Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.27799

Image Local Feature Detection and Description with Attention Mechanism

Yi Liu

School of Information Engineering, Wuhan University of Technology, Wuhan, China
13980643949@163.com

Abstract. To address the limitations of existing deep learning-based image matching algorithms, such as restricted receptive field in feature extraction networks and insufficient robustness against repetitive textures and large disparity scenarios. A novel image feature extraction algorithm combining attention mechanisms with dual-precision feature fusion is proposed. The method first utilizes a self-supervised pre-trained Vision Transformer DINOv2 to extract coarse-grained feature maps containing global semantic information. Subsequently, a fully convolutional network extracts fine-grained feature maps preserving spatial details. In the keypoint detection head, the fused feature map predicts keypoint positions, descriptors, and confidence levels. Experimental results on the HPatches dataset demonstrate that this feature extraction module significantly improves subsequent matching accuracy.

Keywords: Feature extraction, Attention, Fully convolutional, Transformer.

1. Introduction

Robust detection and description of image local features serve as the foundation and critical component for various computer vision tasks. Ideal image features should exhibit high repeatability , high discriminative power and high efficiency. Traditional methods primarily rely on manually designed feature operators such as SIFT [1], SURF [2], and ORB [3]. While effective in specific scenarios, these approaches demonstrate limited feature representation capabilities, showing significant degradation in robustness under complex conditions like extreme lighting, weak textures, or large perspective changes [4]. Recent deep learning-based feature extraction methods [5-7] have gradually outperformed traditional approaches through data-driven learning of robust feature representations. However, existing convolutional neural network (CNN) based methods still face two major challenges: (1) The inherent local receptive field characteristics of CNNs restrict their ability to model long-range contextual dependencies, leading to ambiguous matches in repetitive texture regions or large disparity scenarios; (2) Most methods depend on supervised pre-training models, resulting in limited feature transferability and requiring costly acquisition of large-scale annotated datasets. To address these issues, this paper proposes a novel image feature extraction algorithm. The core innovation lies in introducing a dual-precision feature fusion framework: The self-supervised pre-trained Vision Transformer DINOv2 [8] is utilized to extract coarse-grained features containing global semantic information, while a fully convolutional network is employed to

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.27799

capture fine-grained features preserving local spatial details. A specially designed fusion module effectively combines these two approaches. Evaluated on the HPatches standard dataset [9], our method demonstrates significant improvements in feature matching accuracy, particularly excelling in scenarios with perspective changes and repetitive textures.

2. Method

2.1. Network structure

The feature extraction network architecture consists of three main components: (1) coarse-grained feature extraction module; (2) fine-grained feature extraction module; and (3) feature fusion and keypoint detection head. The input is an image $I \in R \hat{} \{H \times W \times 3\}$, with final outputs including a keypoint coordinate map $K \in R \hat{} \{H \times W \times 2\}$, a 128-dimensional dense descriptor map $X \in R \hat{} \{H \times W \times 128\}$, and a keypoint confidence map $C \in R \hat{} \{H \times W \times 1\}$.

2.1.1. Coarse feature extraction module

This module is designed to capture global semantic information and scene-level context in images. The backbone architecture employs the Vision Transformer model DINOv2.

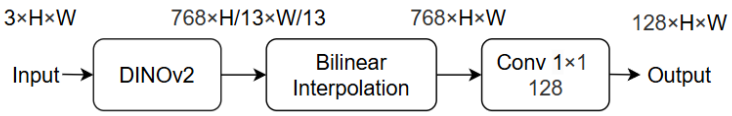

Figure 1. Structure of coarse feature extraction module

The input image is resized to meet DINOv2's specifications. Through DINOv2's Patch Embedding layer, the image is divided into fixed-size blocks and linearly projected into a sequence of embedded vectors. These vectors are overlaid with positional encodings before being fed into a multi-layer Transformer encoder, where global dependencies are captured through multi-head self-attention mechanisms. The process ultimately yields feature maps as output. The feature map has low resolution but contains high-level semantic information. To integrate with fine-grained features, the $Feature_{coarse} \in R^{\wedge} \{H_c \times W_c \times D_c\}$ is bilinearly up sampled to its original image size HW .A lightweight convolutional layer is applied for dimensionality reduction, downgrading the channel count to 128 dimensions to obtain the final coarse-grained feature map $Feature_{coarseFinal} \in R^{\wedge} \{H_c \times W_c \times 128\}$.

