# Depth-Aware Multi-Grid Deep Homography Estimation with Contextual Correlation

论文地址：https://arxiv.org/abs/2107.02524

代码：<https://github.com/nie-lang/Multi-Grid-Deep-Homography>

## 摘要

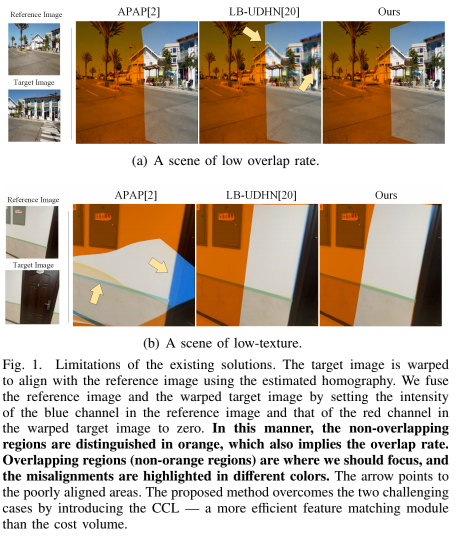
单应性估计是计算机视觉应用中的一项重要任务，例如图像拼接、视频稳定和相机校准。传统的单应性估计方法严重依赖特征对应的数量和分布，导致在低纹理场景中的鲁棒性较差。相反，学习解决方案试图学习强大的深度特征，但在重叠率低的场景中表现出不令人满意的性能。在本文中，我们通过设计上下文相关层 (CCL) 同时解决这两个问题。CCL 可以有效地捕获特征图中的远程相关性，并且可以灵活地用于学习框架中。此外，考虑到单个单应性不能代表具有视差的深度变化图像中的复杂空间变换，我们提出从全局到局部预测多网格单应性。此外，我们通过引入一种新颖的深度感知形状保留损失来为我们的网络配备深度感知能力。大量实验证明了我们的方法在合成基准数据集和现实世界数据集中优于最先进的解决方案。代码和模型将在 https://github.com/nie-lang/Multi-Grid-Deep-Homography 上提供。

## 介绍

单应性是从一个图像平面到另一个具有 8 自由度的可逆映射，其中 2 个用于平移，2 个用于旋转，2 个用于缩放，2 个用于无穷远处的线。它已被广泛用作各种计算机视觉应用的重要组成部分，例如图像拼接 [1]、[2]、视频稳定 [3]-[5]、相机校准 [6] 以及同时定位和映射 (SLAM) ) [7]。

传统算法以两种方式估计单应性。基于像素的方法通过以迭代方式优化两个图像之间的像素级对齐误差来更新单应性参数[8]，通常无法处理低重叠率的场景。基于特征的方法使用不同的特征提取器[9]-[12]和不同的鲁棒估计策略[13]-[15]基于稀疏特征对应来搜索最佳单应性模型。然而，它们在很大程度上依赖于特征对应的数量和分布，导致在低纹理场景中的鲁棒性较差。

与传统算法相比，提出了基于学习的解决方案 [16]-[20] 来处理具有挑战性的低纹理场景，以实现其鲁棒的深度特征提取。然而，通过简单地堆叠卷积层来探索深度特征图之间的匹配关系并不高效。因此，深度学习解决方案通常在重叠率较低的场景中表现出不令人满意的性能。如图1（a）所示，当重叠率较低时，在现有的深度单应性解决方案[20]的结果中可以发现明显的错位。



此外，对于单一的估计单应性，无论是传统解决方案还是深度学习解决方案，都无法将图像与由不同深度平面和不同基线引起的视差完美对齐。为了解决这个限制，提出了非线性多单应性估计算法[2]、[21]-[23]来处理具有挑战性的视差。通常，这些方法将图像划分为密集网格，并且可以使用稀疏特征对应关系为每个网格估计不同的单应性。然而，这些多网格方法对特征对应的数量和分布有更高的要求。在低纹理或低分辨率场景中，它们往往表现出比单个单应性模型更差的性能。如图 1 (b) 所示，现有的传统解决方案 [2] 在低纹理室内案例中失败。

为了同时克服传统算法在低纹理场景中的鲁棒性差和深度学习算法在低重叠率情况下失败的问题，我们设计了一个上下文相关层（CCL）来显式地捕获深度特征图之间的匹配关系。所提出的层是一个灵活的模块，可以很容易地嵌入到其他匹配相关的网络中，使深度学习框架在低纹理和低重叠率场景中都具有鲁棒性。与成本量[24]、[25]相比，所提出的CCL在深度单应性领域具有更高的精度、更快的速度和更低的内存消耗等优点。此外，为了打破单个单应性无法将图像与视差对齐的限制，我们提出了一种多网格深度单应性网络，用于在端到端框架中从全局到局部估计多网格单应性。具体来说，所提出的网络首先预测全局单应性，以粗略地对齐图像并获得网格的原始形状。然后估计残余网格变形，以使我们的网络能够将图像与视差对齐。

所提出的框架可以使用像素级内容对齐损失以无监督的方式轻松训练。但是，单独的内容约束往往会导致网格不自然的变形，例如自相交。现有的方法[26]、[27]通过添加形状保留约束来处理这个问题，鼓励网格中的所有网格都是矩形并且相邻的网格具有相似的形状。然而，这种措施本质上是内容对齐和自然网格形状之间的权衡。在本文中，我们从视差的角度重新思考了这种权衡。提出了一种深度感知形状保留损失，以同时促进内容对齐和自然网格形状。形式上，我们为所有网格分配不同的深度级别，并且只有相同深度级别的网格才会被强制执行以保持形状一致。对于不同深度的网格，不会有形状约束，尽可能鼓励内容对齐。

