

# View-graph Selection Framework for SfM

Rajvi Shah, Visesh Chari, and P J Narayanan

CVIT, Kohli Center for Intelligent Systems, IIIT Hyderabad, India

**Abstract.** View-graph selection is a crucial step for accurate and efficient large-scale structure from motion (sfm). Most sfm methods remove undesirable images and pairs using several fixed heuristic criteria, and propose tailor-made solutions to achieve specific reconstruction objectives such as efficiency, accuracy, or disambiguation. In contrast to these disparate solutions, we propose an optimization based formulation that can be used to achieve these different reconstruction objectives with task-specific cost modeling and construct a very efficient network-flow based formulation for its approximate solution. The abstraction brought on by this selection mechanism separates the challenges specific to datasets and reconstruction objectives from the standard sfm pipeline and improves its generalization. This paper mainly focuses on application of this framework with standard sfm pipeline for accurate and ghost-free reconstructions of highly ambiguous datasets. To model selection costs for this task, we introduce new disambiguation priors based on local geometry. We further demonstrate versatility of the method by using it for the general objective of accurate and efficient reconstruction of large-scale Internet datasets using costs based on well-known sfm priors.

**Keywords:** View-graph · Structure from Motion · Disambiguation

## 1 Introduction

View-graph is a crucial input structure for large-scale structure from motion (sfm). Nodes in this graph represent images (also called cameras/views) and edges represent relative motion or epipolar geometries (EGs) between the nodes. View-graphs help in ‘organizing’ unordered image collections useful to, (i) select a core set of images for reconstruction, and (ii) identify noisy EGs that might degrade the quality of reconstruction. State-of-the-art sfm methods like incremental [22, 32, 17], hierarchical [3, 7, 28, 18], or global [21, 2, 14, 13, 1], all rely on a view-graph based initial pruning step for efficient and accurate reconstruction.

Large community photo collections often display point-of-view bias, as some viewing angles are more popular than others. Using the full view-graph for sfm is often computationally expensive and unnecessary due to high redundancy. Also, for closely clustered images, narrow baselines increase the uncertainty of triangulation, causing large re-projection errors. While isolated erroneous EGs can be overcome by robust averaging (global sfm) and repeated bundle adjustment (incremental sfm), large number of incorrect EGs can degrade reconstruction. Hence, it is crucial to select ‘good’ images and more importantly ‘good’ pairs for

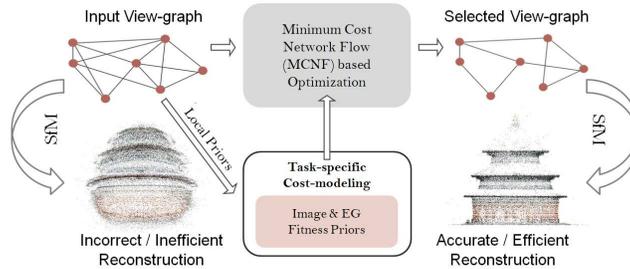


Fig. 1: Outline of the proposed selection framework. With appropriately modeled costs, the framework can select view-graphs that meet desired reconstruction objectives.

accuracy. Standard SfM methods apply fixed heuristics such as inliers, baselines, homography, loop closure for conservative selection/pruning of images and pairs. While these heuristics work well for a large variety of datasets, they are insufficient to identify and remove consistent noise that arise in ambiguous scenes.

Ambiguity in pairwise matching arises because man-made structures often comprise of repetitions (windows, arches), symmetries (circular structures, similar facades), and large duplicate elements (minarets, domes). While standard SfM pipelines are robust to handle a large number of inaccurate EGs in isolation, for scenes with high ambiguity, such wrong EGs form consistent sets, resulting in mis-registered cameras, ‘phantom’ structures, or structures incorrectly folded along symmetry. Previous works propose tailor-made solutions to handle such scenes with local/global steps for ‘reasoning’ ambiguity [34, 16, 11, 29, 10, 33]. While these methods show good results on challenging datasets, they do not operate within the framework of standard SfM pipelines. We believe that these challenges can be addressed under the same umbrella of ‘selecting’ a complete, consistent and noise-free view-graph for accurate and efficient reconstruction.

The main motivation of our work is to formalize the core problem of view-graph selection to meet different objectives within a unified framework. We pose this as an optimization problem with image and pair selection costs and propose a novel network-flow based approximation for its efficient solution. This abstraction allows different objectives to be achieved by plugging in task-specific costs while keeping the overall selection and reconstruction framework the same. Figure 1 shows an outline of this framework. We mainly focus on the application of this framework for accurate and ghost-free reconstructions of highly ambiguous datasets and show how pairwise selection cost can be modeled using local priors, and also introduce a new context-based prior. Additionally, we also show its usefulness for accurate and efficient reconstruction of general large-scale Internet landmarks datasets with costs modeled using commonly used local heuristics.

To the best of our knowledge, this is the first attempt at systematizing view-graph selection. The proposed framework brings greater flexibility and generalization to standard SfM pipelines and its application is not limited to the specified use-cases. A unified framework also lands optimal view-graph selection as a learning problem if and when task-specific ground-truth data becomes available.

## 2 Related Work

In sfM literature, view-graph selection is not dealt with in a unified sense, instead, specific methods are designed to meet specific selection objectives. For efficiency and completeness, incremental/hierarchical sfM methods compute spanning sub-graphs [24, 15, 8, 28], with different criteria and selection strategies. For global methods, optimality of a subgraph is often related to consistency and robustness goals [13, 35, 5, 27, 14, 31]. Our framework is not a one-size-fits-all replacement to specialized methods, but provides a mechanism to express different objectives using task-specific image and pairwise selection costs. Here, we revisit prior works related to the general objective of accuracy and the specific objective of disambiguation, and discuss their relevance to our work.

