中国光学（中英文） Chinese Optics 核心期刊

基于轻型自限制注意力的结构光相位及深度估计混合网络

朱新军 赵浩淼 王红一 宋丽梅 孙瑞群

A hybrid network based on light self-limited attention for structured light phase and depth estimation

ZHU Xin-jun, ZHAO Hao-miao, WANG Hong-yi, SONG Li-mei, SUN Rui-qun

引用本文:
朱新军, 赵浩淼, 王红一, 宋丽梅, 孙瑞群. 基于轻型自限制注意力的结构光相位及深度估计混合网络[J]. 中国光学, 2024, 17(1): 118-127. doi: 10.37188/CO.2023-0066
ZHU Xin-jun, ZHAO Hao-miao, WANG Hong-yi, SONG Li-mei, SUN Rui-qun. A hybrid network based on light self-limited attention for structured light phase and depth estimation[J]. *Chinese Optics*, 2024, 17(1): 118-127. doi: 10.37188/CO.2023-0066

在线阅读 View online: https://doi.org/10.37188/CO.2023-0066

您可能感兴趣的其他文章

Articles you may be interested in

基于线结构光传感器的轨道板几何形貌检测方法

A detection method based on line-structured light sensor for geometrical morphology of track slab

中国光学（中英文）. 2018, 11(5): 841 https://doi.org/10.3788/CO.20181105.0841

基于结构字典学习的图像复原方法

Image restoration approach based on structure dictionary learning

中国光学（中英文）. 2017, 10(2): 207 https://doi.org/10.3788/CO.20171002.0207

结构光照明超分辨光学显微成像技术与展望

Structured illumination super-resolution microscopy technology: review and prospect

中国光学（中英文）. 2018, 11(3): 307 https://doi.org/10.3788/CO.20181103.0307

深海光学照明与成像系统分析及进展

Analysis and research progress of deep-sea optical illumination and imaging system

中国光学（中英文）. 2018, 11(2): 153 https://doi.org/10.3788/CO.20181102.0153

基于小波神经网络的光纤陀螺误差补偿方法

A fiber optic gyro error compensation method based on wavelet neural network

中国光学（中英文）. 2018, 11(6): 1024 https://doi.org/10.3788/CO.20181106.1024

光程补偿近红外光透射反射干涉重构微结构内部形貌

Internal profile reconstruction of microstructures based on near-infrared light transmission reflection interferometry with optical path compensation

中国光学（中英文）. 2019, 12(2): 395 https://doi.org/10.3788/CO.20191202.0395

文章编号　2097-1842(2024)01-0118-10

基于轻型自限制注意力的结构光相位及深度估计混合网络

朱新军*, 赵浩淼, 王红一, 宋丽梅, 孙瑞群
（天津工业大学 人工智能学院, 天津 300387）

摘要: 相位提取与深度估计是结构光三维测量中的重点环节, 目前传统方法在结构光相位提取与深度估计方面存在效率不高、结果不够鲁棒等问题。为了提高深度学习结构光的重建效果, 本文提出了一种基于轻型自限制注意力(Light Self-Limited-Attention, LSLA)的结构光相位及深度估计混合网络, 即构建一种 CNN-Transformer 的混合模块, 并将构建的混合模块放入 U 型架构中, 实现 CNN 与 Transformer 的优势互补。将所提出的网络在结构光相位估计和结构光深度估计两个任务上进行实验, 并和其他网络进行对比。实验结果表明: 相比其他网络, 本文所提出的网络在相位估计和深度估计的细节处理上更加精细, 在结构光相位估计实验中, 精度最高提升 31%; 在结构光深度估计实验中, 精度最高提升 26%。该方法提高了深度神经网络在结构光相位估计及深度估计的准确性。

关　键　词: 结构光; 深度学习; 自限制注意力; 相位估计; 深度估计
中图分类号: TP394.1;TH691.9　　**文献标志码:** A　　**doi:** 10.37188/CO.2023-0066

A hybrid network based on light self-limited attention for structured light phase and depth estimation

ZHU Xin-jun*, ZHAO Hao-miao, WANG Hong-yi, SONG Li-mei, SUN Rui-qun
(*School of Artificial Intelligence*, *Tiangong University*, *Tianjin* 300387, *China*)
* *Corresponding author*, *E-mail*: *xinjunzhu@tiangong.edu.cn*

Abstract: Phase retrieval and depth estimation are vital to three-dimensional measurement using structured light. Currently, conventional methods for structured light phase retrieval and depth estimation have limited efficiency and are lack of robustness in their results and so on. To improve the reconstruction effect of structured light by deep learning, we propose a hybrid network for structured light phase and depth estimation based on Light Self-Limited Attention (LSLA). Specifically, a CNN-Transformer hybrid module is constructed and integrated into a U-shaped structure to realize the advantages complementary of CNN and Transformer. The proposed network is experimentally compared with other networks in structured light phase estimation and structured light depth estimation. The experimental results indicate that the proposed network achieves finer detail processing in phase and depth estimation compared to other networks. Specifically, for structured light phase and depth estimation, its accuracy improves by 31% and 26%, respectively. Therefore,

收稿日期: 2023-04-14; **修订日期:** 2023-05-15
基金项目: 国家自然科学基金 (No. 61905178); 天津市教委科研计划项目 (No. 2019KJ021)
Supported by National Natural Science Foundation of China (No. 61905178); Science & Technology Development Fund of Tianjin Education Commission for Higher Education (No. 2019KJ021)

the proposed network improves the accuracy of deep neural networks in the structured light phase and depth estimation areas.