大量实验表明，我们的算法在合成基准数据集和现实世界数据集中都通过了最先进的单应性估计方法。此外，还提供了提议的 CCL 和成本量之间的综合比较，以支持我们的优势。

总而言之，我们将我们的贡献总结如下：

• 我们设计了一个新的上下文相关层（CCL）来明确地探索长期特征相关性。它在深度单应性领域的准确性、参数数量和速度上都优于传统的成本量。

• 我们提出了一个端到端的多网格单应性估计网络，使我们的网络能够将图像与视差对齐。

• 为了避免对齐和形状之间的权衡，提出了一种深度感知形状保留损失，以同时改善内容对齐和自然网格形状。

## 相关工作

### 2.1 线性单应性估计

基于像素的解决方案迭代地估计最优同系以最小化对齐误差，例如 L1 范数、L2 范数和负归一化相关 [8]。但处理重叠率低的场景具有挑战性。

与基于像素的解决方案不同，基于特征的方法通过基于稀疏特征对应关系最小化重投影误差来估计单应性。这些方法的第一步是使用特征点提取器检测稀疏特征点，例如 SIFT [9]、ORB [10]、BRISK [11] 和 SOSNet [12]。在特征匹配之后，将采用具有异常值拒绝的鲁棒估计算法来解决特征对应的单应性，例如 RANSAC [13]、MAGSAC [14] 和 MAGSAC++ [15]。基于特征的解决方案在重叠率较低的场景中效果很好。尽管如此，它们的性能在很大程度上取决于特征对应的质量，通常在低纹理、低光或低分辨率场景中失败。

与基于特征的方法相比，基于学习的解决方案以其强大的特征提取能力解决了低纹理场景的挑战。 DeTone 等人。 [16] 首次将深度学习应用于单应性估计，开发了 VGG 风格的单应性回归网络。在他们的开创性工作中，网络预测了 4 个顶点的 8 个运动，而不是直接求解单应性。通过将空间变压器网络（STN）[28] 应用于单应网络，Nguyen 等人。 [17] 建议以无监督的方式使用像素级光度损失来训练回归网络。张等人。 [18] 和 Le 等人。 [29] 建议通过学习内容感知注意掩码来拒绝前景和动态对象。张等人。 [30] 设计了一个级联的 Lucas-Kanade 网络，以逐步细化多尺度深度特征中的估计单应性。为了解决卷积特征匹配效率低的问题，Nie 等人。 [19] 连接特征金字塔和代价量来回归残差单应性。虽然成本量显着提高了网络的感受野，但内存和时间消耗却显着增加。在本文中，所提出的 CCL 可以作为深度单应性估计中成本量的更好替代方案，具有提高的准确性和降低的计算复杂度。

### 2.2 Non-Linear Multi-Homography Estimation

传统的多单应性估计在重叠区域追求精确的局部对齐。萨拉戈萨等人[2] 建议在图像域上放置一个网格，并为每个网格计算一个局部单应性，以实现尽可能投影 (APAP) 扭曲。为了获得更自然且投影失真更少的结果，Lin 等人[21]通过将单应性模型与相似变换联系起来来细化局部扭曲。为了获得更好的对齐性能，Li 等人[31]提出了一种稳健的弹性翘曲（robust ELA），可以看作是基于网格的模型和直接变形策略的结合。在 [32] 中，Liao 和 Li 提出了两个单透视扭曲，以保持透视一致性并减少投影失真。李等人[22] 将图像划分为超像素，并根据从变形残差向量计算出的最佳单应性对它们进行自适应变形。此外，贾等人[33]利用线点一致性约束来保留宽视差图像拼接的结构。

这些传统的多单应性解决方案可以实现比单一单应性解决方案更好的对齐效果，尤其是在有视差的场景中。然而，为了准确计算不同的局部扭曲，它们对特征点的质量有更严格的要求，在低纹理或低分辨率场景中的性能通常比单一单应性解决方案差。

现有的基于学习的多单应性解决方案与视频稳定有关，可以看作是一个小的基线对齐应用程序。王等人[26] 提出学习一个多网格变换网络来对齐相邻的视频帧。为了稳定地训练该网络，采用网格间一致性项和网格内规则性项来保持网格的形状。在 [34] 中，提出了 MeshFlow [5] 的深度学习替代方案，它具有内容感知能力，可以拒绝小基线场景中的动态对象。这些方法消除了对特征点的依赖，并在低纹理场景中表现出显着提高的鲁棒性。