*Implicit view-graph filtering in SfM methods* Pruning undesirable images/pairs from input view-graph is implicit to both incremental and global sfM methods and is often done using thresholds on various criteria. Accuracy of incremental sfM hinges on seed pairwise reconstruction and next best view selection. Wide baseline seed pair selection is ensured using criteria based on epipolar inliers (that don't fit a homography) [22, 23], or inlier ratios (to detect pairwise motion as planar, rotational, or general) [17]. For next best view selection, triangulation angle, inliers, or correspondence distribution are used as criteria [22, 17].

Global sfM methods first use relative rotations to estimate global rotations followed by global translations estimation [21, 2, 13, 1, 30, 27]. In [2], to reduce the state space of camera parameters for MRF based estimation, relative twist and unusual aspect ratios are used as view filtering criteria. Methods that use lie-algebraic averaging of relative rotations [4, 5, 35, 14, 13, 1, 27] often discard images with unknown calibration and handle outlier EGs with loop consistency checks [35, 14, 13] and robust cost functions.

Many of these heuristics are easy to incorporate as image and pairwise selection costs into our framework. Other heuristics that employ global reasoning (such as loop consistency) can be potentially remodeled as a pairwise heuristic.

*SfM methods for Disambiguation* Initial methods for disambiguation focused on inferring missing correspondences [34] and using it as a prior with global objectives of camera pose inference [16] or consistent view-graph expansion [11]. Other methods consist of social network principles based track refinement [29], geometry-aware feature matching [19], triplet-consistent graph expansion [20], and geodesic manifold based ambiguity detection and correction [33]. A post-reconstruction approach to disambiguation uses back-projected 3D points to identify conflicting observations [9, 10]. In contrast to these specialized algorithms, we use our selection framework with costs modeled as a combination of disambiguation priors. We also propose a new context based pairwise selection prior that is based on both, missing correspondences and conflicting observations and unlike [9], it can be computed pre-reconstruction. Our solution is more general and very efficient as compared to prior methods and still recovers correct structures for a variety of ambiguous datasets.

### 3 Posing view-graph selection as optimization

Let the input view-graph be  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where the set of vertices (nodes)  $\mathcal{V}$  represents the images and the set of edges  $\mathcal{E}$  represents the pairwise epipolar geometries (EGs). The goal is to select a subset of nodes  $\mathcal{V}'$  and a subset of edges  $\mathcal{E}'$  (a subgraph  $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ ) that meets a desired objective. We denote the indicator variables for image selection as  $\delta_i$  and pair selection as  $\delta_{ij}$ . Here, each  $\delta_i, \delta_{ij}$  correspond to the image vertex  $v_i \in \mathcal{V}$  and the pairwise edge  $e_{ij} \in \mathcal{E}$ . Intuitively, this problem can be represented as a minimization of the form,

$$\arg \min f(\delta) = \sum_i q_i \delta_i + \sum_i \sum_j q_{ij} \delta_{ij} \quad (1)$$

$$\text{subject to } \sum \delta_i \leq N, \sum \delta_{ij} \leq M \quad (2)$$

$$\delta_{ij} \leq \delta_i, \delta_{ij} \leq \delta_j \quad (3)$$

$$\text{where } \forall i \in \mathcal{V}, \forall (i, j) \in \mathcal{E}, \delta_i, \delta_{ij} \in \{0, 1\}, \\ N < |\mathcal{V}|, M < |\mathcal{E}|$$

Activation of indicator variables  $\delta_i$  and  $\delta_{ij}$  imply selection of corresponding image  $v_i$  and pair  $e_{ij}$ .  $q_i$  is the cost of selecting the view  $v_i$  and  $q_{ij}$  is the cost of selecting the edge  $e_{ij}$  and these costs assumed to be negative in the minimization sense.  $N$  and  $M$  indicate thresholds on maximum number of selected images and pairs and ensure that the subgraph selection is non-trivial (eqn. 2). Selection of image pair  $(i, j)$  requires that both images  $i$  and  $j$  constituting the pair also must be selected (eqn. 3).

Modeling these costs appropriately can express different objectives for view-graph selection. In this paper, we discuss cost modeling for accurate and ghost-free reconstructions using a few relevant priors. However, this can be used to achieve other objectives using a variety of other priors known in the literature or even learned costs.

This minimization problem can be formulated as a binary integer linear program (BILP), which is NP-complete. To find an exact optimal solution, stock solvers for BILP use branch-and-bound like techniques to intelligently iterate through all possible solutions. However, owing to the NP-complete nature of the problem, it is infeasible to compute the exact optimal solution for many problems of our interest. In fact, in our experiments, the computation time for a branch-and-bound based ILP solver was very high even for toy-sized problems. A standard trick to achieve efficient solutions with some approximation is to use a linear programming (LP) relaxation with rounding to obtain an integer solution. However, in preliminary experiments, the solutions obtained using this approach seemed too inaccurate to be useful. Furthermore, this formulation does not take into account connectivity of the selected subgraph. To tackle both these problems, we pose view-graph selection as a minimum cost network-flow (MCNF) problem (see supplementary material for more discussion) that guarantees us a binary solution in polynomial time and encourages connectivity in the selected sub-graph. The proposed MCNF formulation is explained in the next section.

