Occ2Net: 一种基于3D 占据估计的有效且稳健的带有遮挡区域图像匹配方法

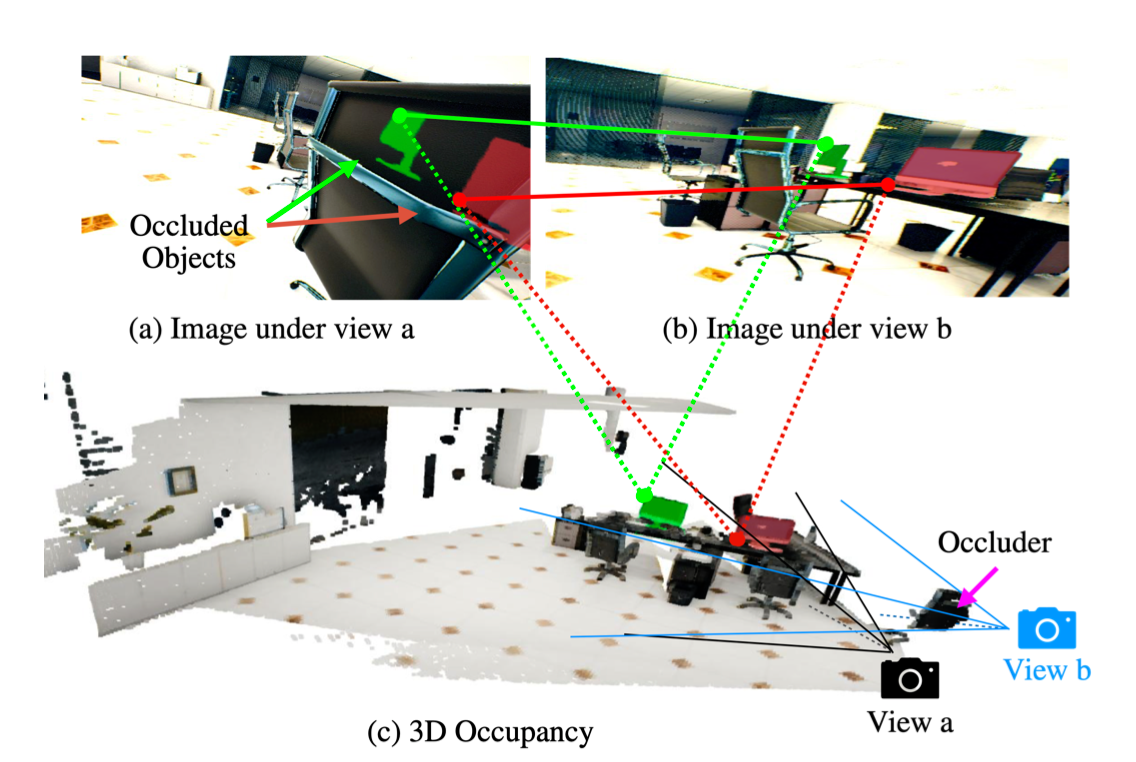
论文链接：https://arxiv.org/abs/2308.16160

1. 总体思路：

图像匹配是各种视觉应用中的基础和关键任务，如同时定位和映射（SLAM）和图像检索，这些任务都需要精确的姿态估计。然而，大多数现存的方法忽视了由相机运动和场景结构引起的对象间的遮挡关系。在本文中，我们提出了一种新颖的图像匹配方法Occ2Net，该方法使用3D占位图模型来描述遮挡关系，并推断出被遮挡区域的匹配点。借助占位估计（OE）模块编码的归纳偏差，它大大简化了构建一个多视图一致的3D表示的过程，该表示能够整合多视图信息。再结合遮挡感知（OA）模块，通过引入注意力层和旋转对齐，实现了被遮挡点和可见点的匹配。我们在真实世界和模拟数据集上评估了我们的方法，结果显示其在多项指标上，尤其是在遮挡场景下，优于当前最先进的方法。

1. 任务背景以及解决思路：

图像匹配是各种视觉应用中的基础且关键任务，例如SLAM（同时定位与地图创建）和图像检索。它的目标是在两幅或更多的图像中识别和对应相同或相似的结构/内容。图像匹配可以分为两类：基于特征的方法[34,20,10,35]和密集式方法[38,14]。基于特征的方法从图像中提取稀疏的关键点和描述符，然后基于相似性指标进行匹配；而密集式方法是估计图像像素或区块之间的密集对应关系。