Key words: structured light; deep learning; self-limited attention; phase estimation; depth estimation

1　引　言

光学三维测量是光学计量和信息光学中最重要的研究领域和研究方向之一[1]。结构光三维测量技术是光学三维测量的重要方式[2], 具有速度快和精度高等优点, 在机器人引导、虚拟现实、人机交互、文物保护、机器人视觉、生物医学等领域有相当广阔的应用前景[3]。在过去的研究中, 可以把传统的结构光相位提取技术主要分为两类: 以傅立叶变换法为代表的单幅图相位提取方法[4] 和相移方法[5]。傅立叶算法只需要单个条纹图案并可直接获得包裹相位图。该方法受物体运动影响的程度较低但存在相位提取精度较低和计算时间长等问题。相移算法至少需要 3 幅条纹图像[6], 通过相移计算获得相对相位图。该方法对于投影光栅的标准性和相移量准确度的要求较高。其精度高但受运动环境影响较大。以上方法获得的相位通常为包裹相位, 需要通过一些展开方法将相对相位图变换为绝对相位[7]。此外, 在结构光深度估计方面, 传统方法需要进行系统标定、相位深度映射完成深度测量与三维测量。

综上所述, 准确的条纹相位估计与深度估计是结构光测量的主要挑战。近年来, 在结构光三维重建领域中出现了许多基于深度学习的方法。Feng 等人通过实验证明深度神经网络可以显著提高单个条纹图案的相位估计精度[8], 具体来说, 通过卷积网络预测中间结果, 最终得到高精度相位图。Nguyen 等人提出了一种端到端的方法, 使用全卷积网络由条纹图得到深度图[9]。Jeught 等人提出了一种完全基于深度学习的从单个变形条纹图案中提取深度信息的方法[10]。张钊等人提出了一种多阶段深度学习单帧条纹投影的三维测量方法, 通过分阶段学习方式依次获得物体的深度信息[11]。其中, 大部分结构光相位估计方法都基于卷积网络, 不利于上下文信息的特征捕获, 而 Transformer 的优势是利用注意力的方式捕获全局的上下文信息, 以对目标建立远距离依赖, 提取有力的特征。文献 [12-14] 对这两种网络进行优势互补[12-14]。Zhu 等人开发了一种混合 CNN-Transformer 模型, 该模型通过条纹级次预测进行相位展开[15], 在条纹级次预测方面取得了较好的效果。

本质上, 基于深度学习的相位估计和深度估计都属于回归预测问题。目前上述的模型虽然取得了较好的效果, 但存在较大的误差, 仍有可提升的空间。为此, 本文提出了一种 CNN- Transformer 相结合的模块, 分 4 个阶段应用在 U 型总体架构中, 将局部特征与全局特征相融合, 并更新了注意力机制的算法, 在网络中使用轻型自限制注意力(Light Self-Limited-Attention, LSLA)机制以节省计算成本和参数数量。最后, 将所提出的网络在结构光深度估计和相位估计两个任务中都进行了实验。

2　条纹投影轮廓术原理

典型的条纹投影轮廓术(Fringe Projection Profilometry , FPP)系统由投影仪和相机组成[16], 图 1 为 FPP 系统原理图, 投影仪将条纹图案投影到目标物体上, 变形的条纹包含了物体的特征信息, 然后由相机捕获并输出数字图像。在多频率的 FPP 中, 投影仪会投射多个不同频率的条纹图案, 相机捕获多张图片, 然后通过相位解算得到物体表面的三维形状信息。

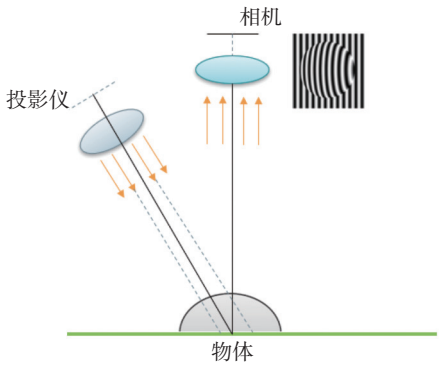

图 1　FPP 系统原理图

Fig. 1　Schematic diagram of the FPP system

通过分析捕获的变形条纹, 解出相位跳变的包裹相位。图像中的条纹可以表示为:

$$I_n(x,y)=A(x,y)+B(x,y)\cos\left[\varphi(x,y)+\frac{2\pi n}{N}\right] \quad , \tag{1}$$

其中，$I_n(x,y)$为某一像素点(x,y)的光强，$A(x,y)$为该像素点的背景光强，$B(x,y)$为该点的调制光强，$\varphi(x,y)$为该点的相位值，$n=0,1,\cdots,N-1$，为相移的步数。根据公式 (1) 计算出包裹相位：

$$\varphi(x,y)=\arctan\left(\frac{\sum_{n=0}^{N-1}I_n\sin\left(\frac{2\pi n}{N}\right)}{\sum_{n=0}^{N-1}I_n\cos\left(\frac{2\pi n}{N}\right)}\right) \quad . \tag{2}$$

包裹相位存在2π的相位跳变。为得到连续的相位信息，需要将相位跳变的包裹相位展开，相位展开公式如下：

$$\phi(x,y)=\varphi(x,y)+2\pi k(x,y) \quad , \tag{3}$$

其中，$\phi(x,y)$是展开相位，$k(x,y)$是条纹级数。

根据标定相位与三角坐标的关系，可以得到绝对相位，根据绝对相位可以直接获取三维高度信息。

3 基于轻型自限制注意力的混合网络

3.1 总体结构

网络的总体结构是 U 型结构，网络结构图如图 2 所示。其由一个编码器-解码器组成。该模型的输入是结构光图像，在编码器网络中，通过下采样方式将输入图像的分辨率降低，提取低分辨率特征，经过 4 次下采样后，进入解码器网络，再进行相同次数的上采样操作，将编码器学习的低分辨率特征映射到高分辨率像素空间中。该模型还使用了跳连接，在解码器网络中将浅层特征和深层特征相融合，从而提高边缘信息的精细度。

