MESA: Matching Everything by Segmenting Anything

Yesheng Zhang   Xu Zhao*

Department of Automation, Shanghai Jiao Tong University

Abstract

*Feature matching is a crucial task in the field of computer vision, which involves finding correspondences between images. Previous studies achieve remarkable performance using learning-based feature comparison. However, the pervasive presence of matching redundancy between images gives rise to unnecessary and error-prone computations in these methods, imposing limitations on their accuracy. To address this issue, we propose MESA, a novel approach to establish precise **area** (or **region**) matches for efficient matching redundancy reduction. MESA first leverages the advanced image understanding capability of SAM, a state-of-the-art foundation model for image segmentation, to obtain image areas with implicit semantic. Then, a multi-relational graph is proposed to model the spatial structure of these areas and construct their scale hierarchy. Based on graphical models derived from the graph, the area matching is reformulated as an energy minimization task and effectively resolved. Extensive experiments demonstrate that MESA yields substantial precision improvement for multiple point matchers in indoor and outdoor downstream tasks, e.g. +13.61% for DKM in indoor pose estimation.*

1. Introduction

Feature matching aims at establishing correspondences between images, which is vital in a broad range of applications, such as SLAM [5], SfM [39] and visual localization [38]. However, achieving exact point matches is still a significant challenge due to the presence of matching noises [33], including scale variations, viewpoint and illumination changes, repetitive patterns, and poor texturing.

The underlying goal of feature matching is to find correspondences for points in the source image from a highly redundant candidate set, *i.e.* all the points of the target image. Therefore, feature matching essentially involves reducing the **matching redundancy**, which is caused by troublesome matching noises. Recent years have witnessed significant

---
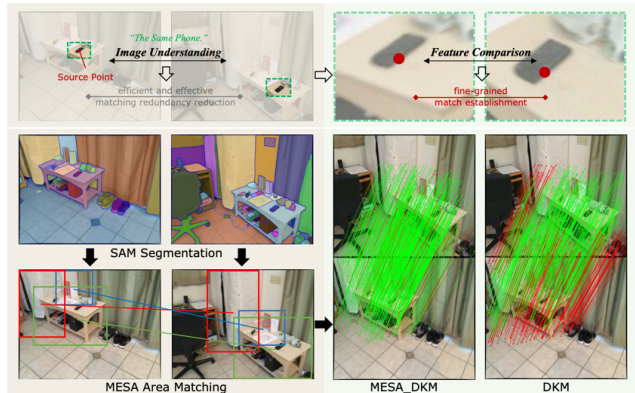
* Corresponding author: Xu Zhao.

Figure 1. **The matching redundancy reduction in MESA.** High-level image understanding enables efficient matching redundancy reduction, allowing for precise point matching by dense feature comparison. Therefore, MESA effectively reduces the matching redundancy by area matching based on SAM [22] segmentation, significantly improving the accuracy of DKM [17].

advancements in learning-based methods for feature matching. Especially the semi-dense [8, 40] and dense methods [17], which search for matches densely by deep feature comparison, obtain an impressive precision gap over sparse methods [14]. However, as they heavily rely on dense feature comparison of the entire image, limited input resolution and error-prone computation in matching-redundant parts of images decrease their accuracy. Thus, effective matching redundancy reduction is urgently required for them.

On the other hand, most matching redundancy can be effectively and efficiently identified through high-level image understanding and only strongly correlated local **areas** (also known as **regions**) need dense feature comparison to determine precise matches. (Fig. 1). Therefore, recent methods [9, 15] perform learning-based overlap segmentation before matching. However, implicit learning needs non-reusable computational overhead and the redundancy still exists in the overlap. To address these issues, some work turn to explicit semantic prior [18, 52]. Unlike manually specifying a topic number to group feature patches [18], *Area to Point Matching* (A2PM) framework proposed by SGAM [52] provides a more intuitive way to reduce matching redundancy. Specifically, this framework establishes

semantic area matches first where matching redundancy is largely removed, and then obtains accurate correspondences within these areas utilizing an state-of-the-art point matcher. Nevertheless, SGAM relies on semantic segmentation to determine area matches. Its performance, thus, decreases when encountering inexact semantic labeling and semantic ambiguity [52]. Furthermore, the benefits of A2PM cannot be extrapolated to more general scenes due to the close-set semantic labels. Hence, reducing matching redundancy through semantic segmentation suffers from impracticality.

Recently, the Segment Anything Model (SAM) [22] has gained notable attention from the research community due to its exceptional performance and versatility, which has the potential to be the basic front-end of many tasks [32, 49, 50]. This suggests that the foundation model can accurately comprehend image contents across various domains. Taking inspiration from this, we realize that the image understanding from the foundation model can be leveraged to reduce matching redundancy. Thus, we propose to establish area matches based on SAM segmentation to overcome limitations of SGAM [52]. Similar to the semantic segmentation, the image segmentation also provides multiple areas in images, but without semantic labels attached to these areas. However, the general object perception of SAM ensures that its segmentation results inherently contain implicit semantic information. In other words, a complete semantic entity is always segmented as an independent area by SAM. Hence, matching these areas effectively reduces matching redundancy and promotes accurate point matching within areas [52]. Furthermore, the absence of explicit semantics alleviate issues of inaccurate area matching caused by erroneous labeling. The generic limitations due to semantic granularity are also overcome. Nevertheless, area matching can not be simply achieved by semantic labels but requires other approaches under this situation.

In this work, we propose Matching Everything by Segmenting Anything (MESA, Fig. 2), a method for precise area matching from SAM segmentation. MESA focuses on **two main aspects**: *constructing area relations* and *finding area matches based on these relations*. Since SAM areas only provide local information, matching them independently can lead to inaccurate results, especially in scenes with scale variation and repetitiveness. To address this, MESA first constructs a novel graph structure, named *Area Graph* (AG), to model the global context of areas for subsequent precise matching. The AG takes areas as nodes and connects them with two types of edges: undirected edges for adjacent areas and directed edges for areas that include one another. Both edges capture global information, and the latter also enables the construction of hierarchy structures in AG similar to [43] for efficient matching. Afterwards, MESA performs area matching by deriving two distinct graphical models from the AG: *Area Markov Ran-*

dom Field (AMRF) and *Area Bayesian Network* (ABN). The AMRF involves global-informative edges in AG, thus allowing global-consistent area matching through energy minimization on the graph. The ABN, furthermore, is proposed to facilitate the graph energy calculation, through the hierarchy structure of AG. Specifically, the graph energy is determined based on the similarity and spatial relation between areas, making this energy minimization effectively solvable through the *Graph Cut* [23]. The area similarity calculation is accomplished by a novel learning-based model and accelerated by the ABN. Finally, we propose a global matching energy to refine accurate area matches, addressing the issue of multiple solutions in *Graph Cut*, leading to effective matching redundancy reduction.

