中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-13

论文引用格式:Zhou Guangbin, Yang Yanwei, Xiao Zhaolin. Image Defocus Estimation by Learning from a Novel Light Field Autofocus Dataset[J/OL]. Journal of Image and Graphics, XXXX:1-13. DOI: 10.11834/jig.260267. (周广彬, 杨延伟, 肖照林. 光场自动对焦数据集构建及图像离焦程度估计方法[J/OL]. 中国图象图形学报,XXXX:1-13. DOI: 10.11834/jig.260267.)[DOI:10.11834/jig.260267]

光场自动对焦数据集构建及图像离焦程度估计方法

周广彬[1,2],杨延伟[1],肖照林[1,3*]

1. 西安理工大学计算机科学与工程学院,西安　710048; 2. 陕西理工大学数学与计算机科学学院,汉中　723000; 3. 陕西省网络计算与安全技术重点实验室,西安　710048

摘　要:**目的**　自动对焦模型训练通常需要带有深度标注的多焦点数据,但真实场景中精准深度标注获取困难。针对该问题,本文提出一种基于Blender的光场重聚焦堆栈数据合成与标注方法,并在此基础上构建单帧图像离焦程度估计网络。**方法**　首先利用Blender物理渲染与光场重聚焦生成多焦点堆栈数据,并同步获得对应的深度标注,构建包含深度信息的多焦点堆栈数据集。随后,针对单张散焦图像推断对焦步长时存在的镜头移动方向二义性问题,设计行进方向判别器,通过图像由外围到中心的深度变化趋势和散焦梯度估计弥散圆斜率符号,以判断镜头移动方向。在确定方向后,构建步长估计器,由单张焦点切片预测到目标聚焦索引所需的移动步长。同时,引入初始透镜对焦索引的轴向位置编码,增强模型对镜头对焦位置的感知能力。**结果**　实验结果表明,在误差容忍度为±4片索引的条件下,本文方法的聚焦索引预测准确率达到96.8%,相比基线方法提高4.6%。在推理效率方面,本文方法无需遍历完整焦点堆栈,推理速度高于传统全栈遍历算法。**结论**　本文所提出的光场自动对焦数据集构建及图像离焦程度估计方法,在数据合成与标注过程中避免了真实采集中机械标定误差和焦点呼吸效应的影响,缓解了自动对焦模型训练中精准深度标注不足的问题,并提升了聚焦索引预测的准确率与对焦控制的执行效率。

关键词:自动对焦;光场重聚焦;行进方向判别;弥散圆斜率;步长估计

Image Defocus Estimation by Learning from a Novel Light Field Autofocus Dataset

Zhou Guangbin[1,2], Yang Yanwei[1], Xiao Zhaolin[1,3*]

1. School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China; 2. School of Mathematics and Computer Science, Shaanxi University of Technology, Hanzhong 723000, China; 3. Shaanxi Key Laboratory of Network Computing and Security Technology, Xi'an 710048, China

Abstract: **Objective** Autofocus aims to move the lens to an appropriate focal position so that the region of interest can be clearly imaged. In data-driven autofocus methods, a network is usually trained to predict the target focus position or lens movement from defocused images. However, such training requires focus stacks with reliable labels, while accurate depth or focus annotations are difficult to obtain in real imaging. Public datasets constructed from stereo disparity estimation or analytical defocus simulation may be affected by calibration errors, focus breathing, and the mismatch between simplified degradation models and the actual imaging process. In addition, single-image autofocus has an axial ambiguity problem.

收稿日期:2026-05-13;**修回日期**:2026-07-17

*通信作者:肖照林　*xiaozhaolin@xaut.edu.cn

基金项目:国家自然科学基金项目(62371389);陕西省教育厅科学研究计划项目(25JK0382)

Supported by: National Natural Science Foundation of China(62371389); Scientific Research Program Funded by Shaanxi Provincial Education Department(25JK0382)

Similar defocus appearances may occur on both sides of the true focal plane, making it difficult to determine whether the lens should move forward or backward from a single defocused image. To address these problems, this paper proposes a light field autofocus dataset construction method and an image defocus estimation network for single-slice focus prediction. **Method** A synthetic light field autofocus dataset, named SLFAF, is constructed using Blender-based physical rendering and light field refocusing. The dataset contains 50 static virtual scenes, including indoor, outdoor, forest, lake, street, and architectural environments. For each scene, a 9×9 virtual camera array is used to render 81 sub-aperture images with a resolution of 1920×1920 pixels, while the corresponding camera intrinsic parameters, extrinsic parameters, and Z-channel depth maps are also recorded. The 2−100 m depth range is converted into disparity according to binocular geometry and divided into 50 focal levels for light-field refocusing and focus labeling. For each patch stack, the local ground-truth depth is calculated by averaging the 3×3 neighborhood around the patch center in the depth map, reducing the influence of single-pixel noise and geometric deviation. The averaged depth is mapped to the nearest focal depth level, and the corresponding level is used as the ground-truth focus index of the patch stack. Finally, 447 original focus stacks were obtained from the 50 scenes to construct the SLFAF dataset. Based on this dataset, a lightweight image defocus estimation network was designed. The network consists of two modules: a movement direction discriminator and a step estimator. The movement direction discriminator is used to reduce the axial ambiguity in single-image autofocus. It takes a 128×128 RGB image patch as input and first extracts edge-related information to suppress redundant background content. Then, the network compares the defocus and structural responses between peripheral and central image regions, implicitly estimating the sign of the circle-of-confusion slope. The output is a binary movement direction label, trained with cross-entropy loss. After the movement direction is determined, the step estimator predicts the absolute movement step from the current focus slice to the target focus index. This module is implemented as a lightweight convolutional regression network and trained with mean square error loss. Since the same defocus appearance may correspond to different initial lens positions, an axial positional encoding strategy, termed Lens-PE, is introduced. The initial focal index is normalized, broadcast to the same spatial size as the image patch, and concatenated with the RGB image as an additional input channel. In this way, the network can jointly use image defocus features and the initial lens position from the early feature extraction stage. **Result** Quantitative and qualitative experiments are conducted on the SLFAF dataset. Mean absolute error, root mean square error, and error-tolerance accuracy are used as evaluation metrics. Under the single-slice input setting, the proposed method achieves accuracies of 37.9%, 75.9%, 91.4%, and 96.8% when the prediction error is equal to 0, within 1, within 2, and within 4 focus indices, respectively. The corresponding MAE and RMSE are 1.21 and 2.98. Compared with the MobileNet-v2 baseline under the same single-image prediction setting, the proposed method improves the zero-error accuracy from 31.0% to 37.9% and the accuracy within four focus indices from 92.2% to 96.8%. The inference time of the proposed network is 63ms, which is much shorter than traditional full-stack traversal methods such as Gradient Laplacian and Intensity Variance, which require 814ms and 784ms, respectively, in the same experimental setting. Qualitative results show that the predicted focus slices are closer to the ground-truth focus indices and present clearer edge structures than the baseline results. Ablation experiments further verify the role of the proposed modules. When the movement direction discriminator is added to the base regression network, the MAE decreases from 2.74 to 1.89, showing that explicit direction discrimination is useful for reducing axial ambiguity. When the initial lens position is introduced by deep scalar concatenation, namely Lens-index, the MAE decreases to 1.53 and the accuracy within four focus indices reaches 92.4%. By contrast, the proposed Lens-PE strategy further reduces the MAE to 1.21 and improves the accuracy within four focus indices to 96.8%. These results indicate that early spatial fusion of the initial lens position is more effective than introducing the lens index only at the final regression stage. **Conclusion** The proposed method avoids the influence of mechanical calibration errors and focus-breathing effects during data synthesis and annotation, alleviates the shortage of accurately depth-annotated autofocus training data, and improves both focus index prediction accuracy and focus control efficiency.

