# LightGlue 论文翻译

摘要

我们引入了LightGlue，这是一种深度神经网络，它学习匹配图像之间的局部特征。我们重新审视了SuperGlue的多个设计决策，即稀疏匹配的最新技术，并得出了简单但有效的改进。累积起来，他们在内存和计算方面使 LightGlue 更有效，更准确，更容易训练。一个关键的特性是LightGlue适应问题的难度:在直观上容易匹配的图像对上，推理要快得多，例如由于更大的视觉重叠或有限的外观变化。这为在诸如 3D 重建之类的延迟敏感应用中部署深度匹配器开辟了令人兴奋的前景。代码和经过训练的模型可在 github.com/cvg/LightGlue 公开获得。

介绍

找到两幅图像之间的对应关系是许多计算机视觉应用的基本构建块，如相机跟踪和3D映射。图像匹配最常见的方法依赖于稀疏兴趣点，这些兴趣点使用编码其局部视觉外观的高维表示进行匹配。在不同的视点和照明下，可靠地描述每个点是具有挑战性的，在表现出对称性、弱纹理或外观变化的情况下。为了拒绝由遮挡和缺失点引起的异常值，这种表示也应该具有区分性。这产生了两个难以满足的相互冲突的目标、鲁棒性和唯一性。

为了解决这些限制，SuperGlue[56]引入了一种新的范式——一个同时考虑两幅图像的深度网络来联合匹配稀疏点并拒绝异常值。它利用强大的 Transformer 模型[74] 来学习匹配来自大型数据集的具有挑战性的图像对。这在室内和室外环境中都产生了稳健的图像匹配。SuperGlue在具有挑战性的条件下对视觉定位非常有效[59,55,58,57]，并能很好地推广到其他任务，如空中匹配[83]、物体姿态估计[69]，甚至鱼类再识别[47]。

这些改进在计算上是十分复杂的，而图像匹配的效率对于需要低延迟(如跟踪)或高处理体积(如大规模映射)的任务至关重要。此外，众所周知，SuperGlue 与其他基于 Transformer 的模型一样难以训练，需要许多从业者无法访问的计算资源。因此，后续工作[8,65]未能达到原始SuperGlue模型的性能。然而，自从最初的发表以来，变形金刚得到了广泛的研究、改进并应用于多种语言[17,51,13]和视觉[18,6,29]任务。在本文中，我们利用这些见解来设计LightGlue，这是一种比SuperGlue更准确、更高效、更容易训练的深度网络。我们重新审视了它的设计决策，并结合了许多简单而有效的架构修改。我们提取一种方法来训练资源有限的高性能深度匹配器，只需几个 GPU 天即可达到最先进的准确性。如图1所示，与现有的稀疏和密集匹配器相比，LightGlue在效率-精度权衡方面是次优的。

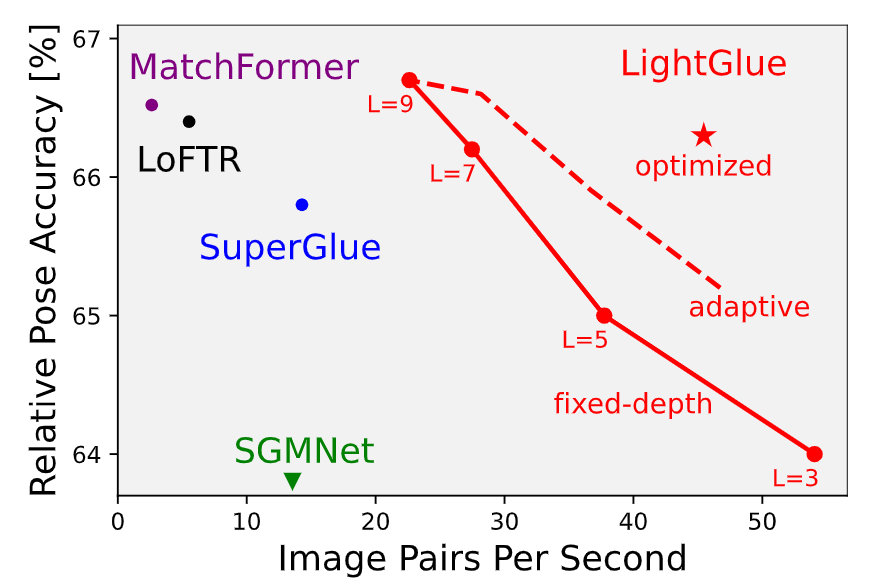


图 1. LightGlue 比 SuperGlue 等现有方法更快、更好。它的自适应停止机制对速度和准确性权衡进行了细粒度的控制。我们最终优化的模型⋆以 8 倍的速度提供了更接近密集匹配器 LoFTR 的准确度，这里在典型的室外条件下。

