LoFTR: Detector-Free Local Feature Matching with Transformers

Jiaming Sun[1,2*] Zehong Shen[1*] Yuang Wang[1*] Hujun Bao[1] Xiaowei Zhou[1†]
[1]Zhejiang University [2]SenseTime Research

Abstract

We present a novel method for local image feature matching. Instead of performing image feature detection, description, and matching sequentially, we propose to first establish pixel-wise dense matches at a coarse level and later refine the good matches at a fine level. In contrast to dense methods that use a cost volume to search correspondences, we use self and cross attention layers in Transformer to obtain feature descriptors that are conditioned on both images. The global receptive field provided by Transformer enables our method to produce dense matches in low-texture areas, where feature detectors usually struggle to produce repeatable interest points. The experiments on indoor and outdoor datasets show that LoFTR outperforms state-of-the-art methods by a large margin. LoFTR also ranks first on two public benchmarks of visual localization among the published methods. Code is available at our project page: https://zju3dv.github.io/loftr/.

1. Introduction

Local feature matching between images is the cornerstone of many 3D computer vision tasks, including structure from motion (SfM), simultaneous localization and mapping (SLAM), visual localization, etc. Given two images to be matched, most existing matching methods consist of three separate phases: feature detection, feature description, and feature matching. In the detection phase, salient points like corners are first detected as interest points from each image. Local descriptors are then extracted around neighborhood regions of these interest points. The feature detection and description phases produce two sets of interest points with descriptors, the point-to-point correspondences of which are later found by nearest neighbor search or more sophisticated matching algorithms.

The use of a feature detector reduces the search space of matching, and the resulted sparse correspondences are sufficient for most tasks, e.g., camera pose estimation. However, a feature detector may fail to extract enough interest points that are repeatable between images due to various factors such as poor texture, repetitive patterns, viewpoint change, illumination variation, and motion blur. This issue is especially prominent in indoor environments, where low-texture regions or repetitive patterns sometimes occupy most areas in the field of view. Fig. 1 shows an example. Without repeatable interest points, it is impossible to find correct correspondences even with perfect descriptors.

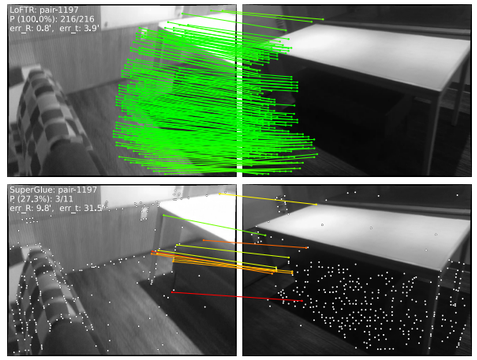

Figure 1: **Comparison between the proposed method LoFTR and the detector-based method SuperGlue [37].** This example demonstrates that LoFTR is capable of finding correspondences on the texture-less wall and the floor with repetitive patterns, where detector-based methods struggle to find repeatable interest points.[1]

Several recent works [34, 33, 19] have attempted to remedy this problem by establishing pixel-wise dense matches. Matches with high confidence scores can be selected from the dense matches, and thus feature detection is avoided. However, the dense features extracted by convolutional neural networks (CNNs) in these works have limited receptive field which may not distinguish indistinctive regions. Instead, humans find correspondences in these indistinctive regions not only based on the *local* neighborhood, but with a larger *global* context. For example, low-texture regions in

*The first three authors contributed equally. The authors are affiliated with the State Key Lab of CAD&CG and ZJU-SenseTime Joint Lab of 3D Vision. †Corresponding author: Xiaowei Zhou.

[1]Only the inlier matches after RANSAC are shown. The green color indicates a match with epipolar error smaller than 5×10^{-4} (in the normalized image coordinates).

LOFTR：探测器的本地功能与变压器匹配

Jiaming Sun[1,2*] Zehong Shen[1*] Yuang Wang[1*] Hujun Bao[1] Xiaowei Zhou[1†]
[1]Zhejiang University [2]SenseTime Research

Abstract

我们提出了一种对本地图像特征的新方法
匹配。而不是顺序地执行图像特征检测，描述和匹配，我们建议首先在粗级别建立像素明智的匹配，后来在精细水平处改进良好的匹配。与使用成本卷进行搜索相关性的密度方法相比，我们使用传输中的自我和跨关注层来获得在两个图像上调节的特征描述符。传输提供的全球接收领域使我们的方法能够在低纹理区域中产生密集的比赛，其中特征探测器通常是恒星产生可重复的兴趣点。室内和室外数据集的实验表明，LOFTR超越了最先进的方法，通过大边距。LOFTR还在发布的方法中排名第一的视觉局部公共基准。代码可在我们的项目页面上找到：https://zju3dv.github.io/loftr/。

1. Introduction

图像之间的本地特征是角落 -
许多3D计算机视觉任务的石头，包括来自运动（SFM）的结构，同时定位和映射（SLAM），视觉本地化等。考虑到两个要匹配的图像，大多数现有匹配方法包括三个单独的阶段：功能检测，功能描述和特征匹配。在检测阶段中，首先将凸起的突出点作为来自每个IM的感兴趣点。然后在邻居周围提取本地描述符

这些兴趣点的障碍地区。特征模式和描述阶段产生两组利用描述符，其稍后通过最近的邻近搜索或更复杂的匹配算法找到的点对点对应关系。

使用特征检测器可以减少搜索空间
匹配，导致稀疏的对应关系对于大多数任务，例如摄像机姿势估算是如此。但是，特征检测器可能无法提取足够的兴趣点

这是由于诸如差的质地，重复模式，视点变化，照明变化和运动模糊等各种因素之间的可重复图像。这个问题在室内环境中突出，低纹理区域或重复模式有时占据了大部分视野中的区域。图。图1示出了示例。如果没有重新攻击的兴趣点，即使使用完美的描述符也无法找到正确的心律申请。

最近的几项作品[34,33,19]已经尝试了
通过建立像素 -
明智的密集匹配来eDY这个问题。具有高置信区分的匹配可以从密集的匹配中选择，因此避免了特征检测。然而，在这些作品中卷积的Neu-Ral网络（CNNS）提取的密集特征具有有限的接收领域，其可能不会区分空域区域。虽然，人类不仅基于当地社区，但人类不仅可以基于本地社区的对应关系，而且在全球范围内找到更大的环境。例如，低纹理区域

★前三名作者同等贡献。作者是附属的
与CAD＆CG的状态重点实验室和3D Vision
ZJU-Sensetime联合实验室。 通讯作者：Xiaowei Zhou。

表示具有小于5×10-4的eMipolar误差的匹配（在NOR-苹果化图像坐标中）。

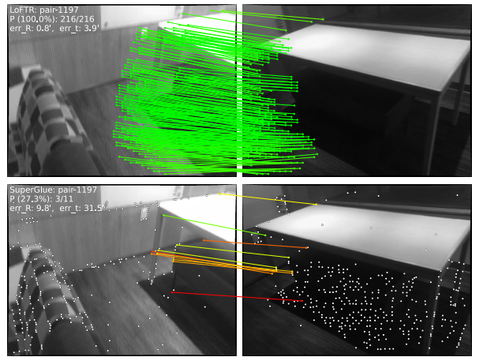

图1：所提出的方法LOFTR与基于探测器的方法SuperGlue的比较[37]。该示例演示了Loftr能够在纹理墙壁和地板上找到关于重复模式的对应关系，其中基于探测器的方法难以找到可重复的兴趣点。

Fig. 1 can be distinguished according to their relative positions to the edges. This observation tells us that a large *receptive field* in the feature extraction network is crucial.

Motivated by the above observations, we propose Local Feature TRansformer (LoFTR), a novel detector-free approach to local feature matching. Inspired by seminal work SuperGlue [37], we use Transformer [48] with self and cross attention layers to process (transform) the dense local features extracted from the convolutional backbone. Dense matches are first extracted between the two sets of transformed features at a low feature resolution ($1/8$ of the image dimension). Matches with high confidence are selected from these dense matches and later refined to a sub-pixel level with a correlation-based approach. The global receptive field and positional encoding of Transformer enable the transformed feature representations to be context- and position-dependent. By interleaving the self and cross attention layers multiple times, LoFTR learns the densely-arranged globally-consented matching priors exhibited in the ground-truth matches. A linear transformer is also adopted to reduce the computational complexity to a manageable level.

We evaluate the proposed method on several image matching and camera pose estimation tasks with indoor and outdoor datasets. The experiments show that LoFTR outperforms detector-based and detector-free feature matching baselines by a large margin. LoFTR also achieves state-of-the-art performance and ranks first among the published methods on two public benchmarks of visual localization. Compared to detector-based baseline methods, LoFTR can produce high-quality matches even in indistinctive regions with low-textures, motion blur, or repetitive patterns.

2. Related Work

Detector-based Local Feature Matching. Detector-based methods have been the dominant approach for local feature matching. Before the age of deep learning, many renowned works in the traditional hand-crafted local features have achieved good performances. SIFT [26] and ORB [35] are arguably the most successful hand-crafted local features and are widely adopted in many 3D computer vision tasks. The performance on large viewpoint and illumination changes of local features can be significantly improved with learning-based methods. Notably, LIFT [51] and MagicPoint [8] are among the first successful learning-based local features. They adopt the detector-based design in hand-crafted methods and achieve good performance. SuperPoint [9] builds upon MagicPoint and proposes a self-supervised training method through homographic adaptation. Many learning-based local features along this line [32, 11, 25, 28, 47] also adopt the detector-based design.

The above-mentioned local features use the nearest neighbor search to find matches between the extracted interest points. Recently, SuperGlue [37] proposes a learning-based approach for local feature matching. SuperGlue accepts two sets of interest points with their descriptors as input and learns their matches with a graph neural network (GNN), which is a general form of Transformers [16]. Since the priors in feature matching can be learned with a data-driven approach, SuperGlue achieves impressive performance and sets the new state of the art in local feature matching. However, being a detector-dependent method, it has the fundamental drawback of being unable to detect repeatable interest points in indistinctive regions. The attention range in SuperGlue is also limited to the detected interest points only. Our work is inspired by SuperGlue in terms of using self and cross attention in GNN for message passing between two sets of descriptors, but we propose a detector-free design to avoid the drawbacks of feature detectors. We also use an efficient variant of the attention layers in Transformer to reduce the computation costs.

Detector-free Local Feature Matching. Detector-free methods remove the feature detector phase and directly produce dense descriptors or dense feature matches. The idea of dense features matching dates back to SIFT Flow [23]. [6, 39] are the first learning-based approaches to learn pixel-wise feature descriptors with the contrastive loss. Similar to the detector-based methods, the nearest neighbor search is usually used as a post-processing step to match the dense descriptors. NCNet [34] proposed a different approach by directly learning the dense correspondences in an end-to-end manner. It constructs 4D cost volumes to enumerate all the possible matches between the images and uses 4D convolutions to regularize the cost volume and enforce neighborhood consensus among all the matches. Sparse NCNet [33] improves upon NCNet and makes it more efficient with sparse convolutions. Concurrently with our work, DRC-Net [19] follows this line of work and proposes a coarse-to-fine approach to produce dense matches with higher accuracy. Although all the possible matches are considered in the 4D cost volume, the receptive field of 4D convolution is still limited to each matches' neighborhood area. Apart from neighborhood consensus, our work focuses on achieving global consensus between matches with the help of the global receptive field in Transformers, which is not exploited in NCNet and its follow-up works. [24] proposes a dense matching pipeline for SfM with endoscopy videos. The recent line of research [46, 45, 44, 15] that focuses on bridging the task of local feature matching and optical flow estimation, is also related to our work.