Key words: autofocus; light field refocusing; movement direction discrimination; circle of confusion slope; step estimation

论文引用格式:DOI:10.11834/jig.260267

0 引 言

光学成像系统广泛应用于医疗诊断、科学研究及军事监测等领域，而自动对焦（auto focus，AF）作为确保成像质量的关键功能，是实现高精度成像的前提。其本质是相机自动搜索最优镜头位点，以使感兴趣区域（region of interest，ROI）准确成像。若未能精准对焦，成像系统会产生离焦模糊；由于此类模糊常受光圈参数、径向畸变等复杂物理因素调制，其图像退化在后期处理中难以被精确补偿。因此，在采集端实现精准对焦是成像链路中的关键环节（Zhu等，2025）。

传统AF技术主要分为主动式与被动式两类（Ray，2002）。主动式AF利用红外、激光或超声波传感器获取深度，并驱动音圈马达完成闭环控制，虽在低照度或低对比度环境下具有一定优势，但易受探测距离和目标反射率限制。被动式AF主要包括相位检测对焦（phase detection auto focus，PDAF）和对比度检测对焦（contrast detection auto focus，CDAF）。PDAF通过微透镜或双核传感器量化相位差，实现单步位移预测，但在低光噪声干扰下精度下降明显；CDAF则基于图像梯度评估清晰度，并通过搜索算法迭代寻优，虽稳态精度较高，但多次往复搜索导致收敛速度慢，难以满足动态场景的实时性需求。

针对传统AF技术的局限，研究人员开始利用卷积神经网络（convolutional neural networks，CNN）强大的非线性表征能力，构建从离焦图像或焦点堆栈中预测最佳焦点的AF模型（Ho等，2020）。目前深度对焦研究主要分为两类：（1）基于焦点预测的AF方法。此类方法旨在建立图像特征与镜头移动的映射关系。早期研究多将对焦建模为离散的分类任务，通过对比多帧离焦序列的特征来预测目标位置（Chen等，2010；Han等，2011）。为减少采样帧数，研究重点逐渐转向基于单帧图像的连续焦点预测。例如，Ren等人（2018）利用CNN直接从全息图中提取深度分布，Wang等人（2021）通过建模图像离焦程度与物理距离的关联以动态调整焦深。进一步地，为突破单帧空间信息的局限，全像素双核（dual pixel，DP）传感器（戴玉超等，2022）的相位线索及双目视差被广泛引入，通过分析左右视图差异预测焦点，深入挖掘DP数据的空间位置特性（Choi等，2023），或引入焦点一致性损失以提升预测稳定性（Lee等，2025）；同时，也有研究基于联合立体匹配构建双目编码网络以指导对焦（Xu等，2023；Liu等，2025），甚至利用专家轨迹正则化的强化学习策略来抑制对焦过程中的振荡现象（Zhu等，2025）。然而，此类数据驱动模型高度依赖精确的深度标注。现有的大规模公共数据集（如通过视差估计生成的DP数据集（Herrmann等，2020）或基于理论散焦模型合成的数据（Wang等，2021）），常因机械标定误差、焦点呼吸效应或散焦退化模型与真实物理过程存在偏差，导致网络在真实场景中的泛化受限。此外，单张散焦图像自动对焦存在镜头行进方向的二义性问题，若缺乏有效的行进方向判别机制以及对初始透镜位置的全局感知，系统往往难以实现高精度对焦。（2）基于去模糊与重聚焦的AF方法。此类方法侧重于图像采集后的信息补偿。部分研究尝试摆脱机械马达控制，通过盲估计模糊核（Schuler等，2015）或构建端到端的抗模糊网络（Nimisha等，2017），直接从失焦图像中恢复清晰细节。另一种主流思路是利用焦点堆叠技术，通过网络提取不同对焦切片的特征并融合为全聚焦画面（Liu等，2017；Guo等，2018），针对其庞大的数据采集量，后续衍生出了基于块级统计或场景动态评估的自适应采样策略以减少无效捕获帧数（Li等，2018）。尽管此类补偿方法能在一定程度上重构图像细节，但严重失焦造成的物理信息丢失是不可逆的，且堆栈扫描与多帧融合往往伴随高昂的计算开销与时间延迟，根本无法满足动态场景的实时对焦需求。

针对上述情况，本文提出了一种光场自动对焦数据集构建及图像离焦程度估计方法。主要贡献如下：1）针对AF缺乏精准训练深度标注数据的难点，本文利用Blender模拟真实光学系统的光场特性，通过虚拟摄像机的内参（焦距、光圈、景深），构建出高精度多焦点堆栈数据集（SLFAF）。该方法有效避免了真实场景拍摄中存在的机械标定误差、焦点呼吸效应以及散焦退化模型偏差问题，为深度学习模型提供了高置信度的监督样本。2）在SLFAF数据集的基础上，提出了一种图像离焦程度估计网络，网络包括两个子模块，其一为行进方向判别器。针对镜头移动中存在的“轴向二义性”问题（即无法仅凭模糊程度判断应前进还是后退），该网络通过捕捉图像外围到中心的散焦梯度变化拟合弥散圆斜率符号，