Our method makes several contributions. **1)** In pursuit of effective matching redundancy reduction, we propose MESA to achieve accurate area matching from results of SAM, an advanced image segmentation foundation model. MESA realizes the benefit of A2PM framework in a more practical manner, leading to precise feature matching. **2)** We introduce a multi-relational graph, AG, to model the spatial structure and scale hierarchy of image areas, contributing to outstanding area matching. **3)** We propose graphical area matching by converting AG into AMRF and ABN. The AMRF is employed to formulate the area matching as an energy minimization, which is effectively solvable through the *Graph Cut*. The energy calculation is efficiently performed by the property of ABN and the proposed learning area similarity model. **4)** MESA is evaluated in indoor and outdoor downstream tasks of feature matching, obtaining remarkable improvements for multiple point matchers.

2. Related Work

Sparse, Semi-Dense and Dense Matching. There are three types of feature matching methods: sparse, semi-dense and dense. Classical sparse feature matching [31] involves detecting and describing keypoints in images, followed by keypoint matching. The learning counterpart of this framework utilizes neural networks to perform feature detection [2, 14], description [16, 36, 53] or matching [37, 51]. To avoid the detection failure in sparse methods, semi-dense methods [20, 26, 40] are proposed, also known as the detector-free methods. These methods [8, 42, 43] perform dense feature matching over the entire image and then select confident matches, which achieve impressive matching precision. Dense matching methods [17, 35, 44] output a dense warp with confidence map for the image pair. Recent DKM [17] reformulates this framework as a Gaussian Process and achieves *state-of-the-art* performance. Our method, however, focus on area matching between images, which can be combined with both semi-dense and dense point matching methods to increase their precision.

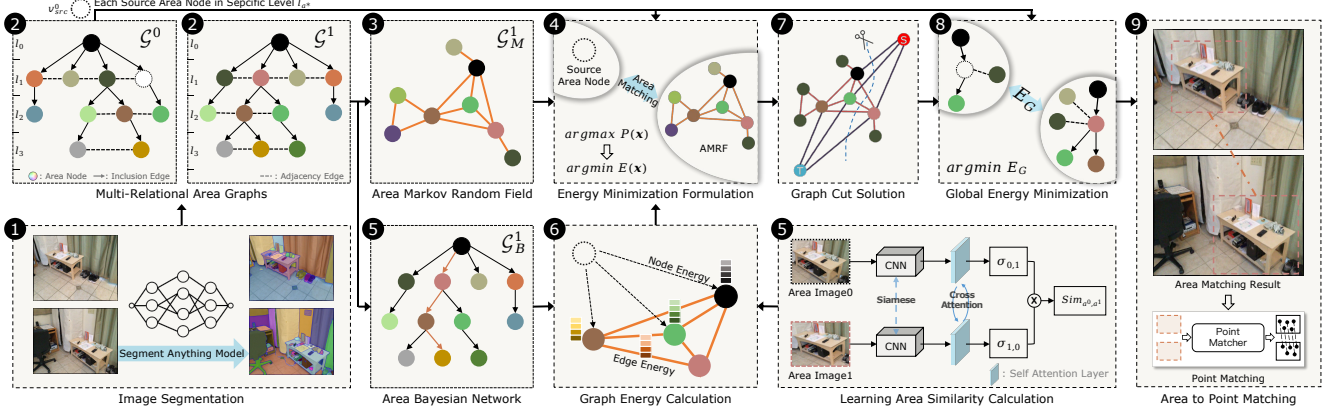**Matching Redundancy Reduction.** Matching redundancy

Figure 2. **Overview of MESA.** Based on ❶ *SAM segmentation*, we first construct ❷ *Area Graph*s. Then the graph is turned to two graphical models based on its two different edges. Through ❸ *Area Markov Random Field*, area matching is formulated as an ❹ *Energy Minimization*. Then, leveraging ❺ *Area Bayesian Network* and our ❺ *Learning Area Similarity Calculation*, ❻ *Graph Energy* can be efficiently calculated. Therefore, ❼ *Graph Cut* is utilized to obtain putative area matches. Finally, ❽ *Global Energy Minimization* determines the best area match, which serves as the input of subsequent point matcher for precise feature matching, following the ❾ *Area to Point Matching* framework [52].

is evident in non-overlapping areas between images, motivating several works [9, 12, 20] to focus on covisible areas extraction. They predict overlaps between images by iterative matching [20] or overlap segmentation [9, 12]. Nevertheless, matching redundancy still exists in the overlapping area, when it comes to detailed point matching. TopicFM [18] proposes to divide image contents into topics and then restrict matching to the same topic to avoid redundant computation. The learning-based topic inference, however, is implicit, suffering from lack of versatility and generalize issues. SGAM [52] explicitly finds area matches between images, which are fed into point matchers for redundancy reduction. The Area to Point Matching framework is simple yet effective. Our method further builds on its advantages.

Area to Point Matching. Establishing area matches is an effective way to reduce the matching redundancy. Thus, SGAM [52] proposes the Area to Point Matching (A2PM) framework. This framework obtains area matches as the input of point matchers, which possess higher resolution and less matching noise than original images, contributing to accurate point matching. However, SGAM heavily relies on explicit semantic prior, leading to performance sensitivity of area matching to semantic labeling precision and scene semantic granularity. In contrast, MESA utilizes pure image segmentation for area matching, which is more practical and remedies drawbacks associated with explicit semantic.

3. Area Graph

In this section, we introduce the concept of an Area Graph (AG) and explain how to construct the AG from the SAM [22] results of an image. The main reason to adopt the AG is that direct area matching on SAM results is unacceptable, as global information is ignored in independent areas. Also, fixed area sizes hinder robust point matching under

scale changes. Therefore, AG is proposed to capture the global structural information of these areas and construct scale hierarchy for them, contributing to accurate and robust area matching.

