# SAG-GNN: Semantic-Aware Guided GNN for Descriptor-Free 2D-3D Matching

Supplementary Material

## A. Additional Details

**Datasets.** MegaDepth [23] is a popular outdoor dataset containing images from 196 scenes worldwide, with sparse 3D reconstructions generated using COLMAP [35]. This dataset offers a rich set of 2D images, corresponding 3D point clouds, and valuable pose information, making it a comprehensive resource for image matching and 2D-3D matching research. We selected 25,624 images from 99 scenes as the training set and 12,399 images from 49 scenes as the test set, using SIFT [25] as the keypoint detector. Cambridge-Landmarks [17] is a mid-sized outdoor dataset consisting of 6 urban scenes. The 3D models are generated by COLMAP, and pose data is also provided. We used the SuperPoint [10] detector for keypoint detection and evaluated our model on four of these scenes. 7Scenes [37] is a compact indoor dataset consisting of RGBD images and corresponding camera poses, which are provided by a depth SLAM system. This dataset is designed for evaluating localization accuracy in indoor environments. By projecting SIFT keypoints from RGB images onto depth maps, we can extract depth information corresponding to the 2D keypoints, which helps in constructing 3D models. We tested our model on the standard test sequences of the seven scenes provided in the dataset to gauge its performance in indoor localization tasks.

**Pipeline.** For 2D-3D matching, we first select the top-k database images with the highest co-visibility for each query image. The 3D points to be matched are obtained through the correspondence between these co-visible images and the 3D model, while the 2D points are extracted from the query image using the keypoint detector. To improve localization accuracy, the localization process typically matches 3D point clouds from multiple reference images [31]. For the MegaDepth dataset, we used top-1 for training and reported matching and localization results for top-1 and top-10; for the generalization tests on the Cambridge-Landmarks and 7Scenes datasets, we used top-10 matching and reported the localization results accordingly. During training, we limit the maximum number of keypoints per image to 1024 and manually control the inlier ratio to 0.5, ensuring the number of keypoints input to the matcher remains between 100 and 1024. For the construction of GT (Ground Truth) match pairs, we project the 3D points onto the query image based on the known pose and filter matches based on the bearing vector (BV) distance between the 3D projection points and the query image keypoints, with a threshold of 0.001. During the testing phase, we no longer manually control the inlier ratio and ensure the number of keypoints remains between 10 and 1024.

## B. Visualization of Ablation

We removed two key modules proposed in the paper: the Bidirectionally-Aligned Fusion Block (Fusion) and the Semantic-Guided Interaction (Guided), and visualized the resulting 2D-3D matching outcomes in Fig. 6. The results clearly show that after removing these carefully designed semantic modules, the model's semantic perception capability significantly deteriorates, with dynamic objects such as pedestrians and carriages in the query image leading to incorrect matches with the point cloud, resulting in reduced localization accuracy. In contrast, our SAG-GNN effectively avoids these errors, maintaining high-quality matches and better localization accuracy.

## C. Ablation of Semantic-Guided Interaction

Our method originally calculates the pure semantic similarity (Sem.) and applies it as a weight to the attention scores in the Semantic-Guided Interaction module. We then replaced the semantic features with pure geometric features (Geo.) and geometric-semantic concatenated features (Geo.+Sem.), and recalculated the weighting based on these new feature representations. This change allows us to assess the impact of using different feature types (pure geometric, geometric-semantic, and pure semantic) for guidance. As shown in Tab. 5, the performance with pure semantic feature guidance yields the best results.