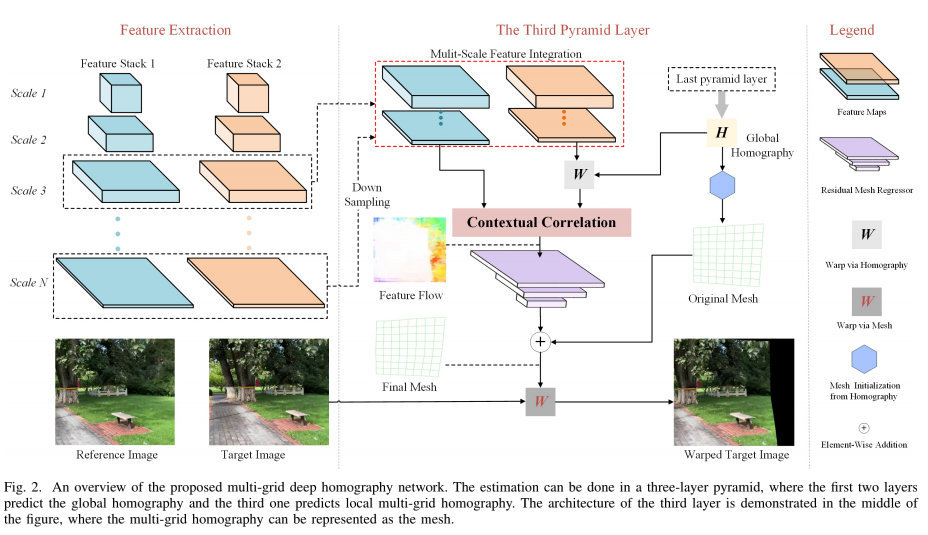
然而，在图像拼接等低重叠率的应用中，这些学习框架并不健壮，甚至无法工作。 相比之下，所提出的网络可以很好地解决这个问题。

## 方法

### 3.1 网络总览

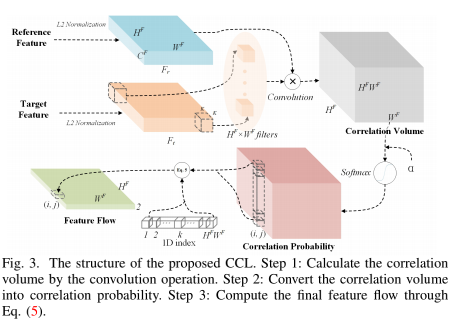
所提出的算法是一种基于学习的多网格深度单应性解决方案，其架构如图2所示。给定参考图像 和目标图像，我们的目标是回归可以扭曲目标图像以与参考图像对齐的多网格单应性。继之前的工作 [16]、[17] 之后，我们将单应性表示为四个顶点的八个运动，可以将其表示为大小为 2×2×2 的矩阵。然后，U×V 多网格单应性可以表示为大小为 (U + 1) × (V + 1) × 2 的矩阵。

我们采用一系列具有共享权重的卷积池块来提取不同尺度的输入图像的深度特征（, , k = 1, 2, ..., N）。然后我们对这些不同尺度的特征进行resize和堆叠，形成一个特征金字塔，其中第l层整合了从尺度l到尺度N的多尺度特征。接下来，我们使用特征金字塔的顶部（l = 1 , 2, 3) 从集成的多尺度特征中预测多网格单应性。在我们的设计中，第 l 层将第 (l-1) 层的预测单应性作为当前层的已知先验信息，并预测残差单应性。具体来说，前两个金字塔层预测全局单应性，而第三个金字塔层预测 U × V 多网格单应性。通过这种方式，我们预测从全局到局部的多网格单应性，使我们的网络能够处理具有视差的图像。



### 3.2 上下文相关性

现有的深度单应性解决方案在重叠率较低的场景中效果不佳，因为卷积层无法有效地探索远程相关性。为了解决这个问题，Nie 等人[19] 将代价量引入网络以显式增强特征匹配能力。然而，代价量的应用显着增加了空间和时间成本。为了缓解这个问题，我们提出了 CCL，其结构如图 3 所示。它以 、 的特征图作为输入并输出密集的特征流。我们将特征流定义为包含垂直和水平运动的特征对应的运动。



与成本体积相比，所提出的模块具有以下优点：（1）在深度单应性方面具有更好的性能，（2）更快的速度，（3）更少的内存消耗，（4）由于patch-to而在相关匹配方面更加稳健补丁匹配而不是点对点匹配。前三个优点将在第 IV-D 节的实验中得到证明，最后一个将在下面实现的第一步中解释。

接下来，我们介绍实现的三个步骤：

1. **相关量（Correlation Volume）**

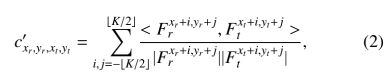
我们首先使用通道维度上的 范数对多尺度特征（如图 3 顶部所示）进行归一化，如 ， ∈ ××，然后提取它们之间的潜在相关关系。

代价体积[25]从，中提取全局相关关系为形状为××(2+1)(2+1)的3D体积，用余弦相似度表示相关性如下：



其中 () 和 () 分别表示 ， 的空间位置。c 表示 和 的余弦相似性。

与计算点对点相关性的成本体积相比，我们在这一步计算任意 K × K 区域之间的补丁对补丁相关性。如图 3 顶部所示，我们首先从 中提取密集块 (K × K)，步幅设置为 1。然后我们将这些块堆叠为卷积滤波器，并使用它们对 执行卷积操作。我们将卷积的输出称为相关体积，其形状为 ××。 该体积中的每个值表示一对任意区域之间的相似性，可以表示为等式(2):

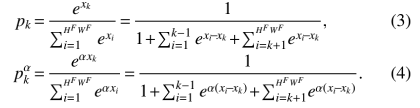


其中 wpsoffice 表示向下取舍的操作。对比代价量的形状，本文提出的相关量仅是其四分之一。

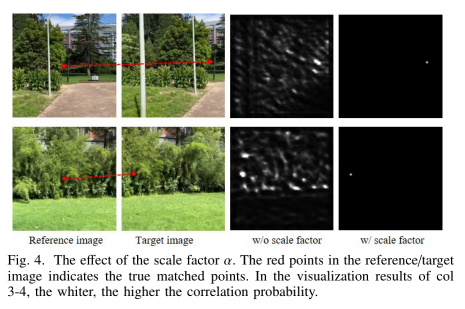
1. **Scale Softmax**

相关量中的每个位置都可以看成一个长度为 的向量。 我们设置 K = 3，这个向量中的每个值都在 0 到 9 之间。然后我们使用 softmax 函数来激活这些向量，将特征匹配转换为分类问题（ 类别）。以这种方式，相关量被转换为相关概率，如图 3 右侧所示。受 [35] 的启发，这些向量将在激活之前乘以恒定比例因子 α (α > 1) 以增加类间距离。