Transformers in Vision Related Tasks. Transformer [48] has become the *de facto* standard for sequence modeling in natural language processing (NLP) due to their simplic-

图。图1可以根据其对边缘的相对型锯齿来区分。该观察告诉我们，特征提取网络中的大型接收领域至关重要。

通过上述观察，我们提出了LO−
CAL功能变压器（LOFTR），一种用于本地特征匹配的新型探测器方法。灵感来自精湛的工作SuperGlue
[37]，我们使用变压器[48]与自我和跨关注层来处理（转换）从卷积骨架中提取的密集局部功能。首先在低特征分辨率（图像维度的1/8）的两组变换特征之间提取密集匹配。具有高置信度的匹配从这些密集的匹配和后来通过基于相关的方法精制到子像素水平。变换器的全局接收领域和位置编码能够成为上下文和位置的变换特征表示。通过多次交织自我和跨关注层，LOFTR了解在地面真实匹配中展出的密集被安排的全球同意匹配前沿。线性变压器也是

采用将计算复杂性降低到一个人衰老的水平。

我们在几个图像上评估所提出的方法
使用室内和室外数据集的匹配和相机姿态估算任务。实验表明，LOFTR通过大边缘进行基于探测器和无检测器的特征匹配基线。LOFTR还实现了最先进的性能，并在视觉本地化的两种公共基准中排名第一的方法。与基于探测器的基线方法相比，即使在具有低纹理，运动模糊或重复模式的神秘区域中，LOFTR也可以产生高质量匹配。

2. Related Work

基于探测器的本地特征匹配。基于探测器的方法是局部有序匹配的主导方法。在深入学习的年龄之前，很多

在传统的手工制作的当地功能中的着名作品取得了良好的表现。SIFT [26]和

ORB
[35]可以说是最成功的手工制作的本地特征，并且在许多3D Culer Vision任务中被广泛采用。大型观点的性能

通过基于学习的方法可以显着改善本地特征的照明变化。尤其，

升降机[51]和魔术点[8]是基于第一个基于学习的本地特征。它们采用了探测器的设计，手工制作方法，实现了良好的性能。SuperPoint
[9]在魔术点构建，并通过同型适应提出自我监督的培训方法。许多基于学习的本地功能

沿着这条线[32,11,25,28,47]还采用了基于探测器的设计。

上述当地功能使用最近的
邻居搜索以查找提取的否则点之间的匹配项。最近，SuperGlue [37]提出了一种基于学习的本地特征匹配方法。
SuperGlue
AC−CEPTS与他们的描述符作为输入，并使用图表神经网络（GNN）的匹配来学习它们，这是变压器的一般形式[16]。由于可以使用数据驱动方法学习特征匹配的前沿，因此SuperGlue在本地特征匹配中实现了令人印象深刻的性格，并将新技术设置了新的技术。然而，作为依赖验样的方法，它具有无法检测到非确定区域中可重复兴趣点的根本缺点。
SuperGlue中的内容范围仅限于检测到的兴趣点。我们的工作受到SuperGlue在GNN中使用SuperGlue的启发，以便在两套描述符之间传递的消息，但我们提出了一种无探测器设计，以避免特征结果的缺点。我们还使用变压器中的注意事项的有效变体来降低计算成本。

探测器的本地特色匹配。没有探测器
方法删除特征检测器相位，直接采用密集的描述符或密集特征匹配。密集功能匹配日期的思想返回SIFT流程[23]。[6,39]是基于学习的基于学习的方法，用于以对比丢失学习像素的特征描述符。类似于基于探测器的方法，最近的邻居搜索通常用作匹配密度描述符的后处理步骤。NCNET
[34]通过直接以端到端的方式直接学习密集信念提出了不同的方法。它构建了4D成本卷来枚举 −

在图像之间进行所有可能的匹配，并使用4D卷积来规范成本卷，并在所有匹配中强制执行邻居共识。疏

NCNet
[33]改进了NCNet，并使其更加迅速冗余卷积。与我们同时同时

工作，DRC−Net
[19]遵循这一行作品，提出了一种粗细的方法，以产生更高的准确性的密集匹配。尽管所有可能的匹配都是在4D成本体积中持有的，但是4D概率的接受领域仍然仅限于每个匹配的邻域区域。除了邻居的共识之外，我们的工作侧重于在变压器的全球接受领域的帮助下实现与匹配之间的全球共识，这在没有利用NCNet及其后续作品中。[24]提出了一种具有内窥镜视频的SFM的密集匹配管道。最近的研究线[46,45,44,15]，其侧重于桥接本地特征匹配和光学流量估计的任务，也与我们的工作有关。

视觉相关任务中的变形金刚。变压器[48]已成为自然语言处理（NLP）的序列建模的事实标准，因为它们是简单的 −

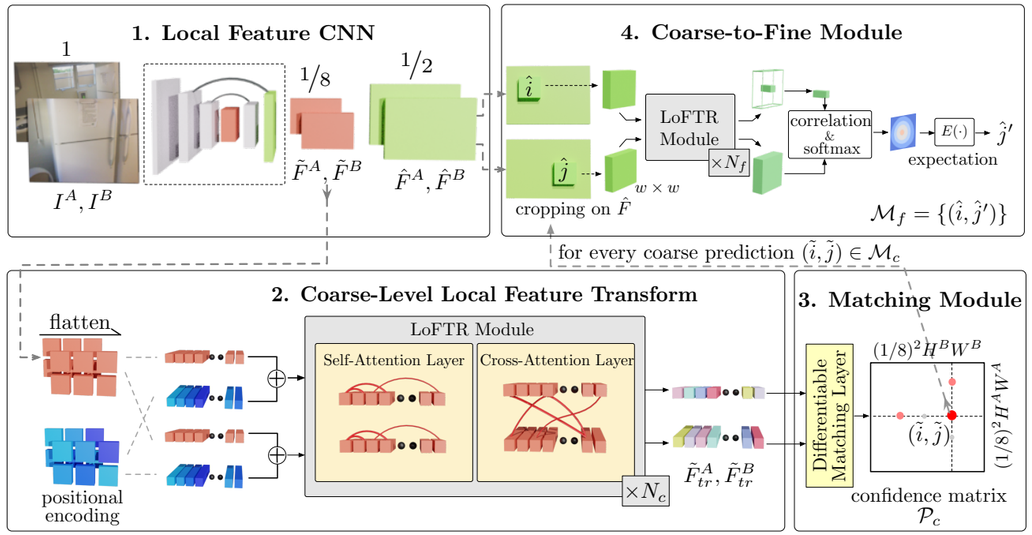

Figure 2: **Overview of the proposed method.** LoFTR has four components: **1.** A local feature CNN extracts the coarse-level feature maps $\tilde{F}^A$ and $\tilde{F}^B$, together with the fine-level feature maps $\hat{F}^A$ and $\hat{F}^B$ from the image pair I^A and I^B (Section 3.1). **2.** The coarse feature maps are flattened to 1-D vectors and added with the positional encoding. The added features are then processed by the Local Feature TRansformer (LoFTR) module, which has N_c self-attention and cross-attention layers (Section 3.2). **3.** A differentiable matching layer is used to match the transformed features, which ends up with a confidence matrix $\mathcal{P}_c$. The matches in $\mathcal{P}_c$ are selected according to the confidence threshold and mutual-nearest-neighbor criteria, yielding the coarse-level match prediction $\mathcal{M}_c$ (Section 3.3). **4.** For every selected coarse prediction $(\tilde{i}, \tilde{j}) \in \mathcal{M}_c$, a local window with size $w \times w$ is cropped from the fine-level feature map. Coarse matches will be refined within this local window to a sub-pixel level as the final match prediction $\mathcal{M}_f$ (Section 3.4).

ity and computation efficiency. Recently, Transformers are also getting more attention in computer vision tasks, such as image classification [10], object detection [3] and semantic segmentation [49]. Concurrently with our work, [20] proposes to use Transformer for disparity estimation. The computation cost of the vanilla Transformer grows quadratically as the length of input sequences due to the multiplication between query and key vectors. Many efficient variants [42, 18, 17, 5] are proposed recently in the context of processing long language sequences. Since no assumption of the input data is made in these works, they are also well suited for processing images.

3. Methods

Given the image pair I^A and I^B, the existing local feature matching methods use a feature detector to extract interest points. We propose to tackle the repeatability issue of feature detectors with a detector-free design. An overview of the proposed method LoFTR is presented in Fig. 2.

3.1. Local Feature Extraction

We use a standard convolutional architecture with FPN [22] (denoted as the local feature CNN) to extract multi-level features from both images. We use $\tilde{F}^A$ and $\tilde{F}^B$ to denote the coarse-level features at $1/8$ of the original image dimension, and $\hat{F}^A$ and $\hat{F}^B$ the fine-level features at $1/2$ of the original image dimension.

Convolutional Neural Networks (CNNs) possess the inductive bias of translation equivariance and locality, which are well suited for *local* feature extraction. The downsampling introduced by the CNN also reduces the input length of the LoFTR module, which is crucial to ensure a manageable computation cost.

3.2. Local Feature Transformer (LoFTR) Module

After the local feature extraction, $\tilde{F}^A$ and $\tilde{F}^B$ are passed through the LoFTR module to extract position and context dependent local features. Intuitively, the LoFTR module transforms the features into feature representations that are easy to match. We denote the transformed features as $\tilde{F}^A_{tr}$ and $\tilde{F}^B_{tr}$.

Preliminaries: Transformer [48]. We first briefly introduce the Transformer here as background. A Transformer encoder is composed of sequentially connected encoder layers. Fig. 3(a) shows the architecture of an encoder layer.

The key element in the encoder layer is the attention

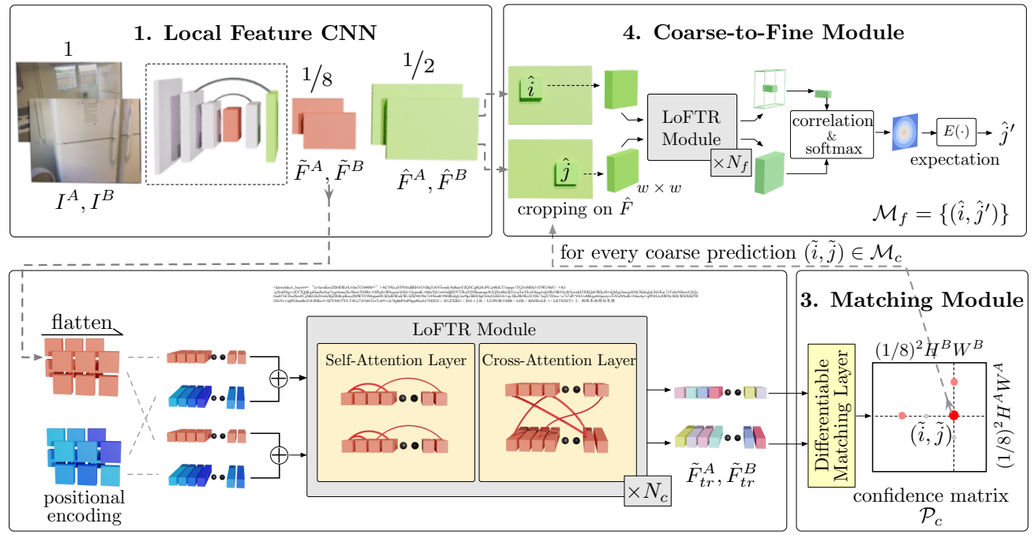

图2：所提出的方法概述。 LOFTR有四个组件：1。本地特征CNN提取粗级特征映射~F a和~f b，与图像对Ia和1b（第3.1节第3.1节）以及微级特征映射F a和f b。
2.粗略特征贴图趋于平坦于1-D矢量，并添加了位置编码。然后通过本地特征变压器（LOFTR）模块处理附加的功能，该模块具有NC自我关注和横向层（第3.2节）。
3.可分辨率匹配层用于匹配变换的特征，该特征最终置于置信矩阵PC。根据置信阈值和相互最近邻标准选择PC中的匹配，产生粗级匹配预测MC（第3.3节）。
4.对于每个选定的粗略预测（~i，~j）∈mc，从细级别的特征映射裁剪具有尺寸W×W的本地窗口。粗略匹配将在本地窗口中精制到子像素级别作为最终匹配预测MF（第3.4节）。

