基于显著图的 2D 转 3D 技术研究

翁高奕[1] 姚剑敏[1,2] 陈恩果[1] 严 群[1]
WENG Gaoyi YAO Jianmin CHEN Enguo YAN Qun

摘 要 为进一步提升图像处理的精准度和高效性，在对现有算法进行详尽分析的基础上，致力于研究一种创新的基于显著图的二维至三维转换技术。因此，针对动态显著性图的提取，利用空洞卷积实现多尺度空间特征抽取，通过时空域特征抽取模块融合空间与时间信息形成显著性图谱序列，并采用卷积长短时存储器（ConvLSTM）单元用于三维动态内容生成。在 DAVIS 和 FBMS 数据集上，量化指标优于 FCNS、DSS 等算法，2D-3D 转换的立体舒适度得分也更高。该技术有效提升了 2D 转 3D 的准确性和效率，具有应用价值。

关键词 显著图；2D 转 3D；深度信息；视觉重要性

doi：10.3969/j.issn.1672-9528.2025.02.043

0 引言

随着计算机视觉技术的日益成熟，2D 转 3D 技术凭借其独特的技术优势，在影视制作、游戏开发、虚拟现实（VR）与增强现实（AR）等前沿领域中得以深度应用，成为推动这些行业数字化变革与创新发展的关键支撑技术之一[1]。

基于显著图的 2D 转 3D 技术，依托显著图携带的图像特征，运用专门算法模型实施深度估计。通过挖掘像素点空间位置、物体轮廓等信息，精确计算深度值，完成三维重建，实现 2D 向 3D 的转换[2]。该技术大幅提升转换准确性，有效规避传统方法深度估计偏差造成的图像失真，还极大减少手动调整工作量，降低人为误差。因此，在影视制作、虚拟现实、游戏开发等领域，展现出广阔的应用前景。

然而，现有的基于显著图的 2D 转 3D 技术仍面临诸多挑战。首先，显著图的生成方法有待进一步优化，以提高其准确性和稳定性。其次，如何在显著图的基础上有效提取深度信息，实现精确的三维重建，亦为亟待解决的问题[3]。

为应对上述挑战，本文提出了一种新颖的基于显著图的 2D 转 3D 技术。该技术首先采用深度学习模型生成高质量的显著图，利用显著图提取深度信息，并结合多视角图像进行三维重建。此外，本文还引入了一种基于视觉重要性的优化策略，旨在提高三维重建的精度和逼真度。大量实验结果表明，本文方法有效提升了 2D 转 3D 的准确性和高效性，为相关领域应用提供了有力支持。

1 动态图像的显著性检测

由于画面流畅，所以在各种场合下，视频都很受人喜爱；因此，要获得满意的视频图像，首先要解决的一个关键问题是如何从视频中提取出具有时间连续特征的图像。本文拟采用空腔卷积实现多个尺度的空域特征提取，并结合长短时记忆基元自身的特点，实现对特征的时空同步抽取，从而得到兼具时空信息的显著预测图谱[4]。在此基础上，提出了一种基于运动图像的显著性检测方法。如图 1 所示。

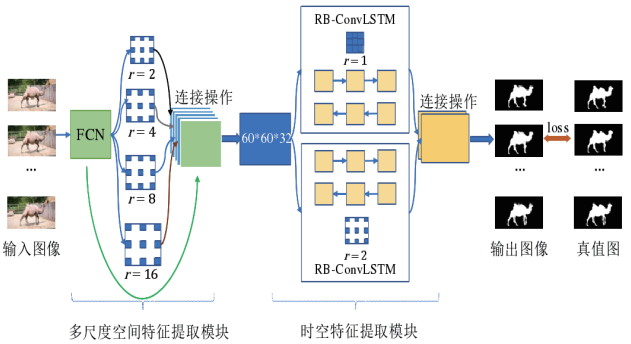


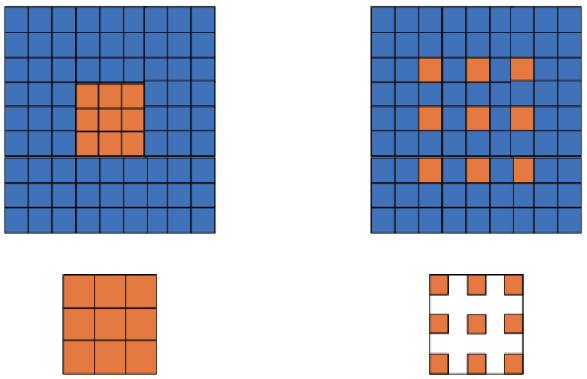
图 1 动态显著性检测模型框架

1.1 关键模型多尺度空间显著性特征提取

传统的卷积神经网络是由一系列的卷积和沉淀层组成的。其中，池化主要是通过增加感受野，降低模型中的参数数量和数据，从而增强模型的推广性能[5]。但是，采用池化运算对图像进行降采样，会导致图像的局部细节损失，并损