将相关量向量的每个位置表示为 ，且假设 。接着就可以通过如下公式计算得到带有或者没有比例因子 α的匹配 (k的概率：



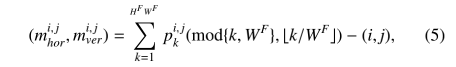
比较公式 3 和公式 4 的分母，我们可以观察到，这个比例因子可以使匹配概率低的 更低（当 k 远小于 时），而匹配概率最高的 更高（当 k = ）。 换句话说，**弱相关被抑制，而最强相关被增强**。为了进一步验证这个观点，图 4 可视化了在 softmax 之后相关体积中心的向量，无论是否具有二维比例因子。可以看出，在没有比例因子的情况下，目标图像域（col 3）中的许多区域与参考图像的中心点相似。相比之下，大多数不匹配的区域都被这个因素拒绝（第 4 列）。



1. 特征流（Feature Flow）

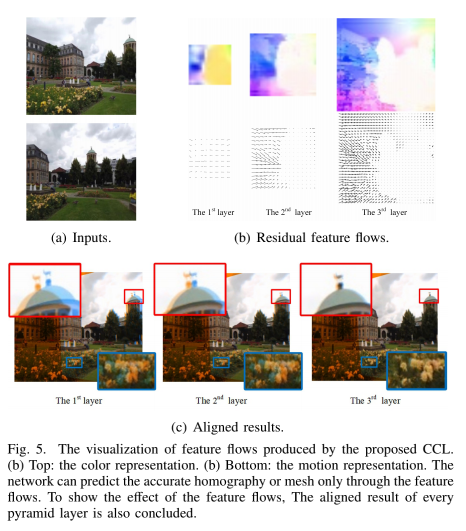
在这一步中，我们重新思考深度单应性估计的本质——回归目标图像的 2 个正交方向上的 4 个顶点的 8 个运动。基于这种理解，我们认为，如果我们从表示特征图中密集相关性的特征运动中回归单应运动，预测单应运动的难度会降低。因此，我们建议将相关量转换为特征流——**特征级别上特征对应之间的密集特征运动**。

设 表示在位置 (i, j, k) 激活 scale softmax 的相关量值，位置 (i, j) 处的特征运动 (, ) 可以计算为如下：

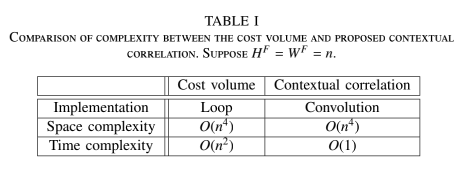


其中 mod{,} 表示模运算。

最终，我们得到从参考特征到目标特征的特征流（××2）作为后续回归网络的输入。我们将图 5 中的特征流可视化，其中中间特征流和每个金字塔层中的最终对齐结果成对显示。



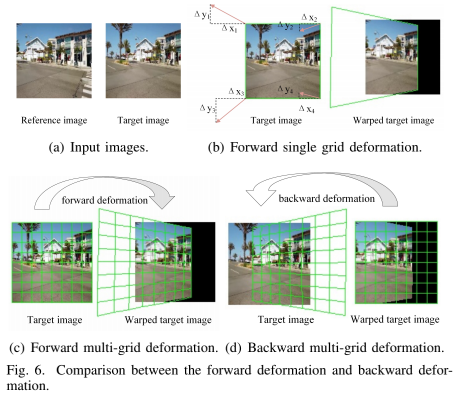
此外，表 I 中提供了成本量和我们的 CCL 之间复杂度的比较。与成本量相比，CCL 通过**拒绝大多数不匹配的位置来输出特征流的轻量级表示**，从而实现更有效的预测。



### 3.3 反向多网格变形(Backward Multi-Grid Deformation)

现有的深度单应性解决方案是预测目标图像中4个顶点的8个运动，而不是直接求解单应性中的8个未知参数。通过这 8 个动作，我们可以求解相应的单应性并将目标图像扭曲以与参考图像对齐，如图 6 (a)(b) 所示。我们将这种**从目标图像域到参考图像域的运动方向称为前向变形**。然后，通过在目标图像上放置一个网格，我们可以将单个网格变形（1×1）扩展到多网格变形（U×V），如图 6（c）所示。与每个像素共享相同单应性的单个网格扭曲不同，多网格扭曲必须为扭曲的目标图像中的每个像素分配不同的单应性。换句话说，在多网格变形中，我们需要弄清楚变形后的目标图像中的像素属于哪个网格。由于扭曲目标图像中的网格形状在前向变形后是不规则的，因此在**深度学习实现中很难使用快速有效的方式获得扭曲目标图像中每个网格和每个像素之间的对应关系**。因此，如果继续采用前向变形，可以大幅度降低速度。

为了避免这个问题，我们设计了一个向后变形的解决方案。具体来说，如图6（d）所示，我们在扭曲的目标图像中放置一个规则网格，然后预测从参考图像域到目标图像域的网格运动。与前向变形相比，**在所提出的后向变形中，扭曲的目标图像中的网格形状是规则的**。因此，我们可以很容易地为扭曲的目标图像的每个像素批量分配不同的单应性，从而在 GPU 中实现并行加速。