2.1.2. Fine feature extraction module

This module is designed to accurately capture local texture details and spatial structural information in images. By adopting a fully convolutional network architecture, it prevents spatial information loss caused by down sampling and achieves pixel-level feature extraction. The process begins with input passing through a 3×3 standard convolution to extract basic local features at 32 output channels. Multiple ResBlocks are connected in series during the intermediate stage, each containing: a 3×3 convolution, batch normalization, and ReLU activation.

DOI: 10.54254/2755-2721/2025.27799

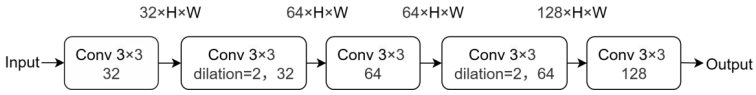

Figure 2. Structure of fine feature extraction module

2.2. Feature fusion and key point detection head

2.2.1. Feature fusion and attention optimization

The coarse-grained feature map and fine-grained feature map from different paths are concatenated and fused:

$$Feature_{fuse} = Concat(Feature_{coarseFinal}, Feature_{fine}) \in R^{\wedge}\{H_c \times W_c \times 256\} \quad (1)$$

2.2.2. Key point detection head

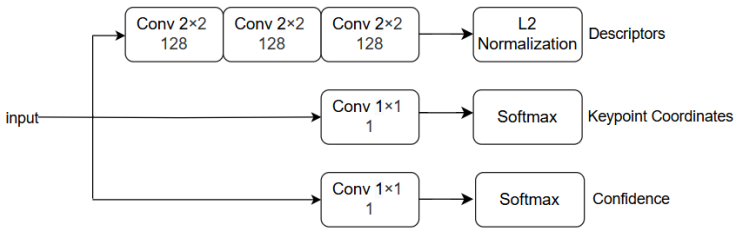

Figure 3. Structure of key point detection head

After the input image passes through the coarse feature extraction module and the fine feature extraction module, the coarse feature map $Feature_{coarse}$ and the fine feature map $Feature_{fine}$ are obtained respectively. The same size coarse feature map and fine feature map are concatenated as the input of the keypoint detection head.

The keypoint detector employs a decoupled design with three distinct outputs. The first output is a 3D tensor $x \in R^{128 \times H \times W}$, representing a dense set of 128-dimensional descriptors for each pixel. The second output is a feature map $S \in [0,1]_{H \times W}$ that maps the extracted keypoints to their coordinates. To enhance detection performance while controlling the number of detected keypoints, this approach divides the feature map S into $N \times N$ cells, selecting the maximum value within each cell as the keypoint. The third output is an association reliability map $R \in [0,1]_{H \times W}$, indicating confidence levels for each descriptor X_{ij} across pixels, where $i = 1 \ldots W$ and $j = 1 \ldots H$.

2.3. Loss function

The network's loss function consists of two components: coordinate position loss and confidence loss for keypoints.

For coordinate position loss, let I and I' be two images from the same scene where $U \in R^{H \times W \times 2}$ represents their true match. Specifically, if pixel (i, j) in image I corresponds to

DOI: 10.54254/2755-2721/2025.27799

pixel $(i^{'}, j^{'})$ in image $I^{'}$, then $I^{'}$ is synthesized through a known homography transformation, allowing precise determination of U. Let S and $S^{'}$ denote the keypoint feature maps of images I and $I^{'}$. The goal is to ensure all local maxima in correspond to their counterparts. This paper introduces cosine similarity to measure the degree of correspondence between S and $S_U^{'}$ local maxima. When keypoint positions align between feature maps, the cosine similarity $cos(S, S_U^{'})$ between them also reaches its maximum. We localize the concept by averaging cosine similarity across multiple small regions $N \times N$. We define as the overlapping set of all sub-blocks and formulate the loss function as:

$$L_{cos} = \frac{1}{|P|} \sum_{p \in P} cos(S[p], S_U^{'}[p]) \tag{2}$$

Here, $S[p]$ represents the vectorized representation of a small block p extracted from S, and $S_U^{'}[p]$ similarly represents the corresponding small block extracted from $S_U^{'}$. According to experimental results, this paper sets N to 16.