ITY和计算效率。最近，变形金刚在计算机视觉任务中也得到更多的关注，例如图像分类[10]，对象检测[3]和Seman-Tic分段[49]。与我们的工作同时，[20]建议使用变压器进行差异估计。由于查询和密钥向量之间的乘法，Vanilla变压器的计算成本作为输入序列的长度。最近在处理长语言序列的背景下提出了许多有效的变量[42,18,17,5]。由于在这些工作中没有假设输入数据，因此它们也非常适合处理图像。

3. Methods

鉴于图像对IA和IB，现有的本地FEA-

Ture匹配方法使用特征检测器来提取最终的点。我们建议用无探测器设计来解决特征探测器的重复性问题。所提出的方法LOFTR的概述在图2中示出。2。

3.1. Local Feature Extraction

我们使用标准的卷积架构

FPN [22]（表示为局部特征CNN）以提取

来自两个图像的多级别功能。我们使用~f a和~f b

为了表示原始倍数的1/8处的粗级特征，F a和f b在原始图像尺寸的1/2处的细级别特征。

卷积神经网络（CNNS）拥有in-

平移等级和地区的耐心偏见，非常适合局部特征提取。由CNN引入的井下电压也降低了LOFTR模块的输入长度，这是至关重要的，以确保管理能力计算成本。

3.2。本地特征变压器（LOFTR）模块

局部特征提取后，通过~f A和~f B通过

通过LOFTR模块提取位置和上下文相关的本地功能。直观地，LOFTR模块将功能转换为易于匹配的特征表示。我们表示转换的功能如~f a

and $\tilde{F}^B_{tr}$.

预备：变压器[48]。我们首先简要介绍

在这里作为背景消除变压器。变压器编码器由顺序连接的编码器层组成。图。图3（a）示出了编码器层的架构。

编码器层中的关键元素是注意力

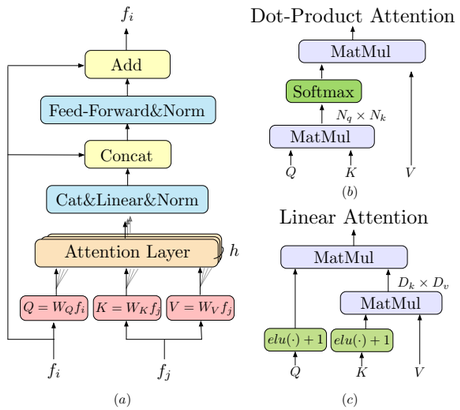

Figure 3: **Encoder layer and attention layer in LoFTR.** (a) Transformer encoder layer. h represents the multiple heads of attention. (b) Vanilla dot-product attention with $O(N^2)$ complexity. (c) Linear attention layer with $O(N)$ complexity. The scale factor is omitted for simplicity.

layer. The input vectors for an attention layer are conventionally named query, key, and value. Analogous to information retrieval, the query vector Q retrieves information from the value vector V, according to the attention weight computed from the dot product of Q and the key vector K corresponding to each value V. The computation graph of the attention layer is presented in Fig. 3(b). Formally, the attention layer is denoted as:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T)V.$$

Intuitively, the attention operation selects the relevant information by measuring the similarity between the query element and each key element. The output vector is the sum of the value vectors weighted by the similarity scores. As a result, the relevant information is extracted from the value vector if the similarity is high. This process is also called "message passing" in Graph Neural Network.

Linear Transformer. Denoting the length of Q and K as N and their feature dimension as D, the dot product between Q and K in the Transformer introduces computation cost that grows quadratically ($O(N^2)$) with the length of the input sequence. Directly applying the vanilla version of Transformer in the context of local feature matching is impractical even when the input length is reduced by the local feature CNN. To remedy this problem, we propose to use an efficient variant of the vanilla attention layer in Transformer. Linear Transformer [17] proposes to reduce the computation complexity of Transformer to $O(N)$ by substituting the exponential kernel used in the original attention layer with an alternative kernel function $\text{sim}(Q, K) = \phi(Q) \cdot \phi(K)^T$, where $\phi(\cdot) = \text{elu}(\cdot) + 1$. This operation is illustrated by the computation graph in Fig. 3(c). Utilizing

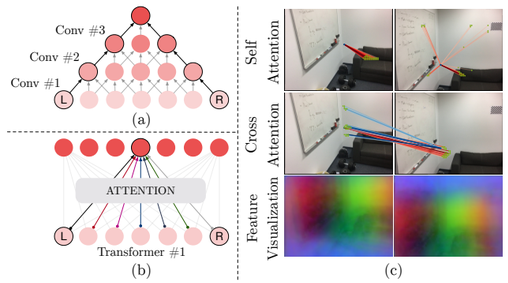

Figure 4: Illustration of the receptive field of **(a)** Convolutions and **(b)** Transformers. Assume that the objective is to establish a connection between the L and R elements to extract their joint feature representation. Due to the local-connectivity of convolutions, many convolution layers need to be stacked together in order to achieve this connection. The global receptive field of Transformers enables this connection to be established through only one attention layer. **(c)** Visualization of the attention weights and transformed dense features. We use PCA to reduce the dimension of the transformed features $\tilde{F}^A_{tr}$ and $\tilde{F}^B_{tr}$ and visualize the results with RGB color. Zoom in for details.

the associativity property of matrix products, the multiplication between $\phi(K)^T$ and V can be carried out first. Since $D \ll N$, the computation cost is reduced to $O(N)$.

Positional Encoding. We use the 2D extension of the standard positional encoding in Transformers following DETR [3]. Different from DETR, we only add them to the backbone output once. We leave the formal definition of the positional encoding in the supplementary material. Intuitively, the positional encoding gives each element unique position information in the sinusoidal format. By adding the position encoding to $\tilde{F}^A$ and $\tilde{F}^B$, the transformed features will become position-dependent, which is crucial to the ability of LoFTR to produce matches in indistinctive regions. As shown in the bottom row of Fig. 4(c), although the input RGB color is homogeneous on the white walls, the transformed features $\tilde{F}^A_{tr}$ and $\tilde{F}^B_{tr}$ are *unique* for each position demonstrated by the smooth color gradients. More visualizations are provided in Fig. 6.

Self-attention and Cross-attention Layers. For self-attention layers, the input features f_i and f_j (shown in Fig. 3) are the same (either $\tilde{F}^A$ or $\tilde{F}^B$). For cross-attention layers, the input features f_i and f_j are either ($\tilde{F}^A$ and $\tilde{F}^B$) or ($\tilde{F}^B$ and $\tilde{F}^A$) depending on the direction of cross-attention. Following [37], we interleave the self and cross attention layers in the LoFTR module by N_c times. The attention weights of the self and cross attention layers in LoFTR are visualized in the first two rows of Fig. 4(c).

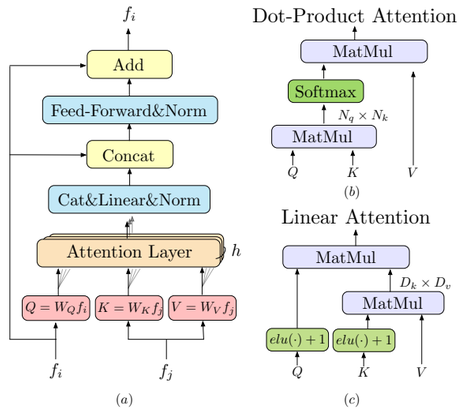

图3：leoftr中的编码器层和注意层。（A）变压器编码器层。h代表了多个关注头。（b）香草圆点－产品注意力o（n 2）复杂性。（c）具有O（n）复杂性的线性注意层。为简单起见，省略了比例因子。

层。注意层的输入向量是赋值，键和值。类似于信息检索，查询向量Q根据从Q的点乘积计算的注意重量和对应于每个值v的键矢量k的注意重量，从值矢量v检索来自值矢量v的信息。注意层的计算图在图2中示出。3（b）。正式地，注意层表示：

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T)V.$$

直观地，注意操作通过测量查询单元和每个关键元素之间的相似性来选择相关的信息。输出向量是由相似性分数加权的值矢量的总和。结果，如果相似度高，则从值矢量提取相关信息。该过程也称为图形神经网络中的“消息传递”。

线性变压器。表示Q和K的长度和它们的特征尺寸为D，变压器中的点产品Q和K和k在变压器中引入了从输入序列的长度逐渐增加的计算成本（O（n 2））。在本地特征匹配的上下文中，直接应用Vanilla版本的变压器也是可以实现的，即使当本地特征CNN减少输入长度也是如此。为了解决这个问题，我们建议在转换器中使用Vanilla注意层的有效变体。线性变压器[17]提议减少

通过替代核函数SIM（Q，k）=ϕ（q）ϕ（k）ϕ（k）ϕ（k）ϕ（k）ϕ（·）= ELU（·）+ 1.该操作由图3（c）中的计算图说明。利用

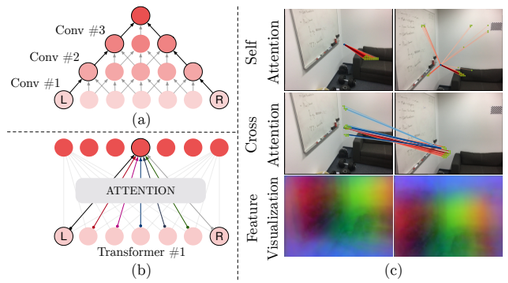

图4：（a）卷积和（b）变压器的接受领域的插图。假设目标是建立L和R元素之间的连接以提取它们的联合特征表示。由于卷积的局部连接，许多卷积层需要堆叠在一起以实现这一连接。变压器的全球接收领域使得能够仅通过一个注意层建立。（c）可视化注意力和转化的密集特征。我们使用PCA来减少转换功能的尺寸~f a

TR并可视化结果

用RGB颜色。放大详情。

矩阵产品的关联性特性，可以首先进行ϕ（k）t和v之间的多重阳离子。由于D n，计算成本减少到O（n）。

位置编码。我们使用的2D扩展

DETR

[3]后变压器中的标准位置编码。与DETR不同，我们只将它们添加到骨干输出一次。我们留下了辅助材料中位置编码的正式定义。位置，位置编码在正弦格式中给出了每个元素的唯一位置信息。通过将编码的位置添加到~f a和~f b，转化的功能将变为位置依赖性，这对于Loftr在不确定的再园中产生匹配的能力至关重要。如图1的底行所示。如图4（c）所示，虽然输入的RGB颜色在白色壁上均匀，但是变换的特征~f a