3.1. Area Graph Definition

The AG ($\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$) takes image areas as nodes and contains two kinds of edges to model inter-area relations (Fig. 3), thus making it a *multi-relational graph* [1]. The graph nodes include both areas provided by SAM and areas generated by graph completion (cf. Sec. 3.2). We divide these areas into L levels according to their sizes, corresponding to different image scales, to serve as the foundation of area scale hierarchy. On the other hand, the graph edges ($\mathcal{E} = \mathcal{E}_{in} \bigcup \mathcal{E}_{adj}$) represent two relations between areas, *i.e.* inclusion ($\mathcal{E}_{in}$) and adjacency ($\mathcal{E}_{adj}$). The inclusion edge $e_{in} \in \mathcal{E}_{in}$ is directed, pointing from an area to one of its containing areas. It forms a hierarchical connection between graph nodes, enabling robust and efficient area matching especially under scale changes. The adjacency edge $e_{adj} \in \mathcal{E}_{adj}$ is undirected, indicating the areas it connects share common parts but without the larger one including the smaller one. This edge captures the spatial relations between areas. By the above two edges, AG models both the spatial and scale structure of image areas, thus contributing to exact area matching.

3.2. Area Graph Construction

The construction of AG includes collecting areas as nodes and connecting them by proper edges. Firstly, not all SAM areas can function as nodes, since some are too small or have extreme aspect ratios, rendering them unsuitable for point matching. Thus, *Area Pre-processing* is performed first to obtain initial graph nodes. We then approach the
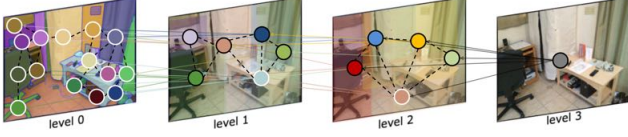
Figure 3. **The proposed Area Graph.** The graph nodes (circles with masks representing rectangle areas) includes both areas from SAM results (white boundaries) and our graph completion algorithm (black boundaries). They are divided into four levels according to their sizes. The adjacency edges (dashed lines) and inclusion edges (arrows) connect all nodes. Only adjacency edges within the same level are shown for better view.

edge construction as a *Graph Link Prediction* problem [34]. Afterwards, the preliminary AG is formed, but it still lacks matching efficiency and scale robustness. Thus, we propose the *Graph Completion* algorithm, which generates additional nodes and edges to construct the scale hierarchy.

Area Pre-processing. To filter unsuitable areas, we set two criteria: the acceptable minimal area size (T_s) and maximum area aspect ratio (T_r). Any area that has smaller size than T_s or larger aspect ratio than T_r, gets screened out. The remaining areas are added into the candidate areas set. For each filtered area, we fuse it with its nearest neighbor area in the candidate set. We repeat the filtering and fusion on the candidate set until no areas get screened out. Then, we assign a level l_a to each candidate area a based on its size, by setting L size thresholds ($\{TL_i \mid i \in [0, L-1]\}$):

$$l_a = i \mid TL_i \leq W_a \times H_a < TL_{i+1}. \qquad (1)$$

The size level is the basis of scale hierarchy in AG.

Graph Link Prediction. The edge construction is treated as a link prediction problem [34]. Given two area nodes (v_i, v_j), the edge between them (e_{ij}) can be predicted according to the spatial relation of their corresponding areas (a_i, a_j). This approach adopts the ratio of *the overlap size* (O_{ij}) to *the minimum area size* between two areas ($\delta = O_{ij}/\min(W_i \times H_i, W_j \times H_j)$) as the score function:

$$e_{ij} \in \begin{cases} \mathcal{E}_{in} & , \delta >= \delta_h \\ \mathcal{E}_{adj} & , \delta_l < \delta < \delta_h \ , \\ \varnothing & , \delta \leq \delta_l \end{cases} \qquad (2)$$

where δ_l, δ_h are predefined thresholds.

Graph Completion. Initial AG is achieved by connecting all processed nodes with different edges. However, since SAM inherently produces areas containing complete entity, there are few inclusion relations among areas. Consequently, initial AG lacks the scale hierarchy, which reduces its robustness at scale variations and makes accessing nodes inefficient. To address the issue, we propose the *Graph Completion* algorithm, which generates additional nodes and edges to ultimately construct a tree structure in the original graph. The core of this algorithm is to generate parent nodes for each orphan node, *i.e.* the node has

no parent node in the next higher level, from small to large size levels. The algorithm begins at the smallest level, collects all orphan nodes, and clusters them based on their center coordinates. Nodes in the same cluster have their corresponding areas fused with their nearest neighbors. It is noteworthy that these generated areas containing multiple objects preserve the internal implicit semantic. Based on our level thresholds, the resulting areas correspond to new higher level nodes. If a node remains single after clustering, we increase its area size to the next level to allow for potential parent nodes. We repeat the above operations on each level and connect generated nodes to others by suitable edges. More details can be found in the Supp. Mat.

4. Graphical Area Matching

In this section, we describe our area matching method, which formulates area matching on the graph, utilizing two graphical models derived from the AG. Given two AGs ($\mathcal{G}^0, \mathcal{G}^1$) of the input image pair (I_0, I_1) and one area ($a_{src}^0 \in I_0$) corresponding to node $v_{src}^0 \in \mathcal{G}^0$ (termed as the *source node*), area matching involves finding the node $v_j^1 \in \mathcal{G}^1$ with the highest probability of matching its corresponding area a_j^1 to the source node area a_{src}^0. However, treating this problem as independent node matching is inadequate, as which disregards global information from adjacency and inclusion relations between areas modeled in AG by its two edges. Meanwhile, these two edges respectively convert AG into two graphical models, *i.e.* Markov Random Fields (undirected edges) and Bayesian Network (directed edges). Thus, the area matching can be naturally formulated inside a framework of graphical model.

4.1. Area Markov Random Field

The adjacency edge in AG implies the correlation between probabilities of two nodes matching the source node. By considering the general adjacency relation, which includes the inclusion relations as adjacency too, the $\mathcal{G}^1$ is transformed into an undirected graph. Next, random variables ($\boldsymbol{x}$) are introduced for all nodes to indicate their matching status with the source node. The binary variable $x_i \in \boldsymbol{x}$ is equal to 1 when v_i^1 matches v_{src}^0 and 0 otherwise. Therefore, Area Markov Random Field (AMRF, $\mathcal{G}_M^1 = \langle \mathcal{V}, \mathcal{E}_{adj} \rangle$) is obtained and area matching can be performed by maximizing the joint probability distribution over the AMRF:

$$\arg\max_{\boldsymbol{x}} P(\boldsymbol{x}). \qquad (3)$$

Based on the *Hammersley-Clifford* theorem [10], the probability distribution defined by AMRF belongs to the *Boltzmann distribution*, which is an exponential of negative energy function ($P(\boldsymbol{x}) = \exp(-E(\boldsymbol{x}))$). Therefore, the area matching can be formulated as an energy minimization.