然而，这两种方法都无法很好地应对遮挡情况，这在真实世界环境中是常见的。上图展示了这些挑战的一个例子。由于相机运动，两个图像的视差很大。尽管存在大量的重叠区域，但大的视差导致了遮挡，大大减少了可见匹配对的数量。此外，在这个例子中，场景中的地面和墙壁纹理都比较低，两个可以辨别的显示器被标记为绿色和红色的遮罩，指示在图像(b)中可见但在图像(a)中被遮挡的区域。这些因素使得现有的算法难以提取出足够的匹配对进行相机姿态估计。类似的情况在室内导航或自动驾驶中常见。为了解决这些问题，我们提出了一种新的图像匹配方法，称为Occ2Net，它不仅匹配可见的点对，还匹配被遮挡的点和可见的点。

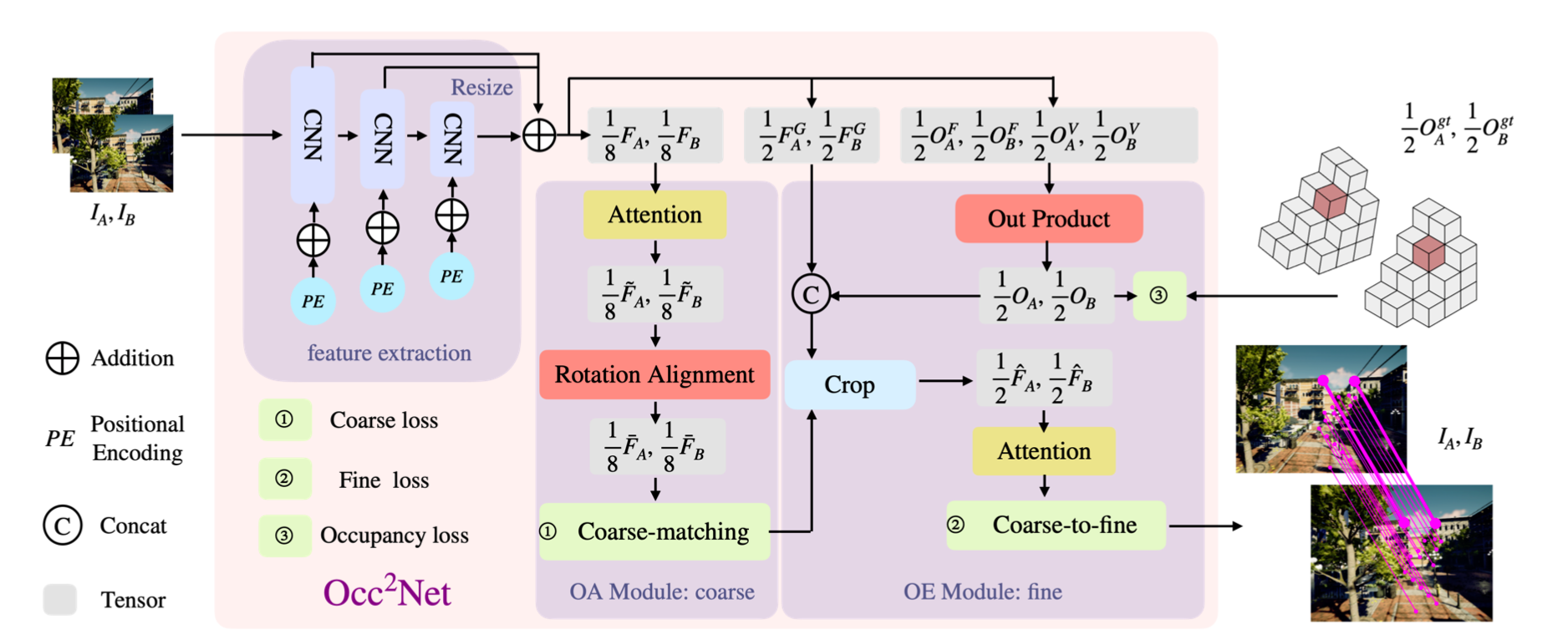
基于这个观察，我们设计了Occ2Net来匹配3D点。参照NeRF[24]，我们将每个像素视为从相应相机发出的一条射线。NeRF通过在射线上等间隔采样来获取3D点，并通过可微分渲染学习它们的信息。然而，在匹配算法中，我们在推理时没有姿态信息，所以我们将射线上的采样简化为两个点：一个可见点和一个被遮挡的点。在训练时，我们使用真实深度和姿态来重新投影并确定一个3D点是被遮挡还是可见。