1. 福州大学物理与信息工程学院 福建福州 350108
2. 晋江博感公司 福建晋江 362200
[基金项目] 国家自然科学基金（62175032）；福建省杰出青年基金项目（2024J010046）；福建省技术创新重点攻关及产业化项目（2024G020）；闽都创新实验室产学研融合发展专项（2024CXY106）

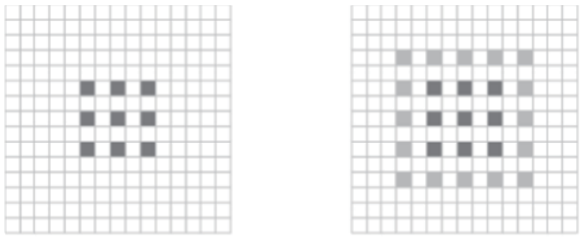
失某些重要的小尺度物体；近年来，黑卷积算法被广泛应用于图像处理领域，其主要原因在于：利用黑卷积算法可以扩大感受野，而无需池化层，从而更好地保持图像的细节信息。在图 2 中，可以看到空洞率为 2 时的卷积核心的感受野与空洞率的关系，随着空洞率的增加，感受野会变大，起到和池化层一样的作用，但是不易造成空间细节丢失。



（a）普通卷积核感受野　　（b）空洞卷积核感受野

图 2 普通卷积核与空洞卷积核感受野对比图

在此基础上，必须解决几个关键问题，即 Gridding Effect。在图 3 中，卷积核心仅能看到“网格”中的信息，因此会失去详细内容。3×3 卷积核内洞比率为 2，多层次运算后，其特征图将发生间断，其原因是相邻像元相互独立且相互依存，导致其结果缺乏关联性。在高空隙比条件下，同步采集的数据间隔大，卷积过少，容易造成局部数据缺失，造成“格点”现象。



（a）第一次卷积操作时的感受野　（b）第二次卷积操作时的感受野

图 3 网格效应示意图

1.2 空间与时序显著性信息融合

1.2.1 介绍 ConvLSTM 网络模型

LSTM 最大特点即引入了一种存储器单元，可以通过存储单元对长期变化的数据进行累加。单元内容的存取、写入及清除操作，均通过受控闸得以实现。每一次新的输入进入，输入门被启动，存储在单元中的信息即会累积到单元中。同样，是否记忆细胞中的过去状态 $\boldsymbol{c}_{a-1}$，也由遗忘门 $\boldsymbol{f}_a$ 控制；输出 Gate 控制最近一个单元状态 $\boldsymbol{c}_a$ 对最后一个隐含层状态 $\boldsymbol{Q}_a$ 的影响 [6]。可以通过公式来表达具体的计算程序：

$$\begin{aligned}
\boldsymbol{i}_a &= \sigma(\boldsymbol{W}_1^X \cdot \boldsymbol{X}_a + \boldsymbol{W}_1^Q \cdot \boldsymbol{Q}_{a-1} + \boldsymbol{W}_1^c * \boldsymbol{c}_{a-1}) \\
\boldsymbol{f}_a &= \sigma(\boldsymbol{W}_f^X \cdot \boldsymbol{X}_a + \boldsymbol{W}_f^Q \cdot \boldsymbol{Q}_{a-1} + \boldsymbol{W}_f^c * \boldsymbol{c}_{a-1}) \\
\boldsymbol{o}_a &= \sigma(\boldsymbol{W}_o^X \cdot \boldsymbol{X}_a + \boldsymbol{W}_o^Q \cdot \boldsymbol{Q}_{a-1} + \boldsymbol{W}_o^c * \boldsymbol{c}_{a-1}) \\
\boldsymbol{c}_a &= \boldsymbol{f}_a * \boldsymbol{c}_{a-1} + \boldsymbol{i}_a * \tanh(\boldsymbol{W}_e^X \cdot \boldsymbol{X}_a + \boldsymbol{W}_e^e \cdot \boldsymbol{Q}_{a-1}) \\
\boldsymbol{Q}_a &= \boldsymbol{o}_a * \tanh(\boldsymbol{c}_a)
\end{aligned} \tag{1}$$