## 4 View-graph Selection as MCMF Problem

The network in MCMF problems is a directed graph with at least one source and one sink nodes. We denote the network as  $\mathcal{T} = (\mathcal{N}, \mathcal{A})$ , where  $\mathcal{N}$  represents the set of nodes and  $\mathcal{A}$  represents the set of directed edges/arcs. Each edge  $(i, j)$  in the network has a cost  $c_{ij}$  associated to let across one unit of flow, and the cost incurred by an edge is proportional to the flow  $(x_{ij})$  through it. Each edge also has a lower and an upper bound  $(l_{ij}, u_{ij})$  on the amount of flow  $(x_{ij})$  that can pass through it, known as capacity constraints (**c.c**). The source sends a certain units of flow that the sink node must receive. At all other nodes, flow must be conserved. Let us denote the total flow as  $F$  and the remainder flow at a node  $i$  as  $b_i$ , then  $b_i = F$  when  $i$  is source,  $b_i = -F$  when  $i$  is sink, and  $b_i = 0$  otherwise. These constraints are known as equal flow constraints (**e.f.c**). The MCMF problem is about sending the total flow from source to sink at a minimum cost, without violating the capacity and flow constraints. This minimization with flow and capacity constraints can be described as,

$$\text{minimize } \sum_{(i,j) \in \mathcal{A}} c_{ij} x_{ij} \quad (5)$$

$$\text{subject to, } \sum_{\{j:(i,j) \in \mathcal{A}\}} x_{ij} - \sum_{\{j:(j,i) \in \mathcal{A}\}} x_{ji} = b_i \quad \forall i \in \mathcal{N} \quad \text{e.f.c} \quad (6)$$

$$l_{ij} \leq x_{ij} \leq u_{ij} \quad \forall (i, j) \in \mathcal{A} \quad \text{c.c} \quad (7)$$

*Network construction* To pose view-graph selection as an MCMF problem, we construct the network as follows. All indicator variables corresponding to image/pair selection  $\{\delta_i\}$ ,  $\{\delta_{ij}\}$  are represented using arcs  $(i, j) \in \mathcal{A}$  in the network, source and sink nodes are auxiliary. Since view selection variables are represented as arcs, each vertex  $i$  in the view-graph corresponds to two nodes,  $(2i - 1, 2i)$  in the network. Each odd node  $(2i - 1)$  corresponding to the vertex  $i$  in the view-graph is connected to the source node and similarly each even node  $(2i)$  is connected to the sink node. The arcs corresponding to the pairwise selection variables  $\{\delta_{ij}\}$ , join the even node of the lower index image with the odd node of the higher index image. This choice prevents cycle formation in the network. Summarizing, the network has  $|\mathcal{N}| = 2|\mathcal{V}| + 2$  nodes and  $|\mathcal{A}| = |3\mathcal{V}| + |\mathcal{E}|$  arcs. These arc connections are summarized in Figure 2 along with a pictorial example.

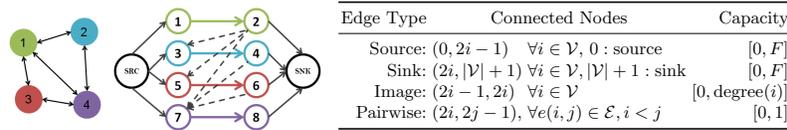


Fig. 2: Network construction for a sample view-graph is shown on the left. Image nodes and selection arcs in are color-coded to match vertices in the view-graph. Pairwise selection arcs are depicted by black dashed lines. Arc connections and their capacities are described in the table shown on the right.

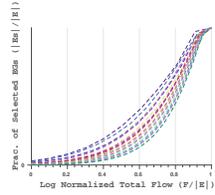


Fig. 3: Effect of total flow on edge selection : selected edges increase monotonically with flow.

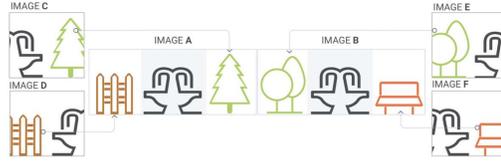


Fig. 4: Image pair (A,B) capture a scene with duplicate elements. Though duplicate elements in the images yield many matches, elements in context regions tend to find matches with non-intersecting sets of images.

*Equal flow and capacity constraints* At source and sink the sent and received flow is equal to the total flow. Equal flow constraints require that in and out flow at every other node remain equal. In our formulation, capacity constraints - lower and upper bounds on flow through an arc are specified based on the edge type as mentioned in Figure 2.

To understand the choice of these capacities, consider the vertex  $v_1$  in the depicted view-graph. The arc corresponding to  $v_1$ 's selection variable  $\delta_1$  in the network is  $a(1, 2)$ . Corresponding to  $v_1$ 's degree in the view-graph, the node  $n_2$  in the network has three outgoing arcs  $a(2, 3)$ ,  $a(2, 5)$ , and  $a(2, 7)$  for pairwise selection variables  $\delta_{12}$ ,  $\delta_{13}$ , and  $\delta_{14}$ . The flow starting from source node, after passing through  $a(1, 2)$ , should plausibly be able to pass through all three outgoing arcs. Since flow can only be divided in integer units, the minimum capacity of image selection arc  $a(1, 2)$  has to be at least 3 ( $deg(v_i)$ ). Under minimum cost solution, the flow at any node will continue to take the path of least resistance (cost). Now, suppose that the cost assigned to  $a(2, 3)$  is the least amongst the three outgoing arcs and its max. capacity is 3 units (or any value  $1 < k < deg(v_i)$ ). In this scenario,  $a(2, 3)$  being the lowest cost arcs will pull all 3 units of flow from  $a(1, 2)$ , starving the other arcs of any flow and preventing the corresponding view-graph edges from ever getting selected. To avoid this, we restrict the maximum flow through pairwise selection arcs to 1.

*Effect of cost normalization and flow on solution* In minimization sense, negative costs provide encouragement for flow to pass through an arc, whereas positive costs provide discouragement. Suppose, all costs are negative, then total flow  $F$  of 1 unit will select the lowest cost chain in the network (often the longest). As we increase the value of total flow, more paths get explored and when  $F = |\mathcal{E}|$  all images and EGs get selected. When costs are both encouraging and discouraging, many positive cost arcs will act as barriers for the flow. As a result, at some value of total flow  $F$  the selection will (nearly) saturate and may never select the full view-graph. For the proposed applications, we use only encouraging costs ( $-1 \leq c_{ij} \leq 0$ ), while flow remains the only free parameter.