基于这些简化，Occ2Net将可见点之间的匹配扩展到可见点与遮挡点之间的匹配。为了实现带有3D显著性的匹配，我们使用了一个3D占用率估计（OE）模块，这大大简化了以文献[5,23]为指导的多视角3D表示方法。由于3D匹配的难度、占用率的大存储空间以及占用率估计的误差，我们没有使用整个图像的占用率来估计匹配，而是采用了粗到精的结构。在粗糙的步骤中，我们使用了遮挡感知（OA）模块来获取每个补丁之间的匹配，OE模块则用于获取每个补丁中的精细匹配。

我们在两个数据集上评估我们的提出的方法：ScanNet[6]和 TartanAir[43]，这些数据集包含了各种程度遮挡的现实和模拟场景。我们使用了几种衡量标准，将我们的方法与几种最先进的基于特征的方法以及密集方法进行了比较。实验结果显示，我们的方法在这两个数据集上都取得了优越的准确度，超过了现有方法很大的边距。而且，我们的方法在处理遮挡情况下展示出了高的稳健性和效率。

总的来说，我们提出了一种能够识别遮挡点的图像匹配算法，该算法在真实世界和合成数据集上均优于最先进的方法。具体来说，我们的贡献如下：我们提出了一种新颖的能够识别遮挡的图像匹配算法Occ2Net，该算法使用三维占用模型来描绘物体之间的遮挡关系，并推断出被遮挡区域内匹配点的位置。我们将一个占用估计（Occupancy Estimation， OE）模块和一个遮挡感知（Occlusion-Aware, OA）模块结合起来，使用具有占用估计的粗到细结构实现可见-被遮挡匹配。我们的实验显示Occ2Net在真实世界数据集ScanNet和模拟数据集TartanAir上都实现了最先进的姿态估计精度。

1. 实现方法



上图展示了我们的Occ2Net的概述，它通过隐式模拟物体-遮挡关系，帮助在遮挡下进行匹配。我们的方法已经在各种可见-可见和可见-遮挡的情况下进行了广泛的测试，取得了与最新技术方法相当的成果。然而，在这一部分里，我们的主要关注点主要是挑战性较高的可见-遮挡匹配问题。

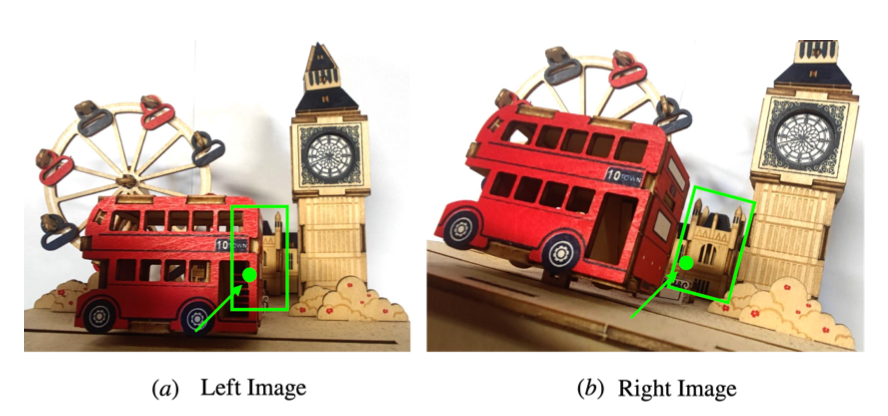
## 特征提取模块

我们的特征提取有三个主要功能：粗特征F的提取，细特征FG的提取，以及生成用于3D占据率估计的张量OF和$OV。为了匹配可见点和遮蔽点，我们使用了一个大感受野的金字塔结构，并为不同的尺度添加位置编码。由于相机距离的不同，金字塔结构有助于识别不同尺度的同一物体。为了成功地匹配遮蔽区域和可见区域，我们扩大了金字塔结构的感知范围，并在匹配过程中使用了更多周围信息。如图所示，图像中包含椅子的部分在椅子部分有所不同，但周围区域非常相似。这是一种常见的现象，即匹配到的遮盖区域和可见区域周围的图像特征是相似的。我们还添加了不同尺度的位置编码，这丰富了特征的位置信息，以便从上述相似的周围区域准确推断出遮蔽点的偏移量。