TR是独一无二的

光滑色梯度展示的位置。图2中提供了更多可视化。6。

自我关注和跨关注层。为了自我－

注意层，输入特征fi和fj（如图3所示）相同（~f a或~f b）。对于跨关注层，根据横向的方向，输入特征Fi和FJ是（~f a和~f b）或（~f b和~f a）。遵循[37]，我们通过NC次交织在LOFTR模块中的自我和跨注意力层。Loftr中的自我和跨注意层的注意力在图4的前两行中可视化。4（c）。

3.3. Establishing Coarse-level Matches

Two types of differentiable matching layers can be applied in LoFTR, either with an optimal transport (OT) layer as in [37] or with a dual-softmax operator [34, 47]. The score matrix $\mathcal{S}$ between the transformed features is first calculated by $\mathcal{S}(i,j) = \frac{1}{\tau} \cdot \langle \tilde{F}^A_{tr}(i), \tilde{F}^B_{tr}(j) \rangle$. When matching with OT, $-\mathcal{S}$ can be used as the cost matrix of the partial assignment problem as in [37]. We can also apply softmax on both dimensions (referred to as dual-softmax in the following) of $\mathcal{S}$ to obtain the probability of soft mutual nearest neighbor matching. Formally, when using dual-softmax, the matching probability $\mathcal{P}_c$ is obtained by:

$$\mathcal{P}_c(i,j) = \mathrm{softmax}\left(\mathcal{S}(i,\cdot)\right)_j \cdot \mathrm{softmax}\left(\mathcal{S}(\cdot,j)\right)_i .$$

Match Selection. Based on the confidence matrix $\mathcal{P}_c$, we select matches with confidence higher than a threshold of θ_c, and further enforce the mutual nearest neighbor (MNN) criteria, which filters possible outlier coarse matches. We denote the coarse-level match predictions as:

$$\mathcal{M}_c = \{(\tilde{i},\tilde{j}) \mid \forall (\tilde{i},\tilde{j}) \in \mathrm{MNN}(\mathcal{P}_c),\ \mathcal{P}_c(\tilde{i},\tilde{j}) \geq \theta_c\}.$$

3.4. Coarse-to-Fine Module

After establishing coarse matches, these matches are refined to the original image resolution with the coarse-to-fine module. Inspired by [50], we use a correlation-based approach for this purpose. For every coarse match $(\tilde{i},\tilde{j})$, we first locate its position $(\hat{i},\hat{j})$ at fine-level feature maps $\hat{F}^A$ and $\hat{F}^B$, and then crop two sets of local windows of size $w \times w$. A smaller LoFTR module then transforms the cropped features within each window by N_f times, yielding two transformed local feature maps $\hat{F}^A_{tr}(\hat{i})$ and $\hat{F}^B_{tr}(\hat{j})$ centered at $\hat{i}$ and $\hat{j}$, respectively. Then, we correlate the center vector of $\hat{F}^A_{tr}(\hat{i})$ with all vectors in $\hat{F}^B_{tr}(\hat{j})$ and thus produce a heatmap that represents the matching probability of each pixel in the neighborhood of $\hat{j}$ with $\hat{i}$. By computing expectation over the probability distribution, we get the final position $\hat{j}'$ with sub-pixel accuracy on I^B. Gathering all the matches $\{(\hat{i},\hat{j}')\}$ produces the final fine-level matches $\mathcal{M}_f$.

3.5. Supervision

The final loss consists of the losses for the coarse-level and the fine-level: $\mathcal{L} = \mathcal{L}_c + \mathcal{L}_f$.

Coarse-level Supervision. The loss function for the coarse-level is the negative log-likelihood loss over the confidence matrix $\mathcal{P}_c$ returned by either the optimal transport layer or the dual-softmax operator. We follow SuperGlue [37] to use camera poses and depth maps to compute the ground-truth labels for the confidence matrix during training. We define the ground-truth coarse matches $\mathcal{M}^{gt}_c$ as the mutual nearest neighbors of the two sets of ⅛-resolution grids. The distance between two grids is measured by the re-projection distance of their central locations. More details are provided in the supplementary. With the optimal transport layer, we use the same loss formulation as in [37]. When using dual-softmax for matching, we minimize the negative log-likelihood loss over the grids in $\mathcal{M}^{gt}_c$:

$$\mathcal{L}_c = -\frac{1}{|\mathcal{M}^{gt}_c|} \sum_{(\tilde{i},\tilde{j}) \in \mathcal{M}^{gt}_c} \log \mathcal{P}_c(\tilde{i},\tilde{j}) .$$

Fine-level Supervision. We use the ℓ_2 loss for fine-level refinement. Following [50], for each query point $\hat{i}$, we also measure its uncertainty by calculating the total variance $\sigma^2(\hat{i})$ of the corresponding heatmap. The target is to optimize the refined position that has low uncertainty, resulting in the final weighted loss function:

$$\mathcal{L}_f = \frac{1}{|\mathcal{M}_f|} \sum_{(\hat{i},\hat{j}') \in \mathcal{M}_f} \frac{1}{\sigma^2(\hat{i})} \left\| \hat{j}' - \hat{j}'_{gt} \right\|_2 ,$$

in which $\hat{j}'_{gt}$ is calculated by warping each $\hat{i}$ from $\hat{F}^A_{tr}(\hat{i})$ to $\hat{F}^B_{tr}(\hat{j})$ with the ground-truth camera pose and depth. We ignore $(\hat{i},\hat{j}')$ if the warped location of $\hat{i}$ falls out of the local window of $\hat{F}^B_{tr}(\hat{j})$ when calculating $\mathcal{L}_f$. The gradient is not backpropagated through $\sigma^2(\hat{i})$ during training.

3.6. Implementation Details

We train the indoor model of LoFTR on the ScanNet [7] dataset and the outdoor model on the MegaDepth [21] following [37]. On ScanNet, the model is trained using Adam with an initial learning rate of 1×10^{-3} and a batch size of 64. It converges after 24 hours of training on 64 GTX 1080Ti GPUs. The local feature CNN uses a modified version of ResNet-18 [12] as the backbone. The entire model is trained end-to-end with randomly initialized weights. N_c is set to 4 and N_f is 1. θ_c is chosen to 0.2. Window size w is equal to 5. $\tilde{F}^A_{tr}$ and $\tilde{F}^B_{tr}$ are upsampled and concatenated with $\hat{F}^A$ and $\hat{F}^B$ before passing through the fine-level LoFTR in the implementation. The full model with dual-softmax matching runs at 116 ms for a 640×480 image pair on an RTX 2080Ti. Under the optimal transport setup, we use three sinkhorn iterations, and the model runs at 130 ms. We refer readers to the supplementary material for more details of training and timing analyses.

4. Experiments

4.1. Homography Estimation

In the first experiment, we evaluate LoFTR on the widely adopted HPatches dataset [1] for homography estimation. HPatches contains 52 sequences under significant illumination changes and 56 sequences that exhibit large variation in viewpoints.

3.3. Establishing Coarse-level Matches

两种类型的可分辨率匹配层可以是ap−

在LoFTR中使用最佳运输（OT）层，如[37]或双软MAX算子[34,47]。转换特征之间的分数矩阵S首先通过S（i，j）= 1计算

使用OT，可用作部分分配问题的成本矩阵，如[37]中。我们还可以在S的两种尺寸上应用Softmax（在下降中被称为双软Max），以获得软互补邻匹配的概率。正式，当使用双软MAX时，匹配概率PC通过：

$$\mathcal{P}_c(i,j) = \mathrm{softmax}\left(\mathcal{S}(i,\cdot)\right)_j \cdot \mathrm{softmax}\left(\mathcal{S}(\cdot,j)\right)_i .$$

匹配选择。基于置信矩阵PC，我们选择具有高于θc阈值的置信度的匹配，并进一步强制实施互联邻（MNN）标准，该标准过滤可能的异常粗匹配。我们表示粗略级别的匹配预测：

$$\mathcal{M}_c = \{(\tilde{i},\tilde{j}) \mid \forall (\tilde{i},\tilde{j}) \in \mathrm{MNN}(\mathcal{P}_c),\ \mathcal{P}_c(\tilde{i},\tilde{j}) \geq \theta_c\}.$$

3.4. Coarse-to-Fine Module

建立粗匹配后，这些比赛是重新的

用粗致细模块被罚款到原始图像分辨率。灵感来自[50]，我们使用基于相关的相关方法来实现此目的。对于每个粗匹配（~i，~j），我们首先在细级特征映射f a和f b处定位其位置（i，j），然后裁剪两组尺寸W×w的本地窗口。然后，较小的LOFTR模块将每个窗口中的裁剪功能转换为NF次数，产生两个变换的本地特征映射F a

分别在I和J.然后，我们将F a的中心向量相关联

TR（i）与F B中的所有vectorTR（J）并因此产生

一种热图，其表示J的邻域中每个像素的匹配概率。通过计算在概率分布上的情况下，我们在IB上获得了具有子像素精度的最终位置J'。收集所有匹配{（i，j'）}生成最终的细级别mf。

3.5. Supervision

最终损失包括粗级的损失

and the fine-level: = c + f.

粗级监督。损失功能

粗级是通过最佳转换层或双软MAX运算符返回的通信矩形PC上的负值对数丢失。我们遵循超级胶水[37]使用相机姿势和深度地图，在培训期间计算置信矩阵的地面真理标签。我们定义了地面真理粗匹配MGT

两组的最近邻居的1/8分辨率

网格。通过其中心位置的重新投影距离来测量两个网格之间的距离。补充提供更多的脱尾。利用最佳传输层，我们使用与[37]中的相同的损耗配方。使用双软MAX进行匹配时，我们将在MGT中的网格上最小化最小化否定的对数似然丢失

$$\mathcal{L}_c = -\frac{1}{|\mathcal{M}^{gt}_c|} \sum_{(\tilde{i},\tilde{j}) \in \mathcal{M}^{gt}_c} \log \mathcal{P}_c(\tilde{i},\tilde{j}) .$$

细级监督。我们使用 2损失进行细级

细化。在[50]之后，对于每个查询点，我们

还通过计算相应热图的总方差σ2（i）来测量其不确定性。目标是为了选择具有低不确定性的精细位置，从而导致最终的加权损失功能：

$$\mathcal{L}_f = \frac{1}{|\mathcal{M}_f|} \sum_{(\hat{i},\hat{j}') \in \mathcal{M}_f} \frac{1}{\sigma^2(\hat{i})} \left\| \hat{j}' - \hat{j}'_{gt} \right\|_2 ,$$

in which jgt是通过从f a翘曲的计算来计算的 (i) to

FTR（J）与地面真理相机姿势和深度。我们

忽略（i，j'）如果我的翘曲位置从b的本地窗口中掉出来

TR（J）计算时。梯度不是

backpropaga

3.6. Implementation Details

我们训练Scannet上的Loftr的室内模型[7]

DataSet和Megadepth [21]的户外模型[21] fol−lowing

[37]。在Scannet上，使用adam训练模型，初始学习速率为1×10−3，批量大小为64.它会在64 GTX 1080TI

GPU上培训24小时后收敛。本地特征CNN使用Reset−18

[12]的修改版本作为骨干。整个型号培训结束于终端，随机初始化的权重。NC设置为4，NF为1. θ c被选为0.2。窗口尺寸w等于5. ~f a

使用F A和F

B之前，在实施中的细级LOFTR之前。双软MAX匹配的完整模型在RTX

2080TI上为640×480映像对运行在116毫秒。在最佳传输设置下，我们使用三个陷阱迭代，而模型在130毫秒运行。我们将读者推荐给补充材料，以便更多地进行培训和时序分析。

4. Experiments

4.1. Homography Estimation

在第一个实验中，我们在广泛的广泛评价Loftr

采用HPAPTES DataSet

[1]以获取同字估计。HPatches含有52个序列，在显着的亮度变化和56个序列中表现出大变化的观点。

Category	Method	Homography est. AUC			#matches
		@3px	@5px	@10px	
Detector-based	D2Net [11]+NN	23.2	35.9	53.6	0.2K
	R2D2 [32]+NN	50.6	63.9	76.8	0.5K
	DISK [47]+NN	52.3	64.9	78.9	1.1K
	SP [9]+SuperGlue [37]	53.9	68.3	81.7	0.6K
Detector-free	Sparse-NCNet [33]	48.9	54.2	67.1	1.0K
	DRC-Net [19]	50.6	56.2	68.3	1.0K
	LoFTR-DS	**65.9**	**75.6**	**84.6**	1.0K

Table 1: **Homography estimation on HPatches [7].** The AUC of the corner error in percentage is reported. The suffix DS indicates the differentiable matching with dual-softmax.