实现对镜头移动方向的精准锁定。其二为步长估计器与镜头轴向位置编码:通过估计图像离焦程度,实现由“单张”失焦切片直接映射到“目标”聚焦索引的预测。受自注意力机制中位置编码(Vaswani等,2017)的启发,在步长估计器中引入镜头轴向位置编码,利用镜头初始位置信息为步长估计器提供位置先验,增强模型对镜头轴向位置的全局感知能力,实现由单张焦点切片直接预测目标聚焦索引。3)在SLFAF数据集上与近年来全栈遍历AF算法和单张预测AF算法比较,证明了本文方法的有效性。

1 方 法

1.1 自动对焦原理与镜头移动轴向二义性问题

1.1.1 自动对焦原理

自动对焦旨在通过分析初始位置图像的对比度、边缘等散焦线索,预测并驱动马达将镜头移动至准确对焦位置。遵循几何光学中的高斯成像公式:

$$\frac{1}{f}=\frac{1}{u}+\frac{1}{v} \tag{1}$$

式中,f、u、v分别为焦距、物距与像距。对于定焦系统,f为常数,对焦本质上是通过调整镜头与成像平面间的距离(即改变像距v)来实现。基于此,若将连续的透镜位置离散化为n个等份的焦点索引。对于深度为u_m($m\in\{1,2,\cdots n\}$)的目标物,仅存在唯一的透镜位置l_m使其处于清晰对焦状态,m则称为焦点堆栈的GT(ground truth)聚焦索引。受此启发,深度学习方法将AF任务简化为从焦点

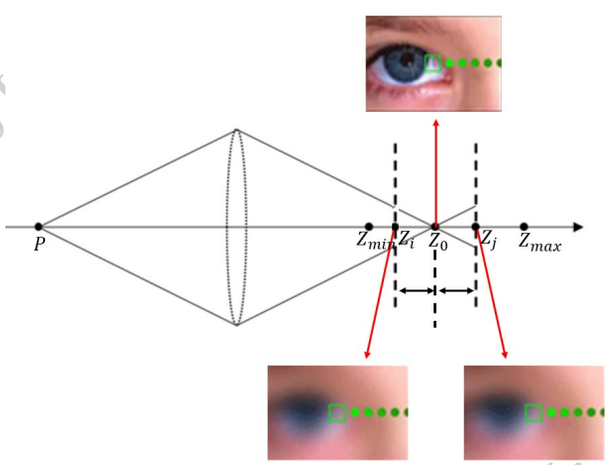

图1 自动对焦过程中镜头移动轴向二义性示例

Fig. 1 Illustration of axial ambiguity of lens movement during autofocus

堆栈中任意选择焦点切片I_k($k\in\{1,2,\cdots n\}$)作为网络输入,预测目标物在堆栈中的GT聚焦索引m。由此,自动对焦问题便可视为从单个焦点切片输入中预测聚焦索引的回归任务。

1.1.2 镜头移动方向轴向二义性问题

行进方向轴向二义性的物理本质源于散焦退化过程的对称性。根据几何光学原理,仅当物点与感光平面完全共轭时,光线方能汇聚为清晰的像点;若感光平面偏离实际焦平面,光线将发散并形成弥散圆(circle of confusion, CoC)。由于CoC的物理尺寸严格取决于感光平面偏离焦点的绝对距离,导致在目标焦点的两侧(即近焦侧与远焦侧)会产生视觉特征高度相似的散焦图像(如图1中的Z_i与Z_j所示)。这种物理属性上的对称性,导致系统仅凭单帧散焦图像难以判别镜头应当靠近还是远离感光元件,从而产生行进方向二义性。

1.2 多焦点堆栈数据集合成与标注

现有AF数据集普遍面临深度估计不准确以及焦点呼吸效应等多重挑战。究其本质,原始数据中并不包含图像中物体的真实深度。为解决这个问题,本小节采用Blender软件进行光场数据的渲染与合成。数据集包含50个静态虚拟场景,覆盖室内户外、森林湖泊、街道建筑等场景,确保了数据的多样性和丰富性。每个场景由9×9的正四边形矩阵相机阵列导出81幅视差等距子孔径图像(分辨率为1920×1920),总计4050幅,同时对每张图像保存其Blender中对应的相机阵列内外参数和深度信息。在此基础上,本文利用合成的子孔径图像,结合逐像素视差分析进一步制作了有焦点深度标注GT的AF堆栈数据集。综合考虑不同场景下的景深范围,本文将焦点堆栈的深度区间设定为2m~100m,并根据双目立体视觉测距公式(Szeliski等,2022)确定深度与视差之间的对应关系。对于光场相机中同一空间点在左右成像平面上的投影点,设其横坐标分别为x_l和x_r,则视差s可表示为$s=x_l-x_r$。根据双目几何关系,视差s与深度z满足:

$$z=\frac{fB}{s} \tag{2}$$

式中,B表示相机基线距离,f表示焦距。在获得深度区间对应的视差范围后,本文对该视差范围进行均匀划分,得到50个视差序列点,并进一步由上述公式换算得到对应的50个焦点深度序列。最后,利

