**Computer Science and Application** 计算机科学与应用, 2025, 15(4), 22-32
Published Online April 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.154074

Hans汉斯

# CDNet：空间向量用于双视图对应学习的研究

**李浩然**

温州大学计算机与人工智能学院，浙江 温州

收稿日期：2025年2月28日；录用日期：2025年3月27日；发布日期：2025年4月3日

## 摘 要

**特征匹配是计算机视觉中的一项基本而重要的任务，目的是在给定的一对图像之间找到正确的对应关系(即内线)。严格地说，特征匹配通常包括四个步骤，即特征提取、特征描述、建立初始对应集和去除虚假对应(即离群值去除)。然而，现有的方法单纯考虑到了对应点之间的联系，而忽视了场景图片中可以获取的视觉信息。在本文中，我们提出了一种新型剪枝框架Context Depth Net (CDNet)来准确识别内线和恢复相机姿态。我们从对应点中提取方向信息作为提示方法指导剪枝操作，并利用向量场更好地挖掘对应之间的深层空间信息，最后设计一组融合模块来使空间信息更好融合。实验表明，所提出的CDNet在室内室外数据集上的测试结果优于先前提出的方法。**

## 关键词

**向量场，上下文，Transformer**

# CDNet: Using Spatial Vectors for Two-View Correspondence Learning Research

**Haoran Li**

College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou Zhejiang

Received: Feb. 28th, 2025; accepted: Mar. 27th, 2025; published: Apr. 3rd, 2025

## Abstract

**Feature matching is a fundamental and crucial task in computer vision, aiming to find the correct correspondences (*i.e.*, inliers) between a given pair of images. Strictly speaking, feature matching typically involves four steps: feature extraction, feature description, establishing an initial set of correspondences, and removing false correspondences (*i.e.*, outlier removal). However, existing methods only consider the connections between corresponding points while neglecting the visual information that can be obtained from scene images. In this paper, we propose a novel pruning**

文章引用: 李浩然. CDNet: 空间向量用于双视图对应学习的研究[J]. 计算机科学与应用, 2025, 15(4): 22-32.
DOI: 10.12677/csa.2025.154074

**framework called Context Depth Net (CDNet) to accurately identify inliers and recover camera poses. We extract directional information from corresponding points as a cue to guide the pruning process, and utilize vector fields to better mine the deep spatial information between correspondences. Finally, we design a set of fusion modules to better integrate the spatial information. Experiments show that the proposed CDNet performs better on indoor and outdoor datasets than previously proposed methods.**

## Keywords

**Vector Field, Context, Transformer**



## 1. 引言

双视图对应学习旨在在两幅图像之间建立可靠的对应/匹配，并准确地恢复相机姿势，这是计算机视觉中的一项基本任务，在许多应用中起着重要作用，如同 Structure from Motion (SfM) [1]、simultaneous localization and mapping [2]、image fusion [3]。由于场景图片存在结构变化、纹理变化、光照变化、遮挡和模糊，初始对应集通常包含大量异常值，严重影响下游任务的性能。综上所述，去除初始对应集中的异常值是必要的，近期的许多研究的工作倾向于去除初始对应集中的错误匹配以提升下游任务的精确度。

本文提出了一个全新的剪枝叶网络模型 CDNet，能更好地识别对应点之间的关系且去除外点影响，准确恢复相机姿态。

## 2. 相关工作

传统方法：RANSAC [4]及其基于迭代采样策略的变体的基本思想是通过随机采样一组数据来估计模型参数，并根据这些参数评估数据与模型之间的拟合程度。具体而言，在一次的迭代过程中，这类算法从数据集中随机选择最小样本集来估计模型参数，再根据随机采样得到的样本集，估计模型的参数，使用估计的模型参数，计算每个数据点到模型的拟合误差，最后统计内点的数量，作为评估模型好坏的指标。对 RANSAC 的优化有很多，例如 GroupSAC [5]、USAC [6]、PROSAC [7]、EVSAC [8]。虽然这些变体有着优于 RANSAC 的性能，但是还是没有解决传统方法在异常值率高的对应集上性能会严重下降的问题。

基于学习的方法：PointCN [9]创新性地提出了一种使用多层感知机(MLP)来处理对应的无序性，将离群点去除和相机姿态估计问题转化为了本质矩阵的回归和二分类问题，同时将全局上下文嵌入每对对应之中。在之后提出的 OANet [10]则是在模型提取上下文信息的性能上做出了优化，通过 DiffPool and DiffUnpool layers 以可学习的方式捕获无序稀疏对应的局部上下文。ACNet [9]则采用注意机制增强网络性能。伴随着图神经网络(GNN)的进一步研究发展，CLNet [11]提出了一个局部到全局的学习网络，通过逐步剪枝的方法去除离群值。MS2DG-Net [11]构建多稀疏语义，通过融合多尺度信息更好地获取上下文。之后的方向中，部分研究注意到了可以在考虑对应的空间信息时加入场景的视觉信息。因此，LMCNet [12]利用单个图像的特征点描述符来增强对应关系的表示能力。