### 3.4 无监督训练

所提出的多网格深度单应性网络以完全无监督的方式进行训练，因为在现实世界场景中没有基本事实。在本文中，我们设计了一个内容对齐项和一个形状保留项来优化我们的网络。内容对齐项用于对齐图像内容中的输入图像，而形状保留项旨在防止网格形状出现不自然的失真。

**1) Content Alignment Loss.**

给定一个参考图片 和带有重叠区域的目标图片 ，网络的目标是对齐重叠区域。令 表示采用特征金字塔中第 k 层预测网格的空间变换操作，则对该层的内容对齐约束如下所示：



其中 E 表示全为 1 的矩阵， 表示点乘操作。

本文提出的网络包含 3 层金字塔，所以内容对齐损失如下所示：



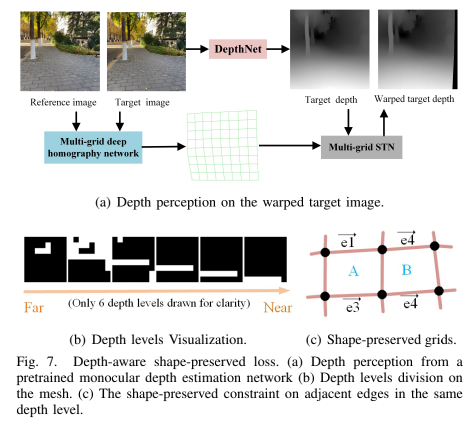
其中 分别表示不同金字塔层的权重。

1. Depth-Aware Shape-Preserved Loss

单独使用内容对齐损失来优化网络会造成不自然的网格失真问题，比如自相交( self-intersection)。为了避免这个问题，在[26,34]中提出了一个形状保持损失，它鼓励相邻网格保持相似形状。然而，这种约束可以很容易地从相邻网格传播到周围环境，强制所有网格保持相似的形状。 这种方式将所有网格视为在同一平面上，以降低内容对齐性能为代价提高形状规则性。

在本文中，我们重新思考了这种形状保持损失：**理想的损失应该只对相同深度的网格施加保形约束**，而不是对不同深度的网格施加保形约束。为了实现它，我们设计了一个深度感知的形状保留损失，它可以估计网格上的不同深度级别并将网格约束在相同的深度级别以保持相似的形状。

计算深度感知形状保留损失的过程如图7所示。首先，我们采用预训练的单目深度估计模型[36]来预测目标图像的深度图。然后，我们使用从我们的网络估计的多网格单应性来扭曲深度图。 获得扭曲深度图后，我们计算扭曲深度图的每个网格中的平均深度值。 接下来，我们将扭曲的深度图划分为具有相同间隔的 M 个不同级别 (, k = 1, 2, ..., M)。 深度级别在图 7（b）中可视化，其中白色表示这些区域处于相同的深度。



假设所有网格都处于相同的深度水平，我们可以用相同的单应性扭曲它们，变形网格中的每条线都是一条直线。基于这一观察，我们在相同深度的相邻网格中约束网格边缘的方向一致。 一个例子如图 7 (c) 所示，其中网格 A 和 B 之间的相似性可以表述如下：



根据等式(8)，我们可以计算变形网格水平和垂直方向的相似矩阵()，它们的大小分别是 U X (V-1)和 (V-1) X U。最后，提出的深度感知形状保持损失如下所示：



其中 表示根据 计算得到的水平和垂直方向的深度一致性矩阵。在 ((V-1)XU) 的每个元素表示相邻网格（水平方向或垂直方向）是否处于同一深度级别。

1. **Objective Function.**

同时考虑内容对齐项和形状保留项，我们得出网络的目标函数：



其中 和 分别表示两个损失的权重。

## 实验结果

### 4.1 数据集和实现细节

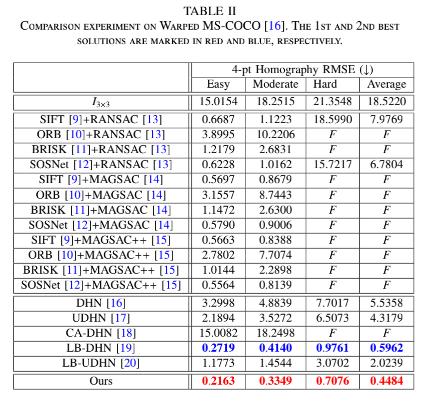
**数据集**。我们在两个公共数据集中验证了所提议网络的性能。第一个是合成基准数据集，称为 Warped MS-COCO [16]。该数据集中的样本是没有视差的图像对，因此可以使用单个单应性将目标图像与参考图像完美对齐。第二个数据集是在无监督深度图像拼接工作 [20] 中提出的真实数据集，其中 10,440 个图像对用于训练，1,106 个图像对用于测试。该数据集称为 UDIS-D，由不同的重叠率、不同程度的视差和可变场景组成，例如室内、室外、夜间、黑暗、雪地和缩放。

**实现细节**。我们的网络使用 Adam [37] 优化器进行训练，该优化器具有指数衰减的学习率，初始化为 进行 500k 次迭代。批量大小设置为 4，我们首先训练网络进行 300k 次迭代，其中 λ 和 μ 分别设置为 1 和 0。对于剩余的 200k 次迭代，我们将 λ 和 μ 设置为 1 和 10。并且 ， 和 α 被指定为 1、4、16 和 10。我们使用 RELU 作为所有卷积层的激活函数，除了回归网络的最后一层不采用激活函数。该实现基于 TensorFlow，网络在具有 NVIDIA RTX 2080 Ti 的单个 GPU 上执行。对齐512×512分辨率的图像大约需要96ms。

