceptron (MLP), $\odot$ is the Hadamard production, $\boldsymbol{P}$ is the assignment matrix, and $\text{Index}(\cdot)$ filters outliers that do not satisfy the certain condition. Different from the above, SemaGlue attempts to incorporate high-level scene object comprehension, *i.e.*, the semantic information into learnable image feature matching, thereby enabling the matcher to access broader ranges of messages beyond local geometric features. Thus, Eq. (1) is modified as:

$$\boldsymbol{F}_d^I = \mathcal{T}(\boldsymbol{C}^I, \boldsymbol{D}^I, \boldsymbol{S}^I). \tag{4}$$

In the common-used multi-layer framework, the features of points are updated from $\boldsymbol{F}_d^{I,0} = \boldsymbol{F}_d^I$ in Eq. (1), and the output in Eq. (4) of each layer can be re-written as:

$$\boldsymbol{F}_d^{I,\ell} = \mathcal{T}(\boldsymbol{C}^I, \boldsymbol{F}_d^{I,\ell-1}, \boldsymbol{S}^I), \ell = 1, \ldots, L. \tag{5}$$

Finally, we transform the matching problem into designing an effective network that can process information from different domains and help the establishment of matching.

Semantic Extractor

We employ the pre-trained SegNext encoder (Guo et al. 2022) to offer a semantic insights. The Semantic Extractor first captures an intermediate semantic feature map $\boldsymbol{S}^I$:

$$\boldsymbol{S}^I = \mathcal{F}_{\text{seg}}(\boldsymbol{\mathcal{I}}^I), \tag{6}$$

where $\boldsymbol{\mathcal{I}}^I \in \mathbb{R}^{H^I \times W^I \times 3}$ is the RGB tensor of image I, $\boldsymbol{S}^I \in \mathbb{R}^{\hat{H}^I \times \hat{W}^I \times C_s}$, $\hat{H}^I = \frac{H^I}{8}$, $\hat{W}^I = \frac{W^I}{8}$, and $\mathcal{F}_{\text{seg}}$ signifies the SegNext encoder, which is frozen during training.

Pre-Processing

It is essential to perform a Pre-Processing on $\boldsymbol{S}^I$ for the shape requirement of input in the CDA block. We first project it with a convolutional network $\text{Conv}(\cdot)$ to transfer the channel length C_s to C, and then further perform $\text{Flatten}(\cdot)$ and bilinear interpolation $\text{Interp}(\cdot)$ thus the shape of the input semantic priors could be the same as $\boldsymbol{F}^{I,\ell}$:

$$\boldsymbol{F}_s^I = \text{Interp}(\text{Flatten}(\text{Conv}(\boldsymbol{S}^I)), \tag{7}$$

where $\boldsymbol{F}_s^I \in \mathbb{R}^{N^I \times C}$ are the captured semantic priors that will be injected into the matching process.

Cross-Domain Alignment Block

In the ℓ-th layer, with local geometric features $\boldsymbol{F}_d^{I,\ell-1}$ and semantic priors $\boldsymbol{F}_s^I$ of the same shape as input, we can directly fuse them using a variety of simple strategies. However, $\boldsymbol{F}_d^{I,\ell-1}$ that focus on spatial structures and $\boldsymbol{F}_s^I$ that perceive high-level semantic information exist in different domains (*i.e.*, geometric and semantic domains). Direct fusion will ignore the domain gaps between them, resulting in sub-optimal descriptions and an unstable training process. Consequently, CDA block attempts to serve as a bridge between the semantic and geometric domains, establishing connections between $\boldsymbol{F}_d^{I,\ell-1}$ and $\boldsymbol{F}_s^I$, yielding

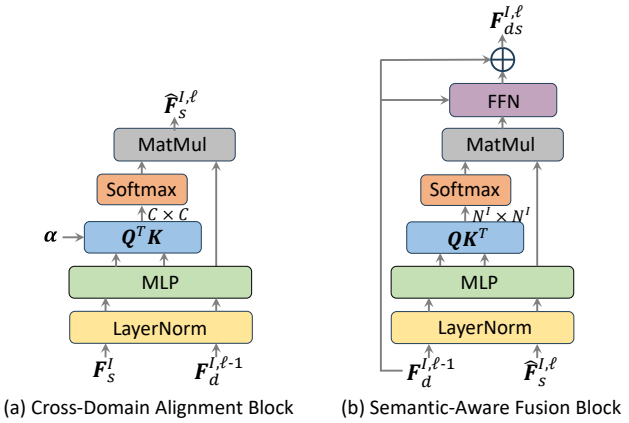

Figure 3: Structure details of Cross-Domain Alignment Block and Semantic-Aware Fusion Block.