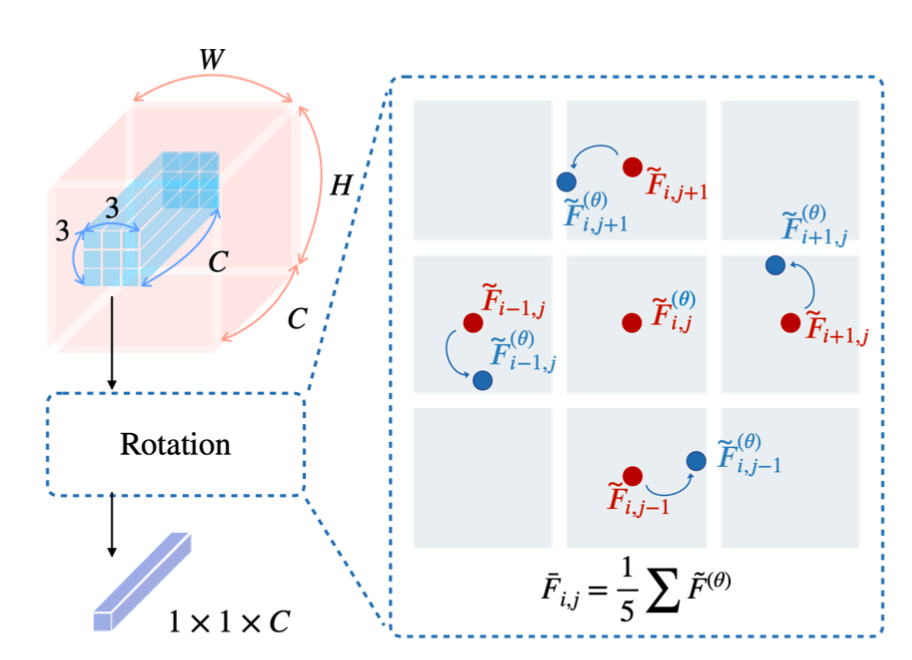
## OA模块

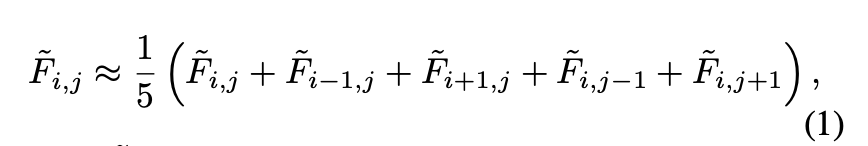
我们设计了OA模块来实现粗略的匹配。OA模块主要由两部分组成：注意力组件和旋转对齐。注意力组件通过自注意力和交叉注意力，加深了对整个图像结构信息的理解，使得特征更有利于粗略匹配。旋转对齐旨在更好地适应不同视角之间的不同旋转，使得可见和被遮挡的部分更易于匹配。

