1. Abstract

Traditional method: potential for real-world applications is restricted

first learnable image matcher

propose a novel key point position-guided attention mechanism

핵심 원칙: Generalization

The generalization capability of image matching models was overlooked.

이미지 매칭 일반화 능력 향상

공중 촬영 이미지 매칭 성능 향상

Based on DINOV2

1. Introduction

2-1 Local image feature matching

Camera pose estimation

3D reconstruction

2-2 Past decades: hand-crafted → learning-based

Highly specialized for the training domain

Drop dramatically on out-of-domain data

Might worse than traditional method in some case

But unrealistic to assume abundant training data will be available

Conclusion: developing architectural improvement to make learnable matching method generalize.

We assess OmniGlue’s generalization across diverse visual domains, spanning synthetic and real images, from scene-level to object-centric and aerial datasets, with small-baseline and wide-baseline cameras.

(2) A new strategy for leveraging positional encoding of keypoints, which avoids an overly reliant dependence on geometric priors from the training domain, boosting crossdomain transfer by up to 6.1% (14.9% relatively).

Domain-agnostic 도메인 불가지

Disentangle 플어주다

Synthetic 합성

Aerial 공중의

Sparse 희박한

~wise : ~별로 따로따로 나눠서 뭔가를 수행한다. ~에 따라 뭔가를 수행한다.

1. Related work

Generalizable Local Feature Matching

1. 전통적인 특징 검출 방법(SIFT 등)과 현대적인 학습 방법(딥러닝 기반 특징 추출기 등)을 결합

SIFT, ORB, SURF

1. 다양한 dataset에서 학습 (계산 효율은 낮을 수 있슴)
2. Robust

Sparse learnable image matching methods

1. 희소한 키포인트를 검출
2. SuperGlue, LightGlue
3. 많은 훈련 데이터를 필요로 하여, 효과적인 특징 기술 및 매칭 전략을 학습합니다.

실시간 응용에 더 적합

(Semi-)Dense Learnable Matching.

LoFTR

Matching with Additional Image Representations.

1. OmniGlue
   1. Overview
      1. image features are extracted using two complementary types of encoders:

SuperPoint, focusing on generic fine-grained matching; (depth)

DINOv2, an image foundation model which encodes coarse but broad visual knowledge (width)

* + 1. Build key-point association graph intra/inter-image with DINOv2

Disentangle positional and appearance signals at this stage

* + 1. Propagate information base on the graph
    2. Optimal matching layers are applied to produce a mapping between the key-points in the two images.

4.2 Feature Extraction

4.2.1 SuperPoint key-point A B

denote the three features of the i^{th} keypoint in set A as d^{A}\_i , p^{A}\_i and g^{A}\_i

1. **d\_{A\_i}​​**：
   * **定义：** 这是集合 A 中第 i 个关键点的 SuperPoint 局部描述符。
   * **解析：** SuperPoint 是一种特征提取器，用于从图像中检测稀疏关键点，并为每个关键点生成局部描述符。d\_{A\_i} ​​ 表示图像 IAI\_AIA​ 中第 i 个关键点的描述符，维度为 C。
2. **p\_{A\_i}**：
   * **定义：** 这是集合 A 中第 i 个关键点的位置嵌入。
   * **解析：** 位置嵌入 p\_{A\_i}​​ 用于编码关键点的空间位置。通常通过将关键点的坐标归一化后输入嵌入层或 MLP 层来获得，使得位置信息在特征匹配过程中被利用。
3. **g\_{A\_i}​​**：
   * **定义：** 这是集合 A 中第 i 个关键点的 DINOv2 描述符。
   * **解析：** DINOv2 是一种基于自监督学习的视觉模型，用于生成图像的全局和局部特征。g\_{A\_i}​​ 表示从 DINOv2 特征图中插值得到的、与第 i 个关键点位置对应的描述符，维度为 C′。

4.2.2 resulting positional features of a keypoint as p∈{R}^c

4.2.3 extract dense DINOv2 feature maps of the two images.

Global Visual Information

4.3 Graph construct

The goal of OmniGlue model is to estimate correspondences between the two key-point sets.

Use top half of key-points in set B with the largest DINOv2 similarities to connect with A\_i

Build GB→A​ and GA→B

Intra-image graphs represent the connectivity between key-points belonging to the same image.

4.4 Information Propagation with Novel Guidance.

DINOv2 guidance: during cross-attention, for key-point Ai, it only aggregates information from the DINOv2- pruned potential matching set selected from B, instead of all its key-points.

The first one updates keypoints based on the intra-image graphs, performing self-attention; the second updates keypoints based on the inter-image graphs, performing cross-attention.

1. Experiment

Interpolate: insert

Prune cut off

# 5

The TF models should take advantage of the GPU automatically, but the PyTorch DINOv2 code needs some modifications to [dino\_extract.py](https://github.com/google-research/omniglue/blob/main/src/omniglue/dino_extract.py):

* [Line 38](https://github.com/google-research/omniglue/blob/main/src/omniglue/dino_extract.py#L38): after this line, add a call self.model.cuda() to send the model to GPU mem
* [Line 113](https://github.com/google-research/omniglue/blob/main/src/omniglue/dino_extract.py#L113): replace with: out = self.model.get\_intermediate\_layers(image.cuda(), n=self.feature\_layer)[0] - i.e., send image to GPU mem with a .cuda() call

After this, hopefully all models are run on GPU and you should see some inference latency improvements.

#18

<https://github.com/google-research/omniglue/issues/18>

LoFTR

本文提出了一种新的局部图像特征匹配方法，通过先在粗层次上建立像素级密集匹配，然后在更细的层次上进行细化，并利用transformer的自注意层和交叉注意力层来获取基于两幅图像的特征描述符。本文的创新点在于Transformer提供的全局感受野使得该方法能够在低纹理区域产生密集匹配。

Off the shelf method 现成的

fine grained 细密度