与以前的方法不同，LightGlue 适应每个图像对的难度，这因视觉重叠量、外观变化或判别信息而不同。图 2 显示，在直觉上比具有挑战性的匹配对上的推理要快得多，这种行为让人想起人类如何处理视觉信息。这是通过 1) 在每个计算块之后预测一组对应关系，以及 2) 使模型能够引入它们并预测是否需要进一步计算来实现的。LigthGlue 还丢弃了无法匹配的早期阶段点，从而将注意力集中在共可见区域。我们的实验表明，LightGlue 是 SuperGlue 的即插即用替代品：它在运行时的一小部分从两组局部特征预测强匹配。这为在SLAM[45,5]等延迟敏感应用中部署深度匹配器或从众包数据中重建更大的场景开辟了令人兴奋的前景[25,60,39,57]。LightGlue 模型及其训练代码将以许可公开发布。

1. 相关工作

图像匹配

描述相同场景或对象的匹配图像通常依赖于局部特征，这些特征是稀疏的关键点，每个关键点都与其局部外观的描述符相关联。虽然经典算法依赖于手工制作的标准和梯度统计[41，23，4，53]，但最近的研究大多集中在设计卷积神经网络(CNN)进行检测[81，16，19，52，73]和描述[42，72]。使用具有挑战性的数据进行训练，CNN 大大提高了匹配的准确性和鲁棒性。局部特征现在有多种风格:有些更好本地化[41]，高度可重复的[16]，存储和匹配[54]便宜，不受特定更改[46]的影响，或忽略不可靠的对象[73]。然后将局部特征与描述符空间中的最近邻搜索进行匹配。由于不匹配的关键点和不完美的描述符，一些对应关系不正确。这些通过启发式方法过滤掉，如 Lowe 的比率测试 [41] 或相互检查、内部分类器 [44, 82]，并通过稳健拟合几何模型 [22, 7]。这个过程需要大量的领域专业知识和调整，并且当条件太具有挑战性时容易失败。这些限制很大程度上是由深度匹配器解决的。

深度匹配器

深度匹配器是深度网络，经过训练可以在给定输入图像对的情况下联合匹配局部特征并拒绝异常值。其第一类SuperGlue[56]将Transformers[74]的表达表示与最优传输[48]相结合，以解决部分分配问题。它学习了关于场景几何和相机运动的强大先验，因此对极端变化具有鲁棒性，并在数据域中很好地泛化。继承了早期Transformers局限性，SuperGlue很难训练，其复杂性随着关键点的数量呈二次增长。随后的工作通过减少注意力机制的大小使其更有效。它们将其限制为一小组种子匹配 [8] 或相似关键点的集群 [65]。这大大减少了大量关键点的运行时间，但在较小的标准输入大小下不会产生任何收益。这也损害了最具挑战性的条件下的鲁棒性，未能达到原始SuperGlue模型的性能。相反，LightGlue 为典型的操作条件（如 SLAM）带来了很大的改进，而不会影响任何难度级别的性能。这是通过动态调整网络规模而不是降低其整体容量来实现的。相反，像 LoFTR [68] 和后续 [9, 78] 这样的密集匹配器匹配分布在密集网格上而不是稀疏位置的点。这提高了对令人印象深刻的水平的鲁棒性，但通常慢得多，因为它处理更多的元素。这限制了输入图像的分辨率，进而限制了对应关系的空间精度。虽然 LightGlue 在稀疏输入上运行，但我们表明，对于一小部分运行时间，公平调整和评估使其与密集匹配器竞争。

Transformer

在语言处理的成功之后，使 Transformer 高效受到了极大的关注。由于注意力的内存占用是处理长序列的一个主要限制，许多工作使用线性公式 [79,32,33] 或瓶颈潜在标记 [35, 30] 来减少它。这使得远程上下文成为可能，但会损害小输入大小的性能。选择性检查点[49]减少了注意力的内存占用，优化内存访问也大大加快了[14]。相反，正交工作通过预测给定层令牌的预测是否最终或需要进一步计算来自适应地调制网络深度 [15, 20, 62]。这主要受到视觉通信为cnn开发的自适应方案的启发。在 Transformers 中，位置编码的类型对准确性有很大影响。虽然绝对正弦[74]或学习编码[17,51]最初很普遍，但最近的工作研究了相对编码[63,67]来稳定训练并更好地捕捉长期依赖关系。LightGlue 将其中一些创新适应 2D 特征匹配，并显示出效率和准确性的提高。

