**深度端到端飞行时间成像**

**摘要**

我们提出了一种用于飞行时间（ToF）摄像机的端到端图像处理框架。现有的ToF图像处理流水线包括一系列操作，包括调制曝光，去噪，相位展开和多径干扰校正。虽然这种级联模块化设计提供了多种好处，例如封闭式解决方案和节能处理，但它也会受到误差累积和信息丢失的影响，因为每个模块只能观察其直接前身的输出，从而导致错误的深度估计。我们偏离了传统的管道模型，并提出了一种深度卷积神经网络架构，可直接从双频原始ToF相关测量中恢复场景深度。为了训练这个网络，我们使用时间分辨渲染器模拟各种场景的ToF图像，设计深度特定的损失，并应用归一化和增强策略以将此模型推广到真实捕获。我们证明了所提出的网络可以有效地利用ToF频率测量的时空结构，并验证联合多路径去除，去噪和相位展开方法在各种具有挑战性的场景中的性能。

1. **介绍**

最近，微软的Kinect One等调幅连续波（AMCW）飞行时间相机不仅在互动商业应用中得到广泛采用，而且还成为计算机视觉中令人兴奋的成像模式结合传统的彩色摄像机，RGB-D数据可以实现高保真场景重建[26]，可以收集大型3D数据集，驱动3D深度学习[9]，用于场景理解[47,24]，动作识别[40] ，面部和姿势跟踪[31]。除了实现这些核心计算机视觉应用之外，RGB-D摄像机在人机交互，机器人技术以及新兴增强或虚拟现实应用中的跟踪方面具有广泛的应用[35]。由于低功率要求，低成本CMOS传感器技术和小型传感器照明基线[20]，AMCW飞行时间相机有可能成为基础成像技术。为简洁起见，我们将在下文中将AMCW飞行时间摄像机简称为ToF摄像机，隐含的理解是它们与其他飞行时间成像技术不同，例如使用SPAD进行直接时间采样（例如[49]）。

ToF相机通过使用周期性调幅泛光灯照亮场景来测量深度，该泛光灯沿直接光路和间接光路反射回相机。然后，摄像机测量入射信号相对于照明信号的相移。为了从这些原始相位测量中提取深度，必须解决许多具有挑战性的重建问题。对于场景中的单个漫反射器，相位测量仅明确地编码深度，直到整数相位包裹，这通过相位展开方法来解决[20]。在存在全局照明的情况下，多个光路径沿着直接和间接路径干扰，导致深度图的严重多径干扰（MPI）失真。最后，由于红外调制的低吸收深度，以及与RGB CMOS图像传感器相比不成熟的传感器技术[31]，原始ToF测量受到严重噪声的影响。

传统上，这三个重建问题，相位展开，MPI减少和去噪，都是在流水线方法中解决的，其中每个步骤孤立地处理单个子问题，如[11,14,34,39]。虽然这种设计有利于分而治之的算法，但它忽略了各个子模块之间的耦合，并在重建流水线中引入了累积误差和信息损失。例如，建立的多频率展开方法[12]在存在MPI或噪声时变得不准确，导致明显的展开错误并且随后形状恢复不准确。

我们提出了一种数据驱动方法，而不是构建重建流水线或依赖其他硬件，而是直接从ToF相机的原始调制曝光生成深度图（参见图1）。

具体来说，我们做出以下贡献：

* 我们通过联合解决相位展开，MPI补偿和原始相关测量的去噪，提出了一种基于学习的端到端飞行时间成像方法。所提出的架构明显优于传统的深度图像流水线，同时在现代GPU上具有高效的交互式帧速率。
* 我们验证了所提出的重建方法在模拟和实验获取的原始双频测量中有效地消除了MPI，相位包裹和传感器噪声。
* 我们引入了一个大规模的原始相关时间 - 关灯数据集，其中包含每个像素的已知地面实况深度标签。将公布数据集和体系结构，以实现所提方法的完全可重复性。

1. **相关工作**

**相位展开** 用于解决相位模糊的已建立的方法获得在两个不同调制频率下的测量[11]，通过用其较低频率对应物展开高频相位来保持长距离范围。虽然对于直接场景有效，但是在存在MPI的情况下，这种双频采集方法变得不准确。当无法获得多频测量时，可以利用统计先验，例如幅度平滑[20]和表面法线约束[10,12]。然而，我们的方法不是建立在这样的手工制作的先验上，它只模拟自然场景的丰富统计数据的子集。相反，我们直接从大量的训练数据中学习空间先验。