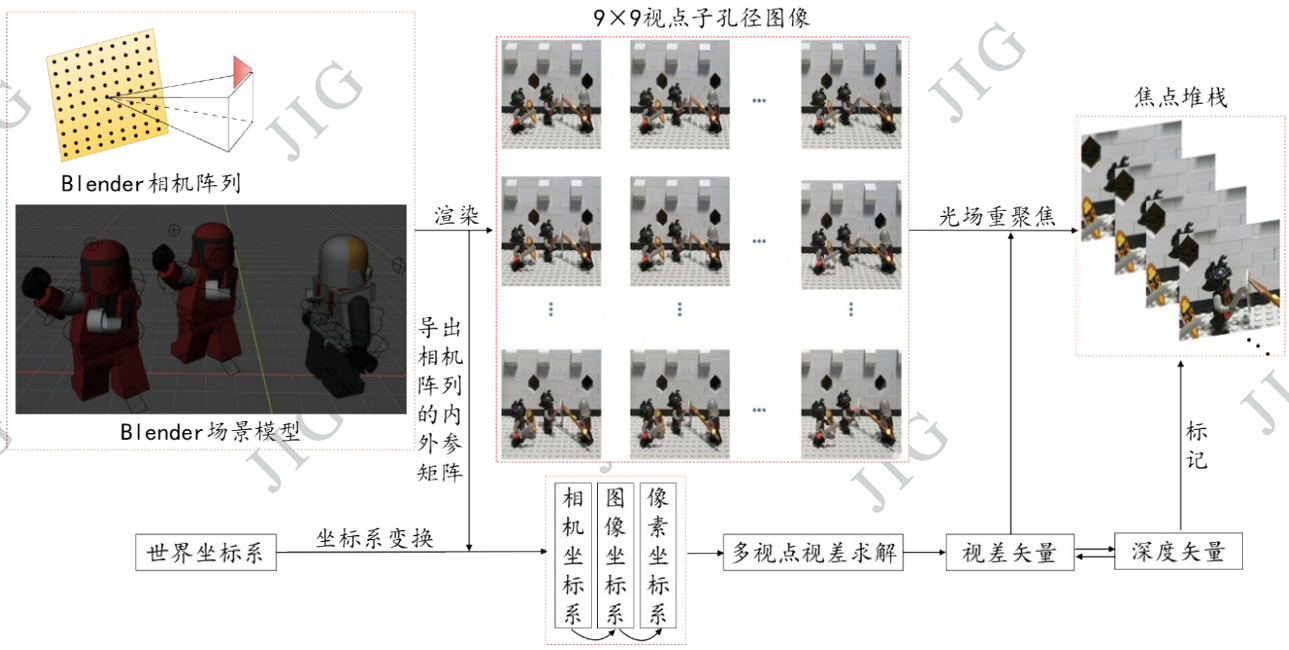

图2　合成光场重聚焦制作焦点堆栈流程

Fig. 2　Process of focus stack fabrication for synthetic light field refocusing

用该深度序列对光场重聚焦生成的焦点堆栈进行GT标注。图2展示了利用逐像素视差与深度对应的光场重聚焦方法制作焦点堆栈的流程。可概括为：首先利用Blender将场景模型渲染为若干幅光场子孔径图像，并得到目标场景的GT深度信息；再通过双目立体视觉测距公式计算出各深度位置的视差矩阵；最后利用基于逐像素视差的光场重聚焦算法（Chlubna等，2021）对平移配准后的子图像进行高分辨率重构，并对堆栈各焦点层次深度进行标注。

由于光学成像受景深限制，单次成像仅在特定深度区间内保持清晰。为构建实际拍摄主体所在区间的局部样本，本文在焦点堆栈中人工裁剪ROI以构建Patch堆栈（如图3）。在ROI选取过程中，本文参考局部纹理丰富度进行覆盖性采样，尽量包含纹理较丰富、纹理适中以及纹理较弱的区域，以提高样本来源的多样性。需要说明的是，该纹理划分仅用于ROI选取过程中的样本覆盖，并未作为独立类别标签参与训练、测试或分组统计。考虑到场景中各个目标物的实际深度存在差异，需对挑选的Patch堆栈进行深度重标注。具体而言，本文利用子孔径图像中所保存的场景真实深度图，根据图像块的中心像素坐标(x, y)访问中心子孔径图像的Z通道所保存的深度图，得到图像块中心像素处的真实深度d。为消除单中心像素采样易引入的量化噪声与几何偏差，本文以ROI中心点(i,j)为原点，取其3×3邻域的平均深度均值作为该局部目标的GT深度$\bar{d}$，计算公式为：

$$\bar{d} = \frac{1}{n^2}\sum_{k=-m}^{m}\sum_{l=-m}^{m} D(i+k, j+l) \tag{3}$$

式中，$n = 3, m = 1, D(i+k, j+l)$代表位置$(i+k, j+l)$处的深度值。在获取$\bar{d}$后，本文在Patch堆栈预设的50个焦点深度序列中进行检索，寻找与$\bar{d}$欧氏距离最近的深度值，该值所对应的索引值$p \in \{1, \cdots, 50\}$即被定义为该图像Patch堆栈的GT聚焦索引。通过这种重标注策略，解决了传统数据集中因目标物深度不一导致的聚焦模糊问题。最终，本文从50个场景堆栈中提取ROI并标注了447个原始焦点堆栈。构建了包含22,350个样本的图像Patch堆栈数据集（SLFAF）。在数据集划分上，训练集涵盖405个焦点堆栈（共得到20,250个图像Patch），测试集涵盖42个焦点堆栈（共得到2,100个图像Patch）。

1.3　图像离焦程度估计网络

深度学习方法将AF任务简化为一个回归任务，输入是一张完整图像中的一个焦点切片，以及当前镜头位置对应的焦点索引（标量），输出是焦点切片正确对焦镜头位置对应的聚焦索引预测数值。考虑到AF系统对实时性与准确性的双重要求，本文设计

