Detector-Free Feature Matching for Visual SLAM: Evaluation of LoFTR and ELFOTR

Ahmet Ajam
Istanbul Technical University
BBL 514E Spring 2025
Student ID: 518231020

TEAM

Ahmet Ajam - 518231020
Responsibilities: Conducting literature research, selecting and preparing datasets, implementing evaluation scripts, running experiments, and writing the final report.

ABSTRACT

This project investigates detector-free feature matching for Visual SLAM systems in challnging conditions, such as motion blur, occlusion, and low-texture scenes. Traditional pipelines like ORB-SLAM often depend on sparse descriptors, which limit their adaptability. This study evaluates two transformer-based detector-free matching models LoFTR and ELFOTR that have recently shows success in end-to-end correspondence prediction without explicit keypoints. Their performance is benchmarked on MegaDepth1500 and ScanNet1500 datasets using standard metrics such as AUC and precision. Our experiments show that ELFOTR maintains high accuracy while significantly improving efficiency, making it suitable for real-time SLAM. code and a small presentation is accessible at https://github.com/ajamiscoding/loftr-elfotr-project-bbl514e

I. INTRODUCTION

Visual SLAM aims to localize a moving agent and map the environment simultaneously. In dynamic environments or those with low texture, sparse keypoint-based methods like ORB, SIFT, or even SuperPoint [1] may fail to maintain stable correspondences. This can degrade both pose estimation and map quality.

Hypothesis: We hypothesize that detector-free matching models using transformers can outperform traditional and sparse learning-based methods by leveraging global context and computing pixel-level correspondences.

Literature Overview:

- **SuperPoint** [1]: A self-supervised model for detecting and describing sparse interest points in an image. It trains a neural network to jointly learn interest point detection and descriptors without labeled data. SuperPoint is often used as the front-end for traditional SLAM or structure-from-motion pipelines.
- **SuperGlue** [2]: A graph neural network-based matcher that improves the accuracy of sparse correspondences. It takes detected keypoints and their descriptors (e.g., from SuperPoint) and performs attention-based matching. Its strength lies in using contextual information across keypoints to resolve ambiguities and improve matching robustness.
- **LoFTR** [3]: Introduced as a detector-free, detector-free correspondence model. It avoids the explicit keypoint detection stage entirely, and instead computes matches using a hierarchical transformer architecture. LoFTR is especially effective in textureless areas or repetitive patterns, which are problematic for traditional keypoint-based methods.
- **ELFOTR** [4]: A recent improvement over LoFTR, ELFOTR focuses on efficiency and reduced inference time. It introduces a lightweight transformer encoder and decoder to maintain detector-free matching capability while making the model faster and more suitable for real-time or embedded applications.

II. MODELS AND DATASET OVERVIEW

In that work we decided to use the detector-free matching models. **LoFTR** uses a two-stage architecture involving coarse and fine transformers to predict pixel-level matches directly between image pairs. It removes the need for keypoint detection, enabling performance in textureless or repetitive environments [3].

ELFOTR improves upon LoFTR by utilizing lightweight transformer modules, leading to faster inference without losing much precision. It is more suitable for deployment on embedded or mobile SLAM systems [4].

Dataset Summary:
MegaDepth is a Structure-from-Motion (SfM) based dataset that provides internet photos of outdoor scenes, including monuments and buildings. It includes estimated depth maps, camera poses, and sparse matching ground truth. It is widely used in evaluating outdoor visual matching models.
ScanNet is an RGB-D dataset captured using depth sensors in indoor environments such as rooms with furniture. It provides depth maps, camera intrinsics and poses, and surface normals. It is commonly used for evaluating indoor SLAM and 3D reconstruction tasks.

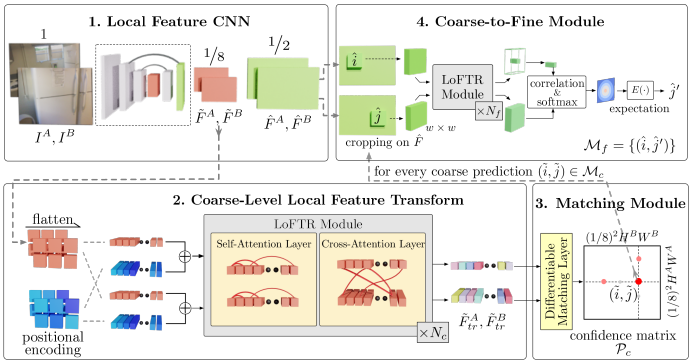

Fig. 1: LoFTR architecture pipeline showing correspondence prediction. Source: [3]