For confidence training, we aim to enhance the reliability of local descriptors X by predicting a confidence value $C_{xy} \in [0,1]$ for each descriptor, which estimates its reliability. This paper adopts a differentiable AP loss proposed by He et al [10]., a standard ranking metric that can be directly optimized during the training process. This paper follows a similar approach. The descriptor of each keypoint (x, y) in the first image is compared with those of all other keypoints in the second image. We can calculate the average AP score for each keypoint (x, y) using the formula.

$$L_{AP} = \frac{1}{B} \sum_{q=1}^{B} L_{AP}(q), L_{AP}\left(q\right) = 1 - AP\left(q\right) \tag{3}$$

Where q represents the number of key points in the first image, and B denotes the total number of all other key points. After calculating the average AP score for each key point, the confidence loss is defined as formula:

$$L_C = 1 - [\lambda(1 - C_{xy}) + AP(x, y)C_{xy}] \tag{4}$$

Here, λ is the threshold of AP score. When the AP score is less than the threshold, the confidence level C_{xy} should be as close to 0 as possible; when the AP score is greater than the threshold, the confidence level C_{xy} should be as close to 1 as possible. According to experimental results, it is found that setting the score threshold λ to 0.5 yields better results. The loss function in the key point extraction network is as follows:

$$L_{total} = \lambda_1 L_{cos} + \lambda_2 L_c \tag{5}$$

In this paper, according to the method proposed by Jerome et al [12]., the cosine similarity loss weight λ_1 and the confidence loss λ_2 are both set to 1.

DOI: 10.54254/2755-2721/2025.27799

3. Experiments & results

3.1. Experimental environment

The experimental setup utilized: hardware configuration (AMD Ryzen 6800H CPU with 16GB) + NVIDIA GTX 4090 GPU (20GB), along with software environment (Ubuntu18.04 operating system + Python 3.8 + PyCharm 2023.2.23+Pytorch 2.1.0+CUDA12.1).

3.2. Comparative experiments and analysis

Experiment 1: Ablation experiment of double precision feature extraction module

Five control experiments are designed in this study: A: Feature extraction using the double-precision feature extraction module; B: Replace the coarse feature extraction module with ResNet50 instead of Dinov2; C: Replace the fine feature extraction module with ResNet50 instead of the feedforward network; D: Feature extraction using only the coarse feature extraction module; E: Feature extraction using only the fine feature extraction module. The data set was selected as HPatches, and the matching accuracy metrics use MMA@3 and MMA@10.

Table 1. Ablation experiment of double precision feature extraction module

Group	Overall		Illumination		viewpoint	
	MMA@3	MMA@10	MMA@3	MMA@10	MMA@3	MMA@10
A	0.7155	0.8744	0.7738	0.9272	0.6615	0.8254
B	0.6438	0.7985	0.7102	0.8614	0.5774	0.7356
C	0.6875	0.8410	0.7448	0.8945	0.6302	0.7875
D	0.1421	0.2513	0.1673	0.2634	0.1169	0.2392
E	0.7042	0.8634	0.7621	0.9148	0.6505	0.8158

This study evaluated the Mean Matching Accuracy (MMA). In Experiment Group D, the average matching accuracy is less than one-fourth when only coarse feature extraction modules are introduced. In Experiment Group E, the module exhibits strong robustness to illumination variations. However, its performance under perspective transformation is significantly weaker, indicating poor cross-region matching effectiveness for large-scale perspective changes without global information from coarse features. When replacing the coarse feature extraction module with ResNet50 in Experiment Group B, accuracy decreases notably. This is because pre-trained ResNet often underperforms DINOv2 in feature extraction. DINOv2 utilizes self-supervised learning through Transformer architecture to establish long-range dependencies, capturing richer local and global features. In Experiment C, the feature maps from the CNN undergo down sampling at the original image size, and only after up sampling can the final feature map be restored to its original dimensions. However, this down sampling process inevitably causes information loss, resulting in reduced accuracy compared to fully convolutional network. These findings validate the effectiveness of the dual-precision feature extraction module proposed in this chapter.

Experiment 2: Comparison experiment between the image matching algorithm and others

To further validate the effectiveness of the algorithm proposed in this chapter, this experiment compares ours with several classical image feature extraction algorithms. The data set remains HPatches, and the matching accuracy metrics use MMA@3 and MMA@10 under various lighting and viewing conditions. The selected algorithms for comparison are as follows:(1) Traditional image

DOI: 10.54254/2755-2721/2025.27799

matching algorithms: SIFT and ORB. (2) Deep learning-based image matching algorithms: SuperPoint [11], R2D2 [12], LF-Net [13], DELF [14], and ReinforcedSP [15].