Category	Method	Pose estimation AUC		
		@5°	@10°	@20°
Detector-based	ORB [35]+GMS [2]	5.21	13.65	25.36
	D2-Net [11]+NN	5.25	14.53	27.96
	ContextDesc [27]+Ratio Test [26]	6.64	15.01	25.75
	SP [9]+NN	9.43	21.53	36.40
	SP [9]+PointCN [52]	11.40	25.47	41.41
	SP [9]+OANet [53]	11.76	26.90	43.85
	SP [9]+SuperGlue [37]	16.16	33.81	51.84
Detector-free	DRC-Net † [19]	7.69	17.93	30.49
	LoFTR-OT†	16.88	33.62	50.62
	LoFTR-OT	21.51	40.39	**57.96**
	LoFTR-DS	**22.06**	**40.8**	57.62

Table 2: **Evaluation on ScanNet [7] for indoor pose estimation.** The AUC of the pose error in percentage is reported. LoFTR improves the state-of-the-art methods by a large margin. †indicates models trained on MegaDepth. The suffixes OT and DS indicate differentiable matching with optimal transport and dual-softmax, respectively.

Category	Method	Pose estimation AUC		
		@5°	@10°	@20°
Detector-based	SP [9]+SuperGlue [37]	42.18	61.16	75.96
Detector-free	DRC-Net [19]	27.01	42.96	58.31
	LoFTR-OT	50.31	67.14	79.93
	LoFTR-DS	**52.8**	**69.19**	**81.18**

Table 3: **Evaluation on MegaDepth [21] for outdoor pose estimation.** Matching with LoFTR results in better performance in the outdoor pose estimation task.

Evaluation protocol. In every test sequence, one reference image is paired with the rest five images. All images are resized with shorter dimensions equal to 480. For each image pair, we extract a set of matches with LoFTR trained on MegaDepth [21]. We use OpenCV to compute the homography estimation with RANSAC as the robust estimator. To make a fair comparison to methods that produce different numbers of matches, we compute the corner error between the images warped with the estimated $\hat{\mathcal{H}}$ and the ground-truth $\mathcal{H}$ as a correctness identifier as in [9]. Following [37], we report the area under the cumulative curve (AUC) of the corner error up to threshold values of 3, 5, and 10 pixels, respectively. We report the results of LoFTR with a maximum of 1K output matches.

Baseline methods. We compare LoFTR with three categories of methods: 1) detector-based local features including R2D2 [32], D2Net [11], and DISK [47], 2) a detector-based local feature matcher, i.e., SuperGlue [37] on top of SuperPoint [9] features, and 3) detector-free matchers including Sparse-NCNet [33] and DRC-Net [19]. For local features, we extract a maximum of 2K features with which we extract mutual nearest neighbors as the final matches. For methods directly outputting matches, we restrict a maximum of 1K matches, same as LoFTR. We use the default hyperparameters in the original implementations for all the baselines.

Results. Tab. 1 shows that LoFTR notably outperforms other baselines under all error thresholds by a significant margin. Specifically, the performance gap between LoFTR and other methods increases with a stricter correctness threshold. We attribute the top performance to the larger number of match candidates provided by the detector-free design and the global receptive field brought by the Transformer. Moreover, the coarse-to-fine module also contributes to the estimation accuracy by refining matches to a sub-pixel level.

4.2. Relative Pose Estimation

Datasets. We use ScanNet [7] and MegaDepth [21] to demonstrate the effectiveness of LoFTR for pose estimation in indoor and outdoor scenes, respectively.

ScanNet contains 1613 monocular sequences with ground truth poses and depth maps. Following the procedure from SuperGlue [37], we sample 230M image pairs for training, with overlap scores between 0.4 and 0.8. We evaluate our method on the 1500 testing pairs from [37]. All images and depth maps are resized to 640×480. This dataset contains image pairs with wide baselines and extensive texture-less regions.

MegaDepth consists of 1M internet images of 196 different outdoor scenes. The authors also provide sparse reconstruction from COLMAP [40] and depth maps computed from multi-view stereo. We follow DISK [47] to only use the scenes of "Sacre Coeur" and "St. Peter's Square" for validation, from which we sample 1500 pairs for a fair comparison. Images are resized such that their longer dimensions are equal to 840 for training and 1200 for validation. The key challenge on MegaDepth is matching under extreme viewpoint changes and repetitive patterns.

Evaluation protocol. Following [37], we report the AUC of the pose error at thresholds $(5°, 10°, 20°)$, where the pose error is defined as the maximum of angular error in rotation and translation. To recover the camera pose, we solve the essential matrix from predicted matches with RANSAC. We don't compare the matching precisions between LoFTR and other detector-based methods due to the lack of a well-

Category	Method	Homography est. AUC			#matches
		@3px	@5px	@10px	
Detector-based	D2Net [11]+NN	23.2	35.9	53.6	0.2K
	R2D2 [32]+NN	50.6	63.9	76.8	0.5K
	DISK [47]+NN	52.3	64.9	78.9	1.1K
	SP [9]+SuperGlue [37]	53.9	68.3	81.7	0.6K
Detector-free	Sparse-NCNet [33]	48.9	54.2	67.1	1.0K
	DRC-Net [19]	50.6	56.2	68.3	1.0K
	LoFTR-DS	**65.9**	**75.6**	**84.6**	1.0K

表1：HPAPTES上的同字估计[7]。报告了百分比的角落错误的AUC。这

后缀DS表示与双模糊的可差匹配。

评估方案。在每个测试序列中，一个参考图像与其余五个图像配对。所有图像都被重大大小，尺寸等于480。对于每个图像对，我们将一组与Megadepth培训的Loftr
[21]提取了一组匹配。我们使用OpenCV与Ransac作为稳健估计计算的同源估计计算。为了对产生不同数量的匹配数的方法进行公平比较，我们将使用估计的H的图像和地面实际H计算图像之间的拐角误差作为[9]中的正确性标识符。在[37]之后，我们在角误差的累积曲线（AUC）下的区域报告到阈值3,5和10像素的阈值，重新观察。我们报告了LOFTR的结果，最多1K输出匹配。

Category	Method	Pose estimation AUC		
		@5°	@10°	@20°
Detector-based	ORB [35]+GMS [2]	5.21	13.65	25.36
	D2-Net [11]+NN	5.25	14.53	27.96
	ContextDesc [27]+Ratio Test [26]	6.64	15.01	25.75
	SP [9]+NN	9.43	21.53	36.40
	SP [9]+PointCN [52]	11.40	25.47	41.41
	SP [9]+OANet [53]	11.76	26.90	43.85
	SP [9]+SuperGlue [37]	16.16	33.81	51.84
Detector-free	DRC-Net † [19]	7.69	17.93	30.49
	LoFTR-OT†	16.88	33.62	50.62
	LoFTR-OT	21.51	40.39	**57.96**
	LoFTR-DS	**22.06**	**40.8**	57.62

表2：扫描仪的评估[7]用于室内姿势计时。百分比的姿势错误的AUC被重新移植。LOFTR通过大幅度提高最先进的方法。 表示在Megadepth培训的模型。后缀OT和DS表示分别与最佳传输和双软MAX的可微差匹配。

Category	Method	Pose estimation AUC		
		@5°	@10°	@20°
Detector-based	SP [9]+SuperGlue [37]	42.18	61.16	75.96
Detector-free	DRC-Net [19]	27.01	42.96	58.31
	LoFTR-OT	50.31	67.14	79.93
	LoFTR-DS	**52.8**	**69.19**	**81.18**

表3：户外姿势估计的Megadepth
[21]评估。与Loftr的匹配导致户外姿势估算任务中更好的性能。

基线方法。我们将Loftr与三个凯特比较 –
方法的理念：1）基于探测器的本地特征包括R2D2
[32]，D2NET
[11]和磁盘[47]，2）基于探测器的本地特征匹配器，即超级点上的SuperGlue
[37][9]功能和3）无液位匹配器，包括稀疏–NCNet
[33]和DRC–Net
[19]。对于本地特征，我们提取最多2K的特征，我们将相互最近的邻居提取为最终匹配。对于直接输出匹配的方法，我们限制了1K匹配的最大值，与Loftr相同。我们在所有基准的原始实现中使用默认的HyperParameters。

结果。标签。图1示出了Loftr显着优于优势
其他基线在所有错误阈值下由一个重要的边缘。具体地，LOFTR和其他方法之间的性能间隙随着更严格的正确性阈值而增加。我们将顶部性能归因于探测器设计提供的更多匹配候选者，并由转映射带来的全球接收领域。而且，粗致细模块也是如此

通过将匹配达到子像素级别来悼念估计精度。

4.2. Relative Pose Estimation

数据集。我们使用Scannet [7]和Megadepth [21]到
展示Loftr对姿势估计的有效性

在室内和室外场景中。

Scannet含有1613个单眼序列
地面真理姿势和深度地图。在Superglue
[37]的过程之后，我们将230M图像对进行培训，重叠分数在0.4和0.8之间。我们在[37]的1500检测对上评估我们的方法。所有图像和深度映射都调整为640×480.此数据集包含具有宽基线的图像对，并延长纹理区域。

Megadepth由196年的1M互联网图像不同 –
欧洲户外场景。作者还提供从ColMap
[40]的稀疏重建，并从多视图立体声计算的深度映射。我们关注磁盘[47]只使用 "Sacre
Coeur" 和 "St." 的场景。彼得广场 "用于验证，从中抽取1500对进行公平的乐趣。修改了图像，使得其较长的尺寸等于840，用于训练和1200用于验证。对Megadepth的关键挑战在特雷特观点变化和重复模式下匹配。

评估方案。在[37]之后，我们在阈值（5 ，10英寸，20英寸）上报告姿势误差的AUC，其中姿势误差被定义为旋转和翻译中的角度误差的最大值。要恢复相机姿势，我们可以通过Ransac解决预测匹配的基本矩阵。由于缺乏良好的缺乏，我们不会比较Loftr和基于探测器的方法之间的匹配精度

Method	Day	Night
	(0.25m,2°) / (0.5m,5°) / (1.0m,10°)	
Local Feature Evaluation on Night-time Queries		
R2D2 [32]+NN	-	71.2 / 86.9 / 98.9
LISRD [31]+SP [9]+AdaLam [4]	-	**73.3** / 86.9 / 97.9
ISRF [29]+NN	-	69.1 / 87.4 / 98.4
SP [9]+SuperGlue [37]	-	**73.3** / 88.0 / 98.4
LoFTR-DS	-	72.8 / **88.5** / **99.0**
Full Visual Localization with HLoc		
SP [9]+SuperGlue [37]	**89.8** / **96.1** / **99.4**	77.0 / **90.6** / **100.0**
LoFTR-OT	88.7 / 95.6 / 99.0	**78.5** / **90.6** / 99.0

Table 4: **Visual localization evaluation on the Aachen Day-Night [54] benchmark v1.1**. The evaluation results on both the local feature evaluation track and the full visual localization track are reported.