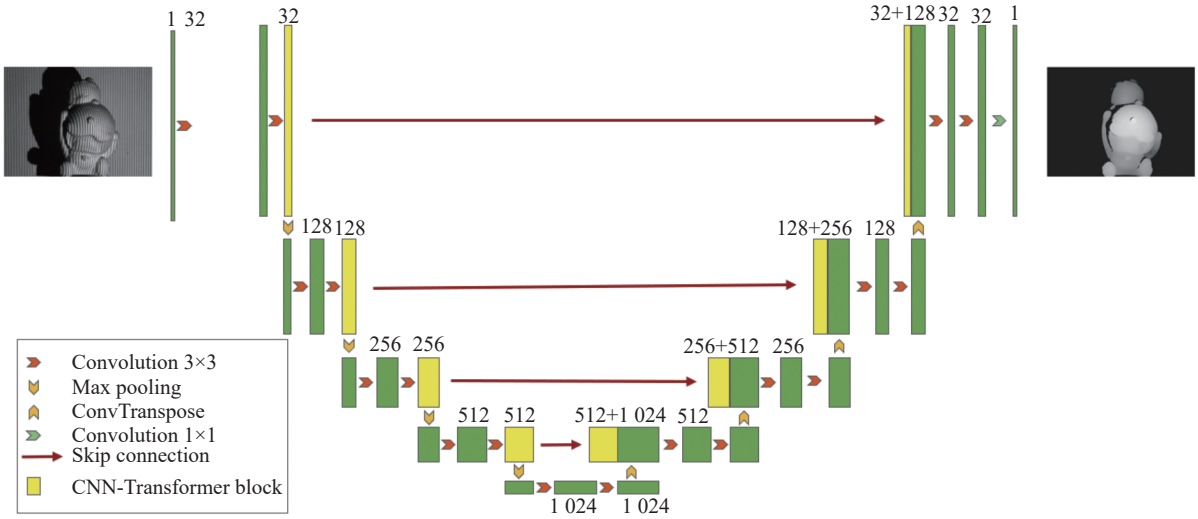

图 2 网络结构图

Fig. 2 Network structure diagram

太窄的网络每一层能捕获的模式有限，此时即使网络再深也不可能提取到足够信息往下层传递。加宽网络可以让每一层学习到更丰富的特征，比如不同方向，不同频率的特征。但宽度增加会导致计算量成平方数增长。考虑到效率问题，要适当调整网络宽度。经实验发现，把网络加宽到 1 024 时效果最优。

3.2 CNN-Transformer 相结合模块

CNN 是一种前馈神经网络，它的特点是卷积层和池化层交替使用，其可以直接处理高维数据，自动提取图像的特征信息，提高模型的计算效率，但不利于上下文信息的特征捕获；Transformer 是一种基于自注意力机制的神经网络模型，可以学习全局信息，提高模型的表达能力。在结构光图像处理过程中，由于图像数据的维度较高，使用传统的 Transformer 模型会导致计算量过大，将 CNN 和 Transformer 结合可以提高模型的计算效率，同时将 CNN-Transformer 的优势相结合，使

用 CNN 提取图像的局部特征, Transformer 学习图像的全局信息, 局部特征与全局特征相融合, 提升模型的表达能力。基于此, 本文在 CMT[17] 的基础上进行了改进, 模块可分为 3 个部分, 模块的结构图如图 3 所示。

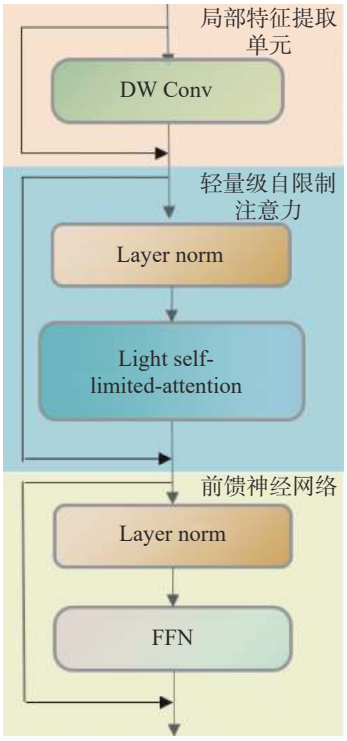

图 3　CNN-Transformer 模块结构图

Fig. 3　Structure of the CNN-Transformer module

3.2.1　局部特征提取单元

由于 Transformer 的特性可能会忽略图像的部分局部信息, 为了更有效地提取图像的局部特征, 使用深度卷积提高网络的非线性能力和网络的表达能力, 如公式 (4) 所示:

$$LFE(\boldsymbol{x}) = DWConv(\boldsymbol{x}) + \boldsymbol{x} \quad , \tag{4}$$

其中, LFE代表局部特征提取单元, $\boldsymbol{x} \in R^{H\times W\times d}$, $H\times W$是当前阶段输入的分辨率, d表示特征的维度, $DWConv$表示深度卷积。深度卷积完全是在二维平面内进行。这种运算对输入层的每个通道独立进行卷积运算, 可以减少卷积的计算量, 高效获取图像的局部信息, 但存在不能有效利用不同通道在相同空间位置上的特征信息的问题, 由于输入的结构光图像是单通道, 故可以避免深度卷积存在的这个问题。