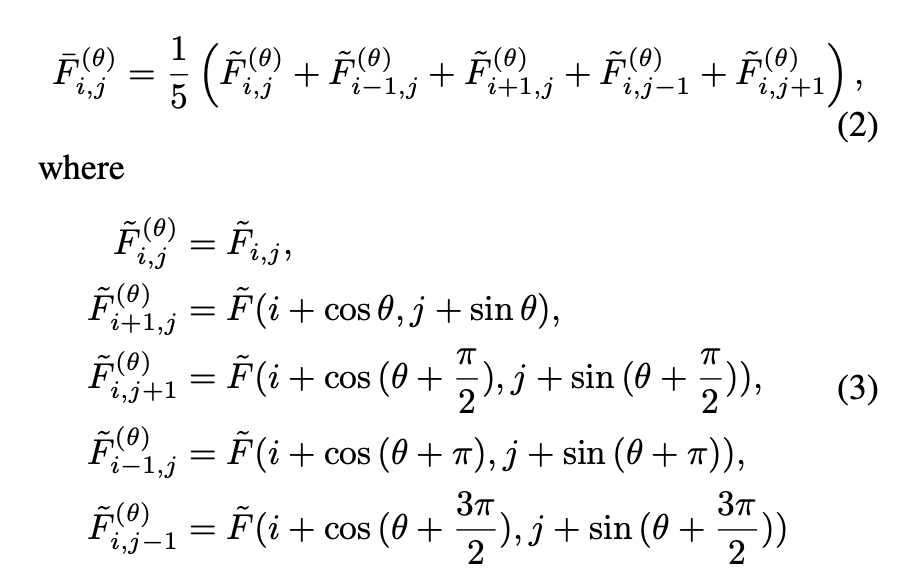
### 旋转对齐



旋转对齐选择具有适当旋转角度的特征。在注意力组件之后，每个特征保留其接受区域的块信息。左图的特征经过适当的旋转后，会更接近于右图的特征，这进一步可以感知遮挡并使块匹配更加容易。例如，如上图所示，我们想要匹配图(a)和图(b)中由箭头指示的点。图(a)中的点被遮挡，在图(b)中可见。在粗略匹配阶段，我们需要匹配两个绿色的块。然而，由于遮挡区域的存在，两个块的特征会稍有不同。尽管我们在特征提取中扩大了接受域，并引入了更多周围信息，但我们不能消除遮挡的影响。旋转对齐的目标是增加两个块的特征相似性。考虑到我们使用2D旋转来补偿3D相机姿态的差异，图像的不同部分的旋转角度应该不同。为了防止旋转改变匹配块的指数，我们开发了一个局部特征旋转算法，对每个块特征进行局部旋转。由于我们不知道适当的旋转角度，我们使用gumbel softmax来选择最佳匹配的旋转角度。



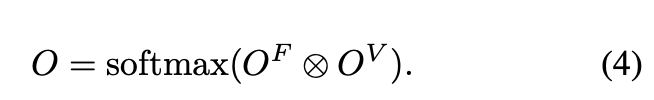




因为我们不知道真实的旋转角度，根据经验我们选择0度和30度作为旋转角度。

## OE模块

在获取粗略级别的补丁匹配后，我们利用OE模块进行细致匹配。对于可见补丁中的点，我们使用小型LoFTR模块生成的热图计算与匹配补丁中心的偏移。对于被遮挡补丁中的点，根据其周围可见内容推断出匹配点的位置，这些内容是通过特征提取和本地3D占位得到的。OE模块被用来计算精确匹配点的位置。首先，我们设计了一个3D占位估计模块。对每一张图，我们通过特征提取得到两个张量OF和OV。我们计算一个4D张量作为OF和OV的外积，以表示估计的3D占位。



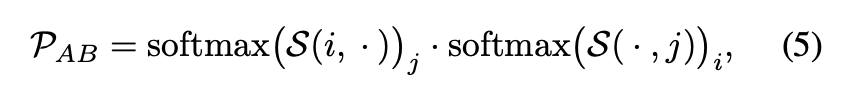
基于真实深度和相机姿势，我们计算真实的3D占用作为监督。我们将在loss部分解释3D占用损失。然后，我们通过注意力模块将估计的3D占用和精细特征相结合。最后，基于两个图像的局部特征和3D占用，我们可以推断出每个匹配点的具体位置。使用粗到精的结构，并且只在精细结构中使用3D占用估计，有两个原因：1. 已经证明粗到精的结构对于可见点之间的匹配是有效的。2. 单眼模型通过3D占用估计是不准确的，而推断遮挡点是困难的。使用粗到精的结构，同时将精细特征与3D占用相结合，使得可以使用从3D占用估计出的3D信息来推断遮挡点的位置。由于粗匹配，推断的误差可以控制在一定范围内。例如，如上图所示，我们试图匹配左图和右图中的绿色框。匹配点的特征应与3D占用一致。我们可以从3D占用推断出，图中的箭头指示了一个精细匹配对。

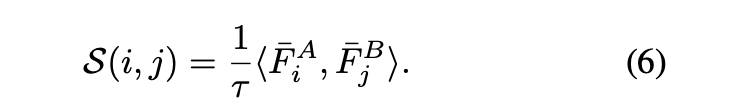
## loss

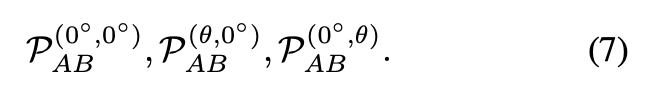
### 粗粒度匹配 loss

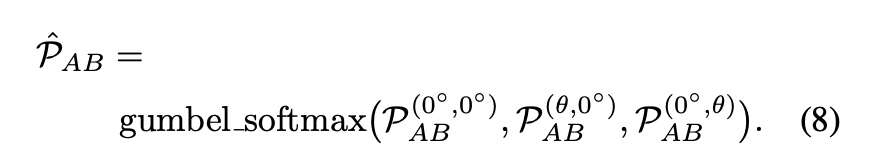
我们的网络将以下情况视为有效匹配：可见与可见的像素块、可见与遮挡的像素块、遮挡与可见的像素块。根据真实的相机位置和深度，我们计算左图像块的再投影。当一个像素块的投影深度大幅度大于右图像对应的深度时，我们将该像素块视为在右图像中被前方的物体遮挡。我们定义图像的遮挡点与像素总数的比率为遮挡比例。我们获取了三组真实的像素块匹配数据：可见-可见匹配、可见-遮挡匹配和遮挡-可见匹配。