amalgamation-applicable semantic priors $\hat{\boldsymbol{F}}_s^{I,\ell}$ in layer ℓ. To achieve this purpose, as shown in Figure 3(a), we first seek the relationship between the feature space of these domains by calculating the feature correlations of $\boldsymbol{F}_d^{I,\ell-1}$ and $\boldsymbol{F}_s^I$:

$$\boldsymbol{\mathcal{A}}_1^{I,\ell} = \text{Softmax}\left(\left(\boldsymbol{Q}^I\right)^T \boldsymbol{K}^I / \boldsymbol{\alpha}^\ell\right), \tag{8}$$

$$\boldsymbol{Q}^I = \boldsymbol{W}_q^\ell \text{LN}(\boldsymbol{F}_s^I), \boldsymbol{K}^I = \boldsymbol{W}_k^\ell \text{LN}(\boldsymbol{F}_d^{I,\ell-1}), \tag{9}$$

where $\boldsymbol{Q}^I, \boldsymbol{K}^I \in \mathbb{R}^{N^I \times C}$ thus $\boldsymbol{\mathcal{A}}_1^{I,\ell} \in \mathbb{R}^{C \times C}$, $\boldsymbol{W}_q^\ell$ and $\boldsymbol{W}_k^\ell$ are learnable weights, $\text{LN}(\cdot)$ indicates the layer norm (Ba, Kiros, and Hinton 2016) and $\boldsymbol{\alpha}^\ell \in \mathbb{R}^C$ are learnable scaling parameters. The obtained $\boldsymbol{\mathcal{A}}_1^{I,\ell}$ is the feature correlation matrix of $\boldsymbol{F}_d^{I,\ell-1}$ and $\boldsymbol{F}_s^I$, which can guide the domain alignment process and lead to eliminating the domain gaps. Thus, we narrow the distance of $\boldsymbol{F}_d^{I,\ell-1}$ and $\boldsymbol{F}_s^I$ with $\boldsymbol{\mathcal{A}}_1^{I,\ell}$:

$$\hat{\boldsymbol{F}}_s^{I,\ell} = \boldsymbol{\mathcal{A}}_1^{I,\ell} \boldsymbol{W}_v^\ell \text{LN}(\boldsymbol{F}_d^{I,\ell-1}), \tag{10}$$

where $\boldsymbol{W}_v^\ell$ are also learnable weights, $\hat{\boldsymbol{F}}_s^{I,\ell}$ are the excavated amalgamation-applicable semantic priors. With Eqs. (8) and (10), CDA block bridges the domain gaps to make full preparations for the applicable amalgamation of $\hat{\boldsymbol{F}}_s^{I,\ell}$ into the matching process, thus the process can perceive the high-level semantic information, improving the performance in complex cases. It is worth noting that we perform an extra CDA block after the Pre-Processing for producing better $\boldsymbol{F}_s^I$ as the input of the CDA block in each layer. The positive effect will be discussed in ablation studies.

Semantic-Aware Fusion Block

After obtaining the amalgamation-applicable semantic priors $\hat{\boldsymbol{F}}_s^{I,\ell}$ that mitigate domain gaps and have the same shape as local geometric features $\boldsymbol{F}_d^{I,\ell-1}$, the amalgamation then could be simply and appropriately conducted. To fully perceive the abundant semantic information to guide the matching process, we propose an attention-based SAF block that

inherently adjusts the amalgamation thereby selectively fusing the most relevant semantic information to the geometric feature representations. As shown in Figure 3(b), SAF block first employs the layer norm and MLP for initial feature manufactures, and draws the correlations between the semantic priors $\hat{\boldsymbol{F}}_s^{I,\ell}$ and the geometric features $\boldsymbol{F}_d^{I,\ell-1}$ using a trainable soft assignment approach like the attention mechanism (Vaswani et al. 2017) to calculate a semantic-aware attention map $\boldsymbol{\mathcal{A}}_2^{I,\ell}$:

$$\boldsymbol{\mathcal{A}}_2^{I,\ell} = \mathrm{Softmax}\left(\boldsymbol{Q}^I(\boldsymbol{K}^I)^T/\sqrt{C}\right), \tag{11}$$

$$\boldsymbol{Q}^I = \boldsymbol{W}_q^\ell \mathrm{LN}(\boldsymbol{F}_d^{I,\ell-1}), \boldsymbol{K}^I = \boldsymbol{W}_k^\ell \mathrm{LN}(\hat{\boldsymbol{F}}_s^{I,\ell}), \tag{12}$$