3.2.2　轻量级自限制注意力

Transformer 可以有效地学习图像的全局特征, 然而在结构光图像处理中, 数据的维度通常很高, 使用全局自注意力的计算成本非常高。本文通过 LSLA 机制[18] 进行全局特征的提取, 将全局自注意力分解为局部和全局两个部分, 并使用位置信息模块以及限制注意力模块来增强位置信息以及控制注意力权重大小。在减少计算量的同时还可以利用图像的空间结构信息, 从而对位置信息进行更好的建模。

在传统的自注意力机制中, 键值对通常表示不同的语义信息。例如, 在自然语言处理中, 键可能表示输入句子中的不同单词, 而值则表示与这些单词相关联的特征向量。在图像领域, 键和值通常表示不同的位置特征和图像特征。然而, 在 LSLA 机制中, 由于涉及图像信息的处理, 因此, 将键值对替换为输入 X, 这样可以有效减少计算成本和模型参数量。此外, 在图像处理时, 相邻像素通常具有相似的特征, 这也使得使用相同的输入 X 作为键和值是可行的。这可以显著减少 LSLA 机制的计算成本, 并使其适用于需要高效处理的应用场景。

LSLA 机制包含位置信息模块和限制注意力模块。位置信息模块使用了一个自适应的位置编码向量, 增强了位置信息的表达, 可以更好地处理输入数据中位置信息的变化。限制注意力模块可以控制注意力权重的大小, 避免出现过于集中的注意力分布, 从而提高了模型的鲁棒性和泛化能力。具体来说, 在 LSLA 机制中, 对于每个输入位置, 首先使用一个局部自注意力模块计算局部上下文信息。然后, 使用一个全局自注意力模块计算全局上下文信息, 再进行融合, 得到最终的特征表示。另外, 限制注意力模块在 softmax 函数之后使用外部位置偏差来限制一些较大的注意力权重值。内外部位置偏差和动态尺度相互配合, LSLA 机制的公式定义为:

$$Attention(\boldsymbol{Q},\boldsymbol{X},\boldsymbol{X}) = (softmax(\boldsymbol{Q}\boldsymbol{X}^{\mathrm{T}} \times \boldsymbol{DS} + \boldsymbol{B}_i) + \boldsymbol{B}_0)\boldsymbol{X} \quad , \tag{5}$$

其中, $\boldsymbol{X}$为原始输入数据, $\boldsymbol{Q}$是查询矩阵, DS 表示查询块附近的块具有较大的动态尺度和内部位置偏差值, B_i和B_0分别是内部位置偏差和外部位置偏差。

3.2.3 前馈神经网络

模块的最后一部分使用比较简单的前馈神经网络 (Feedforward Neural Network, FFN)。FFN 的信号从输入层到输出层单向传递，网络的目标是通过调整权值使预测输出与实际输出之间的误差最小。使用的 FNN 包含两个线性层和一个 GELU 激活函数。第一个线性层将输入的尺寸扩大 4 倍，第二个线性层以相同的比例进行尺寸缩小。这种缩放操作有助于保留有用的信息并去除不必要的信息，中间使用 GELU 激活分离，因为 GELU 在负数输入下更具有平滑性。

4 实验与结果

为了证明所提出的网络对于结构光图像处理的有效性，本文进行了两种实验：结构光相位估计（利用条纹图像预测包裹相位）和深度估计（利用条纹图预测深度图），并且在真实数据集和仿真数据集上分别做了实验。所提出模型的深度学习框架为 Pytorch，实验 GPU 为 24 GB 内存的 NVIDIA GeForce RTX3090。

4.1 结构光相位估计实验

4.1.1 数据准备

在结构光图像到包裹相位的预测实验中，本文使用的真实数据集和仿真数据集都由 1 000 个样本组成。本论文使用的是由 Blender 生成的仿真数据集和 Feng 等人制作的真实数据集[8]，部分数据示例图如图 4 所示。每个样本的大小均为 640 × 480，按照 8∶1∶1 的比例划分训练集、验证集和测试集。实验以输入结构光图像，计算高精度的包裹相位为最终目标。

在大多数相位解调技术中，背景强度 $A(x,y)$ 被视为干扰项，要从总强度中去除。根据公式 (2) 可以将包裹相位公式简化为：

$$\varphi(x,y)=\arctan\frac{M(x,y)}{D(x,y)}=\arctan\frac{cB(x,y)\sin\varphi(x,y)}{cB(x,y)\cos\varphi(x,y)}, \tag{6}$$

其中，c 是取决于相位解调算法的常数（傅立叶方法中 $c=1/2$，N 步相移法中 $c=N/2$），$M(x,y)$ 和 $D(x,y)$ 分别表示分子项 $cB(x,y)\sin\varphi(x,y)$ 和分母项 $cB(x,y)\cos\varphi(x,y)$ 的简写。

从结构光图像预测包裹相位有直接预测包裹相位及先预测出分子项 D 和分母项 M，再通过公式 (6) 对包裹相位进行计算两种方式。由于相位信息通过反正切运算被约束在 −π 到 π 之间，因此，包裹相位会存在 2π 的跳变。直接使用深度学习方法难以精确预测跳变位置，而通过预测中间变量 D 和 M 的方法不存在跳变，因此，利用 D 和 M 可以获得更高质量的包裹相位。实验中，使用 UNet 网络对这两种方式进行训练，比起直接预测包裹相位，通过训练得到 D 和 M 再计算包裹相位的预测精度提升了约 60%。对于单输入双输出任务，需要预测的 D 和 M 有强关联性，在原本网络的基础上添加一个分支对结果准确率略有提升，而且双输出模型只需要训练一次，减少了训练时间，结果见表 1。因此，本文把实验的重点放在同时预测 D 和 M 上。