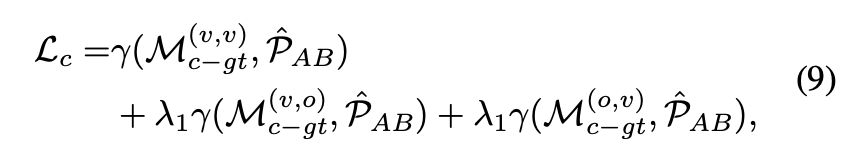
参考LoFTR的loss，有：

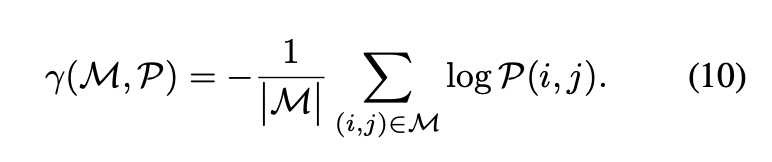






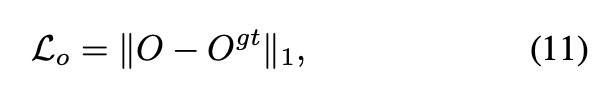






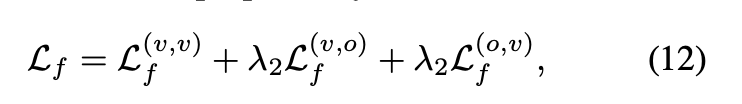
### 3D占位loss

为了从真实深度和相机姿态中生成真实的3D占用率，我们采取了以下步骤：通过结合两张图片的深度和相机姿态，创建一个真实的点云。将点云转化为体素表示，通过将3D空间划分为小立方体单元，并根据其3D占用情况给每个单元分配一个值。使用相机的姿态作为世界坐标系的原点，并将深度分辨率设置为64个体素，将体素表示投影到当前图片上。

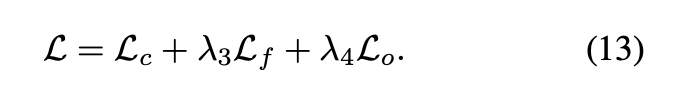


### 细粒度匹配Loss

参考LOFTR



### 整体Loss



1. 实验

我们通过在真实数据集ScanNet和模拟数据集Tartanair上的实验来验证我们方法的有效性。虽然MegaDepth[19]是一个常用于图像匹配的数据集，这也是一个真实的数据集，但我们没有采用它，因为它的深度地图误差比ScanNet的大。TartanAir是一个提供真实相机姿态和深度信息的模拟数据集，包括了像MegaDepth一样的室外场景。我们使用ScanNet和TartanAir数据集来展示在深度不准确的情况下，该算法的泛化能力和适应性，以及其在室内和室外场景中的有效性。

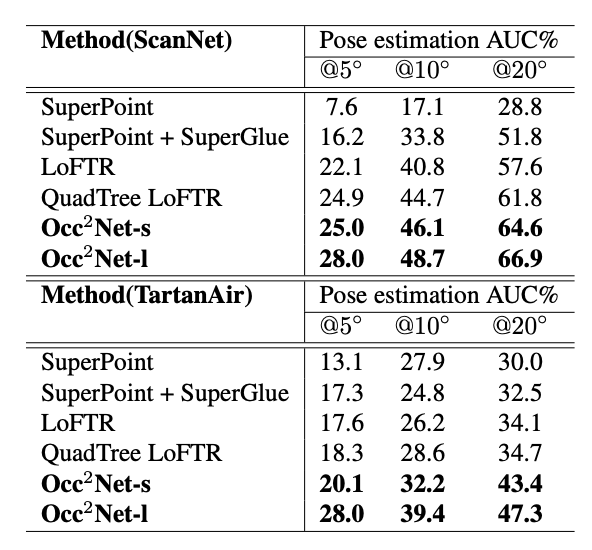
我们使用ScanNet来演示我们的姿态估计方法的有效性。ScanNet是一个RGB-D视频数据集，包含来自1500多次扫描的250万视图，这些视图注有3D相机的姿态、表面重构以及实例级的语义分割。参照SuperGlue的评估程序，我们在训练中抽取了230百万的图像对，它们的重叠得分在0.4到0.8之间。图像对的重叠得分是通过使用基础的深度和姿态将一幅图像重新投影到另一幅图像并计算在投影后不超出边界的像素的比例来定义的。我们在测试集中的1500对图像上评估我们的方法。所有的图像和深度图都被调整到了640x480的大小。