1. 快速特征匹配

问题表述：LightGlue 根据 SuperGlue 在 SuperGlue 之后预测从图像 A 和 B 中提取的两组局部特征之间的部分分配。每个局部特征 i 由 2D 点位置 pi := (x, y)i ∈ [0, 1]2 组成，由图像大小归一化，视觉描述符 di ∈ Rd。图像 A 和 B 有 M 和 N 个局部特征，由 A := {1, ..., M } 和 B := {1,..., N }，分别。我们设计了LightGlue输出一组对应关系M = {(i, j)}⊂A × B。每个点至少匹配一次，因为它源于一个唯一的 3D 点，并且由于遮挡或非可重复性，一些关键点是不可匹配的。与之前的工作一样，我们在 A 和 B 中的局部特征之间寻求软部分分配矩阵 P ∈ [0, 1]M ×N，我们可以从中提取对应关系。

概述——图 3：LightGlue 由一堆 L 个相同的层组成，这些层联合处理两组。每一层都由更新每个点的表示的自我注意单元和交叉注意单元组成。然后，分类器在每一层决定是否停止推理，从而避免不必要的计算。轻量级头部最终从表示集中计算部分分配。

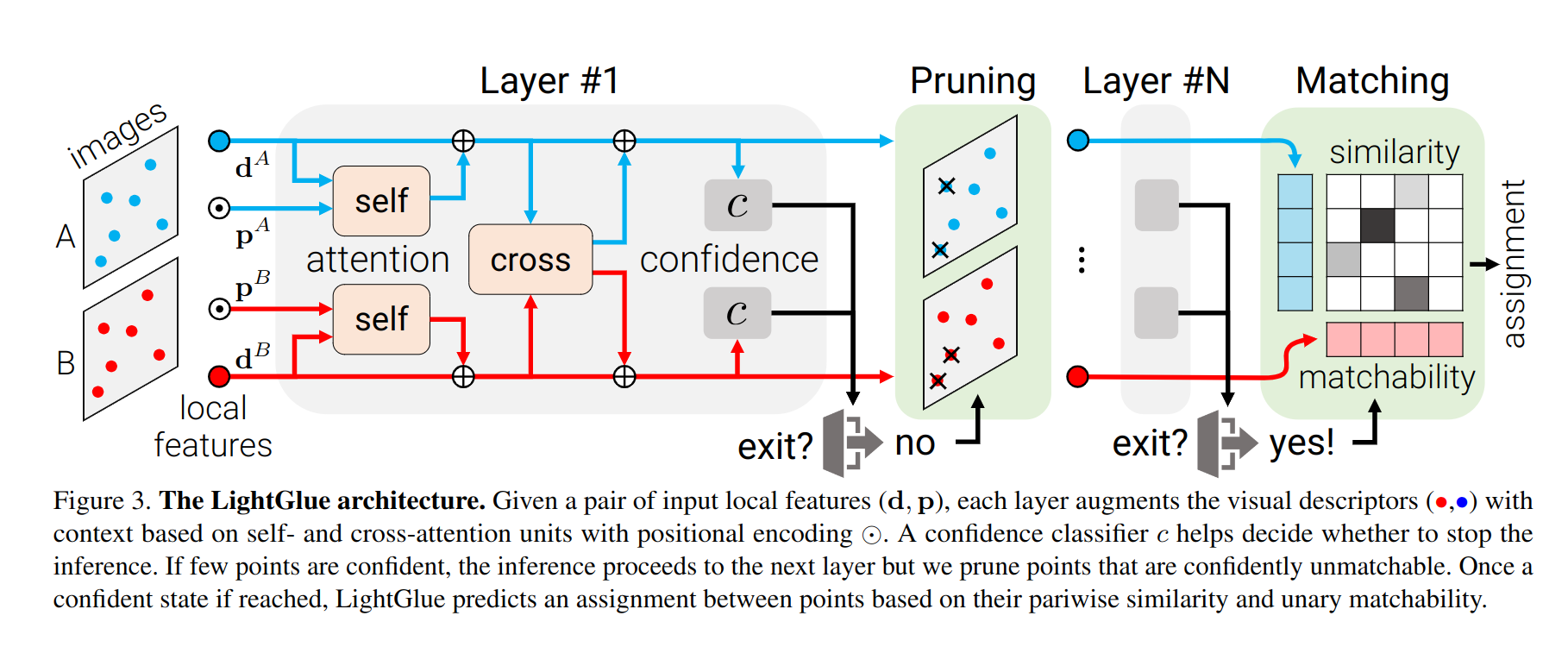


图3。LightGlue架构。给定一对输入局部特征 (d, p)，每一层都使用基于自注意力单元和具有位置编码 ⊙ 的交叉注意力单元的上下文来增强视觉描述符 (•,•)。置信度分类器 c 有助于决定是否停止推理。如果很少有点有信心，则推理继续进行下一层，但我们修剪自信不匹配的点。一旦达到置信状态，LightGlue 根据点的相似性和一元匹配性来预测点之间的分配。