Method	DUC1	DUC2
	(0.25m,10°) / (0.5m,10°) / (1.0m,10°)	
ISRF [29]	39.4 / 58.1 / 70.2	41.2 / 61.1 / 69.5
KAPTURE [14]+R2D2 [32]	41.4 / 60.1 / 73.7	47.3 / 67.2 / 73.3
HLoc [36]+SP [9]+SuperGlue [37]	**49.0** / 68.7 / 80.8	53.4 / **77.1** / 82.4
HLoc [36]+**LoFTR-OT**	47.5 / **72.2** / **84.8**	**54.2** / 74.8 / **85.5**

Table 5: **Visual localization evaluation on the InLoc [41] benchmark**.

Method	Pose estimation AUC		
	@5°	@10°	@20°
1) replace LoFTR with convolution	14.98	32.04	49.92
2) 1/16 coarse-resolution + 1/4 fine-resolution	16.75	34.82	54.0
3) positional encoding per layer	18.02	35.64	52.77
4) larger model with $N_c = 8$, $N_f = 2$	**20.87**	40.23	57.56
Full ($N_c = 4$, $N_f = 1$)	20.06	**40.8**	**57.62**

Table 6: **Ablation study.** Five variants of LoFTR are trained and evaluated both on the ScanNet dataset.

defined metric (e.g., matching score or recall [13, 30]) for detector-free image matching methods. We consider DRC-Net [19] as the state-of-the-art method in detector-free approaches [34, 33].

Results of indoor pose estimation. LoFTR achieves the best performance in pose accuracy compared to all competitors (see Tab. 2 and Fig. 5). Pairing LoFTR with optimal transport or dual-softmax as the differentiable matching layer achieves comparable performance. Since the released model of DRC-Net† is trained on MegaDepth, we provide the results of LoFTR† trained on MegaDepth for a fair comparison. LoFTR† also outperforms DRC-Net† by a large margin in this evaluation (see Fig. 5), which demonstrates the generalizability of our model across datasets.

Results of Outdoor Pose Estimation. As shown in Tab. 3, LoFTR outperforms the detector-free method DRC-Net by 61% at AUC@10°, demonstrating the effectiveness of the Transformer. For SuperGlue, we use the setup from the open-sourced localization toolbox HLoc [36]. LoFTR outperforms SuperGlue by a large margin (13% at AUC@10°), which demonstrates the effectiveness of the detector-free design. Different from indoor scenes, LoFTR-DS performs better than LoFTR-OT on MegaDepth. More qualitative results can be found in Fig. 5.

4.3. Visual Localization

Visual Localization. Besides achieving competitive performance for relative pose estimation, LoFTR can also benefit visual localization, which is the task to estimate the 6-DoF poses of given images with respect to the corresponding 3D scene model. We evaluate LoFTR on the Long-Term Visual Localization Benchmark [43] (referred to as VisLoc benchmark in the following). It focuses on benchmarking visual localization methods under varying conditions, e.g., day-night changes, scene geometry changes, and indoor scenes with plenty of texture-less areas. Thus, the visual localization task relies on highly robust image matching methods.

Evaluation. We evaluate LoFTR on two tracks of VisLoc that consist of several challenges. First, the "visual localization for handheld devices" track requires a full localization pipeline. It benchmarks on two datasets, the Aachen-Day-Night dataset [38, 54] concerning outdoor scenes and the InLoc [41] dataset concerning indoor scenes. We use open-sourced localization pipeline HLoc [36] with the matches extracted by LoFTR. Second, the "local features for long-term localization" track provides a fixed localization pipeline to evaluate the local feature extractors themselves and optionally the matchers. This track uses the Aachen v1.1 dataset [54]. We provide the implementation details of testing LoFTR on VisLoc in the supplementary material.

Results. We provide evaluation results of LoFTR in Tab. 4 and Tab. 5. We have evaluated LoFTR pairing with either the optimal transport layer or the dual-softmax operator and report the one with better results. LoFTR-DS outperforms all baselines in the local feature challenge track, showing its robustness under day-night changes. Then, for the visual localization for handheld devices track, LoFTR-OT outperforms all published methods on the challenging InLoc dataset, which contains extensive appearance changes, more texture-less areas, symmetric and repetitive elements. We attribute the prominence to the use of the Transformer and the optimal transport layer, taking advantage of global information and jointly bringing global consensus into the final matches. The detector-free design also plays a critical role, preventing the repeatability problem of detector-based methods in low-texture regions. LoFTR-OT performs on par with the state-of-the-art method SuperPoint + SuperGlue on night queries of the Aachen v1.1 dataset and slightly worse on the day queries.

Method	Day	Night
	(0.25m,2°) / (0.5m,5°) / (1.0m,10°)	
夜间查询局部特征评估R2D2 [32] + NN	-	71.2 / 86.9 / 98.9
LISRD [31]+SP [9]+AdaLam [4]	-	**73.3** / 86.9 / 97.9
ISRF [29]+NN	-	69.1 / 87.4 / 98.4
SP [9]+SuperGlue [37]	-	**73.3** / 88.0 / 98.4
LoFTR-DS	-	72.8 / **88.5** / **99.0**
HLOC SP [9] + SuperGlue [37]	89.8 / 96.1 / 99.4	77.0 / 90.6 / 100.0
LoFTR-OT	88.7 / 95.6 / 99.0	**78.5** / **90.6** / 99.0

表4：亚琛日夜的视觉本地化评估[54]基准V1.1。报告了本地特征评估轨道和完整的视觉定位轨道的评估结果。

Method	DUC1	DUC2
	(0.25m,10°) / (0.5m,10°) / (1.0m,10°)	
ISRF [29]	39.4 / 58.1 / 70.2	41.2 / 61.1 / 69.5
KAPTURE [14]+R2D2 [32]	41.4 / 60.1 / 73.7	47.3 / 67.2 / 73.3
HLoc [36]+SP [9]+SuperGlue [37]	**49.0** / 68.7 / 80.8	53.4 / **77.1** / 82.4
HLoc [36]+**LoFTR-OT**	47.5 / **72.2** / **84.8**	**54.2** / 74.8 / **85.5**

表5：Inloc上的视觉本地化评估[41]基准。

Method	Pose estimation AUC		
	@5°	@10°	@20°
1) replace LoFTR with convolution	14.98	32.04	49.92
2) 1/16 coarse-resolution + 1/4 fine-resolution	16.75	34.82	54.0
3）每层位置编码18.02 35.64 52.77			
4) larger model with $N_c = 8$, $N_f = 2$	**20.87**	40.23	57.56
Full ($N_c = 4$, $N_f = 1$)	20.06	**40.8**	**57.62**

表6：消融研究。leoftr的五种变种是在Scannet DataSet上培训和评估。

用于检测器图像匹配方法的定义度量（例如，匹配分数或召回[13,30]）。我们认为DRC-Net [19]作为在无检测器的AP - PROACH中的最新方法[34,33]。

室内姿态估计结果。Loftr实现了这一点与所有组合器相比，姿势精度的最佳性能（参见图2和图5）。配对LoFTR与Opti-Mal传输或双软MAX，因为可视匹配层实现了可比性的性能。由于DRC-Net 的释放模型在Megadepth培训，我们提供了Loftr 在Megadepth培训的结果，以进行公平的乐趣。LOFTR 在该评估中通过大型余量（参见图5），展示了DRC-NET ，这展示了我们模型跨数据集的模型的概括性。

户外姿势估计的结果。如标签所示。3，LOFTR在AUC @ 10° 处以61％的探测器的方法DRC-NET优于61％，展示了变压器的有效性。对于SuperGlue，我们使用从开源本地化工具箱HLOC [36]的设置。LOFTR在大型余量（AUC @ 10° 的13％）外，展示了探测器设计的有效性。与室内场景不同，LOFTR-DS在Megadepth上的Loftr-OT更好地执行。更具定性的重新调整可以在图5中找到。5。

4.3. Visual Localization

视觉本地化。除了实现竞争力

相对姿态估计的Formance，Loftr还可以是基于eFIT视觉定位，这是估计相对于对应3D场景模型的给定图像的6-DOF姿势的任务。我们评估长期视觉定位基准测试[43]的Loftr（以下称为VIS-LOC基准）。它侧重于在不同条件下的基准视觉定位方法，例如，日夜更改，场景几何变化，以及具有大量纹理区域的门场景。因此，vi-sual本地化任务依赖于高强度较强的图像匹配方法。

评估。我们评估了两个Visloc曲目的Loftr，包括几个挑战。首先，“手持设备的视觉局部”跟踪需要一个完整的本地化管道。它在两个数据集上基准测试，即关于室外场景和Inloc [41] DataSet有关室内场景的acken-Day-Night DataSet [38,54]。我们

使用Open-Source Loceization Pipeline HLOC [36]，使用Loftr提取的匹配。其次，“长期定位的本地特征”轨道提供了一个固定的本地化管道，以评估本地特征提取器主题和可选的匹配器。这首曲目使用

AACHEN V1.1数据集[54]。我们在辅助材料中提供了在Visloc上测试Loftr的实施细节。

结果。我们在标签中提供Loftr的评估结果。 4和标签。5.我们已经评估了LOFTR配对与最佳传输层或双软MAX运算符，并报告具有更好结果的结果。LOFTR-DS优于本地特征挑战轨道中的所有基准，显示其在日夜变化的鲁棒性。然后，对于手持设备轨道的vi-sual本地化，Loftr-ot-Ot优于所有已发布的方法，即在LOC数据集中具有挑战性，其中包含广泛的外观变化，更少于纹理的区域，对称和重复元素。我们归因于使用变压器和最佳运输层的使用，利用全球信息，并将全球共识与最终比赛共同达成。探测器的设计也起到了批准的作用，防止了低纹理区域中基于探测器的方法的重复性问题。LOFTR-OT与最先进的方法SUPERPOINT + SU-PERGLUE在夜间查询AACHEN V1.1数据集的情况下执行，并且在当天查询时略差。

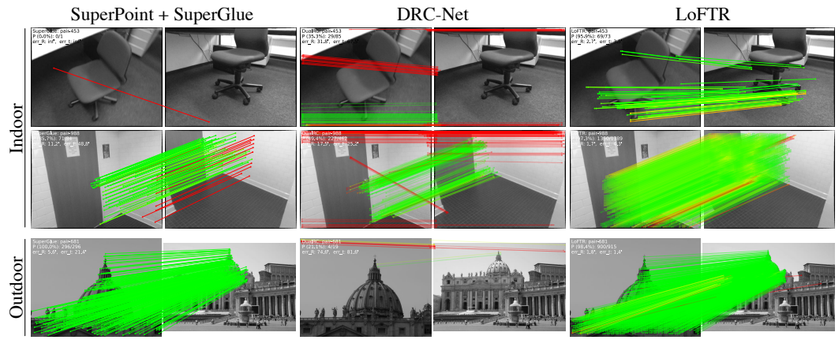

Figure 5: **Qualitative results.** LoFTR is compared to SuperGlue [37] and DRC-Net [19] in indoor and outdoor environments. LoFTR obtains more correct matches and fewer mismatches, successfully coping with low-texture regions and large viewpoint and illumination changes. The red color indicates epipolar error beyond 5×10^{-4} for indoor scenes and 1×10^{-4} for outdoor scenes (in the normalized image coordinates). More qualitative results can be found on the project webpage.