上述方法虽然提高了通信剪枝性能，但这些方法只考虑对应的空间信息作为输入，严重阻碍了深度信息的获取，同时也损害了整体网络性能。在该方向的进一步探索中，有一些研究使用单幅图像的特征点描述符来增强网络输入的表示能力。在本文中，我们基于 CLNet [11]提出了更进一步的设计，并提出以下问题：我们是否可以深度挖掘对应点之间的空间线索？也就是说，如果有序向量矩阵可以更好地帮助模型感知内点比，将有助于网络区分一些模糊对应。

## 3. 方法

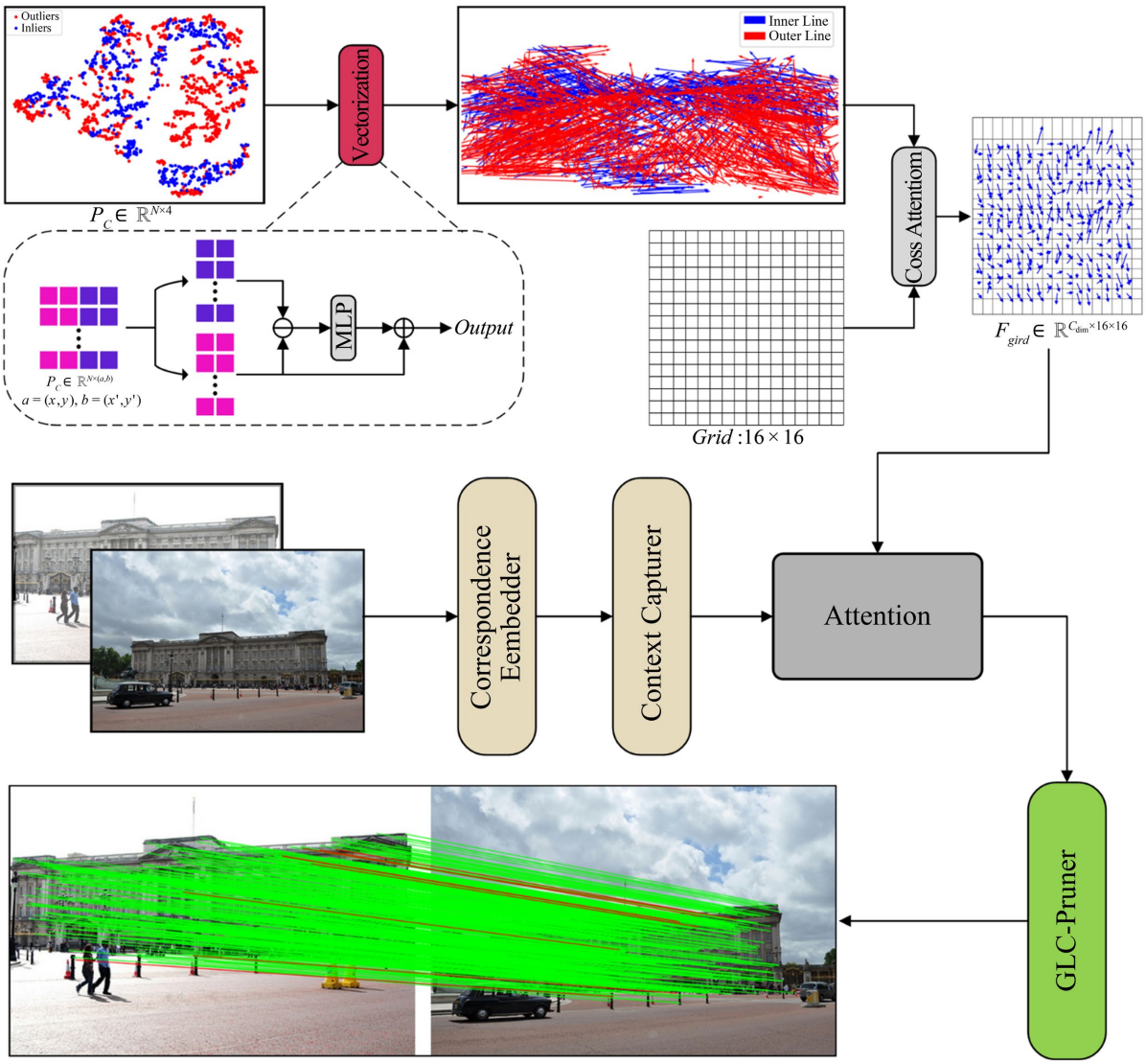


**Figure 1.** CDNet structure
**图 1.** CDNet 结构

特征匹配的任务目标在是从对应集中识别正确对应、去除错误对应。具体来说，在给定的双视图 $(P_A, P_B)$ 和没对双试图对应的场景文本 $P_T$，使用特征检测器(例如 SIFT [13]和 Problem Formulation [14])从双视图 $(P_A, P_B)$ 中提取特征点和描述符，然后根据描述符的最邻近匹配策略，得到初始对应集 $P_C$：

$$P_C = \{c_1, c_2, \cdots, c_N\} \in \mathbb{R}^{N\times 4}, c_i = (x_i, y_i, x_i', y_i') \tag{1}$$