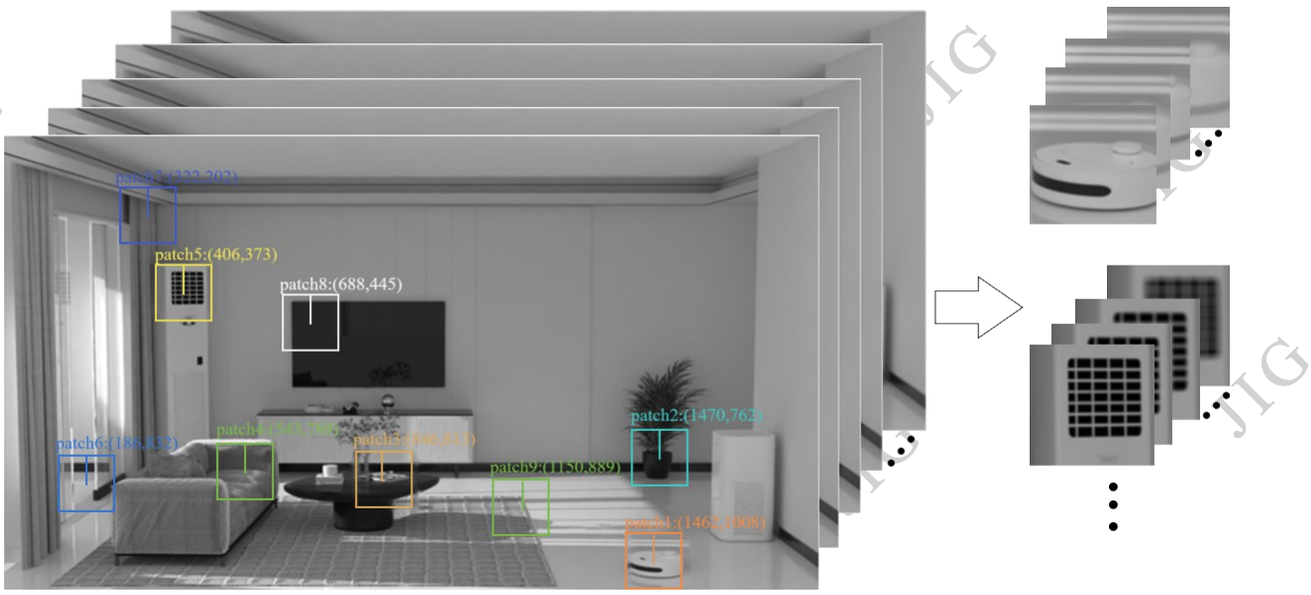

图3　ROI裁剪与Patch堆栈构建示意

Fig. 3　Example of ROI cropping and patch stack construction

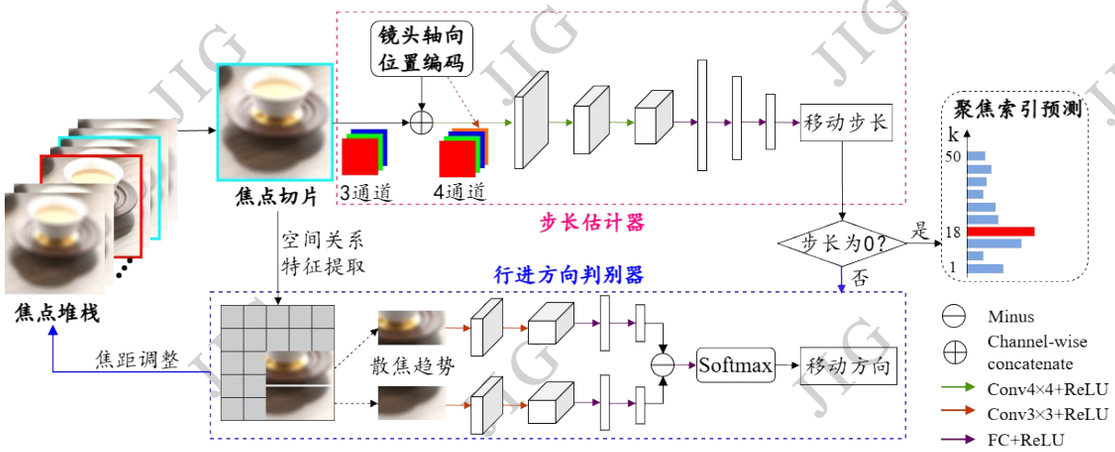

图4　图像离焦程度估计网络框架

Fig. 4　Framework of image defocus estimation network

了一个轻量网络，其架构如图4所示。该网络由两个核心模块协同工作：行进方向判别器与步长估计器。

1.3.1　行进方向判别器

针对透镜行进方向二义性这一问题，本文设计了专用的行进方向判别器网络（见图4）。受图像中心目标弥散圆斜率与镜头移动方向符号相同（Abuolaim等，2021）这一特性启发，并结合散焦模糊检测中局部梯度、纹理和边缘等低层视觉特征可表征散焦程度的研究结论（衡红军等，2021），本文网络通过提取图像局部的深度变化趋势和散焦模糊相关特征，隐式估计CoC斜率符号，从而判断镜头应向正方向或负方向移动。具体判别逻辑如下：

1）斜率为负：表示由近及远扫描时CoC逐渐缩小，说明当前感光平面位于焦点前方，需驱动成像平面向正方向移动（即等效于镜头向负方向移动）。

2）斜率为正：表示由近及远扫描时CoC逐渐扩大，说明当前感光平面已越过实际焦点。此时需驱动成像平面向负方向移动（即等效于镜头向正方向移动）。

由图4可知，行进方向判别器是一个二分类网络。网络输入为尺寸为（128 × 128 × 3）的图像Patch切片，输出为二分类结果，其中输出0表示镜头向负方向移动，输出1表示镜头向正方向移动。

为减少背景冗余信息对方向判别的干扰，首先对输入Patch进行边缘检测，突出图像中的主要结构和散焦边缘信息；随后通过分区对比边缘响应大小，确定图像中的主要结构区域及其相对于图像中心的外围方位。在此基础上，网络将外围区域块与中心区域块分别送入卷积分支进行特征提取，以学习由外围到中心的散焦变化趋势和深度梯度关系。最后，将两路特征进行融合，并通过分类层输出镜头移动方向。该结构并不直接回归焦点位置，而是先判断当前焦平面相对于真实聚焦位置的前后关系，从而为后续步长估计器提供明确的移动方向约束，降低单帧散焦图像中由对称散焦外观引起的轴向二义性问题，提高后续自动对焦控制的稳定性。

1.3.2 步长估计器

在通过行进方向判别网络确定镜头移动方向后，后续对焦控制仍需进一步估计当前焦点切片到目标聚焦位置之间的步长。为此，本文设计了一个轻量卷积回归网络，根据单张焦点切片直接输出其距离目标聚焦索引的步长预测值。该模块的骨干网络以Wang等(2021)提出的AF判别器架构为基准。

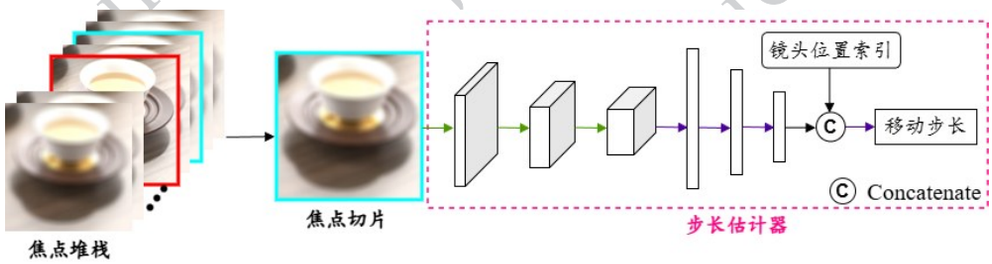

图5 初始透镜位置索引和散焦特征进行维度级联

Fig. 5 Dimensional concatenate of initial lens position index and defocus features