*Tuning flow parameter* To systematically understand the effect of total flow on image and pair selection, we created synthetic view-graphs with varying number of vertices, varying degrees of connectivity simulated using Gaussian assumptions on neighborhood, and randomly assigned (negative) costs (discussed in detail in the suppl. material). We performed sub-graph selection on  $\sim 100$  such synthetic view-graphs using the proposed MCNF approach for increasing values of total flow and observed a consistent pattern in selection behavior. Figure 3 shows the plot for fraction of total edges selected vs. normalized total flow. It can be seen that for negative costs, the relation between total flow and selected vertices and edges is logarithmic in nature. This relation is also observed on real-world (Internet) datasets with well-connected view-graphs (see supplementary material). Depending on the fraction of total vertices and edges we want to be selected, the logarithmic dependence allows us to find the desired sub-graph by a binary search of the flow parameter over the  $[0, 1]$  interval with a logarithmic scale. Such an iterative search is particularly effective in view of the extremely low computational time of the MCNF algorithm. In fact, the combined processing time of the the whole search procedure is still an insignificant fraction of the total time required for SfM, and is thus far from being the computational bottleneck.

*Running time* A crucial advantage of this formulation is that it can be solved very efficiently. Constructing the network and solving for MCNF takes less than a second even for graphs with  $\sim 1000$  nodes and  $\sim 100K$  edges.

## 5 Applications and Cost Modeling

To show that the proposed framework can address different concerns using task-specific costs, we tackle two use-cases. We first discuss cost modeling for reconstruction of highly ambiguous datasets using local disambiguation priors. Later, we discuss cost modeling for the general use-case of accurate and efficient reconstruction of large-scale Internet datasets using common SfM priors.

### 5.1 Cost modeling for ambiguous datasets

We propose that even for highly ambiguous datasets, with conservative selection of an input view-graph such that it consists of a higher fraction of ‘true’ EGs, correct reconstruction can be recovered without any change in the reconstruction pipeline. To achieve this, we use three pairwise measures that act as strong priors for disambiguation and express pair selection cost as a linear combination of these three priors (with uniform weights). These priors are based on local geometry and can be easily computed at the time of initial view-graph construction without significant overhead. Note that  $c^k(e_{ij})$  denotes the contribution of prior  $k$  to the total cost  $c$  for the pair  $e_{ij}$ , and  $g$  denotes a normalization function that distributes raw prior value to the desired cost range.

*Context Similarity* Missing correspondences (matches in a pair, not matching the third image in a triplet) are a useful prior for disambiguation. However, by itself, this prior is not very effective for disambiguation and it is commonly used in an inference framework with other priors or with a global objective [34, 16, 11]. Track covisibility statistics [29] are also useful for disambiguation, but it is not straightforward to apply this prior to describe pairwise fitness. In a post-reconstruction disambiguation approach [10], reconstructed 3D points are back-projected into image pairs and conflicting observations in unmatched regions (context) are identified. However, this measure is also not directly useful as our approach is a pre-process to SfM reconstruction. Motivated by these priors, we propose a new, context based pairwise prior that is suitable to our framework.

For image pair  $(I_i, I_j)$ , the sets of all matched features (matched with any image) are  $S_i, S_j$  and the sets of features that match between  $(I_i, I_j)$  are  $M_i, M_j$ . The difference sets  $U_i = S_i \setminus M_i, U_j = S_j \setminus M_j$  consist of unique features in  $I_i$  and  $I_j$ . Suppose images  $I_i$  and  $I_j$  are looking at a scene with duplicate instances of a structure. The features in the match sets  $M_i, M_j$  will most likely lie on the duplicate elements. We consider the unique feature sets  $U_i, U_j$  to belong to the context regions. If two images are truly looking at the same instance, the context features of both images would have matched similar set of images. On the other hand, if two images are looking at duplicate instances, the context features would be distributed over different sets of images (see Figure 4). We find distribution of features in  $U_i$  and  $U_j$  over all  $N$  images in the collection based on their matches and make an  $N$  dimensional description of the context space. Context feature of image  $I_i$  w.r.t. image  $I_j$  can be described as,

$$\mathbf{w}_{ij} = [w_i^1, w_i^2, \dots, w_i^N] \quad w_i^j = 0 \quad (8)$$

$$w_i^k = |\{u \in U_i \mid u \leftrightarrow u', u' \in S_k, k \neq j\}| \quad (9)$$

This measure is slightly biased against pairs with very low visual overlap, however, combined with the other two measures, it works effectively for selection. We compared the context features of a pair using cosine similarity or hamming distance (after binarization) and found them to be working similarly. The context similarity can be defined as,  $c^k(e_{ij}) = g(\mathbf{w}_{ij}^T \mathbf{w}_{ji})$ .

*Loop consistency* Loop consistency suggests that rotations in an EG triplet when chained should yield identity [6]. We find all triplets in the view-graph and label them as consistent or inconsistent. For each EG, we count the total number of consistent EGs it participates in and use this as a measure of its fitness. Though it is not same as explicitly enforcing loop consistency, it is a convenient way to incorporate a non-local (pairwise) geometric cue as a pairwise prior.  $c^k(e_{ij}) = g\left(\frac{\# \text{ consistent triplets on } (i,j)}{\text{median}(\# \text{ consistent triplets on any } (i,j))}\right)$ .