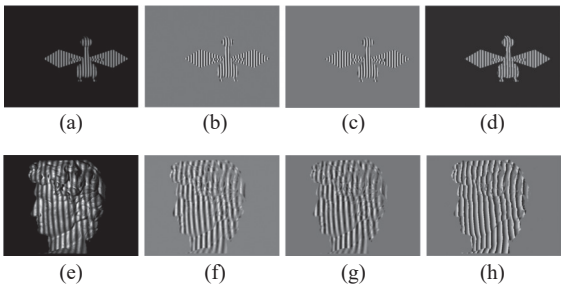

图 4 部分数据示例图。第一行为仿真数据，第二行为真实数据。(a) 仿真条纹图；(b) 仿真条纹图 D；(c) 仿真条纹图 M；(d) 仿真条纹图包裹相位；(e) 真实条纹图；(f) 真实条纹图 D；(g) 真实条纹图 M；(h) 真实条纹图包裹相位

Fig. 4 Sample maps in some datasets. The first lines are simulation data, the second lines are real data. (a) Simulation fringe map; (b) simulation fringe map D; (c) simulation fringe map M; (d) simulation fringe wrapped phase; (e) real fringe map; (f) real fringe map D; (g) real fringe map M; (h) real fringe wrapped phase

表 1 不同包裹相位计算方法比较

Tab. 1 Comparison of the different wrapped phase calculation methods

	MSE	时间 t/s
直接预测包裹相位	0.283 3	5.89
分别预测DM	0.173 9	11.7
同时预测 DM	**0.168 06**	**7.54**

4.1.2　实验结果分析

在计算预测结果的损失时, 背景误差也会算入其中并且对结果有较大影响, 而研究的重点应该在有条纹部分的物体上。所以, 本文提前批量制作了测试集的背景模板, 以便获得更准确的实验结果。使用由同一频率的四步相移获得的条纹图制作背景模板。

相移法的实现一般需要使用投影仪向被测对象投射多张固定位置的光栅条纹图像, 同时使用相机采集。利用 N 步相移法经被测对象调制后的变形条纹公式如式 (7) 所示, 数据均由四步相移所得, $I_0(x,y)$、$I_1(x,y)$、$I_2(x,y)$、$I_3(x,y)$分别为相移 0、π/2、π、3π/2 后的光栅图像, 可得到背景模板$A(x,y)$, 公式如下:

$$A(x,y)=\left[I_0(x,y)+I_1(x,y)+I_2(x,y)+I_3(x,y)\right]/4 \quad . \tag{7}$$

得到背景部分后, 将背景部分设置为一个恒定值(实验中背景部分为 1), 将模板和预测的包裹相位进行比较, 然后去掉包裹相位图的背景部分, 可使得到的包裹相位部分的损失更准确。

本文使用 Unet[19]、DPH[20]、R2Unet[21]、SUNet[22] 等网络在相同数据集上进行训练。得到 D 和 M 后, 根据公式 (7) 得到包裹相位后计算损失值, 结果见表 2。从表 2 可以看出, 本文模型提高了包裹相位预测的精度, 预测时间较 UNet 和 SUNet 长, 较其他网络短。图 5(彩图见期刊电子版)显示了所提出网络与其他网络的比较结果及局部放大图。从局部放大图可以看出, 在相位边缘及物体不连续处, 本文方法比其他方法预测结果更接近标签数据。为了更直观地比较 5 种网络的预测结果, 绘制了预测得到的包裹相位图的第 200 行曲线图, 如图 6(彩图见期刊电子版)所示。可以看出所提出模型得到的结果细节信息比其他网络更加接近标签。

表 2　包裹相位预测方法比较

Tab. 2　Comparison of the wrapped phase prediction methods

	仿真数据		真实数据	
	MSE	时间t/s	MSE	时间t/s
UNet	0.026 58	**6.67**	0.168 06	**7.54**
DPH	0.027 10	11.65	0.129 74	11.78
R2UNet	0.027 34	13.69	0.129 05	14.30
SUNet	0.027 17	7.95	0.143 50	8.29
Ours	**0.023 95**	11.06	**0.116 22**	11.67