其中 $c_i$ 是第 $i$ 个对应点；$(x_i, y_i)$ 和 $(x_i', y_i')$ 是给定两幅图像经过相机本征归一化后的特征点坐标。

在我们的任务中，深度对应点信息从原始场景信息中获取，深度对应点信息获取方式如下：

$$F^{grid} = f_T(P_C) \tag{2}$$

其中 $f_T(\cdot)$ 代表对应点向量转化器，执行深度探索一个空间结构 $P_C$ 得到一个有序向量场 $F^{grid}$ 。

在我们的任务中，对应修剪通常被表述为基本矩阵回归问题和内外点离群分类问题[9]。本文继 CLNet [11]之后，迭代使用 GCL-Pruner 进行对应剪枝，得到最终概率集 $P_f = \{P_1, \cdots, P_i, \cdots, P_M\}$，表示每个候选点作为内线的概率。上述过程可表述如下：

$$(C_f, W_i) = f_{\varnothing}(F^{grid}, P_C) \tag{3}$$

$$P_f = Softmax(W_f) \tag{4}$$

其中 $W_f = \{W_1, \cdots, W_i, \cdots, W_M\}$ 表示最终候选项的权重；$C_f = \{C_1, \cdots, C_i, \cdots, C_M\}$ 表示最终候选集；$f_{\varnothing}(\cdot)$ 表示我们提出的 CDNnet；其中 $\varnothing$ 表示网络参数。

然后，将最终候选集 $C_f$ 和概率集 $P_f$ 作为输入，采用加权八点算法对本质矩阵进行回归。该过程如下：

$$\hat{E} = g(C_f, P_f) \tag{5}$$

式中 $g(\cdot)$ 表示加权八点算法数，矩阵 $\hat{E}$ 表示预测的本质矩阵。

在接下来的内容中，将介绍 CDNet 框架中的对应点向量转化器与全局局部上下文剪除枝模块。CDNet 总体模型如图 1 所示。

### 3.1. 对应点向量转化器

在我们的任务中，我们设计了对应点向量转化器(SExtractor)，矢量变换器通过将无序矢量转换为有序运动，在表达对应关系之间的空间位置关系方面发挥着至关重要的作用。来自 $P_C$ 的每对对应 $c_i = \{(a_i, b_i) | i = 1, \cdots, N, a_i \in R^2, b_i \in R^2\}$，其中 $a_i = (x_i, y_i)$ 和 $b_i = (x_i', y_i')$ 被变换为运动向量 $\{m_i = (a_i, b_i) | i = 1, \cdots, N, m \in R^2\}$，其中 $x_i$ 和 $y_i$ 表示两个对应关键点的坐标，$d_i = b_i - a_i$ 表示位移。具体来说，我们初始化高维空间中的运动矢量和有序运动矢量的坐标。然后，对无序运动矢量进行归一化。首先，我们生成一个网格投影，在本文中设定的网格大小为 $16 \times 16$，这里我们在 $X$ 轴与 $Y$ 轴上生成两个 16 长度的一维张量数值序列来构成网格的基础。$X$ 轴与 $Y$ 轴一维张量通过计算笛卡尔积输出二维张量网格，每一行是一个 $(x, y)$ 坐标对，表示网格中的一个中心点，随后通过变形将结果调整为四维张量，添加批量维度与坐标维度用于后续模型。为了解决这个问题，我们采用 LMCnet [12]的方法将 $m_i$ 转换为高维 $f_i$：

$$f_i = Up_m(m_i),\ i = 1, \cdots, N \tag{6}$$

其中 $Up_m(\cdot)$ 是一个简单而高效的 MLP，在本文中，$m_i$ 从低维空间映射到 128 维空间。

随后，由于我们需要对向量场采样以获得有序数据，因此使用大小为 $K \times K$ 的网格 $X_{grid}$ 对上采样的稀疏向量进行分割。$x_{m,n}^{grid}$ 用于定位采样空间。为了匹配上采样 $m_i$ 空间的维度，$x_{m,n}^{grid}$ 也需要进行相同的维度扩展操作：

$$X_{m,n}^{grid} = Up_{grid}(x_{m,n}^{grid}) \tag{7}$$

其中 $Up_{grid}(\cdot)$，在我们的实现中，$x_{grid}$ 被提升到 128 个维度。

我们的目标是通过插值将稀疏无序的运动矢量转换为密集有序的运动场。随后，通过使用上述网格 $x_{grid}$，我们进行等距采样以获得有序的运动矢量。这些有序的运动矢量明确地表示了不同深度处对应点

的空间信息。我们参考 Convmatch [15]，它使用 GAT [16]来处理无序运动矢量并生成有序运动矢量：

$$m_{m,n}^{grid}\varsigma=\left(\left\{m_i\right\},x_{m,n}^{grid}\right) \tag{8}$$