从图像成像特征来看，离焦程度通常会反映在边缘扩散、纹理衰减和局部梯度变化等视觉线索中。因此，网络首先通过卷积层提取与散焦相关的局部特征；随后通过特征压缩层汇聚不同空间位置的散焦响应，形成整体离焦程度表征；最后利用全连接回归层学习该离焦表征与目标聚焦索引步长之间的映射关系，输出当前焦点切片所需的绝对移动步长。然而，仅依赖单帧图像的散焦外观仍难以充分确定移动步长。原因在于，单帧图像中相似的散焦外观并不一定对应相同的初始透镜位置和目标聚焦索引，其所需的移动步长也可能不同。因此，初始透镜位置是步长估计中的关键先验信息。受自注意力机制中位置编码(positional encoding, PE)(Vaswani等，2017)的启发，本文将初始透镜位置显式引入网络，以增强模型对镜头绝对位置的感知能力。为实现这一先验信息的有效注入，本文设计并对比了以下两种不同策略的网络配置：

1) 镜头轴向位置编码(Lens-PE)：将离散的镜头移动范围归一化为轴向位置编码$f = k/n$(其中$k \in \{1, \cdots, n\}$为初始焦点索引)。将该标量f广播至与输入图像相同的空间分辨率，并作为附加的第4通道与原始3通道图像进行级联，构建4通道输入，参见图4中步长估计器部分，该方式使网络在浅层特征提取阶段即可同时感知图像散焦特征和初始透镜位置。

2) 镜头位置索引(Lens index)：直接将初始透镜位置索引 的标量值与网络深层提取的散焦特征进行维度拼接，网络设计如图5所示，该方式主要在回归末端补充初始位置信息。

消融实验表明，相比于后置的标量拼接方式Lens-index，本文提出的Lens-PE策略能够更早地参与散焦特征提取过程，使模型在初始透镜位置先验约束下，更充分地学习图像离焦程度与移动步长之间的映射关系，从而取得更好的步长估计结果。

1.3.3 损失函数

行进方向判别器的损失函数使用交叉熵损失(cross-entropy, CE)，定义如下：

$$L_{CE} = -\sum_{i=0}^{C} y_i \log(p_i) \tag{4}$$

式中，y_i表示真实标签值为i，p_i为模型预测类别为i的概率，C为类别总数(本文中对应镜头向正向1或负向移动0，即$C = 1$)。步长估计器旨在直接回归当前焦点切片距离目标聚焦索引的绝对位置差，其损失函数使用均方误差(mean squared error, MSE)，定义如下：

$$L_{MSE}=\frac{1}{N}\sum_{j=1}^{N}\left(s_j-\hat{s}_j\right)^2 \tag{5}$$

式中，N为训练的批次大小，s_j为第j个样本的真实移动步长（即地面真实聚焦索引m与当前初始透镜焦点索引k的绝对差值），$\hat{s}_j$为网络回归预测的移动步长。网络的总损失函数L_{total}可由这两个损失加权求和构成：

$$L_{total}=\lambda_1 L_{CE}+\lambda_2 L_{MSE} \tag{6}$$

式中，λ_1与λ_2为可学习的用于平衡分类与回归任务的超参数。

2 实验结果与分析

2.1 数据集与实现细节

数据集：实验采用1.2节中基于Blender渲染合成的光场重聚焦堆栈数据集（SLFAF）。数据集包含447个焦点堆栈。其中，训练集涵盖405个焦点堆栈，测试集涵盖42个焦点堆栈。Blender渲染的关键参数设置中相机焦距为50mm，光圈参数为F/2.8，传感器宽度为36mm，图像水平分辨率为1920，像元尺寸为18.75μm，相邻视点基线为5mm，光场相机阵列设置为9×9，共81个视点。

实现细节：模型训练与测试均在配置为Intel Core i5-11400 CPU（2.60 GHz）与16GB内存的工作站上进行，并搭载单张NVIDIA GeForce RTX 3080Ti GPU提供计算加速。底层算法代码基于TensorFlow深度学习框架实现。在网络训练阶段，采用Adam优化器（设置$\beta_1=0.9,\beta_2=0.999$）进行网络参数的迭代更新。训练过程中的批处理大小设置为8；初始学习率设定为1×10^{-3}，并配合权重衰减策略，每次参数更新后的衰减率设为1×10^{-4}。数据增强方面，对30%的图像进行随机旋转（0-45°），整个网络的最大训练轮数设定为200轮。在样本构建过程中，本文将每个焦点堆栈中的50个焦点切片均作为初始输入样本，而非固定选取某一特定索引的切片。对于同一焦点堆栈，其GT聚焦索引唯一确定，步长标签由当前初始焦点索引与GT聚焦索引之间的绝对距离计算得到；行进方向判别器的监督标签则根据当前初始焦点索引与GT聚焦索引之间的前后关系确定。由于不同初始焦点切片会带来不同的离焦程度和预测难度，因此本文中所有对比方法均

表1 不同网络的预测准确率在SLFAF数据集上的对比结果

Table 1 Comparison results of prediction accuracy of different networks on SLFAF dataset

输入	方法	误差容忍度				MAE	RMSE	Cost(ms)
		=0	≤1	≤2	≤4			
全栈遍历	Gradient Laplacian(Thelen等,2008)	0.751	0.950	0.993	0.995	**0.28**	**0.84**	814
	Intensity Variance(Eric等, 1988)	0.759	**0.973**	**0.993**	**0.996**	0.37	1.43	784
	Normalized SAD(Zabih等,1994)	0.565	0.648	0.753	0.870	2.29	4.47	768
	Ternary Census(Stein等,2004)	0.641	0.706	0.807	0.955	0.84	1.95	682
	Census Transform(Wadhwa等,2018)	**0.773**	0.875	0.951	0.982	0.78	2.19	622
	Normalized Envelope(Birchfield等,2002)	0.595	0.845	0.953	0.970	0.94	2.31	664
	Learning to Autofocus(Herrmann等,2020)	0.241	0.606	0.807	0.954	1.61	3.93	545
	MobileNet-v2(Choi等,2023)	0.273	0.675	0.851	0.972	1.35	3.18	**453**
单张预测	ZNCC Disparity(Wadhwa等,2018)	0.108	0.289	0.428	0.586	6.17	11.35	78
	SSD Disparity(Wadhwa等,2018)	0.167	0.442	0.611	0.770	5.48	10.40	84
	Learned Depth(Garg等,2019)	0.210	0.528	0.709	0.857	3.10	7.46	91
	Learning to Autofocus(Herrmann等,2020)	0.264	0.555	0.753	0.885	2.23	4.65	96
	MCUNet-v2(Choi等,2023)	0.290	0.634	0.827	0.908	1.97	3.92	**44**
	MobileNet-v2(Choi等,2023)	0.310	0.687	0.866	0.922	1.80	3.50	67
	本文(lens-PE)	**0.379**	**0.759**	**0.914**	**0.968**	**1.21**	**2.98**	63