**多路径修正** MPI失真通常在后处理步骤中减少。大量的工作探索了简化的双路径或仅弥散问题的分析解决方案[17,14]，或者试图作为计算成本高昂的优化问题孤立地解决MPI [13,29,7]，并对其进行了很强的场景稀疏性的假设。虽然MPI，相位展开和去噪是耦合的，但是现有的重建方法都没有以联合和计算有效的方式解决它们。

**替代获取** 最近的替代方法试图解决捕获过程中的模糊性。古普塔等人[19]提出使用GHz范围内的高频调制来分离间接照明，由于现有的100MHz范围CMOS技术的限制，这仍然是理论上的。许多工作提出了混合结构光ToF系统[42,39,5]，由于投影机相机基线，需要编码和精心同步的照明，并且显着增大了面积，从而消除了ToF技术的许多固有优点。

**后处理** 我们不是第一个应用深度学习来解决ToF深度重建中的模糊性的人。Son等人[46]使用机械臂从结构光传感器收集具有相应地面实况标签的ToF范围图像，并训练前馈神经网络以消除多径失真。与我们的工作同时，Marco等人[37]训练编码器 - 解码器网络，该网络将ToF范围图像作为输入并预测多径校正版本。然而，这两种方法都不是端到端的，因为它们从特定类型的相机的管道输出后处理深度。原始ToF图像中呈现的大部分信息已经在作为这些方法的输入的深度图像中被破坏。通过忽略ToF重建中许多子问题的耦合性质，他们人为地限制了深度成像性能，正如我们在这项工作中所展示的那样。

**深度图像生成** 深度卷积神经网络已经在监督图像重建问题方面取得了巨大进步，包括去模糊[52]，去噪/修复[51]和超分辨率[28]。虽然这种前馈体系结构对于自然图像上的局部操作很有效，但是对于非局部逆问题，例如MPI移除，通常需要大的感受域。

最近，条件生成对抗网络（cGAN）已经在监督[25]和无监督[53]设置下显示出高质量的图像翻译结果。与传统的GAN [18]不同，在cGAN中，发生器G和鉴别器D都观察输入图像。通过将GAN损失与传统像素损失相结合，人们可以学会惩罚输出图像和目标图像之间的结构差异，而不依赖于领域知识[4]。我们采用这些成功的cGAN策略来训练我们的深度生成网络，并将它们与深度图上的像素丢失和平滑度项及其梯度相结合。

1. **飞行时间成像**

在本节中，我们将回顾ToF深度成像及其核心问题。

**深度获取** ToF相机利用周期性泛光照明和传感器解调信号测量场景的调制曝光。在Lange [32]之后，我们对积分时间T的原始相关测量建模：

其中是辐照度，是具有角频率和相位偏移的可编程参考信号。通常，为零均值，因此测量对环境光是稳健的。在不存在间接光路的理想情况下，通过捕获具有相同但不同的一对调制曝光，可以可靠地恢复每个像素处的场景深度。

是测量的相位，表示光速。虽然计算上简单，但应用式2时，实际数据通常会导致深度估算不佳。这不仅是因为传感器噪声，还因为式3的测量由于相位缠绕和MPI具有固有的模糊性，这需要解决下面描述的不适定重建问题。

**双频相位展开** 由于等式中的相位测量的周期性质。如图3所示，深度估计也“环绕”，并且对于小于调制波长的一半的距离，即在范围内，仅是明确的。双频方法通过在两个不同的频率和 [11]测量来消除与其他相位包裹候选者的真实深度的歧义。这有效地将最大的无模糊深度范围扩展到，其中表示两个频率的最大公约数。为了恢复未知深度，可以在候选深度∈和相位观测值之间创建查找表，并解决以下1D搜索问题[19]。