式中：“*”代表哈达马矩阵相乘；σ 是一个 sigmoid 活化功能，用于对不同的闸进行状态控制；公式中的 $\boldsymbol{c}_a$、$\boldsymbol{i}_a$、$\boldsymbol{o}_a$、$\boldsymbol{f}_a$、$\boldsymbol{W}$、$\boldsymbol{Q}_{a-1}$ 都是一维的张量，$\boldsymbol{W}$ 是一个可学习的参量；tanh 活化功能用于调整每一个单元的数值到 [-1,1] 的间隔。

ConvLSTM 网络模型是将传统的 LSTM 网络与卷积神经网络（CNN）相结合的产物，旨在处理具有空间和时间特征的图像数据。与传统的 LSTM 相比，ConvLSTM 在处理图像序列时具有显著优势，可同时捕捉空间和时间上的相关性。

ConvLSTM 网络模型的核心在于其卷积门控机制，这使得它能够有效处理图像序列中的局部特征。在 ConvLSTM 中，每个门（输入门、遗忘门和输出门）都采用卷积操作，从而使得网络能够学习到图像序列中的空间特征。此外，ConvLSTM 的存储单元也采用卷积操作，使网络能够保持图像序列中的时间特征。

1.2.2 改进的双向卷积长短期记忆单元

在处理图像时，传统的短时记忆方法处理能力会下降。全连通 LSTM（fully connected LSTM layer）可用于求解图像问题，但未对其进行有效的空间关联分析，仅能以单一颜色方式传递给 LSTM，且一次仅能传递一条像素点 [7]。针对这一问题，研究者们将卷积运算与 LSTM 相结合，构建了一种基于卷积的连续短时记忆模型，使得其能够同时实现输入 - 状态与状态两个阶段，并充分利用视频序列中各帧间的时空关联关系。将卷积结构引入网络后，该模型仍可进行端对端输出。

$$\begin{aligned}
\boldsymbol{i}_a &= \sigma(\boldsymbol{W}_1^\rho \odot \boldsymbol{P}_a + \boldsymbol{W}_1^Q \odot \boldsymbol{Q}_{a-1} + \boldsymbol{W}_1^c * \boldsymbol{c}_{a-1}) \\
\boldsymbol{f}_a &= \sigma(\boldsymbol{W}_f^P \odot \boldsymbol{P}_a + \boldsymbol{W}_f^Q \odot \boldsymbol{Q}_{a-1} + \boldsymbol{W}_f^c * \boldsymbol{c}_{a-1}) \\
\boldsymbol{o}_a &= \sigma(\boldsymbol{W}_o^P \odot \boldsymbol{P}_a + \boldsymbol{W}_o^Q \odot \boldsymbol{Q}_{a-1} + \boldsymbol{W}_e^\varepsilon * \boldsymbol{c}_{a-1}) \\
\boldsymbol{c}_s &= \boldsymbol{f}_a * \boldsymbol{c}_{a-1} + \boldsymbol{i}_a * \tanh(\boldsymbol{W}_c^\rho \odot \boldsymbol{P}_a + \boldsymbol{W}_c^Q \odot \boldsymbol{Q}_{a-1}) \\
\boldsymbol{Q}_a &= \boldsymbol{o}_a * \tanh(\boldsymbol{c}_a)
\end{aligned} \tag{2}$$

式中：“⊙”符号表示卷积操作。

2 基于显著图的 3D 内容生成

在传统二维三维影像变换中，研究者往往采用深度图像来完成，其主要步骤包括：视差计算、空洞填充以及三维内容渲染 [8]。本文将显著性图转化为 2D-3D“深度图”，并未对上述转化过程进行简化或删除。图 4 为三维内容产生方法的流程图。尽管显著性图与深度图有本质区别，但要获得具