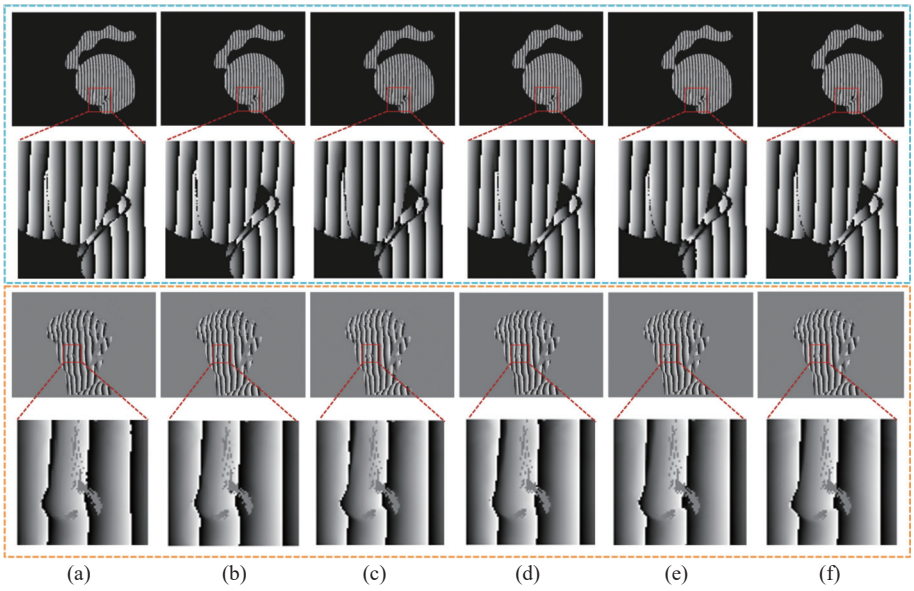

图 5　不同网络仿真和真实数据包裹相位对比。蓝色框为仿真数据, 橙色框为真实数据。(a)UNet; (b)DPH; (c)R2UNet; (d)SUNet; (e)Ours; (f) 标签

Fig. 5　Comparison of different network simulation and real data wrapped phases. The blue boxes are the simulation data, and the orange boxes are the real data. (a) UNet; (b) DPH; (c) R2UNet; (d) SUNet; (e) Ours; (f) Label

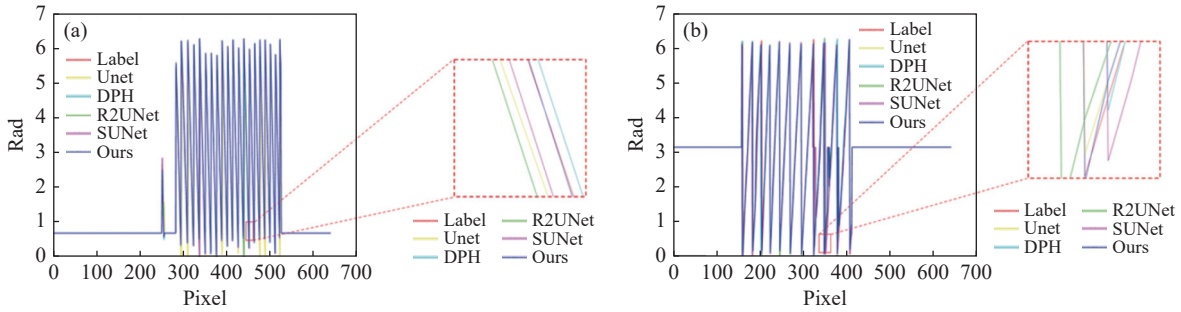

图 6　包裹相位结果曲线图。(a) 仿真数据结果比较; (b) 真实数据结果比较

Fig. 6　Wrapped phase curves.(a) Comparison of simulation data; (b) comparison of real data

4.2　结构光深度估计实验

4.2.1　数据准备

由于公开的结构光深度数据集比较少，本文使用 Blender 制作部分仿真数据集[23]，生成数据集的流程如图 7(彩图见期刊电子版)。Blender 可以通过调整对象模型、相机和投影仪来模拟真实世界的场景，使用相机捕捉和渲染物体图像并输出深度图。

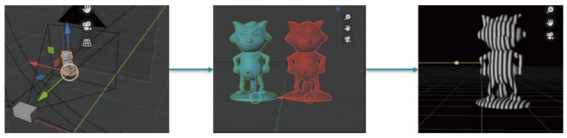

图 7　生成数据集流程图。(a) 模型导入; (b) 调整大小; (c) 投影条纹

Fig. 7　Flowchart of dataset generation. (a) Model import; (b) adjust of the model size; (c) projection fringe

生成数据集的具体步骤如下：将 stl 格式的模型导入到 Blender 中。选中导入的模型，使用变换工具调整它的大小和位置，以适应深度数据集的尺寸和视角。在 Blender 中添加一个平面，将条纹图像贴在平面上，然后将相机对准模型和平面。确定相机的位置、方向、焦距等参数，以保证生成深度数据集的质量。在 Blender 中安装“Structured Light Scanning”插件，设置条纹图像的参数和输出路径，然后点击“扫描”按钮开始生成深度数据集。在插件设置界面可以调整结构光的编码方式、条纹图像的数量和周期、相机参数等。点击“生成数据集”按钮，Blender 会根据这些参数生成深度图像数据集。

本文使用的真实结构光数据集是 Nguyen 等人使用由相机和投影仪组成的 FPP 系统重建的深度图[9]。将具有不同空间频率的一组条纹图案投影到目标物体表面上，捕获变形的结构光图案，计算相移条纹和深度图的相位分布。仿真数据集和真实数据集包含 540 张灰度图像，图像大小是 640 × 480，按照 8∶1∶1 的比例划分训练集、验证集和测试集，部分数据示例图像如图 8 所示。

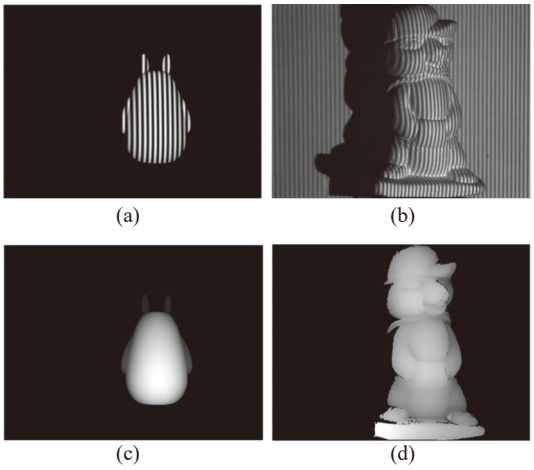

图 8　部分数据示例图。(a) 仿真条纹图; (b) 真实条纹图; (c) 仿真深度图; (d) 真实深度图

Fig. 8　Sample maps in the dataset. (a) Simulated fringe map; (b) real fringe map; (c) simulation depth map; (d) real depth map