*Multiple motions* We model this prior to disambiguate image pairs capturing repetitive or duplicate instances of some scene elements. We remove the correspondences that satisfy the estimated EG from the initial set of matches and

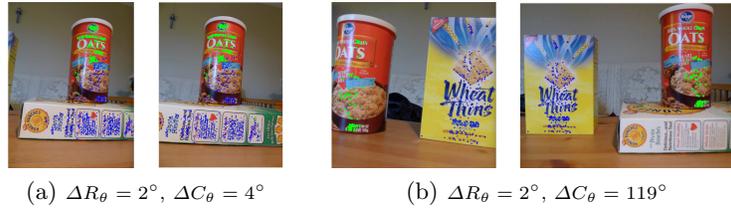


Fig. 5: Multiple motion detection : for correctly matched pair (a) both motions are in agreement, while for pair (b) matches due to ambiguity cause large position difference.

estimate the secondary relative pose using the leftover matches. If sufficient inliers are found, we decompose the relative pose into rotation and translation. We estimate the angular difference between primary and secondary rotations ( $\Delta R_\theta$ ) and also the angle between the both position vectors ( $\arccos(\Delta C_{ij}^T \Delta C_{ij})$ ). If the difference angles are small, the secondary motion is most possibly arising due to threshold sensitivity and measurement drift, otherwise, these indicate presence of correspondences on ambiguous structures. Examples of these scenarios are shown in Figure 5.  $c^k(e_{ij}) = g(\Delta R_\theta, \arccos(\Delta C_{ij}^T \Delta C_{ij}))$ . When secondary motion is detected, the value of this prior is very effective in disambiguating (except for the instances when camera is purely translating along the direction of repetition). However, this prior alone is insufficient for disambiguation, as for many incorrect pairs secondary geometry is not detected. We combine this prior with the other two priors only for the pairs where secondary geometry is detected.

## 5.2 Cost modeling for general datasets

View-graph selection for general datasets is typically done to achieve complete reconstructions with smaller re-projection errors and shorter run-time. For these goals, we use simple priors based on graph-connectivity and local geometry that express common-knowledge selection heuristics and criteria of general SfM methods in form of image and pairwise selection costs, suitable for our framework.

*Image selection priors* For the purposes of accurate and complete reconstruction, we consider three image priors, (i) degree of an image node in the view-graph, (ii) fraction of an image’s features that participates in tracks, (iii) local clustering coefficient of an image node in the view-graph. First two measures favor selecting images with many observations and connections to support longer tracks useful for accurate triangulation. Since Internet photo collections often suffer from point of view bias, using only connectivity based priors could lead to selections within popular components. To compensate for this effect, we include local clustering coefficient (lcc) of a vertex as an image prior to prefer images that provide connections across components than within components. Image selection cost is a linear combination of these priors with uniform weights.

*Pairwise selection priors* Fitness of a pair in traditional sfm pipelines is measured using two common criteria, baseline and number of EG inliers. We use four pairwise priors, (i) number of inliers, (ii) median triangulation angle of pairwise reconstructed features, (iii) overlap (area of a convex hull of the matched features), and (iv) infinite homography, that are reflective of this selection criteria. While the usefulness of first three priors is evident, detecting infinite homography is useful in discouraging pairs with panoramic motion (rotation around a fixed center). Despite a high overlap, such pairs are undesirable as they lead to degenerate or ill-conditioned EG. Simply using homography inliers also rejects valid EGs due to planar regions. We use the fact that calibration normalized infinite homographies are basically rotations [6]. Hence,  $\mathbf{H}^T\mathbf{H}$  should be close to identity for such pairs and we model this prior as,  $\|\mathbf{H}_{ij}^T\mathbf{H}_{ij} - \mathbf{I}\|_F$ .

A more detailed expression of the priors discussed in this section is provided in supplementary material along with other implementation details such as choice of thresholds and normalization ( $g$ ) for different priors.

## 6 Results and Discussion

We apply our view-graph (henceforth mentioned as VG for brevity) selection framework for the two use-cases discussed and show results on a variety of datasets. Implementation details are provided in the supplementary material along with additional analysis and results. Code can be found at, <https://cvit.iit.ac.in/research/projects/cvit-projects/viewgraphselection>.

### 6.1 Ambiguous datasets reconstruction

We show successful reconstruction results on 12 highly ambiguous datasets consisting of small-scale lab-style scenes [16], and large-scale urban scenes [10, 20] with the standard incremental sfm pipeline and VGs selected using our approach. Details of these datasets and selection statistics are given in Table 1a.

*Qualitative comparison* Figure 7 shows reconstruction results for small-scale ambiguous scenes [16]. With selected VGs, we are able to recover true structures for all datasets. Figure 8 shows reconstruction results for large-scale urban ambiguous scenes [10, 20]. Our method is able to recover comparable splits to the method of [10] and successful reconstruction for TOH dataset. Our result for ANC dataset is incomplete as compared to [10, 33] (shown in supplementary material).

*Ablation of priors* To highlight the effectiveness of disambiguation specific priors, we also reconstruct these datasets for VGs selected using random costs (fails on all datasets), baseline priors based costs (fails on all but the ‘books’ dataset), and combination of all costs based on priors for general reconstruction and disambiguation (fails on ‘cup’ and ‘cereal’ datasets). Among the disambiguation priors, context similarity based prior is the most effective standalone. To further

Sr.	Dataset	$ \mathcal{V} $	$ \mathcal{E} $	F	$ \mathcal{V}_s $	$ \mathcal{E}_s $	$t_{sel}$		1	2	3	4	5	6	
1	Cereal	25	228	4	25	49	0.0137	Other Methods	Roberts et al. [16]*	×	×	×	×	×	×
2	Cup	64	1217	16	64	265	0.0268		Jiang et al. [11]	✓	×	✓	✓	✓	✓
3	Oats	24	220	4	24	49	0.0137		Wilson and Snavely [29]	×	×	×	✓	×	×
4	Street	19	95	2	19	26	0.0089		Heinly et al. [10]	✓	-	-	-	-	✓
5	Books	21	161	4	21	41	0.0135		Shen et al. [20]	-	✓	✓	✓	-	×
6	Desk	31	261	4	31	64	0.0138		Yan et al. [33]	-	✓	-	-	✓	-
7	ANC	448	5037	512	416	2497	0.1288	Ours with diff. priors	Baseline	×	×	×	×	✓	×
8	ADT	381	3627	512	340	2148	0.0940		General+Disamb.	×	×	✓	✓	✓	✓
9	RDC	271	3378	128	258	1498	0.0717		Context	✓	✓	✓	✓	✓	✓
10	BG	161	2003	128	129	848	0.0522		Loop	×	✓	✓	✓	✓	✓
11	CSB	277	5191	128	233	1380	0.0743		MM	×	×	×	×	✓	✓
12	ToH	341	50332	32	341	1990	0.2372		Context+Loop	✓	✓	✓	✓	✓	✓
								Context+MM	✓	✓	✓	✓	✓	✓	
								Loop+MM	✓	✓	✓	✓	✓	✓	
								Context+Loop+MM	✓	✓	✓	✓	✓	✓	