然而，在存在噪声的情况下，由于缺乏空间先验，这种理想化的逐像素方法经常失败。最近，Lawin等人[33]提出核密度函数估计作为手工制作的先前的更强大的相位展开。

**多路径干扰** 应用式2时的第二个核心挑战是，实际场景不是几个直接反射的路径，具有不同几何和材料属性的场景可以使多个光路的光线在单个传感器像素中线性组合，如图2所示。对于普通的正弦调制，这些路径混合导致测量与较长直接路径的测量相同，导致固有的模糊测量。将零差模式下的强度调制光源正式化为，成为的许多衰减和相移副本的叠加，沿着所有可能的相同行进时间的路径：

当在式1中被替换时，我们在[23]中模拟相关积分：

其中场景无关函数和已经折叠成，其仅取决于成像装置，并且可以预先校准。基本上，式6探测潜在的，依赖于场景的时间点扩散函数（TPSF，其第一峰值表示真实深度），，传感器观测值，，这促使我们设计学习框架，将在下一节中描述。

1. **飞行时间成像的学习**

在本节中，我们描述了所提出的重建架构和学习损失函数，它们允许我们直接从原始ToF测量中估计深度。为了建立这种端到端方法的直觉，我们使用方程式合成并分析角落场景的相关图像。 MPI引入了每像素相位偏移，不仅取决于场景特定属性，如距离，几何和材料[16]，还取决于调制信号。我们证明了通过在大量合成训练数据集中利用原始ToF测量的时空结构，可以学习从相关图像到深度图的逆映射。具体地，我们将深度生成视为多通道图像融合问题，其中期望深度图是在多个照明传感器配置处测量的相同场景的加权组合。我们的重建网络经过培训，融合了这些时空结构，共同执行MPI去除，去噪和相位展开，同时通过新的损失函数惩罚得到的深度图中的伪像。

* 1. **深度估算网络**

所提出的深度生成网络架构考虑了原始ToF测量的相关性质。与传统的RGB或灰度级强度图像相比，中的像素值对场景和相机设置更敏感，例如，照明信号的频率，相位偏移和功率。因此，理想的网络应该学习跨通道相关性，以及对反照率，幅度和尺度变化不变的空间特征。此外，输入相关性测量和输出深度图像都应与基础场景几何形状一致。虽然两者应该共享深度梯度，但反照率梯度不一定与深度边缘对齐，应该被拒绝。

有了这些动机，我们设计了一个多尺度网络TOFNET，遵循编码器 - 解码器网络架构，跳过连接[44]和ResNet [21]瓶颈层（见图4）。具体地，网络将一叠调制曝光，作为输入，以产生相位展开和MPI补偿的距离图像。然后，我们使用校准的相机内在函数将网络输出转换为深度图。

发生器G的编码器（F1 1至D2）在空间上压缩输入高达其原始分辨率的1/4，同时生成具有增加的感受域的特征图。瓶颈处的ResNet块保持了功能的数量，同时通过多个通道细化其残差，以便在上采样后重建更精细，更清晰的深度。我们还通过逐元素求和在F1 2-U2和F2-U1之间设计对称连接的跳过层。这些跳过连接的设计围绕着场景结构应该在输入和输出之间共享的概念[25]。鉴别器网络D由3个向下卷积层组成，在重叠的片中对G的预测进行分类。在训练期间，我们还通过一些粗略和精细的水平随机增加输入图像的比例，以学习尺度稳健的特征。

我们提出了许多输入/输出配置以及数据规范化和扩充策略以配合网络设计。具体而言，我们不是依靠网络来学习幅度不变特征，而是将像素方向归一化应用于相关输入及其相应的幅度。这有效地改善了模型对照明功率和场景反照率的鲁棒性，从而减少了所需的训练时间，因为不需要增幅。归一化方案的一个缺点是输入可能在反射率低或由于平方反比定律导致距离太大的区域中包含显着放大的噪声。为此，我们引入了边缘感知平滑项，以通过将幅度图馈送到下一节中描述的TV正则化层来利用未使用的幅度信息。

* 1. **损失函数**

由于深度和RGB图像数据的图像统计数据差异很大，传统的1/2范数像素丢失在RGB生成任务中运行良好，导致深度重建性能较差，图像输出模糊。在下文中，我们设计针对深度图像统计量定制的特定于域的标准。