有较好立体视觉效果的三维视频，关键部件必须位于人眼的“舒适区”，而显著性检测正是基于这一点[9]。

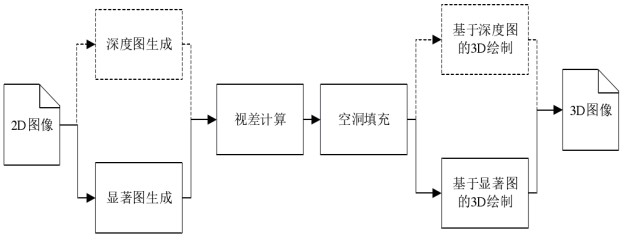


图 4 网格效应示意图基于显著图的 3D 内容生成方法流程图

2.1 视差计算

视差指当观测到相同物体时，因观测角度、距离等因素而引起的图像差别。人们利用眼睛的细微差异来获得真实生活中的三维信息，从而得到三维空间。所以，要获得三维图像，首先需要得到满足人类视觉特征的视差图[10]。在二维三维变换过程中，视差和深度可以相互变换，在本文算法中，视差和重要值相互变换公式为：

$$\text{disparity}(i,j) = V \times (1 - \frac{\text{saliency}(i,j)}{128}) \quad (3)$$

式中：V 为最大视差，其计算公式为：

$$V = \frac{J+O}{J+D} \quad (4)$$

2.2 空洞填充

在二维三维变换中，如何高效、快速的实现空穴的填补是该领域的研究重点和难点。当前，按照充填过程的不同，充填空穴的方式可分为预充填和后充填两种。预处理技术通过对左右视图进行预处理，控制左右视图的生成，防止出现空洞。后处理规则是对已经生成的左右视图分别执行相应的运算，从而实现了空洞的填充[11]。

3 3D 图像生成方法实验结果与分析

在实验过程中，选取动态图上显著性检测的两个对象——DSS 和 FCNS。该算法采用嵌套两层全卷积网络检测动态图像中的显著性，其中 1 个模糊神经网络先提取出初始空间特征，再将其与原始视频帧及后一帧构成的帧对连接，再将其送入下一个模糊神经网络中进行时序特征提取，得到包含动态信息的显著图。试验数据结果如表 1 和图 5 所示。

表 1 不同算法在两种数据集上的量化分析

算法	DAVIS		FBMS	
	MAE ↓	F-measure ↑	MAE ↓	F-measure ↑
本方法	0.038	0.811	0.073	0.801
FCNS	0.053	0.729	0.100	0.735
DSS	0.062	0.717	0.083	0.764

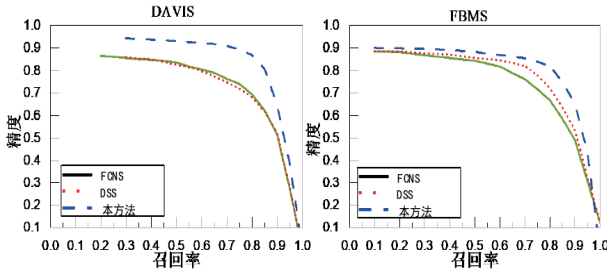


图 5 三种算法在 DAVIS 和 FBMS 数据集上的 *P-R* 曲线

为验证所提出动态 3D 图像生成方法的有效性，选择两个广泛使用的数据集进行实验：DUTS 和 MSRA-B。DUTS 数据集提供了精确的像素级标注，可以评估显著性检测算法在动态场景中的性能；MSRA-B 数据集不仅提供了静态图像的标注，同时还提供了视频序列的标注，使其能够更好地评估算法在处理具有复杂背景变化的场景时的表现。

在实验环节，将本文所提出的基于显著图的 3D 内容生成方法，与当前业内几种主流方法进行了全面对比。其方法主要包括：DSS（dynamic scene segmentation），一种基于深度学习的动态场景分割算法，可以有效处理视频序列中的动态变化；FCNS（fully convolutional network for semantic segmentation），一种全卷积网络，用于语义分割任务，能够处理图像中的静态和动态特征。

从表 1 可以看出，该算法在 DAVIS 和 FBMS 上主观性能较好。此外，由图 5 可知，此模式的 *P-R* 曲线大部分在前两者上方，表明本文所提出方法的有效性。其根本原因是：FCNS 对于时序信息的处理不够充分，无法有效利用时序信息进行分析。但在 DSS 中，不具备对时域特征进行抽取的能力。图 6 显示了在动态图形显著性检测试验中的可视效果反差的结果，因其使用多尺度的凹式卷积来进行空域特性的抽取，可以更好地保留影像的空域细节。