采用相同的样本构建方式、训练/测试划分和初始焦点切片集合，使不同方法在相同初始输入条件下进行比较，从而保证实验结果的可比性。

2.2 定量评估与分析

为验证本文算法的有效性，本文在SLFAF数据集上进行了全面的定量评估。在评价指标方面，除采用常规的平均绝对误差（mean absolute error，MAE）和均方根误差（root mean squared error，RMSE）外，本文参考近年来AF算法采用的评估标准（Herrmann等，2020；Choi等，2023），引入误差容忍度机制来衡量聚焦索引预测准确率，即分别统计预测索引与真实标签绝对误差在0、1、2或4片以内的样本比例。需要说明的是，≤4索引误差仅表示较宽松误差容忍度下的聚焦索引预测准确率，并不作为跨相机、跨数据集通用的景深判据。根据高斯成像公式，设焦距为f，光圈数为N，孔径直径为$A = f/N$，真实目标深度为Z，预测焦点索引对应的对焦距离为g，像元尺寸为p，则像素单位下的弥散圆直径可表示为：

$$c_{px} = \frac{c}{p} = \frac{Af}{\left[p\left(1 - f/g\right)\right]}\left|\frac{1}{g} - \frac{1}{Z}\right| \tag{7}$$

在本文数据集中，$f = 50mm$，$N = 2.8$，$p = 18.75\mu m$，

焦点深度范围为2～100m，并在逆深度空间中均匀

划分为50个焦点索引。当预测索引与GT聚焦索引相差4片时，对应的弥散圆直径约为2 pixels。因此，本文将≤4索引误差作为当前成像参数和焦点采样设置下的较宽松近焦预测容忍指标。

在对比基线的设置上，本文按照输入数据量的差异，将对比算法严格划分为两类：一类是依赖全焦点堆栈遍历算法，另一类是仅需单帧切片输入的深度学习模型。在单帧输入方法的对比实验中，本文未固定选取某一特定焦点索引作为初始输入，而是将测试集中每个焦点堆栈的50个焦点切片分别作为单帧输入样本进行评价。对于同一焦点堆栈，

其GT聚焦索引保持不变，模型根据当前输入切片预测目标聚焦索引。所有单帧深度学习对比方法均采用相同的初始焦点切片集合和相同的评价指标，最终结果为完整测试切片集合上的统计结果，以保证不同方法之间比较的公平性。

如表1所示，本文选取了Choi等人（2023）提出的算法作为对比基准，并与近年来的全栈遍历AF方法和单张预测AF方法进行比较。实验结果表明，在单帧预测这一极具挑战的设定下（即仅依赖单张焦点切片预测目标聚焦索引），本文模型相较于现有深度学习方法实现了显著的精度提升。具体而言，

相较于MobileNet基准模型，本文方法在零误差

表2 行进方向判别器的消融实验结果

Table 2 Ablation experimental results of the travel direction discriminator

输入	网络	判别器	误差容忍度				MAE	RMSE
			= 0	≤1	≤2	≤4		
单张预测	Base		0.075	0.221	0.352	0.584	2.74	4.61
	Base	√	0.224	0.579	0.726	0.834	1.89	3.78
	Base + lens-PE		0.129	0.298	0.434	0.621	1.48	3.34
	Base + lens-PE	√	**0.379**	**0.759**	**0.914**	**0.968**	**1.21**	**2.98**

表3 初始透镜位置索引消融实验结果

Table 3 Initial lens position index ablation experimental results

输入	网络	判别器	误差容忍度				MAE	RMSE
			= 0	≤1	≤2	≤4		
单张预测	Base	√	0.224	0.579	0.726	0.834	1.89	3.78
	Base + lens-index	√	0.357	0.697	0.864	0.924	1.53	3.27
	Base + lens-PE	√	**0.379**	**0.759**	**0.914**	**0.968**	**1.21**	**2.98**