**L1损失** 我们最小化了生成器输出深度和目标深度之间的平均绝对误差，因为它对异常值的鲁棒性：

**深度梯度损失** 为了实施局部平滑的深度图，我们在深度梯度上引入L1惩罚项，即总变差损失，其通过边缘感知方式的图像梯度进一步加权[4]。将表示为相关输入的幅度，我们得到了

**对抗性损失** 为了进一步适应深度统计，我们引入了补丁级条件对抗性损失[53]，最小化了模型生成的深度与地面实况深度之间的结构差距。我们采用最小二乘的GAN [36]来稳定训练过程。

**总体损失** 我们的最终损失是加权组合：

在训练期间，和D交替优化，使得G逐渐细化其产生的深度以说服D假设结果是正确的（标记1），而D通过最小化中的式9平方距离来更好和更好地区分正确和不正确的深度估计。

* 1. **训练和执行**

G和D都在128×128补丁上训练。我们首先在[0.6,1]缩放范围内随机下采样原始240×320图像并应用随机裁剪。这种多尺度策略有效地增加了感受域，并提高了模型对空间尺度的稳健性。图4中的每个卷积块包含空间卷积和ReLU / Leaky ReLU（在D中）非线性层，省略批量归一化以保持跨通道相关性。在我们的所有实验中，我们将方程式10中的损失权重设置为和。我们使用ADAM优化器训练我们的模型，前50个时期的初始学习率为0.00005，然后在另外100个时期内将其线性衰减为0。在单个Titan X GPU上完成培训需要40个小时。

1. **数据集**

由于不存在具有地面实况深度的大型原始ToF数据集，因此我们使用已知的基础事实来模拟合成测量以训练所提出的架构。为了验证合成训练结果是否映射到真实的相机传感器，我们评估了使用具有原始数据访问的ToF开发板获得的实验测量。

* 1. **合成数据集**

为了模拟真实的ToF测量，我们扩展了pbrt-v3 [43]用于时间分辨渲染。具体地，我们根据采样路径的路径长度执行具有直方图合并的双向路径跟踪[41]。对于每个场景模型和摄像机灯配置，我们的渲染器会合成一系列瞬态图像，这些图像由每个像素处的离散化TPSF组成。然后可以通过将瞬态像素与频率相关的相关矩阵相关来模拟原始ToF图像（参见式6）。在训练期间，我们随机地将加性高斯噪声应用于原始图像，这很好地推广了各种噪声水平的真实ToF数据，因为Poisson和Skellam [8]噪声在高光子计数下都被高斯噪声很好地近似。

我们选择了许多公开的室内和室外场景模型[2]，其中包括真实世界3D尺度上的各种几何结构（见图5a和5b）。使用Blender的Z pass渲染器生成地面实况深度图。通过将虚拟摄像机飞越多个观察点和角度来观察每个场景，这些观察点和角度沿着物理上合理的路径[38]。为了将我们的模型推广到真实世界的反射率变化，我们还增加了每个物体的表面反照率以进行训练。总的来说，我们的合成ToF数据集包含100,000个相关深度图像对，大小为320×240，包括从250个观察点观察到的10个反射率变化的场景和8个传感器镜像/方向。

我们通过比较合成数据集和真实数据集之间的深度范围分布来进一步验证我们的合成数据集。我们的合成数据集的平均深度为2.35米，是室内场景的合理范围，它与测量的经验深度分布相匹配（见图5d）。

* 1. **真实数据集**

我们使用现成的Texas Instrument OPT8241-CDKEVM摄像机捕获物理验证ToF测量结果，如图5c所示，默认情况下以48MHz运行。我们通过VoxelSDK [3]调整相应的板载寄存器来修改频率设置。我们选择40和70MHz作为实际和合成测量的调制频率，因为我们的相机原型在此范围内实现了高调制对比度。注意，所提出的架构本身不限于该范围，并且我们的网络可以推广到任何一对/一组调制频率。我们还校准了两个频率的相位非线性[1]，之后我们将测量信号视为正弦波。

我们评估了在受控和野外条件下收集的各种场景的拟议框架，包括墙角，凹形物体以及办公室，厨房，卧室，浴室，起居室等日常环境。有关示例，请参见图8。请注意，真实的场景更加混乱，包括有不规则的形状和复杂的反射的皮肤，布料，织物和镜子，这些在训练过程中不会呈现。