### 4.2 在合成基准数据集的比较

由于 Warped MS-COCO 是一个没有视差的合成数据集，我们只将我们的解决方案与其他单应性解决方案进行比较。此外，为了公平，我们修改了回归网络的第三层金字塔来预测单个单应性而不是多单应性。由于GT的可用性，我们以有监督的方式训练我们的网络。

我们首先将我们的方法与传统的基于特征的解决方案进行比较。如表 II 所示，我们选择了各种特征描述符，例如SIFT[9]、ORB [10]、BRISK [11] 和 SOSNet [12]，以及不同的异常值拒绝算法，例如RANSAC [13]、MAGSAC [14]、MAGSAC++ [15]，组成 12 种不同的解决方案。在这些特征描述符中，采用 SOSNet 将 SIFT 的局部描述符替换为深度学习特征。



然后我们将我们的与基于深度学习的解决方案进行比较，例如DHN [16]、UDHN [17]、CA-DHN [18]、LB-DHN [19] 和 LB-UDHN [20]。包括我们在内的所有学习解决方案都在这个数据集中进行了训练。

我们采用来自 [17]、[20] 的 4-pt 单应 RMSE 作为评估上述所有解决方案的指标。形式上，我们根据性能将测试结果分为 3 个级别——简单（top 0-30%）、中等（top 30-60%）和hard（top 60-100%）。 将 3×3 单位矩阵作为参考的“无扭曲”单应性。

结果如表 2 所示，其中 F 表示该解决方案的性能比 差。 从该表中观察，很明显可以得出以下结论：

(1) 我们的网络在所有情况下都优于其他解决方案，包括传统和深度学习算法（“简单”、“中等”和“困难”）。

(2) 深度学习解决方案更稳健，因为它们在“硬”列中优于传统解决方案。这一优势得益于 CNN 强大的特征提取能力。

在实验中，我们没有将我们的与[30]、[38]进行比较，因为它们本质上是模板匹配算法。 具体来说，他们将在这些方法中将一个大的参考图像和一个小的模板图像作为输入，其中输入之间的重叠率可以是 100%。 相比之下，这些比较方法和我们的方法处理重叠率相对较低的输入。

### 4.3 在真实数据集的比较

**定量比较**

包含丰富深度级别的真实世界图像对是在不同的相机基线下捕获的，这表明由于视差，它们无法使用单个单应性对齐。因此，在这个真实世界数据集（UDIS-D）中，我们添加了与多单应性解决方案的比较——APAP [2]、AANAP [21]、鲁棒 ELA [31]、SPW [32] 和 LCP[33] .由于我们网络中的网格是 8 × 8（我们的网络中采用 8×8 网格的原因将在第 4.5 节中解释），将我们的网格与这些将网格设置为 8 × 8 的方法进行比较是公平的。为了进一步突出我们的性能，我们还将它们与设置为 100 × 100 的网格进行比较。

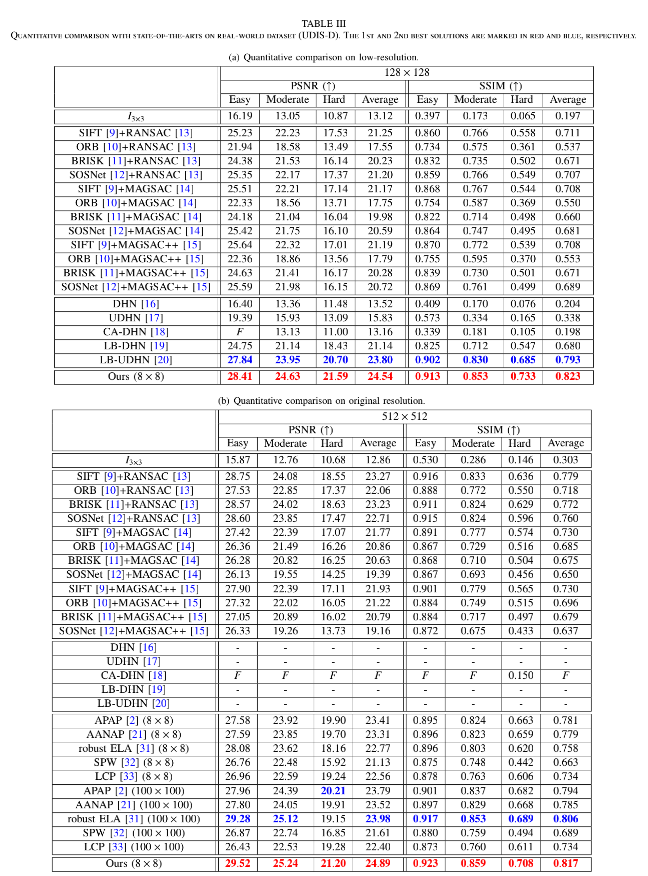
此外，我们在 UDIS-D 中以两种不同的分辨率进行实验，如表3 所示。形式上，UDIS-D中图像的分辨率为512×512，我们将它们调整为128×128以获得低分辨率版本，原因有两个：

1. 低分辨率图像可以模拟具有较少特征点的具有挑战性的场景。
2. 2）由于全连接层的存在，大多数现有的深度学习解决方案只能在固定分辨率（128×128）下工作。