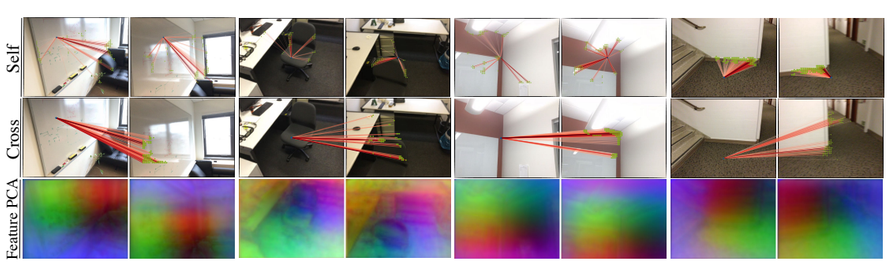

Figure 6: **Visualization of self and cross attention weights and the transformed features.** In the first two examples, the query point from the low-texture region is able to aggregate the surrounding global information flexibly. For instance, the point on the chair is looking at the edge of the chair. In the last two examples, the query point from the distinctive region can also utilize the richer information from other regions. The feature visualization with PCA further shows that LoFTR learns a position-dependent feature representation.

4.4. Understanding LoFTR

Ablation Study. To fully understand the different modules in LoFTR, we evaluate five different variants with results shown in Tab. 6: 1) Replacing the LoFTR module by convolution with a comparable number of parameters results in a significant drop in AUC as expected. 2) Using a smaller version of LoFTR with 1/16 and 1/4 resolution feature maps at the coarse and fine level, respectively, results in a running time of 104 ms and a degraded pose estimation accuracy. 3) Using DETR-style [3] Transformer architecture which has positional encoding at each layer, leads to a noticeably declined result. 4) Increasing the model capacity by doubling the number of LoFTR layers to $N_c = 8$ and $N_f = 2$ barely changes the results. We conduct these experiments using the same training and evaluation protocol as indoor pose estimation on ScanNet with an optimal transport layer for matching.

Visualizing Attention. We visualize the attention weights in Fig. 6.

5. Conclusion

This paper presents a novel detector-free matching approach, named LoFTR, that can establish accurate semi-dense matches with Transformers in a coarse-to-fine manner. The proposed LoFTR module uses the self and cross attention layers in Transformers to transform the local features to be context- and position-dependent, which is crucial for LoFTR to obtain high-quality matches on indistinctive regions with low-texture or repetitive patterns. Our experiments show that LoFTR achieves state-of-the-art performances on relative pose estimation and visual localization on multiple datasets. We believe that LoFTR provides a new direction for detector-free methods in local image feature matching and can be extended to more challenging scenarios, e.g., matching images with severe seasonal changes.

Acknowledgement. The authors would like to acknowledge the support from the National Key Research and Development Program of China (No. 2020AAA0108901), NSFC (No. 61806176), and ZJU-SenseTime Joint Lab of 3D Vision.

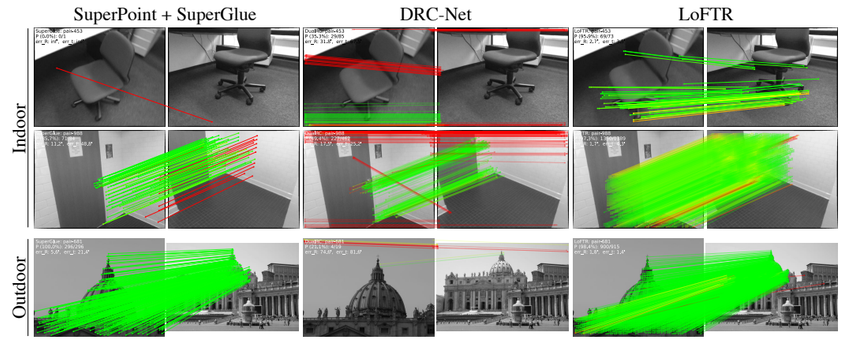

图5：定性结果。将LOFTR与SuperGlue [37]和DRC−Net [19]进行比较。LOFTR获得更正确的匹配和更少的不匹配，成功地应对低纹理区域和大观点和照明变化。红颜色表示室内场景超过5×10−4的末极误差，1×10−4

对于户外场景（在归一化图像坐标中）。项目网页上可以找到更多的定性结果。

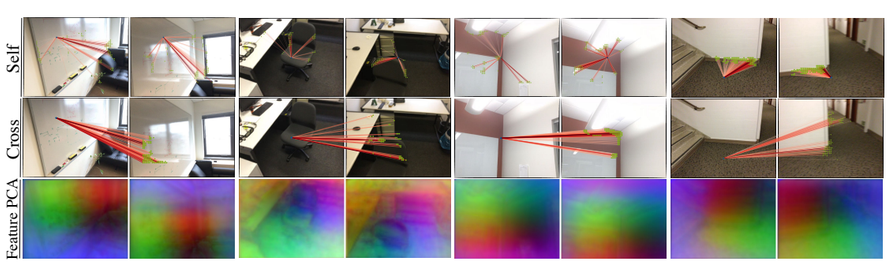

图6：自我和跨注意力的可视化和转换特征。在前两个示例中，来自低纹理区域的查询点能够灵活地聚合周围的全局信息。例如，椅子上的点看着椅子的边缘。在最后两个示例中，来自独特区域的查询点还可以利用来自其他区域的更丰富的信息。具有PCA的功能可视化进一步显示LOFTR学习依赖于位置的特征表示。

4.4. Understanding LoFTR

消融研究。要在Loftr中完全理解不同的模块，我们会评估包含选项卡中显示的结果的五种不同的变体。6：1）通过使用可比较数量的参数进行概览更换LOFTR模块导致AUC的显着下降如预期。2）使用较小版本的Loftr分别在粗略和精细级别的1/16和1/4分辨率的特征映射，导致运行时间为104ms和降级的姿态估计精度。3）使用在每层具有位置编码的DETR式[3]变压器架构，导致显着脱模的结果。4）通过将Loftr层的数量加倍增加模型容量= 8，NF =
2几乎没有改变结果。我们使用相同的培训和评估协议进行这些实验，作为剪刀上的室内姿势估计，具有匹配的最佳传输层。

可视化注意力。我们可视化图6中的注意重量。6。

5. Conclusion

本文提出了一种免费探测器的匹配AP−
Proach，命名Loftr，可以建立准确的半密度与变压器在粗良好的人身上。所提出的LOFTR模块使用变压器中的自我和跨关注层来改造局部功能，以进行上下文和位置依赖，这对于Loftr来说至关重要，以获得具有低质量或重复模式的神秘区域的高质量匹配。我们的专题表明Loftr在多个数据集上的相对姿势估计和视觉本地化上实现了最先进的性能。我们认为Loftr在本地图像特征匹配中提供了无检测器方法的新方向，并且可以扩展到更具挑战性的场景，例如，匹配具有严重季节性变化的图像。确认。作者想诚实 −

边缘来自中国国家重点研发计划的支持（No.20AAA0101901），NSFC（No.61806176）和3D
Vision的Zju−Sensetime联合实验室。

References

[1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *CVPR*, 2017. 5

[2] JiaWang Bian, Wen-Yan Lin, Yasuyuki Matsushita, Sai-Kit Yeung, Tan-Dat Nguyen, and Ming-Ming Cheng. GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence. In *CVPR*, 2017. 6

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3, 4, 8

[4] Luca Cavalli, Viktor Larsson, Martin Ralf Oswald, Torsten Sattler, and Marc Pollefeys. Handcrafted Outlier Detection Revisited. In *ECCV*, 2020. 7

[5] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *ICLR*, 2021. 3

[6] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. *NeurIPS*, 2016. 2

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 5, 6

[8] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Toward geometric deep slam. *arXiv:1707.07410*. 2

[9] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *CVPRW*, 2018. 2, 6, 7

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 3

[11] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A trainable cnn for joint detection and description of local features. *CVPR*, 2019. 2, 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[13] Jared Heinly, Enrique Dunn, and Jan-Michael Frahm. Comparative evaluation of binary features. In *ECCV*, 2012. 7

[14] Martin Humenberger, Yohann Cabon, Nicolas Guerin, Julien Morat, Jérôme Revaud, Philippe Rerole, Noé Pion, Cesar de Souza, Vincent Leroy, and Gabriela Csurka. Robust Image Retrieval-based Visual Localization using Kapture. *arXiv:2007.13867*. 7

[15] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. COTR: Correspondence Transformer for Matching Across Images, 2021. 2

[16] Chaitanya Joshi. Transformers are Graph Neural Networks. https://thegradient.pub/transformers-are-graph-neural-networks/, 2020. 2

[17] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *ICML*, 2020. 3, 4

[18] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *ICLR*, 2020. 3

[19] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. *NeurIPS*, 2020. 1, 2, 6, 7, 8

[20] Zhaoshuo Li, Xingtong Liu, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting Stereo Depth Estimation From a Sequence-to-Sequence Perspective with Transformers. *arXiv:2011.02910*. 3

[21] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 5, 6

[22] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature Pyramid Networks for Object Detection. *CVPR*, 2017. 3

[23] Ce Liu, Jenny Yuen, and Antonio Torralba. SIFT Flow: Dense correspondence across scenes and its applications. *T-PAMI*, 2010. 2

[24] X. Liu, Y. Zheng, B. Killeen, M. Ishii, G. D. Hager, R. H. Taylor, and M. Unberath. Extremely Dense Point Correspondences Using a Learned Feature Descriptor. In *CVPR*, 2020. 2

[25] Yuan Liu, Zehong Shen, Zhixuan Lin, Sida Peng, Hujun Bao, and Xiaowei Zhou. GIFT: Learning transformation-invariant dense visual descriptors via group cnns. *NeurIPS*, 2019. 2

[26] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 2, 6

[27] Zixin Luo, Tianwei Shen, Lei Zhou, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. ContextDesc: Local Descriptor Augmentation with Cross-Modality Context. *CVPR*, 2019. 6

[28] Zixin Luo, Lei Zhou, Xuyang Bai, Hongkai Chen, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. ASLFeat: Learning local features of accurate shape and localization. In *CVPR*, 2020. 2

[29] Iaroslav Melekhov, Gabriel J Brostow, Juho Kannala, and Daniyar Turmukhambetov. Image Stylization for Robust Features. *arXiv:2008.06959*. 7

[30] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *T-PAMI*, 2005. 7

[31] Rémi Pautrat, Viktor Larsson, Martin R Oswald, and Marc Pollefeys. Online Invariance Selection for Local Feature Descriptors. In *ECCV*, 2020. 7

[32] Jerome Revaud, Philippe Weinzaepfel, César De Souza, Noe Pion, Gabriela Csurka, Yohann Cabon, and Martin Humenberger. R2D2: repeatable and reliable detector and descriptor. *NeurIPS*, 2019. 2, 6, 7

[33] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Efficient neighbourhood consensus networks via submanifold sparse convolutions. In *ECCV*, 2020. 1, 2, 6, 7