4.2.2　消融实验

为了证明所提出的网络结构确实有效，本文在真实数据集上进行了消融研究，逐步修改模块结构并评估结果。首先，在 CMT 模块进行实验并把它作为基线模型，将 CMT 模块中的注意力机制改为 LSLA 机制；然后，将 CMT 模块中最后的部分换为较为简单的 FFN；将 CMT 模块放到 U 型结构中，分为 4 个阶段的上采样和下采样，每个阶段的 CMT 模块重叠两次。

实验结果如表 3 所示，通过替换注意力机制和改善网络的总体结构，网络性能得以逐渐提

高。具体地, 将注意力机制替换为 LSLA 后, MSE 下降了 18.9%, 模型预测时间也略有减少; 将 CMT 原本的反向残差前馈网络替换为更为简单的 FFN 后, MSE 几乎没有下降, 但是预测时间减少了 19.5%; 最后把 CMT 模块应用在 U 型结构中, 分为编码器和解码器逐步提取图像特征, 结果表明放入 U 型结构后精度提升了 21%。

表 3　消融实验结果比较

Tab. 3　Comparison of ablation experiment results

	MSE	时间t/s
CMT	11.32	6.89
CMT替换LSLA	9.17	6.45
CMT替换FFN	11.34	**5.54**
CMT+U形结构	**8.94**	9.68

4.2.3　实验结果分析

对于结构光图像的深度估计性能评价, 本文选用了几个有代表性的卷积网络和混合网络与本文提出的网络进行比较, 在仿真数据集和真实数据集上进行实验, 比较结果如表 4 所示。从表 4 可以看出, 所提模型的精度较其他几种网络高, 模型预测的时间较 DPH 和 R2Unet 少, 总体预测效率较高。图 9(彩图见期刊电子版)显示了本文方法与其他网络在仿真数据集与真实数据上的视觉比较效果。前两行是仿真数据及结果, 后两行是真实数据及结果。由图 9 可以看出: 相较其他网络, 本文所提出的网络在边缘处理和细节处理方面更优, 预测结果更接近真实标签数据。

表 4　不同方法深度估计结果比较

Tab. 4　Comparison of the depth estimation results by different methods

	仿真数据		真实数据	
	MSE	时间t/s	MSE	时间t/s
Unet	8.78	**5.98**	9.97	**6.44**
DPH	8.03	8.66	9.86	10.59
R2UNet	7.57	8.73	8.72	10.92
Ours	**6.43**	8.09	**7.64**	8.44

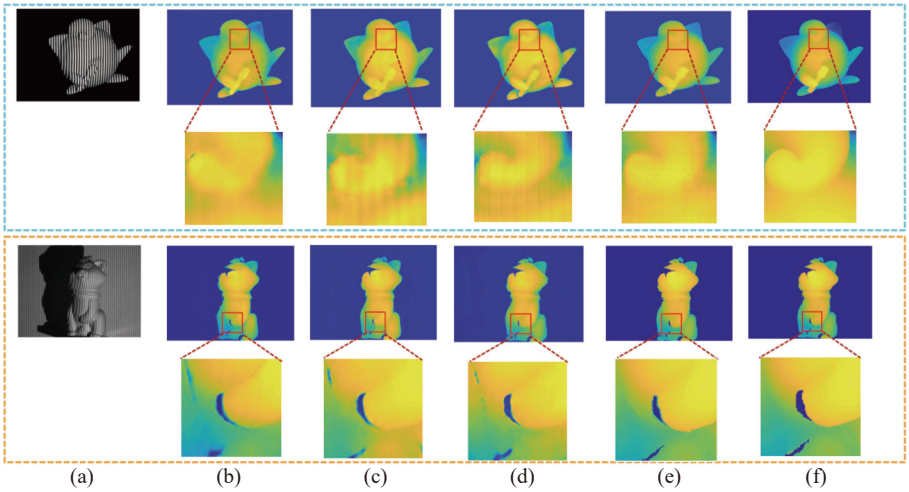

图 9　不同方法深度估计视觉结果比较。蓝色框为仿真数据, 橙色框为真实数据。(a) 输入数据; (b) UNet; (c) DPH; (d) R2UNet; (e) Ours; (f) 标签

Fig. 9　Comparison of the visual results of depth estimation by different methods. The blue boxes are the simulation data, and the orange boxes are the real data. (a) Input data; (b) UNet; (c) DPH; (d) R2UNet; (e) Ours; (f) Label

5　结　论

本文提出一种基于 LSLA 的结构光估计混合网络, 用于处理结构光图像的相关任务, 如由结构光图像预测包裹相位、对结构光图像进行深度估计。所提出的网络使用 U 型结构架构, 分 4 个阶段对结构光图像进行上采样和下采样, 每个阶段都使用了两个重复的 CNN-Transformer 模块。实验证明将 LSLA 机制应用在结构光预测时可以减少预测时间, 提高预测精度。为了评估所提出网络的性能, 挑选了几个有代表性的网络分别在真实数据集和仿真数据集上做对比实验。结果表

