XFeat-Enhanced: Lightweight and Robust Image Matching

Aditya Raghuvanshi (2021114009), Yash Bhaskar (2021114012)

I. INTRODUCTION

Our project is based on the paper 'XFeat: Accelerated Features for Lightweight Image Matching,' which presents a lightweight CNN architecture for detecting, extracting, and matching local features in images. *XFeat consists of two components: Sparse Matching (XFeat), which identifies salient keypoints efficiently, and Semi-Dense Matching (XFeat*), which refines matches at the pixel level while balancing accuracy and computational cost.

As a crucial step for many higher-level vision tasks, local image feature extraction remains a highly active topic of research. Despite the recent advancements, the large improvements achieved from recent image matching methods [1, 2, 3, 4] mostly come at the cost of high computational requirements and increased implementation complexity.

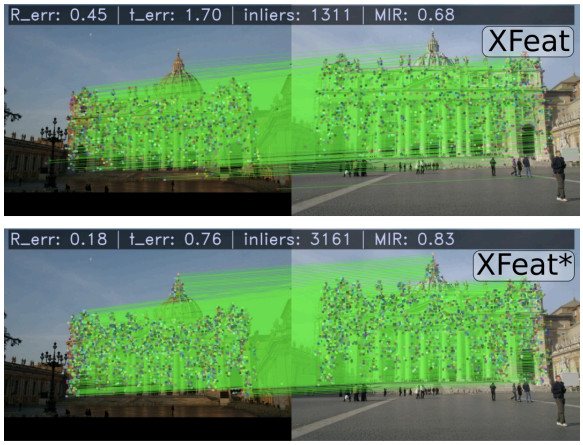
XFeat introduces semi-dense matching through a refinement module that enhances sparse keypoint matches using local descriptors and an MLP-based offset prediction mechanism. XFeat is optimized for efficiency, using a CNN backbone with progressive downsampling, channel depth management, and multi-scale integration, making it up to 5x faster than some deep learning-based methods while maintaining competitive accuracy.

For homography estimation, XFeat was evaluated on the HPatches dataset [2], which consists of image sequences with varying viewpoints and lighting conditions. MAGSAC++ [3] was used to estimate the homography transformation based on feature correspondences.

II. PROJECT PLAN

A. General Project Plan

We aim to replicate the results of this paper. We would train the whole model by ourselves and then get the results. An important point to note is that the paper's code is available online on GitHub[1]. We would be doing the experiments on the Megadepth [5] dataset and a synthetically warped subset of the COCO [6] dataset, similar to the training setup in the original paper.



The paper trained the model using both datasets in a 6:4 ratio, meaning 60% of the training data comes from Megadepth and 40% from the synthetically warped COCO dataset. The training process as described by the original paper states that Convergence is attained after 160,000 iterations on batches of 10 image pairs. That indicates the total number of image pairs used for training the original model is 1,600,000.

The Megadepth dataset provides real-world image pairs with depth information, while the synthetic warps from COCO improve generalization by introducing controlled transformations.

We also plan to run some ablations and experiments to check further the cases where the model fails to identify the objects correctly, which will allow us to get a better understanding of the working of the model and ways to improve it further.

B. New Things Planned

As part of our project proposal, we aim to improve XFeat's performance by addressing outlier matches and enhancing its robustness. Specifically, we seek to make XFeat more resilient to perspective and geometric variations, and to improve the quality and reliability of semi-dense matching by specifically addressing outlier matches.

To achieve robustness against perspective and geometric variations, we will leverage homography transformations of the original image during feature detection and matching. By applying

1. https://github.com/verlab/accelerated_features

these transformations during feature detection and matching, XFeat should become more capable of extracting repeatable features, even under varying perspectives, scaling, and geometric distortions. This homography-guided approach will help XFeat adapt to challenging scenarios and better handle images captured from different viewpoints or with distortions.

Furthermore, we will enhance the semi-dense matching (XFeat*) by implementing an outlier removal strategy. We plan to use DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to cluster the initial set of feature matches based on their spatial proximity. DBSCAN will allow us to identify and remove outliers, those matches that do not belong to any significant cluster. By filtering out these outlier matches before the refinement step, we anticipate a significant improvement in the overall quality and accuracy of the semi-dense matching results. This clustering will allow us to refine the high confidence matches and provide a more accurate and robust representation of the feature correspondences between images.

Beyond these specific areas, we will explore additional ways to improve XFeat in other aspects.

III. COMPUTE RESOURCES AVAILABLE

We both have access to the ADA Cluster of IIIT Hyderabad [11]. Both of us are Dual-Degree Students, and thus, by the research account privileges, we have access to unlimited minutes of usage with up to 4 GPUs allowed.

REFERENCES

[1] 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. In ECCV, pages 20–36. Springer, 2022.

[2] Johan Edstedt, Ioannis Athanasiadis, Marten Wadenb ˚ ack, ¨ and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In CVPR, pages 17765–17775, 2023

[3] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In CVPR, pages 8922–8931, 2021.

[4] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network. IEEE TPAMI, 2023.

[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] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014.

1. https://github.com/verlab/accelerated_features