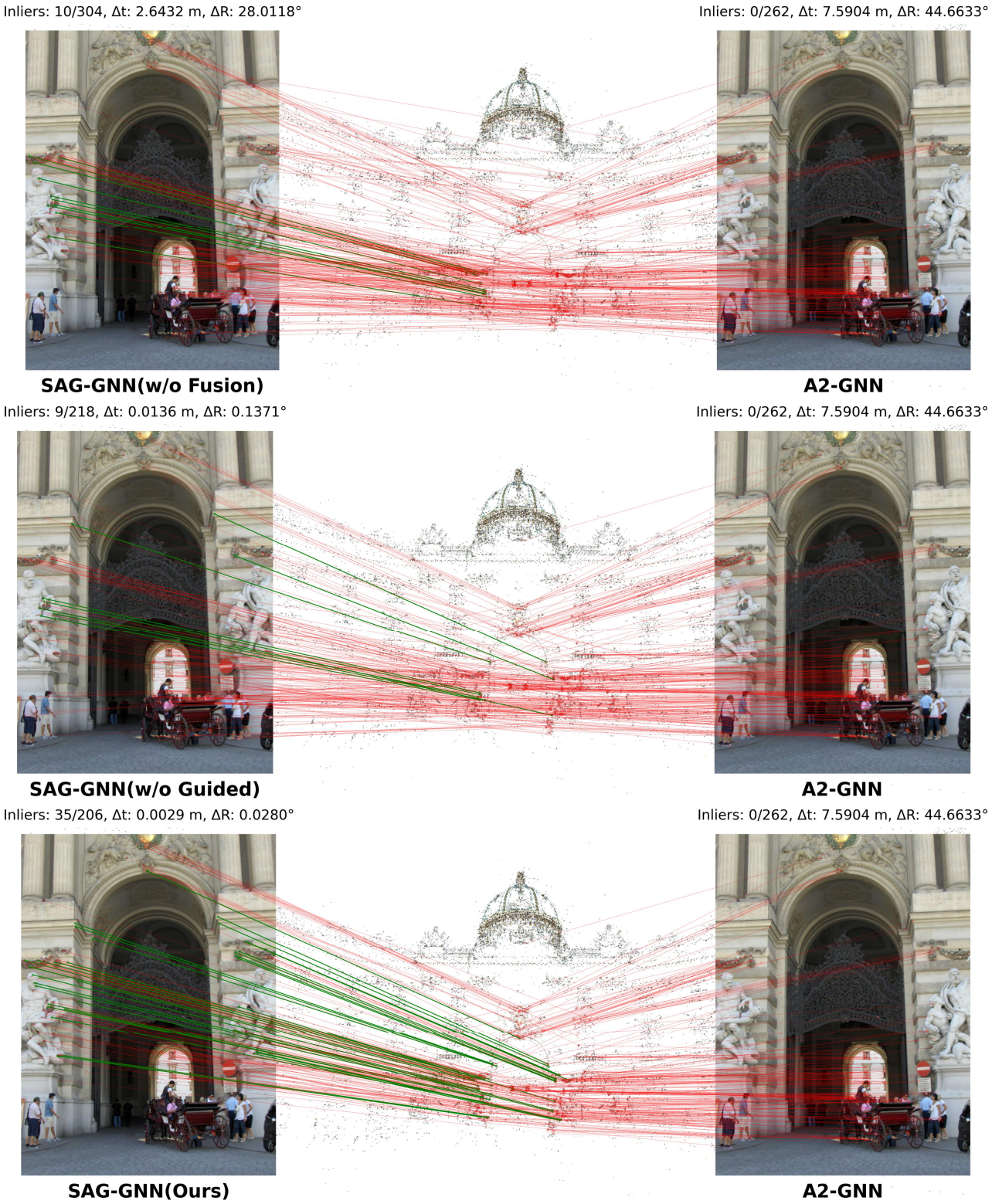


Figure 6. Comparison of 2D-3D matching results of our ablation models.

Table 5. Ablations of our Semantic-Guided Interaction Block.

| Methods | OR Input | Semantic | | Reproj. AUC (%) | Rotation (°) | Translation (m) |
|---|---|---|---|---|---|---|
| | | Fusion | Guided | @1 / 5 / 10px (↑) | Quantile@25 / 50 / 75% (↓) | |
| Baseline | BV | ✗ | ✗ | 12.72 / 41.84 / 48.02 | 0.12 / 0.79 / 26.37 | **0.01** / 0.08 / 2.80 |
| Variants | Feat. | Bidirectional | Geo. | 13.54 / 45.93 / 52.80 | 0.10 / 0.43 / 19.24 | **0.01** / 0.04 / 2.09 |
| | Feat. | Bidirectional | Geo.+Sem. | 14.65 / 48.77 / 55.76 | 0.09 / 0.33 / 16.04 | **0.01** / 0.03 / 1.79 |
| **SAG-GNN** | Feat. | Bidirectional | Sem. | **16.35 / 53.16 / 60.56** | **0.08 / 0.25 / 9.11** | **0.01 / 0.02 / 0.94** |