明：本文提出的网络可以提高结构光图像深度估计和相位估计的性能。在结构光深度估计实验中，精度最高提升 26%，在结构光相位预测实验中，精度最高提升 31%。

参考文献：

[1] 左超，张晓磊，胡岩，等. 3D 真的来了吗?—三维结构光传感器漫谈[J]. 红外与激光工程，2020，49(3)：0303001.
ZUO CH, ZHANG X L, HU Y, *et al*.. Has 3D finally come of age? ——An introduction to 3D structured-light sensor[J]. *Infrared and Laser Engineering*, 2020, 49(3): 0303001. (in Chinese)

[2] 王永红，张倩，胡寅，等. 显微条纹投影小视场三维表面成像技术综述[J]. 中国光学，2021，14(3)：447-457.
WANG Y H, ZHANG Q, HU Y, *et al*.. 3D small-field surface imaging based on microscopic fringe projection profilometry: a review[J]. *Chinese Optics*, 2021, 14(3): 447-457. (in Chinese)

[3] 冯世杰，左超，尹维，等. 深度学习技术在条纹投影三维成像中的应用[J]. 红外与激光工程，2020，49(3)：0303018.
FENG SH J, ZUO CH, YIN W, *et al*.. Application of deep learning technology to fringe projection 3D imaging[J]. *Infrared and Laser Engineering*, 2020, 49(3): 0303018. (in Chinese)

[4] SU X Y, CHEN W J. Fourier transform profilometry: a review[J]. *Optics and Lasers in Engineering*, 2001, 35(5): 263-284.

[5] ZHENG D L, DA F P, KEMAO Q, *et al*.. Phase-shifting profilometry combined with Gray-code patterns projection: unwrapping error removal by an adaptive median filter[J]. *Optics Express*, 2017, 25(5): 4700-4713.

[6] AN Y T, HYUN J S, ZHANG S. Pixel-wise absolute phase unwrapping using geometric constraints of structured light system[J]. *Optics Express*, 2016, 24(16): 18445-18459.

[7] GHIGLIA D C, ROMERO L A. Robust two-dimensional weighted and unweighted phase unwrapping that uses fast transforms and iterative methods[J]. *Journal of the Optical Society of America A*, 1994, 11(1): 107-117.

[8] FENG SH J, CHEN Q, GU G H, *et al*.. Fringe pattern analysis using deep learning[J]. *Advanced Photonics*, 2019, 1(2): 025001.

[9] NGUYEN H, WANG Y Z, WANG ZH Y. Single-shot 3D shape reconstruction using structured light and deep convolutional neural networks[J]. *Sensors*, 2020, 20(13): 3718.

[10] VAN D J S, DIRCKX J J J. Deep neural networks for single shot structured light profilometry[J]. *Optics Express*, 2019, 27(12): 17091-17101.

[11] 张钊，韩博文，于浩天，等. 多阶段深度学习单帧条纹投影三维测量方法[J]. 红外与激光工程，2020，49(6)：20200023.
ZHANG ZH, HAN B W, YU H T, *et al*.. Multi-stage deep learning based single-frame fringe projection 3D measurement method[J]. *Infrared and Laser Engineering*, 2020, 49(6): 20200023. (in Chinese)

[12] RANFTL R, BOCHKOVSKIY A, KOLTUN V. Vision transformers for dense prediction[C]. *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, 2021.

[13] YANG G L, TANG H, DING M L, *et al*.. Transformer-based attention networks for continuous pixel-wise prediction[C]. *Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, IEEE, 2021.

[14] QI F, ZHAI J Z, DANG G H. Building height estimation using Google Earth[J]. *Energy and Buildings*, 2016, 118: 123-132.

[15] ZHU X J, HAN ZH Q, YUAN M K, *et al*.. Hformer: hybrid CNN-transformer for fringe order prediction in phase unwrapping of fringe projection[J]. *Optical Engineering*, 2022, 61(9): 093107.

[16] GENG J. Structured-light 3D surface imaging: a tutorial[J]. *Advances in Optics and Photonics*, 2011, 3(2): 128-160.

[17] GUO J Y, HAN K, WU H, *et al*.. CMT: convolutional neural networks meet vision transformers[C]. *Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, 2022.

[18] CHEN ZH ZH, HANG W, ZHAO Y X. ViT-LSLA: Vision Transformer with Light Self-Limited-Attention[J]. *arXiv:*2210.17115.

[19] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]. *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2015.

[20] WANG L, LU D Q, QIU R W, *et al*.. 3D reconstruction from structured-light profilometry with dual-path hybrid

network[J]. *EURASIP Journal on Advances in Signal Processing*, 2022, 2022(1): 14.

[21] 袁梦凯, 朱新军, 侯林鹏. 基于 R2U-Net 的单帧投影条纹图深度估计[J]. 激光与光电子学进展, 2022, 59(16): 1610001.
YUAN M K, ZHU X J, HOU L P. Depth estimation from single-frame fringe projection patterns based on R2U-Net[J]. *Laser & Optoelectronics Progress*, 2022, 59(16): 1610001. (in Chinese)

[22] FAN CH M, LIU T J, LIU K H. SUNet: swin transformer UNet for image denoising[C]. *Proceedings of 2022 IEEE International Symposium on Circuits and Systems*, IEEE, 2022.

[23] ZHU X J, ZHANG ZH ZH, HOU L P, *et al*. . Light field structured light projection data generation with Blender[C]. *Proceedings of 2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications*, IEEE, 2022.

作者简介：

朱新军(1985—), 男, 山东临沂人, 博士, 副教授, 硕士生导师, 2008 年于临沂师范学院获得学士学位, 2011 年于山东理工大学获得硕士学位, 2015 年于天津大学获得博士学位, 主要从事光学三维测量与智能计算成像的研究。E-mail: xinjunzhu@tiangong.edu.cn