TartanAir[43] 是一个用于评估SLAM算法的具有挑战性的合成基准。该数据集包含室内和室外场景，涵盖了各种各样的场景和运动模式。数据是在各种光照条件、天气以及移动物体存在的情况下，在逼真的模拟环境中收集的。与使用物理数据收集平台收集数据的ScanNet不同，TartanAir数据集中的深度和位姿完全准确。我们按照与Droid-SLAM相同的测试切分方式构建了测试集[40]，该测试集包含32个场景。在每个场景中，我们随机选取了满足重叠分数在0.4和0.8之间且遮挡比率大于0.3的50对图像，从而得到了包含1600对图像的测试集。

### 实验结果

上图展示了ScanNet、TartanAir-室内和TartanAir-室外的匹配示例。我们将QuadTree LoFTR [39]与我们的结果进行了比较。绿色和黄色的线分别表示QuadTree LoFTR和Occ2Net的正确匹配，而红色的线则代表错误的匹配（误差超过10像素）。我们用紫色的框突出显示了一些匹配对。蓝色的遮罩标注了正确的匹配区域。图(a)的左图和右图分别包含了讲台和椅子的不同视角。图(b)的左图包含了墙壁和，而且部分被显示屏遮挡。图(b)的右图仅包含墙壁。图(c)的左图只有灯光和，而右图包含了灯光、和。

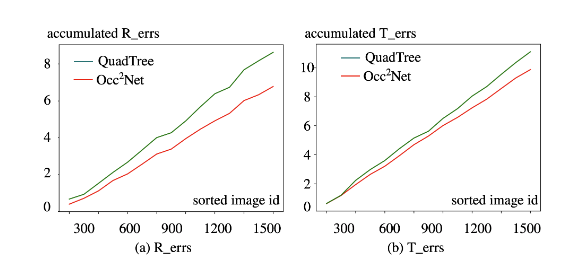
我们遵循SuperGlue的研究方法superglue，报告姿态错误在几个阈值5度， 10度, 20度下的百分比AUC。这里的姿态错误被定义为旋转中的角度误差和平移中的位移误差的最大值。为了恢复相机的姿态，我们解出预测匹配中的本质矩阵，使用RANSAC方法[8]。



如上表所示，我们的方法在ScanNet和TartanAir上的姿态估计精度上都超过了所有竞争对手，展示了我们的方法在室内外、真实和模拟场景中的有效性。Occ2Net-s是Occ2Net-l的一个缩略版本，它没有明确区分可见和遮挡的点，而且所有匹配都共享权重和张量，参数比QuadTree LoFTR少。

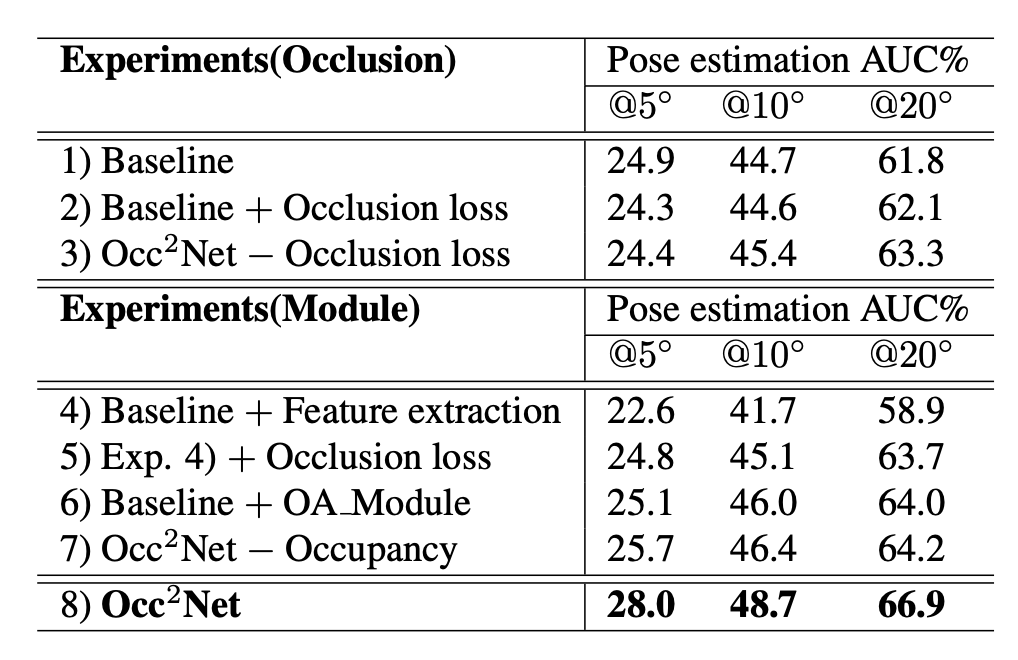
共享权重对最终性能有一定的影响。如我们之前解释过的，准确估计有遮挡的匹配点的确切位置是困难的，所以5度的提升效果并不明显。然而，没有共享权重的模型相当于将有遮挡的匹配点和可见的匹配点加在一起。这两种匹配不会互相影响，而且精度的提升是明显的。TartanAir测试集包含了更多遮挡比例大于40%$的图像对，所以匹配效果得到了更大的提升。