References

[1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krys-田mikolajczyk。HPatches：由手工制作和学习的本地描述符的基准和评估。在CVPR，2017年.5

[2] JiaWang Bian, Wen-Yan Lin, Yasuyuki Matsushita, Sai-Kit Y EU那个, tan−DAT Nguyen, and Ming−Ming Cheng. GMs: 基于网格的运动统计，用于快速，超强特征功能。在CVPR，2017年6月6日

[3]尼古拉斯·索利，弗朗西斯科萨马，加布里埃尔·斯诺伊，尼古拉斯 Usunier，Alexander Kirillov和Sergey Zagoruyko。具有变压器的端到端对象检测。在ECCV，2020年3,4,8

[4] Luca Cavalli, Viktor Larsson, Martin Ralf Oswald, Torsten Sattler和Marc Pollefeys。被手绘异常检测重新审视。在ECCV，2020年.7

[5] Krzysztof Choromanski, Valerii Likhosherstov, David Do-韩，兴友歌，安德烈·戈纳，塔玛萨罗斯，彼得霍金斯，贾德戴维斯，阿昔德·莫希迪因，卢卡斯凯瑟，等。表演者重新思考。ICLR，2021. 3

[6] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. *NeurIPS*, 2016. 2

[7] Angela Day，Angel Q. Chang，Manolis Sail，Maize Hull−Ber，Thomas Funkhouser和Matthiasnie ner。SCANNET：室内场景的丰富注释的3D重建。在 *CVPR*, 2017. 5, 6

[8]丹尼尔撒上了，Tomasz Malisiewicz和Andrew Rabi−诺奇。走向几何深猛击。arxiv：1707.07410。2

[9]丹尼尔撒上了，Tomasz Malisiewicz，Andrew Rabi−诺奇。超级点：自我监督的兴趣点检测和描述。在CVPRW，2018. 2,6,7

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn，Xiaohua Zhai，Thomas Unterthiner，Mostafa dehghani，Matthias Minderer，Georg Heigold，Syl−Vain Gelly，等。图像值为16x16字：用于图像识别的转换器。ICLR，2021. 3

[11] Mihai Dusmanu，Ignacio Rocco，Tomas Pajdla，Marc抛光 − Feys，Josef Sivic，Akihiko Torii和Torsten Sattler。D2−NET：用于联合检测和局部特征描述的可培训CNN。CVPR，2019. 2,6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 图像识别的深度剩余学习。在CVPR，2016. 5

[13] Jared Heinly，Enrique Dunn和Jan−Michael Frahm。com二元特征的比例评估。在ECCV，2012年.7

[14] Martin Humenberger, Yohann Cabon, Nicolas Guerin, Julien Morat，J'erome Revaud，Philippe Rerole，No'e Pion，Cesar de Souza，Vincent Leroy和Gabriela Csurka。强壮的 基于图像检索的视觉本地化使用kapurture。arxiv：2007.13867。7.

[15] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, 和kwang moo yi。COTR：用于跨图像匹配的函件变压器，2021. 2

[16] Chaitanya Joshi。变形金刚是图形神经网络。https://thegradient.pub/transformers-are-graph-neural-networks/, 2020. 2

[17] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and Franc oisfleuret。变形金刚是RNNS：具有线性关注的快速自动预测变压器。在ICML，2020年.3，4

[18] Nikita Kitaev， ukaszKaiser和Anselm Levskaya。关于−former: The efficient transformer. *ICLR*, 2020. 3

[19] xing会l i, Kai Han, Shu的Ali, and Victor PR ISA car IU. dual−resolution correspondence networks. *NeurIPS*, 2020. 1, 2, 6, 7, 8

[20] Zhao说L i, xing痛l IU, Francis X Creighton, Russell H 泰勒和马蒂亚斯·迪伯拉斯。重新审视立体声深度与变压器的序列到序列透视图估计。arxiv：2011.02910。3.

[21]郑琪李和诺亚仍然嗅之以鼻。Megadepth：学习单身 − 查看来自互联网照片的深度预测。在CVPR，2018年5,6

[22] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan，以及Serge J. Ippie。具有用于对象检测的金字塔网络。CVPR，2017. 3

[23]刘，珍妮元和安东尼奥托拉尔巴。筛选流程：跨场景及其应用的密集对应。T− PAMI，2010年.2

[24] X. Liu, Y. Zheng, B. Killeen, M. Ishii, G. D. Hager, R. H. 泰勒，和M. unberath。使用学习功能描述符的极其密度的点符号。在CVPR，2020年.2

[25] Yuan Liu, Zehong Shen, Zhixuan Lin, Sida Peng, Hujun Bao, 和小威周。礼品：通过组CNNS学习变换 − 不变的密集视觉描述符。Neurips，2019. 2

[26] David G Lowe。规模的独特图像特征 − invariant keypoints. *IJCV*, 2004. 2, 6

[27] Zixin Luo, Tianwei Shen, Lei Zhou, Jiahui Zhang, Yao Yao, 史伟李，田芳，龙泉。ContextDESC：使用跨模型上下文扩大的Lo−Chal描述符增强。CVPR，2019. 6

[28] Zixin Luo, Lei Zhou, Xuyang Bai, Hongkai Chen, Jiahui 张，姚瑶，石威李，田芳，龙泉。ASLFeat：学习局部特征精确的形状和电池化。在CVPR，2020年.2

[29] Iaroslav Melekhov, Gabriel J Brostow, Juho Kannala, and Daniyar Turmukhambetov。稳健的图像程式化 Features. *arXiv:2008.06959*. 7

[30] Krystian Mikolajczyk and Cordelia Schmid. A performance 评估本地描述符。T−PAMI，2005. 7

[31] R'emi Pautrat，Viktor Larsson，Martin R Oswald和Marc 波尔菲伊斯。对本地特征去编程器的在线不变性选择。在ECCV，2020年.7

[32] Jerome Revaud, Philippe Weinzaepfel, C esar De Souza, Noe Pion，Gabriela Csurka，Yohann Cabon和Martin Humen−Berger。R2D2：可重复可靠的探测器和描述。Neurips，2019. 2,6,7

[33] Ignacio rocco，Relja Arandjelovi'c和Josef Sivic。高效的邻居共识网络通过Submanifold稀疏卷曲。在ECCV，2020年。1,2,6,7

[34] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. *NeurIPS*, 2018. 1, 2, 5, 7

[35] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *ICCV*, 2011. 2, 6

[36] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 7

[37] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 1, 2, 4, 5, 6, 7, 8

[38] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt. Image Retrieval for Image-Based Localization Revisited. In *BMVC*, 2012. 7

[39] Tanner Schmidt, Richard Newcombe, and Dieter Fox. Self-supervised visual descriptor learning for dense correspondence. *RAL*, 2016. 2

[40] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-Motion revisited. In *CVPR*, 2016. 6

[41] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. InLoc: Indoor visual localization with dense matching and view synthesis. In *CVPR*, 2018. 7

[42] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *arXiv:2009.06732*. 3

[43] Carl Toft, Will Maddern, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, et al. Long-Term Visual Localization Revisited. *T-PAMI*, 2020. 7

[44] Prune Truong, Martin Danelljan, L. Gool, and R. Timofte. Learning Accurate Dense Correspondences and When to Trust Them. *ArXiv*, abs/2101.01710, 2021. 2

[45] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. GOCor: Bringing Globally Optimized Correspondence Volumes into Your Neural Network. In *NeurIPS*, 2020. 2

[46] Prune Truong, Martin Danelljan, and Radu Timofte. GLU-Net: Global-Local Universal Network for dense flow and correspondences. In *CVPR*, 2020. 2

[47] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. DISK: Learning local features with policy gradient. *NeurIPS*, 2020. 2, 5, 6

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 2, 3

[49] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-Deeplab: Stand-alone axial-attention for panoptic segmentation. In *ECCV*, 2020. 3

[50] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. Learning feature descriptors using camera pose supervision. In *ECCV*, 2020. 5

[51] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. LIFT: Learned invariant feature transform. In *ECCV*, 2016. 2

[52] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *CVPR*, 2018. 6

[53] Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning Two-View Correspondences and Geometry Using Order-Aware Network. *ICCV*, 2019. 6

[54] Zichao Zhang, Torsten Sattler, and Davide Scaramuzza. Reference Pose Generation for Long-term Visual Localization via Learned Features and View Synthesis. *IJCV*, 2020. 7

[34] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii，Tomas Pajdla和Josef Sivic。邻里Consus网络。Neurips，2018. 1,2,5,7

[35] Ethan Rublee，Vincent Rabaud，Kurt Konolige和Gary Bradski。ORB：筛选或冲浪的有效替代品。在ICCV，2011年.2,6

[36] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk。从粗略精细：大规模的强大的分层定位。在CVPR，2019年.7

[37] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, 和安德鲁·拉比维奇。Superglue：学习功能与图形神经网络匹配。在CVPR，2020年。1,2,4,5,6,7,8

[38] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt。基于图像的本地化的图像检索重新访问。在BMVC，2012年.7

[39] Tanner Schmidt，Richard Newcombe和Dieter Fox。自己-监督视觉描述符学习密集的冗长。RAL，2016. 2

[40] Johannes L Schonberger和Jan-Michael Frahm。结构-from-Motion revisited. In *CVPR*, 2016. 6

[41] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoon，Marc Polilleys，Josef Sivic，Tomas PJDLA和AK-Inloc：用扣边室内视觉本地化匹配和查看合成。在CVPR，2018年.7

[42] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *arXiv:2009.06732*. 3

[43] Carl Toft，Will Maddern，Akihiko Torii，Lars Hammarstrand，Erik Stenborg，Daniel Safari，Masatoshi Okutomi，Marc Pollfeys，Josef Sivic，Tomas Pajdla，等。重新审视了长期振缩本地化。T-PAMI，2020。7

[44] Prune Truong, Martin Danelljan, L. Gool, and R. Timofte。学习准确的密集信念以及何时相信它们。ARXIV，ABS / 2101.01710,2021.2

[45] Prune Truong，Martin Danelljan，Luc Van Gool，以及工作towmofte。Gocor：将全局优化的VoldeShon-Dence卷带入您的神经网络中。在Neurips，2020年.2

[46] Prune Truong, Martin Danelljan, and Radu Timofte. GLU-NET：全球局部通用网络，用于密集流动和通信。在CVPR，2020年.2

[47] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. DISK: 学习具有政策渐变的本地功能。Neurips，2020. 2,5,6

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszko-Reit，Llion Jones，Aidan N Gomez，Lukasz Kaiser和Illia Polosukhin。关注是你所需要的。Neurips，2017. 2,3

[49] Hui与Wang, Y U圆Z虎, Bradley green, Hart wig Adam, 艾伦玉器，梁驰陈。Axial-Deeblab：独立于Panoptic分割的轴向关注。在ECCV，2020年.3

[50] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and 诺亚偷偷摸摸。使用相机学习功能描述符pose supervision. In *ECCV*, 2020. 5

[51]思慕·彝族，Eduard Trulls，Vincent Lepetit和Pascal富亚。升降机：学习不变的功能变换。在ECCV，2016年.2

[52] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann和Pascal Fua。学习找到良好的通信。在CVPR，2018.6

[53] Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, 天威沉，玉龙陈，龙泉，鸿文廖。使用订单感知网络学习双视图对应关系和几何。ICCV，2019. 6

[54] Zichao Zhang, Torsten Sattler, and Davide Scaramuzza. Ref-通过学习的功能和观看综合来实现长期视觉定位的姿势。IJCV，2020年.7