1. **实验和结果**

在本节中，我们提出了一项分割研究，以验证所提出的架构设计，并提供合成和物理实验，以验证与现有方法相比的重建性能。表1和图7示出了在包含从不可见的场景反射率-视图配置采样的9,400个合成相关深度图像的测试集上的合成结果。图8显示了原始ToF测量的物理结果。我们关注Adam等人[6]将像素方式的多径比分类为低，平均和强等级，这使我们能够理解在直接照明时执行的每种方法的性能，例如：平面壁和困难的全局照明情况，例如一个凹角。在下文中，我们用预测深度图的平均绝对误差（MAE）和结构相似性（SSIM）[50]量化深度误差，与地面实况进行比较。

* 1. **分割研究**

我们通过设计一系列具有截断体系结构和不同输入配置的消融实验来评估各个体系结构组件对整体重建性能的贡献。

**架构组件的影响** 表1比较了所提出的网络架构（表示为COMBINED）与四个分割变体的性能，即

* BASELINE：我们从G中删除跳过连接，只是最小化像素损失;
* SKIPCONN：与BASELINE相同，但G现在包括跳过连接，这会促进输入和输出之间的结构相似性;
* TV：与SKIPCONN相同，只是两个损失功能用于训练：和;
* ADV：与SKIPCONN相同，只是和都被最小化。

图6也显示了相应的角点场景扫描线。BASELINE和 SKIPCONN 网络实现了整体3.1厘米和3.0厘米的深度误差, 这已经大大优于传统的管道方法。然而, 生成的深度图在平坦地区受到明显的重建噪声的影响。通过在训练过程中引入完全变异正则化,生成器生成没有此类工件的输出,但仍包含全局深度偏移量。引入ADV网络中的对抗性损失,了解到特定深度的结构性损失, 这种全局偏移量减少了。我们还发现，对抗性网络生成更清晰的深度图，在深度边缘周围的“飞行像素”伪像更少。最后，通过SKIPCONN，TV和ADV组合，我们提出的网络实现了准确性，平滑性和处理速度之间的最佳平衡。

**输入的影响** 尽管原始相关图像是建议的端到端体系结构的自然输入选择, 但也可以从现有方法的输出中对相位或深度估计进行后处理。具体来说, 我们评估以下输入配置

* CORR，其中输入到网络是一个原始的双频相关图像堆栈, 如前面所示;
* 相位, 我们使用式3将ToF数据转换为两个相图;
* 深度，在类似于[37]的情况下,我们首先应用相位展开(式. 4)以获得原始深度,并依靠TOFNET来去除噪声和MPI。

为此, 我们修改了F1层g的输入通道数,并对权重进行了重新训练。所有其他图层和超参数保持不变。

如表格1所示，COMBINED+PHASE网络实现了5.5cm的整体深度误差，最接近COMBINED+CORR变种。与平滑的相关输入不同，COMBINED+PHASE网络必须学会消除由于深度边界导致的相位缠绕导致的边缘消歧，因此在分配深度值时变得不那么自信。

另一方面，COMBINED+DEPTH网络将相位展开深度作为输入，但必须学习从上一步骤中删除新引入的深度误差以及校MPI。因此，它生成的深度图比COMBINED+PHASE噪声更大，但仍然在数量上优于管道方法。请注意，此观察结果与[37]中的观察结果相符，其提交时代码不可用。

* 1. **与顺序方法的比较**

然后,我们将所提出的直接重建网络与具有代表性的顺序管道方法进行比较。具体而言,我们比较的是由原始双边去噪[48]、查找表相位展开[20, 19]和非线性校正组成的ToF管道作为前三个块。我们将此子管道生成的深度图表示为 PHASOR。为了补偿 MPI,我们还应用了最先进的稀疏反射分析技术[13]作为最后一个阶段,表示为 SRS。我们注意到,关于MPI和相位展开[11, 14, 17, 27, 30]的其他作品要么共享相似的图像形成模型,要么需要量身定制的采集策略,例如,比我们的方法更多的相位或频率测量,因此很难绘制直接比较。