使用通过我们的维数增广获得的高维表示 $f_i$ 和 $X_{m,n}^{grid}$ 来替换(8)中的低维信息，我们得到：

$$F^{grid}=\varsigma\left(\left\{f_i\right\},\left\{X_{m,n}^{grid}\right\}\right),\varsigma=\left(F,X^{grid}\right) \tag{9}$$

其中，$F^{grid}=\left\{f_{m,n}^{grid}\right\}$，$f_{m,n}^{grid}$ 表示 $m_{m,n}^{grid}$ 的高维表示。这里我们把 $\varsigma=\left(F,X^{grid}\right)$ 重新定义为 $\varsigma=\left(F,X\right)=Comb\left(X,Aggr\left(X,F\right)\right)$，其中的 $Aggr\left(\cdot,\cdot\right)$ 是一种类注意力机制的处理方式，目的是通过聚合所有已知的运动矢量 $F$ 来估计 $X$ 位置外的运动场：

$$\begin{aligned}Aggr\left(X,F\right)&=Softmax\left(QK^T\right)V,\\ Q&=W_1X+b_1,\\ \begin{bmatrix}K\\V\end{bmatrix}&=\begin{bmatrix}W_2\\W_3\end{bmatrix}F+\begin{bmatrix}b_2\\b_3\end{bmatrix}.\end{aligned} \tag{10}$$

其中的 $W_1$，$W_2$，$W_3$ 是可学习的权重，$b_1$，$b_2$，$b_3$ 是可学习的学习偏差，类似交叉注意力中的图像通过线性变化得到 $Q$、$K$ 和 $V$。在 $Comb\left(\cdot,\cdot\right)$ 中，经过 $Aggr\left(\cdot,\cdot\right)$ 函数处理的结果通过拼接与高维表格 $F$ 拼接，使用一个 MLP 层。

通过上述隐式插值操作，我们将无序运动矢量转换为有序运动矢量场。值得注意的是，该模块允许运动矢量的变换是可逆的，这有助于后续的融合步骤中将融合后的信息恢复到其基本对应关系，从而能够进行下一次修剪操作。

## 3.2. 全局局部上下文剪除枝模块

在我们的任务中，挖掘对应关系中的一致性对于搜索正确的匹配非常重要。为了充分捕捉联合视觉空间对应的上下文信息，本文设计了一种上全局局部上下文剪除枝模块(GLC-Pruner)结构。直观地说，正确的对应关系在它们的局部和全局上下文中应该是一致的，因此 GLC-Pruner 通过堆叠图神经网络和变换器来明确地捕获局部和全局的上下文。

### 3.2.1. 本地上下文捕获器

如[11] [17]所述，实用的 G-er 首先基于每对联合视觉空间对应之间的欧几里德距离构建一个图：

$$\varsigma_i=\left(v_i,\varepsilon_i\right),\ i\in\left[1,N\right] \tag{11}$$

其中 $v_i=\left\{f_{il},\cdots,f_{ik}\right\}$ 表示特 $f_i$ 的 $k$ 个最近邻，$\varepsilon_i=\left\{e_{il},\cdots,e_{ik}\right\}$ 表示连接 $f_i$ 及其邻的有向边集。边缘的构造可以公式化如下：

$$e_{il}=\left[f_i,f_i-f_{ij}\right] \tag{12}$$

其中 $f_i$ 和 $f_{ij}$ 分别表示第 $i$ 个联合视觉空间对应关系及其第 $j$ 个邻居。$\left[\cdot,\cdot\right]$ 是沿信道维度的级联操作。

### 3.2.2. 全局上下文捕获器

我们采用多头自我关注(MHSA)层来捕获全局上下文并将其融合到每次通信中。实际上，距离相似性被引入到MHSA层，结合长度特征一致性来生成空间感知注意力矩阵。它可以从不同方面利用几何关系，从而使网络能够以更全面、更稳健的方式捕获上下文信息。对于第 h 个头，我们的问题中的查询集 $\left\{q_i^h\right\}_{i=1}^N$、关键字集 $\left\{k_i^h\right\}_{i=1}^N$ 和值集 $\left\{q_i^h\right\}_{i=1}^N$ 是通过在 $\left\{k_i^h\right\}_{i=1}^N$ 上使用三个不同的线性投影得到的：

$$\left[q_i^h, k_i^h, v_i^h\right] = f_i\left[W_q^h, W_k^h, W_v^h\right] + \left[b_q^h, b_k^h, b_v^h\right] \tag{13}$$

其中 $W_{(\cdot)}^h \in \mathbb{R}^{128\times128}$ 和 $b_{(\cdot)}^h \in \mathbb{R}^{1\times128}$ 是投影参数，H = 4 是头数。我们再次将查询、键和值打包成矩阵，得到 $Q^h$、$K^h$ 和 $V^h$。然后，第 h 个头的输出如下：

$$Head^h = Attention\left(Q^h, K^h, V^h\right) = soft\max\left(\frac{QK^T}{c}\right)V^h \tag{14}$$