### 遮挡对于位姿估计的有效性



我们为了量化分析遮挡对姿态估计的影响，我们计算了ScanNet测试集中所有图像对的遮挡比率。遮挡比率定义为在一张图像中可见但在另一张图像中被遮挡的像素的比率。接着，我们将累积估计误差R和t作为遮挡比率的函数进行绘图。如图所示，在具有高遮挡比率的图像对中，其他算法在姿态估计方面的误差相对较大，而我们的算法仍能在这些具有挑战性的场景中达到较小的姿态误差。

### 消融实验



以上表格各实验含义：1）基准线：QuadTree LoFTR。2）基准线+遮挡损失：仅添加可见-遮挡匹配以进行监督；3）Occ2Net -遮挡损失：使用Occ2Net，而不进行遮挡监督；4）基准线+特征提取：使用我们的特征提取骨干网络，取代LoFTR特征提取骨干网络；5）实验4)+遮挡损失：使用我们的特征提取骨干网络并添加可见-遮挡匹配以进行监督；6）基准线+OA模块：添加旋转对齐方法；7）Occ2Net -占用：Occ2Net不进行3D占用估计；8）Occ2Net：我们的方法。

如表格1)和2)所示，仅增加对可见-遮挡点的监控是无效的。为了实现可见点和遮挡点的匹配，需要设计一个能够意识到遮挡并推断遮挡位置的网络结构。比较3)和8)，我们的Occ^2网络需要与遮挡的概念配合，否则效果提升不明显。比较1)和4)，我们的特征提取为3D占用率估计提供了张量，并且比QuadTree LoFTR特征提取的参数少，这导致了可见点匹配效果差。比较4)和5)，我们的特征提取对于遮挡点匹配是有效的。6)与1)的比较说明，增加旋转对齐对于图像匹配是有帮助的。如8)所示，与7)相比，3D占用估计改进了带有遮挡监控的遮挡点匹配。此外，通过同时匹配遮挡和可见点，也增强了可见-可见匹配，从而改善了整体匹配性能。需要注意的是，表格中的所有方法都不能直接与Occ2Net-s进行比较，因为它们包含的参数数量不同。

1. 结论

我们提出了一种考虑到被其他物体遮挡的点的图像匹配方法。我们设计了一种网络结构，名为Occ2Net，它能在一定程度上感知到被遮挡点的存在。Occ2Net利用特征提取获取多尺度的全局和位置特征，这有助于猜测遮挡信息。OA模块使用注意力和旋转对齐，这对于在后续的粗到细的过程中获取更多正确的匹配对非常有用。OE模块使用3D占用估计来结合细节特征进行精细匹配。实验表明，我们的方法实现了这个目标，并大大提高了姿态估计的准确性。我们认为，我们的工作为未来的研究者探索遮挡场景下更好的图像匹配铺平了道路。

然而，尽管我们的网络可以识别被遮挡的点并将它们与可见点进行匹配，但很难找出这些隐藏点的确切位置。尽管我们首先提出了由可见点和隐藏点组成的光线用于图像匹配的想法，但由于数据限制，我们将光线退化为只有两点。在未来，应考虑涉及多层遮挡的更复杂的场景。