where $\boldsymbol{Q}^I, \boldsymbol{K}^I \in \mathbb{R}^{N^I \times C}$ thus $\boldsymbol{\mathcal{A}}_2^{I,\ell} \in \mathbb{R}^{N^I \times N^I}$. Then fuse the semantic information to geometric features as:

$$\tilde{\boldsymbol{F}}_d^{I,\ell} = \boldsymbol{\mathcal{A}}_2^{I,\ell} \boldsymbol{W}_v^\ell \mathrm{LN}(\boldsymbol{F}_d^{I,\ell-1}), \tag{13}$$

$$\boldsymbol{F}_{ds}^{I,\ell} = \boldsymbol{F}_d^{I,\ell-1} + \mathrm{FFN}(\boldsymbol{F}_d^{I,\ell-1} \parallel \tilde{\boldsymbol{F}}_d^{I,\ell}), \tag{14}$$

where $\parallel$ denotes the concatenating by channels, and $\mathrm{FFN}(\cdot,\cdot)$ means a feed-forward network (FFN) that compresses the channel length to C. From Eqs. (11) to (14), $\boldsymbol{\mathcal{A}}_2^{I,\ell}$ controls the information flow through each feature point, allowing each point to selectively focus on the messages from the semantic priors $\hat{\boldsymbol{F}}_s^{I,\ell}$. And the output $\boldsymbol{F}_{ds}^{I,\ell}$ as the new descriptions of feature points maintain both semantic and geometric information, enabling the matching process to reasonably perceive the scenes, leading to better matching performance even if in the case of sparse texture and occlusions.

Information Interaction Block

After the exploration of semantic-aided descriptions $\boldsymbol{F}_{ds}^{I,\ell}$ with CDA and SAF blocks, we perform Information Interaction Block to achieve intra- and inter-image communication and discover more robust representations of feature points. Similar to SuperGlue (Sarlin et al. 2020), we construct $\mathcal{G}_{\mathrm{self}}^\ell$ and $\mathcal{G}_{\mathrm{cross}}^\ell$ for all feature points, and we utilize FFN once more in the middle to recover the neglected information:

$$\boldsymbol{F}_{\mathrm{self}}^{I,\ell} = \mathcal{G}_{\mathrm{self}}^\ell(\boldsymbol{F}_{ds}^{I,\ell}, \boldsymbol{F}_{ds}^{I,\ell}), \tag{15}$$

$$\boldsymbol{F}_{\mathrm{ffn}}^{I,\ell} = \boldsymbol{F}_{\mathrm{self}}^{I,\ell} + \mathrm{FFN}(\boldsymbol{F}_{\mathrm{self}}^{I,\ell} \parallel \boldsymbol{F}_{ds}^{I,\ell}), \tag{16}$$

$$\boldsymbol{F}_{\mathrm{cross}}^{I,\ell} = \mathcal{G}_{\mathrm{cross}}^\ell(\boldsymbol{F}_{\mathrm{ffn}}^{I,\ell}, \boldsymbol{F}_{\mathrm{ffn}}^{J,\ell}). \tag{17}$$

The $\mathcal{G}_{\mathrm{self}}$ and $\mathcal{G}_{\mathrm{cross}}$ here indicate the GAT (Veličković et al. 2018) and the general formula is:

$$\mathcal{G}(\boldsymbol{F}^I, \boldsymbol{F}^J) = \boldsymbol{F}^I + \mathrm{FFN}\big(\boldsymbol{F}^I \parallel \boldsymbol{\mathcal{A}}\boldsymbol{V}^J\big), \tag{18}$$

$$\boldsymbol{\mathcal{A}} = \mathrm{Softmax}\left(\boldsymbol{Q}^I(\boldsymbol{K}^J)^T/\sqrt{C}\right), \tag{19}$$

$$\boldsymbol{Q}^I = \boldsymbol{W}_q \boldsymbol{F}^I, \boldsymbol{K}^J = \boldsymbol{W}_k \boldsymbol{F}^J, \boldsymbol{V}^J = \boldsymbol{W}_v \boldsymbol{F}^J. \tag{20}$$

Furthermore, for $\mathcal{G}_{\mathrm{self}}$, we adopt a relative positional encoding $R(\cdot) \in \mathbb{R}^{C\times C}$ (Su et al. 2024). Re-write Eq. (19) as:

$$\boldsymbol{\mathcal{A}} = \mathrm{Softmax}\left(\boldsymbol{Q}^I R(\boldsymbol{C}^J - \boldsymbol{C}^I)(\boldsymbol{K}^J)^T/\sqrt{C}\right). \tag{21}$$

Then the output of layer ℓ or the input of layer $\ell+1$ is received as $\boldsymbol{F}_d^{I,\ell} = \boldsymbol{F}_{\mathrm{cross}}^{I,\ell}$, until the last layer yielding $\boldsymbol{F}_d^{I,L}$. And the final matching results can be obtained by Eqs. (1) to (5) with the ultimate features $\boldsymbol{F}_d^{I,L}$ of all points.