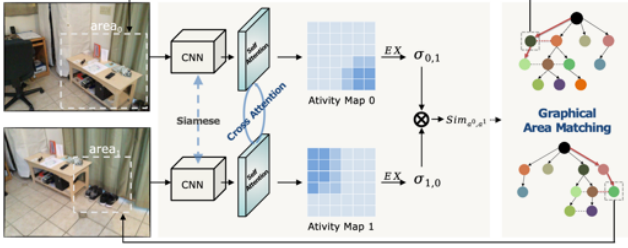$$\arg\min_{\boldsymbol{x}} E(\boldsymbol{x}). \qquad (4)$$

20220

Figure 4. **Learning area similarity.** The area similarity calculation is formed as the patch-level classification. We predict the probability of each patch in one area appearing on the other to construct activity maps. The similarity is obtained by the product of activity expectations, contributing to our exact area matching.

The energy can be divided into two parts, *i.e.* the energy of nodes ($E_\mathcal{V}$) and edges ($E_\mathcal{E}$), based on the graph structure.

$$E(\boldsymbol{x}) = \sum_i E_\mathcal{V}(x_i) + \lambda \sum_{(i,j)\in\mathcal{N}} E_\mathcal{E}(x_i, x_j), \quad (5)$$

where λ is a parameter balancing the terms and $\mathcal{N}$ is the set of all pairs of neighboring nodes. For each graph node v_i^1, its energy is expected to be low when its matching probability is high, which can be reflected by the apparent similarity ($S_{a^0_{src}a^1_i}$) between a^0_{src} and a^1_i.

$$E_\mathcal{V}(x_i) = |x_i - S_{a^0_{src}a^1_i}|. \quad (6)$$

The edge energy aims to penalize all neighbors with different labels, and the Potts model [46] (T) would be a justifiable choice. To better reflect the spatial relation, the Potts interactions are specified by IoU [47] of neighboring areas.

$$E_\mathcal{E}(x_i, x_j) = IoU(a_i^1, a_j^1) \cdot T(x_i \neq x_j). \quad (7)$$

Function $T(\cdot)$ is 1 if the argument is true and 0 otherwise. Finally, the area matching is formulated as an binary labeling energy minimization. By carefully defining the energy function, the energy minimization problem in Eq. (4) is efficiently solvable via the *Graph Cut* algorithm [4, 23]. The obtained minimum cut of the graph $\mathcal{G}_M^1$ is the matched node set ($\{v_h^1 | h \in \mathcal{H}\}$). Although the set may contain more than one area node, the best matching result can be achieved from this set by our refinement algorithm (cf. Sec. 4.4).

4.2. Learning Area Similarity

The proposed *Graph Cut* solution relies on graph energy calculations for both nodes and edges. Unlike easily available area pair IoUs for $E_\mathcal{E}$, determining the area apparent similarity for $E_\mathcal{V}$ is not straightforward. Thus, we turn to the learning-based framework, inspired by recent successes of learning models in point matching [8, 21, 40]. One simple idea is to calculate the correlation of learning descriptors of two areas [21] as the area similarity. However, the descriptor correlation is too rough for accurate area matching

and lacks fine-grained interpretability. To overcome these issues, we decompose the area similarity calculation into two patch-level classification problems as shown in Fig. 4.

Specifically, for each image in the area image pair $\{a_j | j \in \{0, 1\}\}$ reshaped to the same size, we perform binary classification for each $1/8 \cdot W_{a_j} \times 1/8 \cdot H_{a_j}$ image patch p_i^j (*where i is the index of patch and j is the index of area image*) in it, computing the probability of p_i^j appearing on the other area image, termed as the patch activity σ_i^j. To accomplish the classification, we first extract patch-wise features from each area image using a Siamese CNN [24]. Then we update these patch features via self and cross-attention with normalization [45], resulting in patch activities. Utilizing these patch activities, we construct an activity map ($\sigma_m^j = \{\sigma_i^j | j \in \{0, 1\}\}_i$, *where i is the index of element in this set*) for each area image. When two areas are ideally matched, the corresponding 3D point of every pixel in one area is projected onto the other area. Hence, all the patch activities of both areas should be closed to 1, revealing the area similarity can be represented by the product of expectations of two activity maps, $EX(\sigma_m^i) = \sigma_{i,j}$, where $i \neq j, i, j \in \{0, 1\}$.

$$Sim_{a^0, a^1} = EX(\sigma_m^0) \times EX(\sigma_m^1) = \sigma_{0,1} \times \sigma_{1,0}. \quad (8)$$

The training details of this learning model are described in the Supp. Mat.

4.3. Area Bayesian Network

Although the *Graph Cut* can be performed in polynomial time [4], the dense energy calculation over $\mathcal{G}_M^1$ is time consuming. Furthermore, due to the scale hierarchy in AG, this dense calculation is highly redundant. In particular, if the area a^0_{src} is not matched to a_j^1, it won't be matched to any children area of a_j^1. This observation reveals the conditional independence in the similarity calculation, which involves inclusion edges in $\mathcal{G}^1$, thus turning $\mathcal{G}^1$ to a Bayesian Network ($\mathcal{G}_B^1$) [3]. Therefore, the redundancy in the similarity calculation can be reduced. In practice, we calculate the dense similarities by constructing a similarity matrix $M_S \in \mathbb{R}^{|\mathcal{V}^0| \times |\mathcal{V}^1|}$. Note *not* all similarities in M_S need calculation, but any similarity can be accessed in M_S. We first calculate similarities related to all source nodes. Subsequent calculations are saved in M_S as well. For $M_S[i, j]$ that has not been calculated, we achieve it by our learning model (cf. Sec. 4.2):

$$M_S[i, j] = Sim_{a_i^0, a_j^1}. \quad (9)$$

If $M_S[i, j] < T_{as}$, all children nodes $\{v_h^0 | h \in ch^0(i)\}$ and $\{v_c^1 | c \in ch^1(j)\}$ of v_i^0 and v_j^1 are found from $\mathcal{G}_B^0$ and $\mathcal{G}_B^1$, where $ch^j(i)$ is the index set of children indices of node v_i^j from $\mathcal{G}_B^j$. Based on the conditional independence, we have:

$$M_S[h, k] = 0, \ \forall (h, k) \in ch^0(i) \times ch^1(j). \quad (10)$$

This operation effectively reduce the times of similarity calculation, leading to more efficient area matching.