将不同磁头的输出连接起来，生成大小为 N × 128H 的最终输出：

$$MultiHead = Concat\left(Head^1, Head^2, \cdots, Head^H\right) \tag{15}$$

将 MHSA 具体运用于全局上下文捕获中，给定两个对应关系 $c_i = \left(p_i^A, p_i^B\right)$ 和 $c_j = \left(p_j^A, p_j^B\right)$，它们之间的长度相似度计算如下：

$$m_{i,j} = \left\|p_i^A - p_i^B\right\| - \left\|p_j^A - p_j^B\right\| \tag{16}$$

然后，通过 MHSA 层将长度一致性融合到注意力矩阵 $A_3 \in \mathbb{R}^{N\times N}$ 中，同时生成一个空间感知注意力矩阵 $A_3 \in \mathbb{R}^{N\times N}$ 来指导消息传递。此操作可以公式化如下：

$$A_4 = A_3 \odot M_{ls} \tag{17}$$

其中 $M_{ls} \in \mathbb{R}^{N\times N}$ 表示由方程(12)获得的长度相似性矩阵。最后，我们采用了类似于赵等人[11]的嵌入式预测器来处理连接特征。

### 3.3. 损失函数

以下混合函数[18] [19]用于监督我们提出的方法的训练过程：

$$\mathcal{L} = \mathcal{L}_c\left(o_j, y_j\right) + \alpha\mathcal{L}_e\left(\hat{E}, E\right) \tag{18}$$

其中 $\mathcal{L}_c$ 表示分类损失，而 $\mathcal{L}_e$ 表示基本矩阵损失。$\alpha$ 是一个超参数，用于平衡这两种损失。基于赵等人[11]，分类损失 $\mathcal{L}_c$ 可以表示为：

$$\mathcal{L}_c\left(o_j, y_j\right) = \sum_{j=1}^{\lambda} H\left(\omega_j \odot o_j, y_j\right) \tag{19}$$

其中 $H(\cdot)$ 表示二元交叉熵损失函数。$o_j$ 是第 $j$ 次迭代的相关权重。$y_j$ 代表弱监督标签，在可见度距离阈值为 $10^{-4}$ [20]时被选为正样本。$\omega_j$ 是一个自适应温度矢量，$\odot$ 表示哈达玛积。基本矩阵损失 $\mathcal{L}_e$ 可以表示为：

$$\mathcal{L}_e = \frac{\left(p'^T\hat{E}p\right)^2}{\left\|Ep\right\|_{[1]}^2 + \left\|Ep\right\|_{[2]}^2 + \left\|Ep'\right\|_{[1]}^2 + \left\|Ep'\right\|_{[2]}^2} \tag{20}$$

其中 $E$ 是地面真值基本矩阵；$p$ 和 $p'$表示通过本质矩阵 $E$ 获得的虚拟对应关系。$\|\|$ 表示向量第 $i$ 个元素的范数。

## 4. 实验与分析

### 4.1. 数据集

我们在室外和室内基准上测试了我们的方法，以评估相对姿态估计的性能。对于户外场景，采用了雅虎 YFCC100M [20]数据集，该数据集由从互联网上收集的一亿张户外照片组成。对于室内场景，我们采用了 SUN3D [21]数据集，这是一个提供相机姿态信息的大规模室内 RGBD 视频数据集。

根据张等人[10]提出的数据划分方案，所有竞争方法在未知和已知场景上分别进行评估。我们给出了 5˚和 20˚阈值下姿态误差的平均精度(mAP)，其中姿态误差被定义为旋转和平移引起的最大角度误差。

## 4.2. 模型细则

一般来说，SIFT [13]技术用于建立 $N=2000$ 个初始对应关系，信道维数 $C$ 为 128，网络迭代 $\lambda$ 为 2，修剪比 **r** 为 0.5。我们还使用 ORB [22]和 SuperPoint [14]特征匹配算法进行实验。具体测试结果见表 1。此外，为了降低培训成本，TrFormer 仅在第二次迭代中使用。在 SExtractor 中，原始图像的大小被调整为 $H$=16，$W$=16，并且信道尺寸 CF 被设置为 64。在 CVFormer 中，网格大小为 16 × 16，稀疏运动矢量和网格维数增加到 128。在 ContextFormer 中，在两次迭代中，$k$ 个邻居的数量分别设置为 9 和 6，组计数 $\boldsymbol{n}$ 为 3。我们使用 Adam 优化器训练网络，权重衰减为 0，学习率为$10^{-3}$，批处理大小为 32。根据赵等人[11]，在前 20 次迭代中，权重 $\alpha$ 设置为 0，然后在剩余迭代中设置为 0.5。具体参数设置对实验的影响如表 1 所示。