在 [20] 之后，我们使用 PSNR 和 SSIM 在重叠区域中评估性能，可以计算如下：



其中 PSNR 和 SSIM 分别表示对两张图片计算 PSNR 和 SSIM 的操作。



根据表 3，可以观察到：

(1) 我们的方法在不同分辨率（128×128 和 512×512）的每种情况下（“easy”、“moderate”和“hard”）都优于所有单应性估计方法。

(2)**传统解决方案的性能随着分辨率的降低而下降**，因为在低分辨率场景中特征检测的性能会下降。

(3) 大多数基于学习的单应性解决方案效果不佳，因为该数据集中**图像对之间的重叠率很低。卷积层或成本量无法有效地学习远程相关性**。

(4) 最新的传统多单应性解决方案（SPW 和 LCP）在 UDIS-D 中表现出不令人满意的性能。主要原因是它们引入了除点线对齐项之外的其他约束，例如失真项和线保留项，以**牺牲对齐性能为代价来减少投影失真**。同时，点线对齐并不能在每个场景中都很好地发挥作用，尤其是在匹配的点和线不足的情况下。

**定性比较**

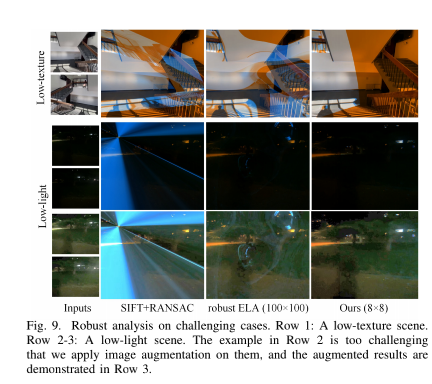
除了表 III，我们还在这个真实世界数据集的原始分辨率中提供了这些方法之间的定性比较。我们选择 SIFT 和 RANSAC 的组合作为传统单应性解决方案的代表作品，因为它在 UDIS-D 中实现了最佳对齐性能。此外，LB-UDHN 被选为基于深度学习的解决方案的代表作品，以获得最佳性能。为了将其应用于具有更高分辨率（512×512）的图像，我们将估计的单应矩阵乘以尺度变换矩阵以适应其他分辨率。至于多单应性解决方案，我们将网格设置为 100×100 来展示它们的结果。图 8 中显示了7个实例，其中每个实例包含不同程度的视差。我们通过将参考图像中蓝色通道的强度和变形目标图像中红色通道的强度设置为零来融合参考图像和变形目标图像。以这种方式，非重叠区域以橙色显示，重叠区域中的错位以不同的颜色突出显示。尽管所提出的方法不能消除所有的错位，但我们的结果中剩余的错位比其他方法要少。



**鲁棒性分析**

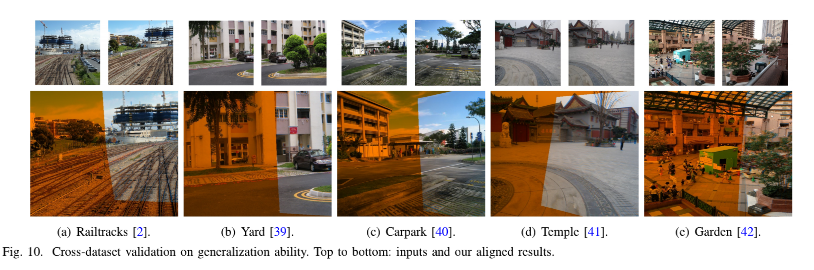
稳健的方法不一定需要良好的平均性能（参见表3 的“平均”列），但要求最差的性能不能很差（参见表3 的“硬”列）。 比较表 3(a) 和表 3(b)，传统解决方案在高分辨率下的性能明显更好，因为随着分辨率的增加，特征点更加丰富。

但是，对于一些具有挑战性的场景，例如低光或低纹理，分辨率的提高并不能提高性能。 图 9 展示了两个具有挑战性的示例。当场景中缺少纹理或光线太低时，手动设计的特征描述符，例如 SIFT、ORB等不适用，导致特征点少或特征对应不匹配。与它们相比，深度学习解决方案通过自适应地学习数据集样本的分布来提取特征图。因此，学习解决方案更加稳健。



**跨数据集验证**

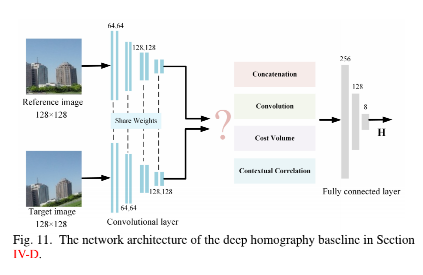
对于大多数深度学习方法来说，泛化是一个缺陷。在本节中，我们验证了所提出方法跨数据集的泛化能力。具体来说，我们在 UDIS-D 中训练我们的网络并在其他数据集中对其进行测试。我们从经典的图像拼接论文 [2]、[39]-[42] 中收集数据集，其中这些数据集是从不同的场景中捕获的，并且包含不同程度的视差。结果如图10所示。即使在其他数据集上进行测试，我们的解决方案仍然具有良好的对齐能力。尤其是这些示例中输入图像之间的重叠率非常低，大多数深度学习解决方案在这些场景中可能会失败。



### 4.4 对比 Cost Volume

除了图像对齐性能外，我们还将提出的 CCL 与单应性估计领域的成本量进行了比较。CCL 和成本量都是**用于在没有学习参数的情况下提取特征对应之间的匹配信息的模块**。它们都可以很容易地插入神经网络。