4.4. Global Matching Energy Minimization

The minimum cut $\{v_h^1 | h \in \mathcal{H}\}$ achieved through the *Graph Cut* may contain more than one area node, which means further refinement is necessary to obtain the best area match. Moreover, the aforementioned graphical area matching, *i.e.* finding the corresponding area node in $\mathcal{G}^1$ for $v_{src}^0 \in \mathcal{G}^0$, only considers the structure information in $\mathcal{G}^1$ and ignores the structure of $\mathcal{G}^0$. To overcome this limitation, we propose a global matching energy E_G for each candidate node v_h^1, which consists of four parts.

$$E_G(v_h^1) = \frac{1}{Z}(\mu \cdot E_{self}(v_h^1) + \alpha \cdot E_{parent}(v_h^1) \\ + \beta \cdot E_{children}(v_h^1) + \gamma \cdot E_{neighbour}(v_h^1)) \quad , \quad (11)$$

where μ, α, β and γ are weights to balance the terms; Z is the partition function. The $E_{self}(v_h^1)$ is the energy related to matching probability between v_{src}^0 and v_h^1:

$$E_{self}(v_h^1) = |1 - Sim_{a_{src}^0, a_h^1}|. \quad (12)$$

The $E_{parent}(v_h^1)$ is the energy related to matching probability between the parent node pairs of v_{src}^0 and v_h^1:

$$E_{parent}(v_h^1) = \\ \min\{|1 - Sim_{a_u^0, a_r^1}| \big| u \in p^0(src), r \in p^1(h)\}, \quad (13)$$

where $p^i(j)$ is the index set of parent nodes of v_j^i in $\mathcal{G}_i$. This energy is the minimum matching energy among all parent node pairs of v_h^1 and v_{src}^0. Same as the E_{parent}, the $E_{children}$ and $E_{neighbour}$ are the energy parts of children and neighbour node matching pairs. Afterwards, the best area match $v_{h^*}^1$ in the set can be found by minimize E_G:

$$h^* = \arg\min_{h \in \mathcal{H}} E_G(v_h^1). \quad (14)$$

If the $E_G(v_{h^*}^1) > T_{E_{max}}$ (a threshold parameter), the source area node v_{src}^0 is considered to have no matches. To further improve the accuracy of final match, we set an energy range threshold T_{Er} to collect all the candidates within a certain energy range.

$$\{v_{\bar{h}}^1 \big| |E_G(v_{\bar{h}}^1) - E_G(v_{h^*}^1)| \le T_{Er}, \bar{h} \in \mathcal{H}\}. \quad (15)$$

Then the final area match is achieved by fusing $v_{h^*}^1$ and all candidates $\{v_{\bar{h}}^1\}_{\bar{h}}$ with E_G as weights, similar to [21]. This refinement completely considers the structure information of both $\mathcal{G}^0$ and $\mathcal{G}^1$ and achieves exact area matches.

Method	AOR↑	AMP@0.6↑	AMP@0.7↑	AMP@0.8↑	AreaNum↑
SEEM [54]+SGAM [52]†	52.96	49.39	29.88	12.32	**2.66**
w/ GAM ($\phi = 0.3$)	57.74	53.86	33.61	13.13	1.21
w/ GAM ($\phi = 1.0$)	**60.59**	**59.43**	36.29	**15.27**	1.55
w/ GAM ($\phi = 5.0$)	60.32	59.34	**36.55**	**15.27**	2.23
MESA	67.98	80.09	57.74	22.73	**3.47**
w/ GAM ($\phi = 0.3$)	**68.99**	**81.11**	**60.75**	**25.77**	1.97
w/ GAM ($\phi = 1.0$)	68.78	81.07	60.06	25.21	3.25
w/ GAM ($\phi = 5.0$)	68.51	80.84	59.22	23.87	3.35

† SGAM with only semantic area matching activated.

Table 1. **Area matching results on ScanNet1500.** We compare the area matching performance between SGAM and MESA, combined with GAM [52] under various ϕ settings. Results are highlighted as **first**, second and third for both series respectively.

5. Experiments

5.1. Area Matching

Since accurate area matching is the prerequisite for the precise point matching, we evaluate our method for this task on ScanNet1500 [11] benchmark.

Experimental setup. The $T_{E_{max}}$ is 0.35. The input size is 320×320 for our learning model. Other scene-independent parameter settings and implementation details can be found in our Supp. Mat. The previous area matching approach, SGAM [52] with SEEM [54] providing the semantic segmentation input, is compared with our method MESA. The impacts of *Geometric Area Matching* (GAM) in [52] to both methods are also investigated, whose parameter ϕ reflects the strictness of outlier rejection. We employ the *Area Overlap Ratio* (AOR, %) and *Area Matching Precision* under three thresholds (AMP@0.6/0.7/0.8, %) [52] as metrics. The average area number per image (AreaNum) is reported.

Results. In Tab. 1, our MESA outperforms SGAM by a large margin (67.98 *vs.* 60.59 in AOR). The poor area matching performance of SGAM is mainly caused by the inaccurate semantic labeling of SEEM, which is avoided in MESA. MESA can also achieve more areas than SGAM, favouring downstream tasks by more point matches. Although GAM brings significant improvement for SGAM, it achieves less enhancement on MESA, because more precise area matching is obtained by MESA. We empirically adopt GAM w/ $\phi = 1.0$ for MESA in the following experiments.

5.2. Indoor Pose Estimation

MESA is also evaluated using the ScanNet1500 benchmark for indoor pose estimation. The benchmark is challenging due to the poor texture and large view variances in scenes.