(a) Datasets and selection details

(b) Comparison and Ablation

Table 1: (a) shows details of ambiguous datasets and selection statistics.  $t_{sel}$  shows the combined running time for all search iterations of MCNF solver. Abbreviated labels correspond to : ANC – Alexander Nevsky Cathedral, ADT – Arc de Triomphe, BG – Brandenburg Gate, CSB – Church on Spilled Blood, RDC – Radcliff Camera, TOH – Temple of Heaven; (b) top rows show results reported by other methods on small-scale ambiguous datasets (1 to 6) where – implies results not reported, \*[16] succeeds on these sets only with time-stamp info. Bottom rows show results of our selection method with various prior based costs for given flow.

evaluate the effectiveness of the disambiguation priors and their combinations, we study the range of flow values for which the given prior based selection results in successful reconstruction (see Figure 6). We observed Context+Loop prior combination to be the most robust in this ablation study but empirically observed benefits of using multiple motions based prior on larger datasets. Qualitative results for the ablation study are provided in supplementary material.

*Runtime comparison* Method proposed in [10] operates post-reconstruction to split incorrectly merged model parts and takes  $\sim 16$  to  $\sim 85$  minutes to process for these datasets. Our framework pre-selects the VG and reconstruction is performed without any additional processing. Recently proposed method of [33] also tackles the disambiguation problem as a pre-process to SfM but their method takes 2–11 minutes on these datasets. Our subgraph selection framework is extremely efficient taking 1–2 seconds for constructing and solving the MCNF problem for these and even larger datasets. Moreover, our framework is intended to be general purpose with disambiguation as one of the specific objectives.

*Criteria for flow parameter search* As can be seen in Table 1a, for the lab-style datasets, the full VG is generally quite dense, with an average vertex degree (edges per vertex)  $\sim 32\%$  of total vertices. However, due to the very high ambiguity in the scenes, large number of pairs are expected to be outliers. Therefore, it makes



Fig. 6: Range of flow values ( $y$ -axis on log scale) for which given prior combinations ( $x$ -axis) lead to correct reconstructions. Priors that lead to correct reconstruction for higher values of flow are better at disambiguation and more robust, since this behavior implies selection of smaller fraction of outlier EGs. Since multiple motions are detected for a smaller fraction of pairs. For most datasets, context prior performs better than loop prior. Combination of all priors performs equivalently for practical purposes.

intuitive sense to expect that a pruned VG comprising a reasonably small fraction ( $\sim 25\%$ ) of the total number edges would lead to a good reconstruction. On the other hand, the VGs for urban ambiguous scenes [10] are not well-connected, as these datasets are already sampled subgraphs (iconics) of the original image set. This is also reflected from the fact that average number of edges per vertex is only  $\sim 5\%$  of total vertices. In view of this sparse connectivity, it makes sense to keep as many vertices as possible and a bigger fraction of edges. Practically, we chose these threshold to be,  $|\mathcal{V}_s| \geq 80\%$  of  $|\mathcal{V}|$ , and  $|\mathcal{E}_s| \geq 5|\mathcal{V}|$  and use the method described in section 4 to get the desired sub-graph by efficiently searching for the appropriate flow value.

## 6.2 General datasets reconstruction

We show that the proposed framework is versatile and can also be used for the general goals of accuracy and efficiency for reconstruction of standard sfM datasets, by modeling image and pair selection costs with well-known and commonly used sfM priors. For this task, we show the results on two datasets, (*i*) MVS benchmark [25], and (*ii*) Internet landmarks [22, 12]. These datasets are reconstructed with incremental [32] and global sfM [26] pipelines, using both, full VGs and VGs selected by our method.

MVS benchmark consists of three toy-sized datasets with ground-truth (GT) camera positions. Table 2a shows that, for both sfM methods, the selected VGs based reconstructions are comparable to the full VGs based reconstructions. Flow parameter for these selections was chosen such that all vertices are selected (for GT comparisons). For large-scale Internet landmarks datasets, we reconstruct the scenes using selected VGs and full VGs with global sfM pipeline (typically slightly less robust than incremental sfM methods) in order to compare the reconstruction accuracy w.r.t. incremental sfM based baseline reconstructions

Dataset	VG	$\mathcal{V}$	$\mathcal{E}$	Incremental SfM			Global SfM		
				$R_{err}$	$T_{err}$	$r_{err}$	$R_{err}$	$T_{err}$	$r_{err}$
Castle	S	30	49	2.44	0.15	0.38	2.21	1.29	1.01
	F	118	118	2.22	0.22	0.34	2.17	7.49	1.14
Fountain	S	11	21	2.90	0.01	0.29	2.82	0.29	0.35
	F	25	25	2.90	0.01	0.72	2.82	0.27	0.59
Herzjesu	S	55	55	2.36	0.03	0.50	2.38	0.75	1.18
	F	25	128	2.38	0.02	0.43	2.39	0.56	1.71