**Table 1.** The influence of parameter values on practical performance
**表 1.** 结果参数值对实际性能的影响

| 网格尺寸 | 学习率 | mAP5˚ | mAP20˚ |
|---|---|---|---|
| **8 × 8** | **$10^{-3}$** | 60.56/61.85 | 79.17/81.09 |
| **16 × 16** | **$10^{-2}$** | 52.77/59.10 | 74.98/78.79 |
| **16 × 16** | **$10^{-3}$** | 61.82/62.82 | 80.76/81.43 |
| **16 × 16** | **$10^{-4}$** | 58.85/61.74 | 78.70/80.85 |
| **32 × 32** | **$10^{-3}$** | 61.03/62.38 | 77.13/80.69 |
| **32 × 32** | **$10^{-4}$** | **58.78/60.25** | **79.14/80.06** |

## 4.3. 比试验结果

如表 2 所示，我们的 CDNet 室内场景中都优于其他的方法。具体来说，在室外室外场景中，与最近没有 RANSAC 的基于 MO 方法(PGFNet)相比，我们的方法在以下方面实现了 4.54%和 1.72%的性能提升：mAP@5˚用于未知场景。同样，与最近基于图的 SOTA 方法(CLNet)相比，我们的方法在没有 RANSAC 的情况下实现了 7.79%和 1.89%的性能提升 mAP5˚。一般来说，我们的提案在所有方法中得分最高。结果表明，所设计的深度空间挖掘和基于全局局部上下文的架构大大增强了网络实践。此外，如图 2 所示，OANet++ [10]和 PGNet 的典型可视化从左到右与我们的网络进行对比。

**Table 2.** Quantitative comparisons of the camera pose estimation
**表 2.** 相机姿态估计的定量比较

| 方法 | YFCC200M (%) | SUN3D (%) |
|---|---|---|
| **OANet++ (2019)** | 41.53 | 22.31 |
| **CLNet (2021)** | 44.88 | 23.83 |
| **MS2DG-Net (2022)** | 45.34 | 23.00 |
| **PGNet (2023)** | 46.28 | 23.87 |
| **Ours** | 48.38 | 24.28 |

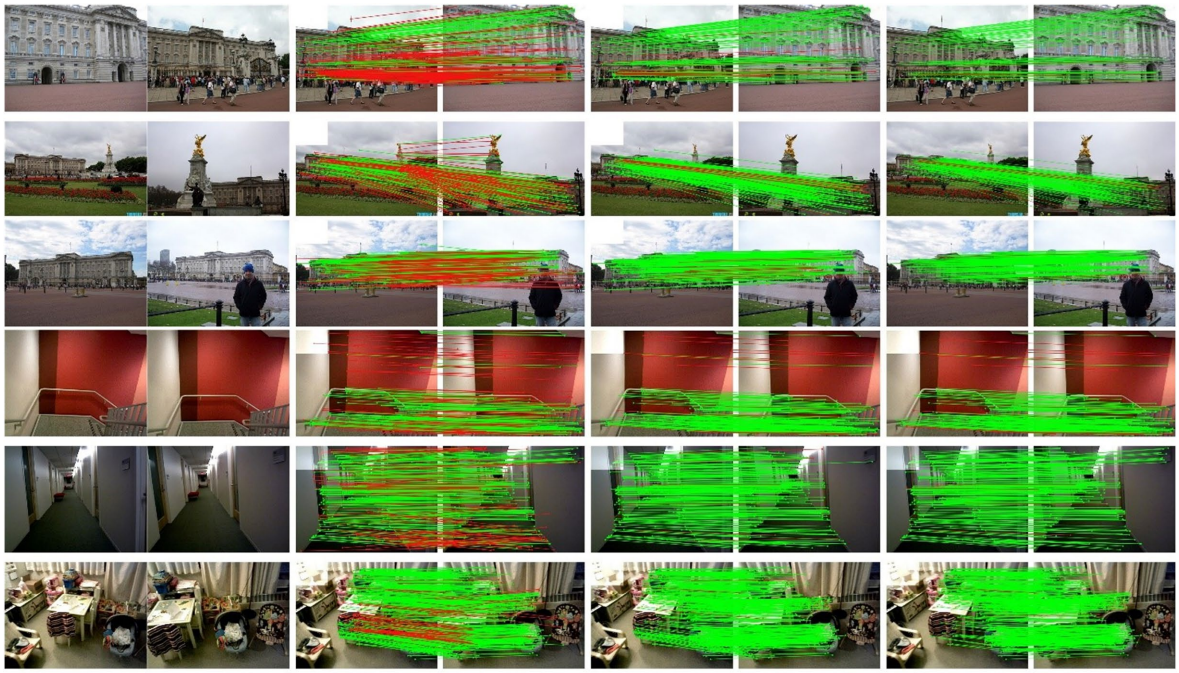


**Figure 2.** Comparative experiments on the YFCC100M dataset
**图 2.** YFCC100M 数据集上的对比实验

## 4.4. 离群点移除

根据提供的信息，表 3 显示了异常值去除任务中比较方法的结果。同样，所有模型都会在室外和室内场景中进行评估。报告了包括精确度、召回率和 $F$ 分数在内的指标。请注意，在我们的实现中，预测的对称极线距离 $d<10^{-4}$ 被视为对应关系的内层。从表 3 的结果来看，CDNet 在户外数据集的精度和 $F$ 分数方面优于所有最先进的方法。此外，与其他基于学习的方法相比，该模型的召回率较低。这主要是由于模型的修剪过程，它不仅修剪了异常值，还修剪了一些内部值，导致预测的几何模型继承了这一特征。同时，修剪模型利用预测的基本矩阵进行恢复处理，并计算预测的极线距离作为过程中的输出。