Experimental setup. The parameters of MESA are the same as our area matching experiment in Sec. 5.1. We adopt the A2PM framework [52] and combine four fine-tuned point matchers with our method, *i.e.* LoFTR [40], ASpan [8], QuadT [43] and DKM [17], to demonstrate our performance and compare with recent methods [6, 19, 21]. The cropped area images are resized to be square with the same side length as the long side of the default size of point

Pose estimation AUC	ScanNet1500 benchmark		
	AUC@5° ↑	AUC@10° ↑	AUC@20° ↑
SP [14]+SGMNet [7] ICCV'21	15.4	32.1	48.3
SiLK [19] ICCV'23	18.0	34.4	50.4
PATS [21] CVPR'23	26.0	46.9	64.3
CasMTR [6] ICCV'23	27.1	47.0	64.4
ASpan [8] ECCV'22	25.6	46.0	63.3
SEEM [54]+SGAM_ASpan [52] arxiv'23	$27.5_{+7.42\%}$	$48.0_{+4.35\%}$	$65.3_{+3.16\%}$
MESA_ASpan	$27.5_{+7.42\%}$	$48.2_{+4.35\%}$	$65.0_{+2.69\%}$
QuadT [43] ICLR'22	24.9	44.7	61.8
SEEM [54]+SGAM_QuadT [52] arxiv'23	$25.5_{+2.53\%}$	$46.0_{+2.95\%}$	$63.4_{+2.59\%}$
MESA_QuadT	$28.7_{+15.26\%}$	$46.5_{+4.03\%}$	$61.9_{+0.16\%}$
LoFTR [40] CVPR'21	22.1	40.8	57.6
SEEM [54]+SGAM_LoFTR [52] arxiv'23	$23.4_{+5.88\%}$	$41.8_{+2.45\%}$	$58.7_{+1.91\%}$
MESA_LoFTR	$22.9_{+3.62\%}$	$41.8_{+2.45\%}$	$58.4_{+1.39\%}$
DKM [17] CVPR'23	29.4	50.7	68.3
SEEM [54]+SGAM_DKM [52] arxiv'23	$\underline{30.6}_{+4.12\%}$	$\underline{52.3}_{+3.10\%}$	$\underline{69.3}_{+1.48\%}$
MESA_DKM	$\mathbf{33.4}_{+13.61\%}$	$\mathbf{55.0}_{+8.48\%}$	$\mathbf{72.0}_{+5.42\%}$

Table 2. **Relative pose estimation results (%) on ScanNet1500 benchmark.** The **best** and second results are highlighted.

matchers. Following [40], the pose estimation AUC at various thresholds are used as metrics. More implementation details can be found in our Supp. Mat.

Results. As we can see in Tab. 2, MESA significantly increases the pose accuracy for all baselines. Due to the presence of more severe matching redundancy in indoor scenes, our approach gains more noticeable enhancement, *e.g.* +15.26% in AUC@5° for QuadT, compared to outdoor scenes in Tab. 3. The improvement obtained by MESA for semi-dense matchers are comparable to those of SGAM [52], because the coarse-to-fine pipeline of these methods makes them insensitive to AOR. However, MESA brings a larger improvement for DKM [17] than SGAM (13.61% *vs.* 4.12%) and boosts the precision to a new *state-of-the-art*, proving the precise area matching by MESA benefits the performance of dense matching.

5.3. Outdoor Pose Estimation

We use the Megadepth1500 [27] which consists of 1500 outdoor image pairs to evaluate the effectiveness of our method for relative pose estimation in outdoor scenes.

Experimental setup. The parameter $T_{E_{max}}$ is set as 0.3. MESA is combined with four baselines to compare with recent methods [18, 29, 48]. We also compare with another covisible-area matching method, OETR [9], combined with our post-processing. The area images are cropped and resized as input of point matcher, same as the indoor experiment. The pose estimation AUC is adopted as the metric.

Results. In Tab. 3, MESA achieves impressive precision improvement for semi-dense baselines, up to +7.58% in AUC@5°. The improvement for the dense method is not as notable as that for semi-dense methods, because lots of repetitiveness in MegaDepth leads to hard area matching, decreasing the accuracy of DKM. However, MESA_DKM does set a new *state-of-the-art* for outdoor pose estimation. MESA also achieves better results than OETR by find-

Pose estimation AUC	MegaDepth1500 benchmark		
	AUC@5° ↑	AUC@10° ↑	AUC@20° ↑
LightGlue [29] ICCV'23	49.9	67.0	80.1
SiLK [19] ICCV'23	43.8	57.7	68.6
ASTR [48] CVPR'23	58.4	73.1	83.8
TopicFM [18] AAAI'23	54.1	70.1	81.6
CasMTR [6] ICCV'23	59.1	74.3	84.8
ASpan [8] ECCV'22	55.4	71.6	83.1
OETR [9]+ASpan AAAI'22	$56.2_{+1.44\%}$	$72.3_{+0.98\%}$	$82.8_{-0.36\%}$
MESA_ASpan	$58.4_{+5.42\%}$	$74.1_{+3.49\%}$	$84.8_{+2.05\%}$
QuadT [43] ICLR'22	54.6	70.5	82.2
OETR+QuadT AAAI'22	$55.4_{+1.47\%}$	$71.2_{+0.99\%}$	$82.1_{-0.01\%}$
MESA_QuadT	$57.9_{+6.04\%}$	$73.1_{+3.69\%}$	$83.9_{+2.07\%}$
LoFTR [40] CVPR'21	52.8	69.2	81.2
OETR+LoFTR AAAI'22	$54.8_{+3.79\%}$	$70.3_{+1.59\%}$	$81.8_{+0.73\%}$
MESA_LoFTR	$56.8_{+7.58\%}$	$72.2_{+4.34\%}$	$83.3_{+2.59\%}$
DKM [17] CVPR'23	60.4	74.9	$\underline{85.1}$
OETR+DKM AAAI'22	$\underline{60.6}_{+0.33\%}$	$\underline{75.1}_{+0.27\%}$	$84.9_{-0.24\%}$
MESA_DKM	$\mathbf{61.1}_{+1.16\%}$	$\mathbf{75.6}_{+0.93\%}$	$\mathbf{85.6}_{+0.59\%}$

Table 3. **Relative pose estimation results (%) on MegaDepth-1500 benchmark.** The **best** and second results are highlighted.

ing more detailed area matches, whereas OETR solely estimates covisible areas.

5.4. Visual Odometry

We also evaluate our method on the visual odometry using the KITTI360 [28] dataset, which densely estimates the camera motion in the self-driving scene. We select four sequences with few moving objects from the total dataset.

Experimental setup. The parameter $T_{E_{max}}$ is set as 0.25. The area image size is 480×480. Our method is combined with four baselines [8, 17, 40, 43] to demonstrate its effectiveness. All baselines are trained on ScanNet [11]. Following [25], we report the relative pose errors (RPE), including the rotational error (R_{err}) and translation error (t_{err}).