Table 2. Comparison test of image matching algorithm and other classical algorithms

Method	Overall		Illumination		viewpoint	
	MMA@3	MMA@10	MMA@3	MMA@10	MMA@3	MMA@10
SIFT	0.2581	0.5427	0.2855	0.5631	0.2307	0.5223
D2-Net	0.4057	0.7864	0.5021	0.8636	0.3162	0.7148
DELF	0.4895	0.7364	0.8956	0.9152	0.1124	0.5704
LF-Net	0.5301	0.5893	0.5731	0.6355	0.4902	0.5464
ReinforcedSP	0.6224	0.7636	0.5735	0.7099	0.6679	0.8135
SuperPoint	0.6946	0.8398	0.7438	0.8903	0.6488	0.7928
Ours	0.7155	0.8744	0.7738	0.9272	0.6615	0.8254

We evaluated the matching accuracy of various feature matching algorithms on the HPATCHES dataset, focusing on three key metrics: Overall performance, illumination variation, and viewpoint variation in Mean Matching Accuracy (MMA). The table data reveals that our proposed algorithm demonstrated the best matching accuracy across all tested indicators, particularly excelling in view variation. Specifically, our algorithm's overall matching accuracy significantly outperformed deep learning models like SuperPoint. In illumination variation experiments, our algorithm far surpasses SuperPoint and D2-Net while slightly trailing DELF. The illumination variation experiments demonstrate that our algorithm maintains higher matching accuracy when handling lighting changes. In viewpoint variation tests, our algorithm significantly outperforms SuperPoint and demonstrating robustness in wide-angle view variations. Our algorithm not only surpasses existing methods in illumination variations and overall performance but also demonstrates significant advantages in viewpoint variation testing, highlighting its enhanced matching capability and generalization ability in wide-angle transformation scenarios.

4. Conclusion

This paper proposes an image feature extraction algorithm based on attention mechanisms and dual-precision feature fusion, aiming to address the limitations of CNN-based feature extraction networks in terms of receptive field constraints and insufficient robustness under repetitive textures and large disparity scenarios. Future work could explore integrating this feature extraction module with more efficient feature matching networks to further optimize end-to-end performance, while also considering deployment on mobile platforms or embedded systems for real-time visual navigation and augmented reality applications.

References

[1] Lowe G D .Distinctive Image Features from Scale-Invariant Keypoints [J].International Journal of Computer Vision, 2004, 60(2): 91-110.

[2] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features [C]. Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9. Springer Berlin Heidelberg, 2006: 404-417.

[3] Rublee E, Rabaud V, Konolige K, et al. ORB: An efficient alternative to SIFT or SURF [C]. 2011 International conference on computer vision. Ieee, 2011: 2564-2571.

DOI: 10.54254/2755-2721/2025.27799

[4] Chum O, Matas J, Kittler J. Locally optimized RANSAC [C]. Joint pattern recognition symposium. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003: 236-243.
[5] Harris C, Stephens M. A combined corner and edge detector [C]. Alvey vision conference. 1988, 15(50): 10-5244.
[6] Abdel-Aziz Y I, Karara H M, Hauck M. Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry [J]. Photogrammetric engineering & remote sensing, 2015, 81(2): 103-107.
[7] Viswanathan D G. Features from accelerated segment test (fast) [C].Proceedings of the 10th workshop on image analysis for multimedia interactive services, London, UK. 2009: 6-8.
[8] Oquab M, Darcet T, Moutakanni T, et al. Dinov2: Learning robust visual features without supervision [J]. arXiv preprint arXiv: 2304.07193, 2023.
[9] Balntas V, Lenc K, Vedaldi A, et al. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors [C].Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5173-5182.
[10] He K, Lu Y, Sclaroff S. Local descriptors optimized for average precision [C].Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 596-605.
[11] DeTone D, Malisiewicz T, Rabinovich A. Superpoint: Self-supervised interest point detection and description [C].Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2018: 224-236.
[12] Revaud J, De Souza C, Humenberger M, et al. R2d2: Reliable and repeatable detector and descriptor [J]. Advances in neural information processing systems, 2019, 32.
[13] Ono Y, Trulls E, Fua P, et al. LF-Net: Learning local features from images [J]. Advances in neural information processing systems, 2018, 31.
[14] Noh H, Araujo A, Sim J, et al. Large-scale image retrieval with attentive deep local features [C].Proceedings of the IEEE international conference on computer vision. 2017: 3456-3465.
[15] Bhowmik A, Gumhold S, Rother C, et al. Reinforced feature points: Optimizing feature detection and description for a high-level task [C].Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 4948-4957.