(a) MVS dataset statistics

Dataset	VG	$\mathcal{V}$	$\mathcal{E}$	$t_{sel}$	$N_c$	$r_{err}$	$R_{err}$	$T_{err}$	$t_{sfm}$
Notre Dame	S	659	16970	1.744	628	1.41	0.072	0.195	1151
	F	714	46746	-	682	1.53	0.089	0.217	1760
Pantheon	S	761	15975	3.721	754	1.06	0.098	0.310	1785
	F	781	139630	-	775	1.31	0.125	0.309	3601
St. Peters	S	1132	39640	2.864	1095	1.341	0.037	0.517	1147
	F	1155	119977	-	1111	1.458	0.028	0.496	1367

(b) Internet landmarks datasets statistics

Table 2: Selection and reconstruction statistics for general datasets. Labels ‘S’ and ‘F’ show selected and full VGs,  $|\mathcal{V}|$ ,  $|\mathcal{E}|$  – vertices and edges,  $N_c$  – #reconstructed cameras,  $R_{err}$ ,  $T_{err}$ , and  $r_{err}$  – median rotation, translation, and reprojection errors,  $t_{sel}$  and  $t_{sfm}$  – runtime for VG selection (cumulative) and SfM reconstruction.

(in absence of ground-truth). Table 2b shows the selection and reconstruction statistics for these datasets. It can be seen that the reconstructions with selected VGs are comparable or more accurate as compared to those with full VGs and SfM run-time with selected VGs is notably shorter. For completeness of the recovered structure, it is desirable to have as many vertices as possible in the subgraph. For efficiency, it is desirable to select fewer edges, however too few edges (low vertex degree) can lead to many short feature tracks and high reprojection errors. With these considerations, we keep  $|\mathcal{V}_s| = 90\%$  of  $|\mathcal{V}|$  and  $|\mathcal{E}_s| = 20|\mathcal{V}|$  as flow search criteria. The selected VG reconstructions are qualitatively similar or better than full VG reconstructions (shown in supplementary material).

## 7 Conclusions and Future Work

We presented a novel and efficient, unified framework for selecting subgraphs from initial view-graphs that can achieve different selection objectives with appropriately modeled image and pairwise selection costs. This mechanism provides an interesting way to separate dataset and task specific challenges from the standard SfM pipeline, thereby improving its generality. We demonstrated utility and potential of this framework by achieving satisfactory results for two objectives. One interesting way to achieve even further abstraction to this problem would be to replace hand-designed costs by a weighted combination ( $c_i = \sum_k \alpha_k f_k(i)$ ) of a number of known and designed priors. Cost formulation of this form would be expressive enough to cater to a wide variety of selection objectives. The problem of modeling costs to meet the desired objective then translates to that of devising new priors to add to the combination and finding the right weights for prior combination. In future, we wish to explore this direction for extending our framework. While this is non-trivial, it can lead to interesting directions of research for searching/learning new priors and the combination weights.

*Acknowledgement* We thank Google India PhD Fellowship and India Digital Heritage Project of Department of Science and Technology, Govt. of India for funding this work.

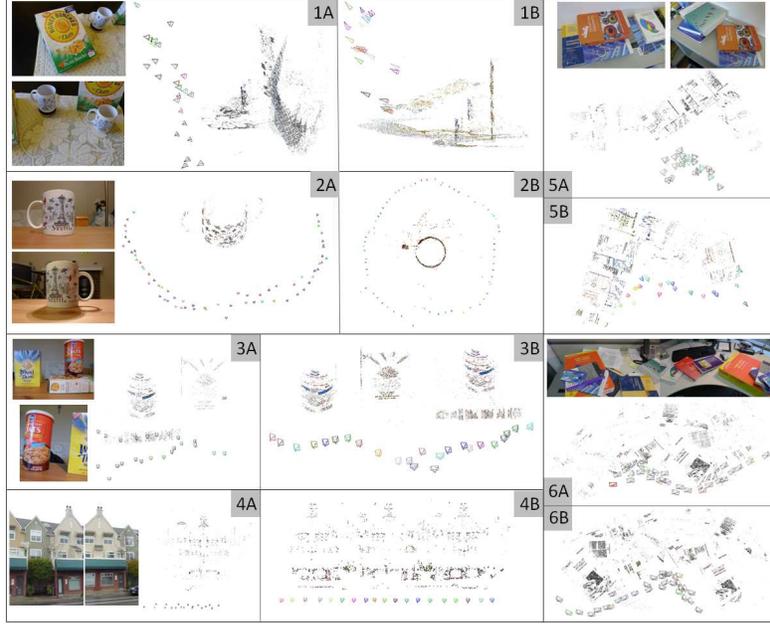


Fig. 7: Reconstructions for small ambiguous datasets (numbered as per Table 1a): (A) indicate full VG based reconstructions, (B) indicate selected VG based reconstructions.

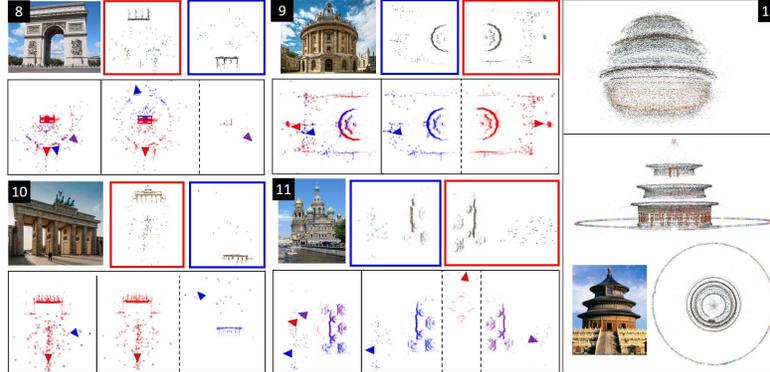


Fig. 8: Comparison of our reconstruction results on large ambiguous datasets (numbered as per Table 1a). For 8 to 11, bottom left – incorrect model with full VG, bottom right – result of correctly split models using the post-reconstruction pipeline of [10], and top row – our result (color-coded to match the splits of [10]). For 12, top – full VG result, bottom – result with our selection.