TABLE I: MegaDepth vs ScanNet: Dataset Characteristics

Feature	MegaDepth	ScanNet
Type	Structure-from-Motion dataset from online photos	RGB-D dataset captured using depth sensors
Scene Type	Outdoor: large monuments and buildings	Indoor: furnished rooms, everyday scenes
Data Modalities	RGB, estimated depth, camera poses	RGB, depth, surface normals, intrinsics and poses
Ground Truth	Sparse depth, camera poses	Accurate depth, trajectory, intrinsics
Applications	Evaluating LoFTR, SuperGlue, JamMa (outdoor)	Indoor SLAM evaluation (ELFOTR, SuperGlue)

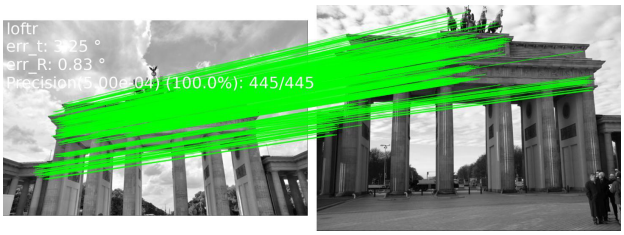

Fig. 2: Sample matching result using LoFTR on MegaDepth. Green lines indicate successful correspondences.

III. Methodology and Evaluation Metrics

The evaluation was conducted using publicly available pre-trained weights for LoFTR [3] and ELFOTR [4]. The same evaluation pipeline was used for both models to ensure fair comparison. Datasets were preprocessed to extract 1500 image pairs from MegaDepth and ScanNet each.

We used the following performance metrics:

- **AUC@5, AUC@10, AUC@20:** Area under the curve for estimated homographies within angular thresholds.
- **Match Count:** Number of predicted correspondences.
- **Precision@5**$\times 10^{-4}$: *Proportion* of accurate correspondences within a tight reprojection error.

IV. Results

This section presents the quantitative results obtained from evaluating LoFTR and ELFOTR on both the MegaDepth and ScanNet datasets. The models are assessed based on matching quality and homography estimation accuracy. Each method was tested in both indoor (ScanNet)
and outdoor (MegaDepth) settings, with key metrics such as AUC at
various thresholds
, number of valid correspondences, and high-precision matching reported below.

A. LoFTR Performance

TABLE II: LoFTR Evaluation Results

Dataset	AUC@5	AUC@10	AUC@20	Matches	Precision
ScanNet	0.1687	0.3363	0.5064	1456.24	0.6267
MegaDepth	0.5193	0.6921	0.8143	3458.37	0.9354

B. ELFOTR Performance

TABLE III: ELFOTR Evaluation Results

Config	Dataset	AUC@5	AUC@10	AUC@20	Matches	Precision
Full	ScanNet	0.1965	0.3718	0.5369	1335.17	0.6505
Opt	ScanNet	0.1684	0.3384	0.5073	1505.35	0.5996
Opt	MegaDepth	0.5586	0.7186	0.8327	3288.18	0.9315
Full	MegaDepth	0.5638	0.7218	0.8348	3288.18	0.9687

V. Discussion and Conclusion

The results show that both LoFTR and ELFOTR are robust matching models, particularly in outdoor scenes (MegaDepth). While LoFTR has a slightly higher match count, ELFOTR achieves higher AUC and precision in the optimized configuration. In indoor scenes (ScanNet), performance is lower overall, but the models remain usable.

ELFOTR stands out due to its reduced inference time, making it highly applicable to real-time visual SLAM pipelines where computational resources are limited. LoFTR, on the other hand, may be preferred in offline settings where matching quality is prioritized. Integrating these models into full SLAM systems like ORB-SLAM3 would be a promising direction for future work.

REFERENCES

[1] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," CVPRW 2018.

[2] P.-E. Sarlin et al., "SuperGlue: Learning feature matching with graph neural networks," CVPR 2020.

[3] J. Sun et al., "LoFTR: Detector-free local feature matching with transformers," CVPR 2021.

[4] Y. Wang et al., "Efficient LoFTR (ELFOTR): Semi-dense matching with transformer backbones," CVPR 2024.

[5] Li, Zhengqi, and Noah Snavely. "Megadepth: Learning single-view depth prediction from internet photos." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[6] Dai, Angela, et al. "Scannet: Richly-annotated 3d reconstructions of indoor scenes." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.