图 6 三种动态显著性检测算法的视觉效果对比图

此外，通过与 FCNS 算法的对比分析，发现 LSTM 算法

在时序数据分析中更具优越性。

尽管本文方法在动态显著性检测方面取得了较好效果，但仍存在一些潜在的改进空间。为了进一步提高模型的性能，可以考虑引入更先进的空间特征提取技术。

其次，尽管 LSTM 单元在处理时间序列特征方面具有优势，但其在处理长序列数据时可能会遇到梯度消失或爆炸的问题。

图 7 显示了基于显著性检测的动态 2D-3D 图像变换方法。基于此，本文方法、Deep3D 算法和文献 [10] 的平均得分分别为 4.37、4.13、4.27，表明基于显著检测的动态三维内容生成方法优于其他两类深度学习方法，主要原因在于，该方法将显著性特征和时序信息纳入到模型的转换模型中。

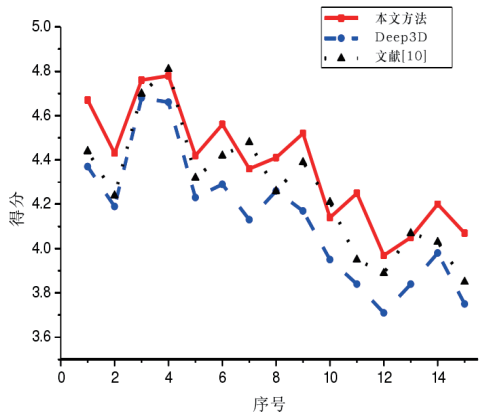


图 7 三种动态图像 2D-3D 转换算法的立体舒适度得分

4 结语

在本文中，研究探索了通过多尺度空间特征提取模块和 ConvLSTM 网络模型结合以生成动态显著图的方法。实验结果表明，将两者有效结合后生成的模型，其客观评价指标优于对比模型，说明该方法达到了预期效果，为生成显著图提供了较好的方法，可以在这一领域中发挥重要作用。

总而言之，本文提出的动态显著图提取的方法具有一定的优点，在客观评价指标得分与视觉对比图视觉效果的基础上，实现了基于显著性检测的三维立体转换方法的可行性与可借鉴性。

参考文献：

[1] HARMAN P.Home based 3D entertainment-an overview[C/OL]//Proceedings 2000 International Conference on Image Processing. Piscataway:IEEE, 2000[2024-06-11].https://ieeexplore.ieee.org/document/900877.

[2] THOMPSON L, JI M H, ROKERS B, et al.Contributions of binocular and monocular cues to motion-in-depth perception[J]. Journal of vision, 2019,19(3):2.

[3] COLTEKIN A, BILAND J.Comparing the terrain reversal effect in satellite images and in shaded relief maps:an examination of the effects of color and texture on 3D shape perception from shading[J].International journal of digital earth, 2019, 12(4):442-459.

[4] 张旭东，李成云，汪义志，等．遮挡场景的光场图像深度估计方法 [J]. 控制与决策 ,2018,33(12):2122-2130.

[5] CHANG J L, WETZSTEIN G. Deep optics for monocular depth estimation and 3D object detection[C/OL]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway:IEEE,2019[2024-02-21].https://ieeexplore.ieee.org/document/9010976.

[6] XU D, WANG W, TANG H,et al. Structured attention guided convolutional neural fields for monocular depth estimation[DB/OL].(2018-03-29)[2024-04-19].https://doi.org/10.48550/arXiv.1803.11029.

[7] FU H, GONG M M, WANG C H,et al.Deep ordinal regression network for monocular depth estimation[C/OL]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Piscataway:IEEE,2018[2024-06-11].https://ieeexplore.ieee.org/document/8578312.

[8] GARG R, BG V K, CARNEIRO G,et al.Unsupervised CNN for single view depth estimation:geometry to the rescue[DB/OL].(2016-07-29)[2024-04-23].https://doi.org/10.48550/arXiv.1603.04992.

[9] GODARD C, AODHA O M, FIRMAN M,et al.Digging into self-supervised monocular depth estimation[DB/OL].(2019-08-17)[2024-02-22].https://doi.org/10.48550/arXiv.1806.01260.

[10] KUZNIETSOV Y, STUCKLER J, LEIBE B.Semi-supervised deep learning for monocular depth map prediction[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE,2017[2024-05-25].https://ieeexplore.ieee.org/document/8099721.

[11] RAMIREZ P Z, POGGI M, TOSI F,et al.Geometry meets semantics for semi-supervised monocular depth estimation[J]. Computer vision,2018,11363(5):298-313.

【作者简介】

翁高奕 (2000—)，男，福建平潭人，硕士研究生，研究方向：深度学习、图像生成等。

姚剑敏 (1978—)，男，福建莆田人，博士，副研究员，研究方向：人工智能、图像处理、计算机视觉等。

（收稿日期：2024-10-21）