**合成数据集的定量结果** 在图7中,我们将我们提出的端到端解决方案与 PHASOR、SRS 和深度后处理变体组合+DEPTH进行了比较,这里表示为深度[37],在测试集中的两个具有代表性的场景上。正如预期的那样, PHASOR 在所有方法中产生的噪声最大,因为它的图像形成中缺乏MPI建模。SRA更好地抑制传感器和多径噪声, 但它并不能显著补偿我们实验中的MPI失真, 可能是由于违反了我们合成的背散射(式. 6)中的稀疏假设[13],其中包含强烈的间接衰变。DEPTH2DEPTH 变型的性能不一致, 并且特别容易产生输入深度质量。最后,我们的方法在噪声抑制和细节保存方面不断产生更接近地面真相的深度。有关一系列详细的其他场景,请参阅补充材料。

**实际数据的定性结果** 为了验证TOFNET对真实摄像机数据的推断，我们在具有挑战性的环境中进行定性实验，如图8所示。特别是，我们评估日常场景，如CONCAVEWALL，KITCHEN，LIVINGROOM，OFFICE和PERSON，传统的管道方法通常会失败，存在噪声，低反射率，远距离和MPI。虽然流水线方法部分或过度补偿MPI并引入高频伪像，但所提出的方法始终如一地生成具有合理形状的分段平滑深度图，证明了所学习的空间相关特征的有效性。

**故障案例** 当测量包含饱和度、建模不充分的材料、低反射率和更精细的几何结构时, ToFNet 会正常失败。然而, 由于深度相关的先验体系结构, 我们的模型将根据本地邻域自适应地估计不可靠的区域, 从而实现比传统技术更稳定的性能。

1. **结论和后续工作**

我们提出了一个端到端TOF成像的学习框架, 并验证了它在联合去噪、相位展开和 MPI 校正方面的有效性, 用于合成和实验捕获的TOF测量。在未来, 我们计划将我们的框架应用于更多类型的TOF相机, 包括基于脉冲的 SPAD 探测器。我们还在探索调制功能和重建方法的共同设计与我们的框架, 有可能使成像模式超越目前的ToF深度相机的能力, 如在散射介质中的成像。

图 1: 顶部: 来自双频测量的给定相位和振幅图像, 传统的TOF摄像机采用了一系列深度地图生成技术, 如去噪(DN)、相位展开(PU)和多径校正 (MP)。这往往会导致不准确的深度估计, 因为低频相位特别容易受到全局照明[19]和各种类型的传感器噪声的影响; 底部: 我们训练一个深卷积网络, 直接从TOF相机的原始相关测量中预测场景深度。该方法对噪声和MPI的鲁棒性大大提高, 并实时运行。

图2:ToF传感器集成了来自角落场景的直接(绿色)和间接(橙色)反射的混合信号。

图 3:具有多路径干扰和不具有多路径干扰的角点场景的双频相关图像示意图。MPI 将依赖场景和频率的偏移量引入到TOF数据(右下角),而TOF数据又被我们的方法视为特征。图像由第5.1节中的方法进行模拟,并进行规范化以进行可视化。

图 4:提出的TOFNET体系结构,包括,顶部:对称滑移连接的编码器-解码器生成器网络G和底部: Patchan 鉴别器网络 D。我们将L通气作为正则化层,在这里表示为电视(总变化)。有关详细的层规格,请参阅补充材料。

图 5:我们通过在公开可用的Blender场景中沿着物理似是出的路径(浴室、早餐、现代浴室、卧室和白色房间[2])的物理似是而非路径对虚拟摄像机进行动画处理, 从而合成瞬态相关图像。可以在合成数据集和真实数据集的深度分布之间观察到合理的对齐。

表 1:针对传统顺序方法对拟议网络及其性能进行定量分割研究。对于测试集的平均深度范围,期测值可作为参考。我们报告每个方案的MAE和SSIM,MAE以米为单位。在最右侧的列中,运行时在 FPS (\* CPU 实现)中报告。

图6:角场景上的比较和切割研究。

图7:合成数据集的结果。顶部:对CONT-BATHROOM中的角落场景减少MPI。底部:在帐篷场景中挑战远程场景，通过我们的方法，去噪,相位展开和MPI是共同解决的。

图8:在实际室内场景中,耦合传感器噪声、深度不连续性(请参阅相位图像中的封装边缘)和多路径模糊的结果必须以联合端到端方式解决。我们的方法忠实地重建更干净的深度,减少多路径扭曲(参见图6和补充扫描线比较)。请注意,与隔离的管道步骤生成的ToF深度