Loss Function

Following the settings of loss function in (Lindenberger, Sarlin, and Pollefeys 2023), we first predict the matchability score $\boldsymbol{\sigma}^{A,\ell}$, $\boldsymbol{\sigma}^{B,\ell}$ with linear projection and then compute the augmented assignment matrix $\boldsymbol{P}^{AB,\ell}$ at each layer:

$$\boldsymbol{\sigma}^{I,\ell} = \mathrm{Sigmoid}(\mathrm{MLP}(\boldsymbol{F}_d^{I,\ell})), \tag{22}$$

$$\boldsymbol{P}^{AB,\ell} = (\boldsymbol{\sigma}^{A,\ell})^T \boldsymbol{\sigma}^{B,\ell} \odot \hat{\boldsymbol{P}}^{AB,\ell}, \tag{23}$$

where $\hat{\boldsymbol{P}}^{AB,\ell}$ is the matching similarity in Eq. (2). And the matching loss is:

$$\mathcal{L} = -\frac{1}{L}\sum_\ell \Bigg(\frac{1}{|M|}\sum_{(i,j\in M)} \log \boldsymbol{P}_{ij}^{AB,\ell} + \frac{1}{2\,|\bar{A}|}\sum_{i\in\bar{A}} \log(1-\boldsymbol{\sigma}_i^{A,\ell}) + \frac{1}{2\,|\bar{B}|}\sum_{j\in\bar{B}} \log(1-\boldsymbol{\sigma}_i^{B,\ell})\Bigg), \tag{24}$$

where M is the GT correspondence with a low reprojection error, and $\bar{A}$, $\bar{B}$ are unmatchable points.

Implementation Details

We set the number of stacked network layers to 9 (*i.e.*, $L = 9$). Following the approach outlined in (Lindenberger, Sarlin, and Pollefeys 2023; Zhang and Ma 2024b), we adopt the same datasets for two-stage training, Oxford and Paris (Radenović et al. 2018) for synthetic homography pre-training and MgedaDepth (Li and Snavely 2018) for fine-tuning. Specifically, in the first stage, images are resized to 640 $\times$ 480 and we extract 512/1024 feature points with SP (DeTone, Malisiewicz, and Rabinovich 2018)/A-LIKED (Zhao et al. 2023). The batch size is set to 48 with a learning rate of 0.0001, which is reduced by 20% every epoch after 20 epochs and the training is terminated after 40 epochs. In the second stage, images are resized to 1024×1024 with zero padding, feature points are extracted up to 2048. Batch size is 16 while the learning rate is 0.0001 for 20 epochs then decayed by a factor of 10 over 10 epochs until 40 epochs. We retain full resolution for the input image of SegNext in the whole training process. Besides, the channel C_s is set to 480 for Eq. (6) and C is 256 for Eq. (7). All processes are conducted with a single RTX3090 GPU.

Experiments

We evaluate SemaGlue on several tasks in this section to demonstrate how significant the role the applicable semantic information plays in image feature matching. Subsequently, through ablation studies, we analyze the overall novelty based on semantic priors and evaluate the effectiveness of the CDA block in resolving domain variances.

Homography Estimation

Homography estimation is a fundamental task in computer vision to determine a linear image-to-image map in a homogeneous space. We conduct this experiment on Hpatches (Balntas et al. 2017) referring to (Sun et al. 2021). Initially, all images are resized such that their smaller dimension is 480 pixels. We employ SP (DeTone, Malisiewicz,

Feature+Matcher		Acc.		AUC			
				DLT		RANSAC	
		@1px	@3px	@1px	@3px	@1px	@3px
SP	MNN	26.8	74.7	0.35	1.90	32.05	51.06
	SuperGlue	32.7	92.7	32.08	64.96	33.48	56.16
	SGMNet	31.9	89.0	17.93	48.39	32.27	53.86
	ResMatch	31.1	87.3	31.23	64.53	32.28	55.05
	IMP	31.2	87.7	24.02	54.32	33.49	56.06
	LightGlue	33.6	94.6	34.67	66.36	**34.87**	56.36
	SemaGlue (Ours)	**34.2**	**95.9**	**35.76**	**67.56**	34.50	**57.10**

Table 1: Homography estimation on Hpatches.