Results Tab. 4 shows that our method achieves prominent precision improvement for DKM [17], ASpan [8] and LoFTR [40] in all sequences. For QuadT [43], its performances increases in most cases with the aid of MESA. Two cases of accuracy declines may caused by its quadtree attention mechanism [43], which evenly splits the area image leading to destruction of semantic integrity, especially in the driving scene with strong symmetry. Our method combined with DKM achieves the best performance, proving the benefits of MESA for the dense matching method.

5.5. Ablation Study

To evaluate the effectiveness of our design, we conduct a comprehensive ablation study for components of MESA.

Area Graph Construction. To justify the AG of MESA, we adopt a naive approach to match areas, which is comparing area similarity densely (CSD). In particular, we first select areas with proper size from all SAM areas of two im-

Method	Seq. 00		Seq. 02		Seq. 05		Seq. 06	
	$R_{err}\downarrow$	$t_{err}\downarrow$	$R_{err}\downarrow$	$t_{err}\downarrow$	$R_{err}\downarrow$	$t_{err}\downarrow$	$R_{err}\downarrow$	$t_{err}\downarrow$
DKM [17]	0.046	0.569	0.102	0.683	0.048	0.623	0.050	0.533
MESA_DKM	**0.039**	**0.499**	**0.051**	**0.611**	**0.041**	**0.546**	**0.044**	**0.488**
ASpan [8]	0.152	2.412	0.149	2.358	0.154	2.523	0.146	1.887
MESA_ASpan	**0.130**	**2.079**	**0.141**	**2.327**	**0.131**	**2.227**	**0.132**	**1.671**
QuadT [43]	0.145	2.307	0.144	2.322	0.135	2.269	0.137	1.762
MESA_QuadT	**0.134**	**2.305**	**0.140**	2.420	**0.123**	**2.131**	**0.135**	1.915
LoFTR [40]	0.132	2.023	0.124	1.913	0.117	1.930	0.126	1.572
MESA_LoFTR	**0.116**	**1.703**	**0.114**	**1.658**	**0.109**	**1.654**	**0.120**	**1.493**

Table 4. **Visual odometry results on KITTI360.** The original results of baselines and better or worse results achieved by our method are highlighted. All baselines are trained on ScanNet [11].

Method	AOR ↑	AMP@0.6 ↑	PoseAUC@5 ↑	AreaNum ↑
MESA_ASpan (**Ours**)	72.75	89.09	27.50	3.47
w/ *CSD*	69.23	84.21	26.78	2.73
w/ *DesSim.* [21]	63.71	62.91	26.05	2.65
w/ *SEEMSeg.* [54]	70.58	85.52	26.18	1.67
w/ $\arg\min E_{self}$	70.98	87.56	26.96	2.90

Table 5. **Ablation study.** Three variants of MESA_ASpan are evaluated for area matching and pose estimation on the ScanNet1500 to demonstrate the importance of various components.

ages. The similarity of each area to all areas in the other images is then calculated and area matches with the greatest similarity is obtained. The comparison results are summarised in Tab. 5. As AG can generate more proper areas for matching, MESA w/ CSD gets less area matches. Thus, the area and point matching performance is also decreased by CSD. Moreover, CSD results in a significant increase in time of area matching ($\sim \times 10$ slower than MESA), due to its inefficient dense comparison.

Area Similarity Calculation. In contrast to our classification formulation for area similarity calculation, another straightforward method [21] involves calculating the distance between learning descriptors of areas. Thus, we replace our learning similarity with descriptor similarity in [21] (*DesSim*) and conduct experiments in ScanNet to investigate the impact. The results are summarized in Tab. 5, including the area number per image, area matching and pose estimation performance. Overall, the performance of *DesSim* experiences a noticeable decline, due to poor area matching precision, indicating the effectiveness and importance of proposed learning similarity calculation.

Image Segmentation Source. We relay on SAM to achieve areas with implicit semantic, whose outstanding segmentation precision and versatility contribute to our leading matching performance. However, areas can also be obtained from other segmentation methods. Therefore, to measure the impact of different segmentation sources, we exchange the segmentation input from SAM [22] with that from SEEM [54] (*SEEMSeg.*) and evaluate the performances. In Tab. 5, MESA with *SEEMSeg.* gets a slight precision decline and fewer areas compared with SAM, leading to decreased pose estimation results. These results in-
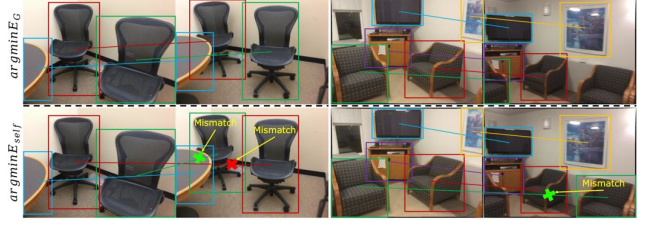

Figure 5. **The qualitative comparison of Global Energy Refinement.** As AG structures of both images are considered by E_G, objects with the same apparent can be distinguished according to their neighbors, which are mismatched by $\arg\min E_{self}$, revealing the robustness of $\arg\min E_G$ under repetitive patterns.

dicates that the advanced segmentation favors our methods. Notably, MESA with *SEEMSeg.* also achieves slight improvement for ASpan, proving the effectiveness of MESA.

Global Energy Refinement. After *Graph Cut*, the proposed global matching energy for the final area matching refinement considers structures of both AGs of the input image pair. To show the importance of this dual-consideration, we replace the global energy with naive E_{self} in Eq. (12) ($\arg\min E_{self}$) and evaluate the performance. In Tab. 5, the refinement relying on E_{self} produces decreased area matching precision and a subsequent decline in pose estimation performance, due to inaccurate area matches especially under repetitiveness. The qualitative results shown in Fig. 5 further indicate the better robustness of global energy under repetitiveness due to dual graph structure capture.

6. Conclusion

In this paper, we propose MESA as an approach for accurate, robust, and practical area matching, aiming at effective feature matching redundancy reduction. To this end, MESA first utilizes the high-level image understanding capability of SAM, an advanced foundation model for image segmentation, to obtain informative image areas. Then, we propose a novel Area Graph (AG) to model the spatial structure and scale hierarchy of image areas. By formulating the area matching on graphical models converted by AG, we adopt the *Graph Cut* algorithm and a proposed learning area similarity model to establish area matches. We further consider the AG structures for both images to introduce the global matching energy refinement, which enables robust and exact area matching, promoting precise point matching. In extensive experiments, MESA achieves impressive area matching accuracy and significantly improves feature matching performance for various point matching methods.