起初，我们将 CCL 替换为我们网络中的成本量。尽管如此，这种替换会带来网络参数数量的急剧增加，超过 GPU 的最大内存导致训练崩溃。因此，我们设计了一个相对简单的网络作为基线，并将提出的 CCL 与该框架上的成本量进行比较。基线架构如图 11 所示，其中采用 8 个共享参数的卷积层和 3 个最大池化层来提取深度特征。然后可以使用某种方法从从不同输入图像中提取的特征图中获得匹配信息。具体来说，直接连接两个卷积层，过滤器数量设置为 256，搜索半径设置为 16，并在不同的实验中采用所提出的 CCL。之后，我们使用三个全连接层来预测目标图像中 4 个顶点的 8 个运动。



实验是在一个 RTX NVIDIA 2080 Ti 上的 Warped MS-COCO 上进行的，所有方案都以监督方式训练了 80 个 epoch。结果如表IV所示，其中对性能、参数数量和速度进行了综合评价。从这张表中，我们可以观察到：

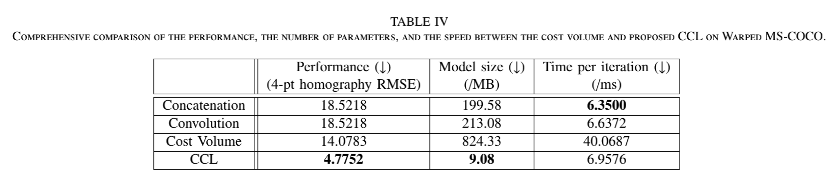
(1) “Concatenation”的性能几乎等于 （18.5220，如表二所示）。如果没有输入特征图之间的匹配关系，网络什么也学不到。

(2) “卷积”的性能与“连接”的性能几乎相同，这表明卷积层对匹配关系一无所知。有限的感受野可以解释这种现象。

(3)代价量确实有助于提取输入特征图之间的匹配关系，但效率不高。此外，参数和速度的成本也显着增加。

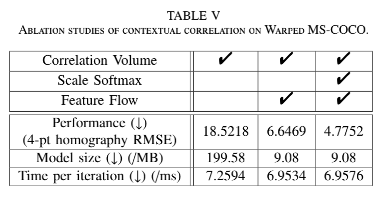
(4) 提议的 CCL 的性能远优于成本体积。此外，与“卷积”相比，这个高效的模块仅增加了 0.32ms 的运行时间。至于模型大小，我们的设计将参数从几百 MB 显着减少到 10 MB。

实际上，模型大小的减小主要归功于特征流。这种特征相关性的轻量级表示（H×W×2）显着减少了后续全连接层的参数。与成本量相比，特征流还拒绝了广泛的冗余匹配信息，从而对单应性进行更有效的预测。

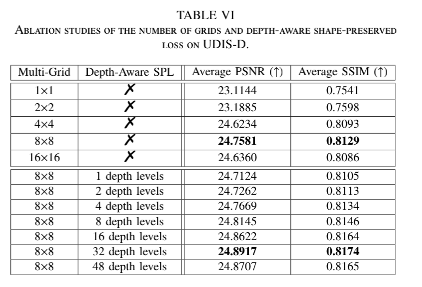


### 4.5 消融实验

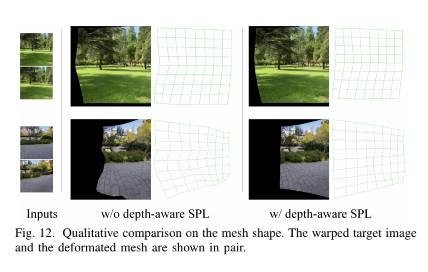
**上下文相关性**。CCL 可以分三个步骤实现，我们对这些步骤进行消融实验。我们在 Warped MS-COCO 上使用基线网络（如图 11 所示）验证每个阶段的有效性，结果如表 V 所示。“**特征流”可以将相关量转换为简单但有效的表示。而“Scale Softmax”可以通过有效抑制低匹配概率点对特征流的干扰来进一步提高匹配性能**。



**网格数**。网格数量的增加可以带来图像对齐的改进。但这不是绝对的，因为网络架构和数据集的大小可能会影响它。因此，我们在工作中探索了最佳网格数，结果如表 VI 的第 2-6 行所示。根据我们的实验，8×8 网格是最好的。



**深度感知形状保留损失**。保形约束可以保持网格的刚性形状。但它认为所有的网格都在同一平面上，在存在视差的情况下降低了网络的对齐性能。将图像网格划分为不同的深度级别可以避免这个问题。我们探索了我们网络中最好的深度层数，结果如表 VI 的第 5、7-13 行所示，其中第 5 行可以视为无形状保留约束，第 7 行可以视为全局形状-保留约束（如 [26]、[34] 所做的）。此外，图 12 中提供了网格形状的定性比较，以证明我们的深度感知形状保留损失的优越性。



## 结论

在本文中，我们提出了一种深度感知的多网格深度单应性估计网络，将图像与视差从全局到局部对齐，突破了现有单一深度单应性估计解决方案的限制。在我们的网络中，我们设计了 CCL 以有效地提取匹配关系，在性能、参数数量和速度方面都优于成本量。此外，还提出了一种深度感知形状保留损失，以同时提高形状规则性和对齐性能。大量的实验证明了我们对现有的单应性和多单应性解决方案的优越性。

但是，**网格的数量可能会受到网络架构和数据集大小的限制**。未来，我们将探索影响网格数量的原因，并在不降低对齐性能的情况下增加网格数量的上限。