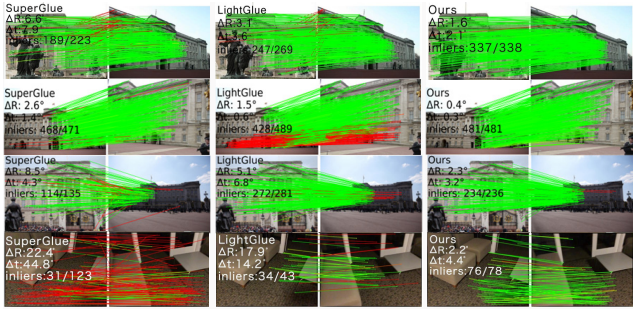

Figure 4: Qualitative illustration of relative pose estimation.

and Rabinovich 2018) to extract up to 1024 feature points for each image, identifying correspondences with image feature matching methods, and finally estimate the homography transformation with both non-robust Direct Linear Transform (DLT) and a robust estimator like RANSAC (Fischler and Bolles 1981). To assess the results, we try to classify a correspondence to be right or not in each image pair and calculate the accuracy (Acc.) at 1 and 3 pixels as the same as (Lindenberger, Sarlin, and Pollefeys 2023). Additionally, the mean reprojection error of the four image corners, as detailed by (DeTone, Malisiewicz, and Rabinovich 2018), is calculated to report the area under the cumulative error curve (AUC) at multiple thresholds (1 and 3 pixels). We choose the Mutual Nearest Neighbor (MNN) as a baseline, comparing SemaGlue with SuperGlue (Sarlin et al. 2020), SGMNet (Chen et al. 2021), ResMatch (Deng et al. 2024), IMP (Xue, Budvytis, and Cipolla 2023), and LightGlue (Lindenberger, Sarlin, and Pollefeys 2023). The results presented in Table 1 demonstrate that SemaGlue consistently outperforms all other methods.

Relative Pose Estimation

Recovering camera relative pose (rotation and translation) from two-view images is a crucial step in numerous vision applications. The accuracy of relative pose estimation serves as an indicator of the image feature matching performance. Following the experimental protocols in (Sarlin et al. 2020), we select MegaDepth-1500 (Li and Snavely 2018) and YFCC100M (Thomee et al. 2016) datasets. We first constrain the maximum dimension of images to 1600 pixels, then detect up to 2048 feature points with both SP (DeTone, Malisiewicz, and Rabinovich 2018) and ALIKED (Zhao et al. 2023) per image. We utilize DLT, RANSAC (Fischler and Bolles 1981) and MAGSAC++ (Barath et al. 2020) as geometric model estimators. The AUC of the maxima error of rotation and translation at different thresholds $(5^\circ, 10^\circ)$ is reported. We choose the same comparative methods as homography estimation. We also report the results in OmniGlue's paper (Jiang et al. 2024) on Megadepth1500 and include dense matching methods LoFTR (Sun et al. 2021) and PDC-Net+ (Truong et al. 2023) on YFCC100M for further comparisons. All results are presented in Tables 2 and 3, and qualitative results are illustrated in Figure 4 using the epipolar error to determine the correspondence accuracy, where a higher proportion of green indicates smaller epipolar error and red indicates a larger error. Additionally, we indicate the estimated rotation and translation errors as well as the percentage of correct matches in the top left corner. SemaGlue consistently outperforms all other methods, demonstrating that incorporating high-level semantic information enables the matcher to better perceive and understand scenes, thereby recovering camera poses accurately.

Feature+Matcher		AUC					
		DLT		RANSAC		MAGSAC++	
		@5°	@10°	@5°	@10°	@5°	@10°
SP	MNN	0.1	0.21	29.31	44.85	27.73	42.22
	SuperGlue	32.34	47.68	48.44	65.70	60.88	75.25
	SGMNet	4.78	11.22	39.95	58.49	47.63	64.56
	ResMatch	26.38	41.01	43.86	61.37	54.26	69.59
	IMP	32.96	48.59	44.94	62.45	56.86	71.87
	OmniGlue	—	—	47.40	65.00	—	—
	LightGlue	39.14	55.22	47.78	65.51	61.45	75.29
	SemaGlue (Ours)	**45.79**	**61.38**	**49.41**	**66.86**	**63.94**	**76.85**
ALIKED	MNN	0.55	1.94	44.62	59.87	47.71	62.67
	LightGlue	45.03	60.40	50.85	67.39	64.17	76.83
	SemaGlue (Ours)	**49.00**	**64.38**	**51.55**	**68.66**	**64.89**	**77.42**

Table 2: Relative pose estimation on MegaDepth-1500.