**Table 3.** Quantitative comparisons of outlier removal on outdoor scene and indoor scene
**表 3.** 室外场景和室内场景异常值去除的定量比较

| 方法 | 室外(%) | | | 室内(%) | | |
|---|---|---|---|---|---|---|
| | P | R | F | P | R | F |
| **RANSAC** | 43.55 | 50.62 | 46.83 | 44.87 | 48.82 | 46.76 |
| **PointNet++** | 46.39 | 84.17 | 59.81 | 46.30 | 82.72 | 59.37 |
| **DFE** | 54.00 | 85.56 | 66.21 | 46.18 | 84.01 | 59.60 |
| **ACENe** | 55.62 | 85.47 | 67.39 | 46.16 | 84.01 | 59.58 |
| **OANet++** | 55.78 | 85.93 | 67.65 | 46.15 | 84.36 | 59.66 |
| **CLNet** | 75.05 | 76.41 | 75.72 | 60.01 | 68.09 | 63.80 |
| **Ours** | **76.03** | **78.28** | **76.88** | **60.65** | **68.05** | **64.62** |

### 4.5. 消融实验

为了分析本文提出的模型中的各个模块对模型整体效果的贡献，我们对 YFCC100M 进行了详细的消融研究，以证明 CDNet 中每个组件的有效性。

#### 4.5.1. 模型的消融

如表 4 所示，我们打算逐步将拟议的组成部分添加到基线中。表中的第一行表示具有修剪策略的 PointCN [9]。考虑到 TrFusion 模块对输入模态完整性的要求，去除某种模态的消融肯定会影响最终的融合效果。具体来说，在第二行中，我们引入了 SExtractor 模块来恢复基线 CLNet。在第三行中，添加了 GLC-Pruner 模块，但没有 RANSAC 后处理和不完整的模态融合。可以看出，与基线相比，mAP5°提高了 43.29%。在第四行中，空间信息之间的融合处理未通过 Attention 层。最后，我们整合了所有模块，在没有 RANSAC 后处理的情况下，mAP5°比基线提高了 45.67%。

**Table 4.** Ablation study of network architecture
**表 4.** 网络架构的消融研究

| SExtractor | GLC-Pruner | Attention | mAP5° | mAP5° |
|---|---|---|---|---|
| | | | 43.25/55.20 | 67.12/75.23 |
| √ | | | 60.57/61.23 | 80.00/80.73 |
| | √ | | 61.80/62.07 | 80.58/80.88 |
| √ | √ | | 60.42/60.77 | 79.95/80.18 |
| √ | | √ | 60.77/61.37 | 80.44/81.76 |
| | √ | √ | 61.15/61.77 | 80.51/80.59 |
| √ | √ | √ | **62.82/62.82** | **80.76/81.43** |

#### 4.5.2. 自回归机制的消融

我们进一步进行消融研究，以验证所提出架构中每个组件的有效性。如表 5 所示，所提出的 GLC-Pruner 的每个组件都有助于进一步提高网络性能。实验证明，我们的 GLC-Pruner 融合模块将性能提高了 3.97%。换句话说，我们的方法能够有效地进行剪枝。

**Table 5.** Ablation study of GLC-Pruner
**表 5.** GLC-Pruner 的消融研究

| Local Context Capturer | Global Context Capturer | mAP5° | mAP5° |
|---|---|---|---|
| | | 60.77/61.37 | 80.44/81.76 |
| √ | | 61.45/61.27 | 79.83/80.93 |
| | √ | 62.02/61.90 | 80.17/80.97 |
| √ | √ | **62.82/62.82** | **80.76/81.43** |

## 5. 结论

本研究从空间向量场的角度探讨了通过利用视觉信息来提高特征匹配任务准确性的可能性。我们挖掘深层视觉信息来指导修剪操作，为此，我们设计了一个对应点向量转化器，我们采用对应关系矢量场

变换器将对应关系转换为密集有序的运动矢量场，进一步利用对应关系之间的空间信息。最后，利用全局上下文捕获器来进行剪枝操作。对比实验和消融研究证明了这种方法的实际有效性。

## 参考文献


[1] Havlena, M. and Schindler, K. (2014) VocMatch: Efficient Multiview Correspondence for Structure from Motion. *Computer Vision—ECCV* 2014, Zurich, 6-12 September 2014, 46-60. https://doi.org/10.1007/978-3-319-10578-9_4