## Bibliography

- [1] A. Chatterjee and V. M. Govindu. Efficient and robust large-scale rotation averaging. In *Proceedings IEEE ICCV*, 2013. 1, 3
- [2] D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. In *Proceedings IEEE CVPR*, 2011. 1, 3
- [3] R. Gherardi, M. Farenzena, and A. Fusiello. Improving the efficiency of hierarchical structure-and-motion. In *Proceedings IEEE CVPR*, 2010. 1
- [4] V. M. Govindu. Lie-algebraic averaging for globally consistent motion estimation. In *Proceedings IEEE CVPR*, 2004. 3
- [5] V. M. Govindu. Robustness in motion averaging. In *Proceedings ACCV*, 2006. 3
- [6] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003. 8, 10
- [7] M. Havlena, A. Torii, J. Knopp, and T. Pajdla. Randomized structure from motion based on atomic 3d models from camera triplets. In *Proceedings IEEE CVPR*, 2009. 1
- [8] M. Havlena, A. Torii, and T. Pajdla. Efficient structure from motion by graph optimization. In *Proceedings ECCV*. 2010. 3
- [9] J. Heinly, E. Dunn, and J.-M. Frahm. Recovering Correct Reconstructions from Indistinguishable Geometry. In *Proceedings 3D Vision (3DV)*, 2014. 3
- [10] J. Heinly, E. Dunn, and J.-M. Frahm. Correcting for Duplicate Scene Structure in Sparse 3D Reconstruction. In *Proceedings ECCV*, 2014. 2, 3, 8, 10, 11, 12, 14
- [11] N. Jiang, P. Tan, and L. F. Cheong. Seeing double without confusion: Structure-from-motion in highly ambiguous scenes. In *Proceedings IEEE CVPR*, 2012. 2, 3, 8, 11
- [12] Y. Li, N. Snavely, and D. P. Huttenlocher. Location recognition using prioritized feature matching. In *Proceedings ECCV*, 2010. 12
- [13] P. Moulon, P. Monasse, and R. Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *Proceedings IEEE ICCV*, 2013. 1, 3
- [14] C. Olsson and O. Enqvist. Stable structure from motion for unordered image collections. In *Proceedings of the Scandinavian conference on Image analysis (SCIA)*, 2011. 1, 3
- [15] R. Raguram, C. Wu, J.-M. Frahm, and S. Lazebnik. Modeling and recognition of landmark image collections using iconic scene graphs. *International Journal of Computer Vision*, 95(3), 2011. 3
- [16] R. Roberts, S. Sinha, R. Szeliski, D. Steedly, and R. Szeliski. Structure from motion for scenes with large duplicate structures. In *Proceedings IEEE CVPR*, 2011. 2, 3, 8, 10, 11

- [17] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Proceedings IEEE CVPR*, 2016. 1, 3
- [18] R. Shah, A. Deshpande, and P. J. Narayanan. Multistage sfm: Revisiting incremental structure from motion. In *Proceedings 3D Vision (3DV)*, 2014. 1
- [19] R. Shah, V. Srivastava, and P. J. Narayanan. Geometry-aware feature matching for structure from motion applications. In *Proceedings IEEE Winter Conference on Applications of Computer Vision*, 2015. 3
- [20] T. Shen, S. Zhu, T. Fang, R. Zhang, and L. Quan. Graph-based consistent matching for structure-from-motion. In *Proceedings ECCV*, 2016. 3, 10, 11
- [21] S. Sinha, D. Steedly, and R. Szeliski. A multi-stage linear approach to structure from motion. In *Proceedings ECCV RMLE Workshop*. 2010. 1, 3
- [22] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3d. *ACM Trans. Graph.*, 25(3), 2006. 1, 3, 12
- [23] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *Int. J. Comput. Vision*, 80(2), 2008. 3
- [24] N. Snavely, S. M. Seitz, and R. Szeliski. Skeletal graphs for efficient structure from motion. In *Proceedings IEEE CVPR*, 2008. 3
- [25] C. Strecha, W. von Hansen, L. J. V. Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Proceedings IEEE CVPR*, 2008. 12
- [26] C. Sweeney. *Theia Multiview Geometry Library: Tutorial & Reference*. University of California Santa Barbara., 2015. 12
- [27] C. Sweeney, T. Sattler, M. Turk, T. Hollerer, and M. Pollefeys. Optimizing the viewing graph for structure-from-motion. In *Proceedings IEEE ICCV*, 2015. 3
- [28] R. Toldo, R. Gherardi, M. Farenzena, and A. Fusiello. Hierarchical structure-and-motion recovery from uncalibrated images. *Computer Vision and Image Understanding*, 140, 2015. 1, 3
- [29] K. Wilson and N. Snavely. Network principles for sfm: Disambiguating repeated structures with local context. In *Proceedings IEEE ICCV*, 2013. 2, 3, 8, 11
- [30] K. Wilson and N. Snavely. Robust global translations with 1dsfm. In *Proceedings ECCV*, 2014. 3
- [31] K. Wilson, D. Bindel, and N. Snavely. When is rotations averaging hard? In *Proceedings ECCV*, 2016. 3
- [32] C. Wu. Towards linear-time incremental structure from motion. In *Proceedings 3D Vision (3DV)*, 2013. 1, 12
- [33] Q. Yan, L. Yang, L. Zhang, and C. Xiao. Distinguishing the indistinguishable: Exploring structural ambiguities via geodesic context. In *Proceedings IEEE CVPR*, July 2017. 2, 3, 10, 11
- [34] C. Zach, A. Irschara, and H. Bischof. What can missing correspondences tell us about 3d structure and motion? In *Proceedings IEEE CVPR*, 2008. 2, 3, 8
- [35] C. Zach, M. Klopschitz, and M. Pollefeys. Disambiguating visual relations using loop constraints. In *Proceedings IEEE CVPR*, 2010. 3