References

[1] Ivana Balazevic, Carl Allen, and Timothy Hospedales. Multi-relational poincaré graph embeddings. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[2] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key. net: Keypoint detection by hand-crafted and learned cnn filters. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5836–5844, 2019. 2

[3] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006. 5

[4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001. 5

[5] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 2021. 1

[6] Chenjie Cao and Yanwei Fu. Improving transformer-based image matching by cascaded capturing spatially informative keypoints. *ICCV*, 2023. 6, 7, 3

[7] Hongkai Chen, Zixin Luo, Jiahui Zhang, Lei Zhou, Xuyang Bai, Zeyu Hu, Chiew-Lan Tai, and Long Quan. Learning to match features with seeded graph matching network. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6301–6310, 2021. 7

[8] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. *ECCV*, pages 20–36, 2022. 1, 2, 5, 6, 7, 8, 3

[9] Ying Chen and et al. Guide local feature matching by overlap estimation. *AAAI*, 36(1):365–373, 2022. 1, 3, 7

[10] Peter Clifford. Markov random fields in statistics. *Disorder in physical systems: A volume in honour of John M. Hammersley*, pages 19–32, 1990. 4

[11] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 6, 7, 8, 1, 3, 4

[12] Kun Dai, Tao Xie, Ke Wang, Zhiqiang Jiang, Ruifeng Li, and Lijun Zhao. Oamatcher: An overlapping areas-based network for accurate local feature matching. *arXiv preprint arXiv:2302.05846*, 2023. 3

[13] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n' pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[14] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 1, 2, 7

[15] Ying Chen Dihe Huang, Jianlin Liu Yong Liu, Wenlong Wu Shang Xu, Fan Tang Yikang Ding, and Chengjie Wang. Adaptive assignment for geometry aware local feature matching. *CVPR*, 2023. 1

[16] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 8092–8101, 2019. 2

[17] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. DKM: Dense kernelized feature matching for geometry estimation. *CVPR*, 2023. 1, 2, 6, 7, 8, 3, 4, 5

[18] Khang Truong Giang, Soohwan Song, and Sungho Jo. Topicfm: Robust and interpretable topic-assisted feature matching. *AAAI*, 37(2):2447–2455, 2023. 1, 3, 7

[19] Pierre Gleize, Weiyao Wang, and Matt Feiszli. Silk–simple learned keypoints. *ICCV*, 2023. 6, 7

[20] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6207–6217, 2021. 2, 3, 4, 7

[21] Yijin Li Junjie Ni, Hujun Bao Zhaoyang Huang, Hongsheng Li, and Guofeng Zhang Zhaopeng Cui. Pats: Patch area transportation with subdivision for local feature matching. *CVPR*, 2023. 5, 6, 7, 8, 4

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *ICCV*, 2023. 1, 2, 3, 8, 5

[23] Vladimir Kolmogorov and Ramin Zabin. What energy functions can be minimized via graph cuts? *IEEE transactions on pattern analysis and machine intelligence*, 26(2):147–159, 2004. 2, 5

[24] Laura Leal-Taixé, Cristian Canton-Ferrer, and Konrad Schindler. Learning by tracking: Siamese cnn for robust target association. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 33–40, 2016. 5

[25] Shenghao Li, Qunfei Zhao, and Zeyang Xia. Sparse-to-local-dense matching for geometry-guided correspondence estimation. *IEEE Transactions on Image Processing*, 32:3536–3551, 2023. 7

[26] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. *Advances in Neural Information Processing Systems*, 33:17346–17357, 2020. 2

[27] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 7, 1

[28] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 7

[29] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. *ICCV*, 2023. 7

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019. 2

[31] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[32] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 2

[33] Jiayi Ma, Xingyu Jiang, Aoxiang Fan, Junjun Jiang, and Junchi Yan. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129(1):23–79, 2021. 1

[34] Víctor Martínez, Fernando Berzal, and Juan-Carlos Cubero. A survey of link prediction in complex networks. *ACM Comput. Surv.*, 49(4), 2016. 4

[35] Iaroslav Melekhov, Aleksei Tiulpin, Torsten Sattler, Marc Pollefeys, Esa Rahtu, and Juho Kannala. Dgc-net: Dense geometric correspondence network. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1034–1042, 2019. 2

[36] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 32, 2019. 2

[37] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[38] Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, et al. Back to the feature: Learning robust camera localization from pixels to pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3247–3257, 2021. 1

[39] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1

[40] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 1, 2, 5, 6, 7, 8, 3, 4

[41] MA Syakur, BK Khotimah, EMS Rochman, and Budi Dwi Satoto. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP conference series: materials science and engineering*, page 012017. IOP Publishing, 2018. 1

[42] Dongli Tan, Jiang-Jiang Liu, Xingyu Chen, Chao Chen, Ruixin Zhang, Yunhang Shen, Shouhong Ding, and Rongrong Ji. Eco-tr: Efficient correspondences finding via coarse-to-fine refinement. *European Conference on Computer Vision*, pages 317–334, 2022. 2

[43] Shitao Tang, Jiahui Zhang, Siyu Zhu, and Ping Tan. Quadtree attention for vision transformers. *ICLR*, 2022. 2, 6, 7, 8

[44] Prune Truong, Martin Danelljan, and Radu Timofte. Glunet: Global-local universal network for dense flow and correspondences. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6258–6268, 2020. 2

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[46] Fa-Yueh Wu. The potts model. *Reviews of modern physics*, 54(1):235, 1982. 5

[47] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 516–520, 2016. 5

[48] Jiahuan Yu, Jiahao Chang, Jianfeng He, Tianzhu Zhang, Jiyang Yu, and Wu Feng. ASTR: Adaptive spot-guided transformer for consistent local feature matching. *CVPR*, 2023. 7

[49] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 2

[50] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. 2

[51] Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning two-view correspondences and geometry using order-aware network. *ICCV*, 2019. 2

[52] Yesheng Zhang, Xu Zhao, and Dahong Qian. Searching from area to point: A hierarchical framework for semantic-geometric combined feature matching. *arXiv*, 2023. 1, 2, 3, 6, 7, 5

[53] Xiaoming Zhao, Xingming Wu, Jinyu Miao, Weihai Chen, Peter CY Chen, and Zhengguo Li. Alike: Accurate and lightweight keypoint detection and descriptor extraction. *IEEE Transactions on Multimedia*, 2022. 2

[54] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in neural information processing systems*, 2023. 6, 7, 8, 5