[2] Mur-Artal, R., Montiel, J.M.M. and Tardos, J.D. (2015) ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics*, **31**, 1147-1163. https://doi.org/10.1109/tro.2015.2463671

[3] Ma, J., Ma, Y. and Li, C. (2019) Infrared and Visible Image Fusion Methods and Applications: A Survey. *Information Fusion*, **45**, 153-178. https://doi.org/10.1016/j.inffus.2018.02.004

[4] Fischler, M.A. and Bolles, R.C. (1981) Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, **24**, 381-395. https://doi.org/10.1145/358669.358692

[5] Ni, K., Jin, H. and Dellaert, F. (2009) GroupSAC: Efficient Consensus in the Presence of Groupings. 2009 *IEEE* 12*th International Conference on Computer Vision*, Kyoto, 29 September-2 October 2009, 2193-2200. https://doi.org/10.1109/iccv.2009.5459241

[6] Raguram, R., Chum, O., Pollefeys, M., Matas, J. and Frahm, J. (2013) USAC: A Universal Framework for Random Sample Consensus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 2022-2038. https://doi.org/10.1109/tpami.2012.257

[7] Fragoso, V., Sen, P., Rodriguez, S. and Turk, M. (2013) EVSAC: Accelerating Hypotheses Generation by Modeling Matching Scores with Extreme Value Theory. 2013 *IEEE International Conference on Computer Vision*, Sydney, 1-8 December 2013, 2472-2479. https://doi.org/10.1109/iccv.2013.307

[8] Ma, J., Jiang, X., Fan, A., Jiang, J. and Yan, J. (2020) Image Matching from Handcrafted to Deep Features: A Survey. *International Journal of Computer Vision*, **129**, 23-79. https://doi.org/10.1007/s11263-020-01359-2

[9] Yi, K.M., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M. and Fua, P. (2018) Learning to Find Good Correspondences. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 2666-2674. https://doi.org/10.1109/cvpr.2018.00282

[10] Zhang, J., Sun, D., Luo, Z., Yao, A., Zhou, L., Shen, T., *et al*. (2019) Learning Two-View Correspondences and Geometry Using Order-Aware Network. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 5844-5853. https://doi.org/10.1109/iccv.2019.00594

[11] Zhao, C., Ge, Y., Zhu, F., Zhao, R., Li, H. and Salzmann, M. (2021) Progressive Correspondence Pruning by Consensus Learning. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 6444-6453. https://doi.org/10.1109/iccv48922.2021.00640

[12] Liu, Y., Liu, L., Lin, C., Dong, Z. and Wang, W. (2021) Learnable Motion Coherence for Correspondence Pruning. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 3236-3245. https://doi.org/10.1109/cvpr46437.2021.00325

[13] Lowe, D.G. (2004) Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, **60**, 91-110. https://doi.org/10.1023/b:visi.0000029664.99615.94

[14] DeTone, D., Malisiewicz, T. and Rabinovich, A. (2018) SuperPoint: Self-Supervised Interest Point Detection and Description. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Salt Lake City, 18-22 June 2018, 224-236. https://doi.org/10.1109/cvprw.2018.00060

[15] Zhang, S. and Ma, J. (2024) ConvMatch: Rethinking Network Design for Two-View Correspondence Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **46**, 2920-2935. https://doi.org/10.1109/TPAMI.2023.3334515

[16] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P. and Bengio, Y. (2017) Graph Attention Networks. arXiv: 1710.10903. https://doi.org/10.48550/arXiv.1710.10903

[17] Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M. and Solomon, J.M. (2019) Dynamic Graph CNN for Learning on Point Clouds. *ACM Transactions on Graphics*, **38**, 1-12. https://doi.org/10.1145/3326362

[18] Hartley, R. and Zisserman, A. (2004) Multiple View Geometry in Computer Vision. 2nd Edition, Cambridge University Press. https://doi.org/10.1017/cbo9780511811685

[19] Ranftl, R. and Koltun, V. (2018) Deep Fundamental Matrix Estimation. *Computer Vision—ECCV* 2018, Munich, 8-14 September 2018, 292-309. https://doi.org/10.1007/978-3-030-01246-5_18

[20] Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., *et al*. (2016) YFCC100M: The New Data in Multimedia Research. *Communications of the ACM*, **59**, 64-73. https://doi.org/10.1145/2812802

[21] Xiao, J., Owens, A. and Torralba, A. (2013) SUN3D: A Database of Big Spaces Reconstructed Using SfM and Object Labels. 2013 *IEEE International Conference on Computer Vision*, Sydney, 1-8 December 2013, 1625-1632. https://doi.org/10.1109/iccv.2013.458

[22] Rublee, E., Rabaud, V., Konolige, K. and Bradski, G. (2011) ORB: An Efficient Alternative to SIFT or SURF. 2011 *International Conference on Computer Vision*, Barcelona, 6-13 November 2011, 2564-2571. https://doi.org/10.1109/iccv.2011.6126544