Feature+Matcher		AUC					
		DLT		RANSAC		MAGSAC++	
		@5°	@10°	@5°	@10°	@5°	@10°
SP	MNN	0.01	0.11	15.72	29.66	15.6	29.19
	SuperGlue	19.06	33.07	39.47	59.75	47.77	66.94
	SGMNet	9.84	19.85	34.22	54.50	35.26	55.75
	ResMatch	18.10	31.00	35.17	55.81	42.81	62.67
	IMP	24.21	39.79	38.68	59.16	47.08	66.11
	LightGlue	21.64	35.98	38.27	58.91	47.75	66.75
	SemaGlue (Ours)	**30.14**	**46.94**	**40.10**	**60.35**	**49.38**	**68.15**
ALIKED	MNN	0.08	0.37	32.34	52.32	32.72	51.72
	LightGlue	27.94	44.18	43.89	63.80	49.89	68.33
	SemaGlue (Ours)	**36.63**	**54.76**	**44.65**	**64.51**	**52.29**	**69.98**

Table 3: Relative pose estimation on YFCC100M.

Visual Localization

Visual localization that estimates the 6-degree of-freedom camera pose of a given reference image concerning its 3D scene model also requires robust and precise matching algorithms. Following (Chen et al. 2021), we incorporate different matching methods into the official Hloc (Sarlin et al. 2019) pipeline and assess them on Aachen Day-Night

Feature+Matcher		Day	Night
		(0.25m, 2°) / (0.5m, 5°) / (5.0m, 10°)	
	COTR	82.4 / 91.9 / 96.8	75.5 / 90.8 / 99.0
	LoFTR	**83.9 / 92.6 / 97.2**	**79.6 / 91.8 / 100.0**
SP	MNN	86.9 / 92.0 / 95.5	73.5 / 79.6 / 88.8
SP	SuperGlue	87.9 / 95.0 / **97.9**	84.7 / **92.9** / **99.0**
SP	SGMNet	86.5 / 93.2 / 97.2	82.7 / 91.8 / **99.0**
SP	ResMatch	86.8 / 93.7 / 97.2	81.6 / 91.8 / 98.0
SP	LightGlue	88.0 / 93.8 / 97.5	84.7 / 91.8 / **99.0**
SP	SemaGlue (Ours)	**88.6** / **95.1** / 97.8	**86.7** / 91.8 / **99.0**

Table 4: Visual localization on Aachen Day-Night v1.0.

Feature+Matcher		DUC1	DUC2
		(0.25m, 10°) / (0.5m, 10°) / (1.0m, 10°)	
SP	MNN	30.3 / 48.5 / 57.1	23.7 / 38.2 / 45.0
SP	SuperGlue	44.9 / 66.2 / 78.8	46.6 / **74.0** / **77.1**
SP	SGMNet	39.9 / 56.6 / 70.2	39.7 / 59.5 / 65.6
SP	ResMatch	42.9 / 61.6 / 73.7	38.2 / 62.6 / 69.5
SP	LightGlue	44.0 / 64.1 / 75.8	42.7 / 67.9 / 73.3
SP	SemaGlue (Ours)	**47.5** / **68.2** / **80.3**	**47.3** / 73.3 / 75.6

Table 5: Visual localization on Inloc.

v1.0 (Sattler et al. 2018) and Inloc (Taira et al. 2018). Specifically, with COLMAP (Schonberger and Frahm 2016), we initially triangulate a 3D point cloud from all reference images with known poses and calibration, then retrieve 20 reference images for each query image on Aachen Day-Night v1.0 and 40 reference images on InLoc with NetVLAD (Arandjelovic et al. 2016), matching the query image and the retrieved ones with image feature matching methods, where the feature points are detected up to 4096 by SP (DeTone, Malisiewicz, and Rabinovich 2018). Finally, camera poses can be estimated by RANSAC (Fischler and Bolles 1981) and a Perspective-n-Point solver. We report the pose recall in Tables 4 and 5 at multiple distance and orientation thresholds. We choose similar comparison methods as relative pose estimation. For Aachen Day-Night v1.0, we further add dense matching methods LoFTR (Sun et al. 2021) and COTR (Jiang et al. 2021), showing SemaGlue's promising performance in visual localization.

Analysis

We analyze SemaGlue by evaluating its zero-shot performance on unseen data, computational efficiency, and ablation experiments, highlighting its practicability and the impact of high-level semantic understanding.

Zero-Shot on Unseen Data Since OmniGlue (Jiang et al. 2024) is designed to show strong generalization in zero-shot missions, we test the generalization ability of different feature matching algorithms on the indoor dataset Scannet (Dai et al. 2017). Results are depicted in Table 6. SemaGlue exhibits powerful robust generalization, surpassing LightGlue with a 6.1% improvement on AUC@5°.

Computational Usage Image feature matching often requires real-time performance. Figure 5 shows runtime and memory usage statistics, highlighting that SemaGlue

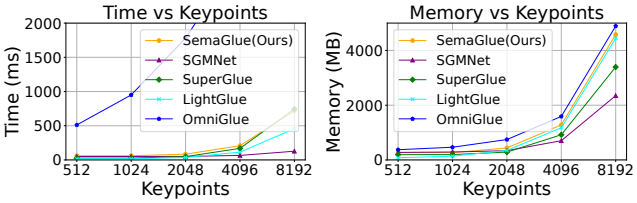

Figure 5: Computational usage.

Matcher	RANSAC		
	@5°	@10°	@20°
MNN	9.70	22.27	37.11
OmniGlue	14.00	28.29	44.30
LightGlue	14.23	30.29	46.87
SemaGlue (Ours)	**15.10**	**31.25**	**48.36**

Table 6: Zero-shot on Scannet.

Num.	Ex.	CDA		SAF	DLT		RANSAC	
		Extra	Inherent		@5°	@10°	@5°	@10°
(a)					38.78	54.39	47.34	65.46
(b)	✓			✓	42.25	58.34	47.35	64.92
(c)	✓		✓	✓	44.88	60.63	48.00	65.66
(d)	✓	✓	✓		42.52	58.10	47.89	65.15
(e)	✓	✓	✓	✓	45.79	61.38	49.41	66.86

Table 7: Ablation on MegaDepth1500. Ex. is the Semantic Extractor. Extra means the extra CDA block after the Pre-Processing. Inherent indicates the CDA block at each layer.

achieves optimal performance across tasks while maintaining real-time nature and competitive resource consumption.

Ablation Studies We conduct ablation studies by performing relative pose estimation on MegaDepth-1500 (Li and Snavely 2018), and reporting AUC with DLT or RANSAC (Fischler and Bolles 1981) in Table 7. (a) serves as the baseline. (b) integrates semantic information with only SAF while ignoring the domain gaps. (c) attempts to add CDA blocks to each layer but overlooks the extra one before the first layer that refines the features to initially bridge the domain gaps. (d) consists of the full CDA blocks but removes the SAF blocks. (e) is the full SemaGlue. Results reveal that SemaGlue benefits from all its ingredients.

Conclusion

We explore a new framework called SemaGlue for image feature matching, which perceives the semantic information during matching, fusing the semantic priors and the local geometric descriptions to enhance the representations of feature points thereby improving the matching performance, especially in the case of sparse texture and occlusions. SemaGlue proposes a novel Cross-Domain Alignment Block to bridge the domain gaps between geometric and semantic features, and utilizes a Semantic-Aware Fusion Block to achieve applicable semantic amalgamation. Extensive experiments demonstrate the impressive performance of SemaGlue and prove its practicability and generalization ability when used as the cornerstone for many visual tasks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (U23B2050 and 62276192), and the Fund of National Key Laboratory of Multispectral Information Intelligent Processing Technology (61421132302).

References

Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; and Sivic, J. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5297–5307.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Balntas, V.; Lenc, K.; Vedaldi, A.; and Mikolajczyk, K. 2017. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5173–5182.

Barath, D.; Noskova, J.; Ivashechkin, M.; and Matas, J. 2020. MAGSAC++, a fast, reliable and accurate robust estimator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1304–1312.

Brown, M.; and Lowe, D. G. 2007. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74: 59–73.

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 *Proceedings of the IEEE International Conference on Computer Vision*, 6301–6310.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 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*, 1501–1509.

DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2018. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 224–236.

Fischler, M. A.; and Bolles, R. C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6): 381–395.

Guo, M.-H.; Lu, C.-Z.; Hou, Q.; Liu, Z.; Cheng, M.-M.; and Hu, S.-M. 2022. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems*, 35: 1140–1156.

Hu, H.; Qiao, Z.; Cheng, M.; Liu, Z.; and Wang, H. 2021. DASGIL: Domain Adaptation for Semantic and Geometric-Aware Image-Based Localization. *IEEE Transactions on Image Processing*, 30: 1342–1353.

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 Conference on Computer Vision and Pattern Recognition*, 19865–19875.

Jiang, W.; Trulls, E.; Hosang, J.; Tagliasacchi, A.; and Yi, K. M. 2021. Cotr: Correspondence transformer for matching across images. In *Proceedings of the IEEE International Conference on Computer Vision*, 6207–6217.

Jiang, X.; Zhang, S.; Zhang, X.-P.; and Ma, J. 2023. Improving sparse graph attention for feature matching by informative keypoints exploration. *Computer Vision and Image Understanding*, 235: 103803.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE International Conference on Computer Vision*, 4015–4026.

Koutsoudis, A.; Vidmar, B.; Ioannakis, G.; Arnaoutoglou, F.; Pavlidis, G.; and Chamzas, C. 2014. Multi-image 3D reconstruction data evaluation. *Journal of Cultural Heritage*, 15(1): 73–79.

Li, Z.; and Snavely, N. 2018. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2041–2050.

Li, Z.; Zhang, S.; and Ma, J. 2023. U-Match: Two-view Correspondence Learning with Hierarchy-aware Local Context Aggregation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1169–1176.

Lindenberger, P.; Sarlin, P.-E.; and Pollefeys, M. 2023. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE International Conference on Computer Vision*, 17627–17638.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60: 91–110.

Lu, X.; Yan, Y.; Kang, B.; and Du, S. 2023. Paraformer: Parallel attention transformer for efficient feature matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1853–1860.

Ma, J.; Jiang, X.; Fan, A.; Jiang, J.; and Yan, J. 2021. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129(1): 23–79.

Ma, J.; Zhao, J.; Jiang, J.; Zhou, H.; and Guo, X. 2019. Locality preserving matching. *International Journal of Computer Vision*, 127: 512–531.

Ma, J.; Zhao, J.; Tian, J.; Yuille, A. L.; and Tu, Z. 2014. Robust point matching via vector field consensus. *IEEE Transactions on Image Processing*, 1706–1721.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; HAZIZA, D.; Massa, F.; El-Nouby, A.; et al. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*.

Radenović, F.; Iscen, A.; Tolias, G.; Avrithis, Y.; and Chum, O. 2018. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5706–5715.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763.

Rublee, E.; Rabaud, V.; Konolige, K.; and Bradski, G. 2011. ORB: An efficient alternative to SIFT or SURF. In *Proceedings of the IEEE International Conference on Computer Vision*, 2564–2571.

Sarlin, P.-E.; Cadena, C.; Siegwart, R.; and Dymczyk, M. 2019. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12716–12725.

Sarlin, P.-E.; DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2020. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4938–4947.

Sattler, T.; Maddern, W.; Toft, C.; Torii, A.; Hammarstrand, L.; Stenborg, E.; Safari, D.; Okutomi, M.; Pollefeys, M.; Sivic, J.; et al. 2018. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8601–8610.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4104–4113.

Schönberger, J. L.; Pollefeys, M.; Geiger, A.; and Sattler, T. 2018. Semantic visual localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6896–6906.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.

Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; and Zhou, X. 2021. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8922–8931.

Taira, H.; Okutomi, M.; Sattler, T.; Cimpoi, M.; Pollefeys, M.; Sivic, J.; Pajdla, T.; and Torii, A. 2018. InLoc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7199–7209.

Thies, J.; Zollhöfer, M.; and Nießner, M. 2019. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics*, 38(4): 1–12.

Thomee, B.; Shamma, D. A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; and Li, L.-J. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2): 64–73.

Truong, P.; Danelljan, M.; Timofte, R.; and Van Gool, L. 2023. Pdc-net+: Enhanced probabilistic dense correspondence network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 10247–10266.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7167–7176.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30: 1–11.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2018. Graph attention networks. In *Proceedings of the International Conference on Learning Representations*, 1–12.

Wang, Y.; Sun, R.; Luo, N.; Pan, Y.; and Zhang, T. 2024. Image-to-Image Matching via Foundation Models: A New Perspective for Open-Vocabulary Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3952–3963.

Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; and Philip, S. Y. 2020. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1): 4–24.

Xue, F.; Budvytis, I.; and Cipolla, R. 2023. Imp: Iterative matching and pose estimation with adaptive pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 21317–21326.

Yi, K. M.; Trulls, E.; Ono, Y.; Lepetit, V.; Salzmann, M.; and Fua, P. 2018. Learning to find good correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2666–2674.

Zhang, S.; and Ma, J. 2024a. ConvMatch: Rethinking Network Design for Two-View Correspondence Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5): 2920–2935.

Zhang, S.; and Ma, J. 2024b. DiffGlue: Diffusion-Aided Image Feature Matching. In *Proceedings of the ACM International Conference on Multimedia*, 8451–8460.

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

Zhang, Y.; Zhao, X.; and Qian, D. 2023. Searching from area to point: A hierarchical framework for semantic-geometric combined feature matching. *arXiv preprint arXiv:2305.00194*.

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.