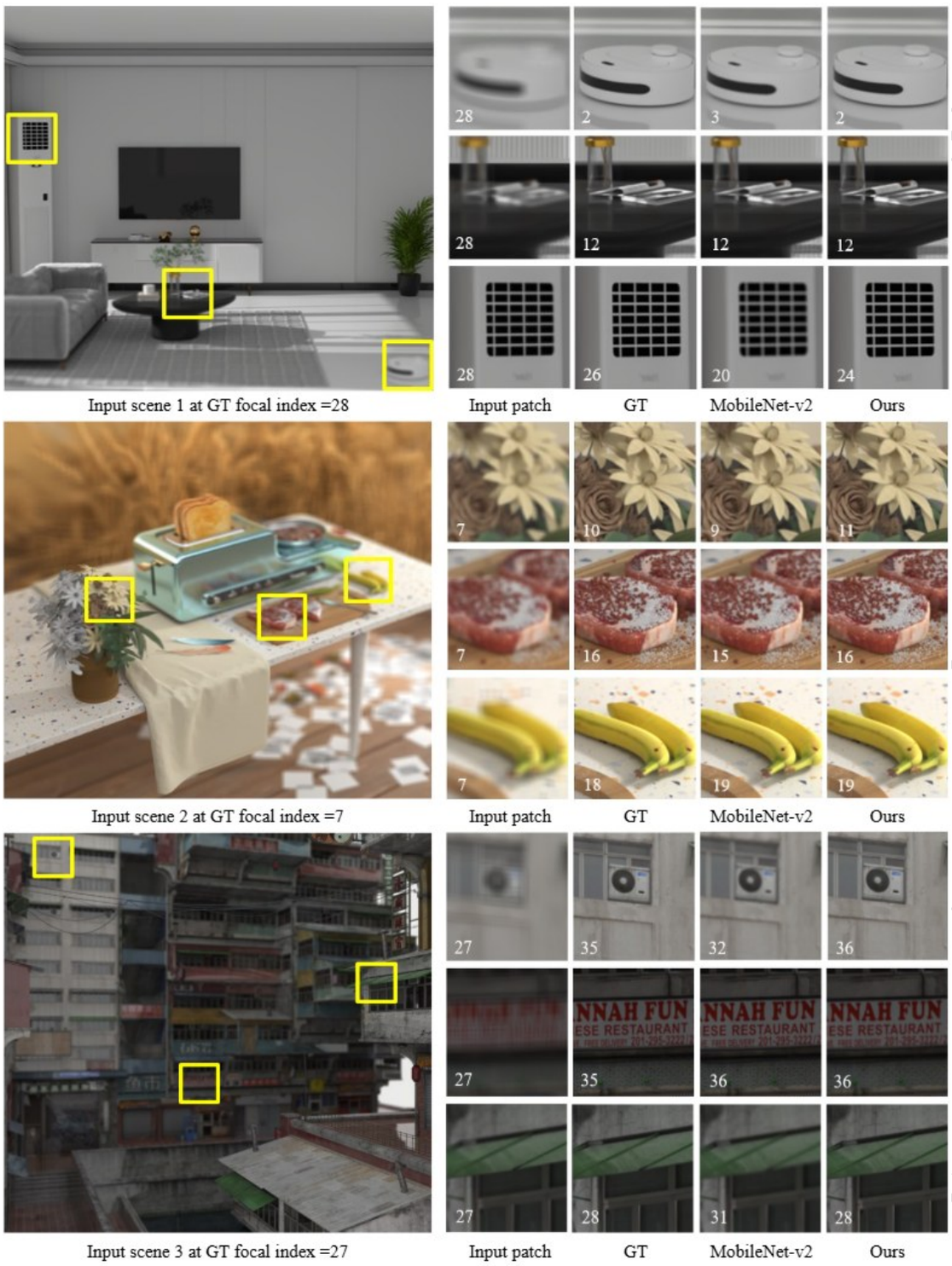

图6　本文模型与基线网络的预测图像结果对比

Fig. 6　Comparison of prediction image results between the proposed model and baseline network

(*Error* = 0)与误差容忍度为4(*Error* ≤ 4)的评估标准下,预测准确率分别大幅提升了6.9%和4.6%,且计算复杂度较Choi等人(2023)的方法明显降低。总体来看,本文方法在≤4片误差容忍度下的综合准确率高达96.8%,这充分验证了本文提出的离焦程度估计的单帧AF算法在步长预测任务上的精准性。需要说明的是,本文模型专为单帧输入设计,不支持全栈遍历预测。尽管传统全栈对比度算法精度略高,但其遍历整个堆栈导致耗时极长,而本文采用轻量级CNN进行单张预测,推理时间仅为传统算法

的1/10。

2.3 定性评估与分析

为直观评估本文模型在SLFAF数据集上的实际对焦性能，图6展示了三个典型场景下本文算法与MobileNet-v2基线模型Choi等人(2023)的可视化对比结果，其中图像左下角数字表示预测的焦点索引。由图6可见，在常规样例中，本文方法预测的焦点索引整体更接近GT索引，对应焦点切片的边缘结构和局部纹理也更为清晰，说明所提出的行进方向判别器与Lens-PE位置编码能够有效辅助单帧离焦图像的步长估计。

此外，图6中也给出了部分纹理较弱或近似单深度平面的样例。由于此类ROI内部可用于判断离焦程度的边缘、纹理和深度变化线索相对有限，焦点索引预测难度较高。对于这些样例，本文方法虽未完全命中绝对GT索引，但预测结果仍与GT索引保持较小偏差，表明模型在有限散焦线索条件下仍具有一定的步长估计能力。需要说明的是，本文中的纹理复杂度划分主要用于ROI采样阶段的样本覆盖，并未作为独立测试类别标签，因此该部分结果仅作为定性可视化分析，具体的分纹理复杂度定量评估将在后续工作中进一步开展。

2.4 消融实验

为验证本文网络架构中核心模块的设计合理性与性能贡献，本节在SLFAF数据集上开展了系统的消融实验。实验以基础的CNN回归网络作为基准模型(Base)，重点评估了行进方向判别器的引入，以及两种初始透镜位置先验输入策略(即镜头轴向位置编码Lens-PE与拼接镜头索引Lens-index)的有效性。表2展示了针对行进方向判别器模块的定量的消融分析结果，在Base的基础上引入行进方向判别器显著提升了网络的容忍度，同时降低了MAE与RMSE，进一步验证了行进方向判别器的有效性。

为进一步验证初始透镜位置先验的作用，本文在保留行进方向判别器的基础上，对比了Base、Lens-index和Lens-PE三种策略。其中，Base不引入初始透镜位置先验，仅根据单张散焦图像进行步长回归；Lens-index将归一化后的初始透镜焦距索引作为一个标量，在深层特征后与全连接层进行拼接，用于辅助最终步长预测；Lens-PE则将初始透镜位置编码为与输入图像空间尺寸一致的附加通道，并与输入图像进行浅层融合，使网络在早期特征提取阶段即可感知初始透镜位置。

实验结果表明，引入初始透镜位置先验能够显著提升步长估计性能。相比Base，Lens-index将±4片容忍度下的准确率由83.4%提升至92.4%，MAE由1.89降至1.53，说明初始透镜位置是步长回归的重要先验信息。进一步地，Lens-PE的准确率提升至96.8%，MAE降至1.21，RMSE降至2.98，优于深层标量拼接的Lens-index。其原因在于，步长估计不仅与当前图像的散焦程度有关，也与图像采集时的初始透镜位置有关。Lens-index仅在回归末端引入位置标量，难以充分调制前端卷积层对散焦纹理、边缘模糊和弥散圆特征的提取；而Lens-PE通过早期空间融合，使网络在低层特征提取阶段即可联合建模图像散焦信息与初始透镜位置先验，从而更有效地学习由当前焦点切片到目标聚焦位置的移动步长映射关系。

3 结 论

本文针对现有自动对焦模型存在训练阶段缺乏高精度物理标注数据的难点，提出了一种光场自动对焦数据集构建及图像离焦程度估计方法，在数据合成与标注过程中避免了真实采集中机械标定误差和焦点呼吸效应的影响，构建了具备高精度的焦点堆栈数据集(SLFAF)。在此数据集的基础上，本文提出了一种图像离焦程度估计方法来实现自动对焦。其中，针对散焦退化过程中的透镜移动轴向二义性难题，本文设计的行进方向判别网络通过隐式建模弥散圆斜率变化与透镜位移的物理映射关系，实现了对镜头移动方向的精准锁定；同时，在基准模型的步长估计器上引入了镜头轴向位置编码(Lens-PE)，在增强网络对全局物理坐标感知能力的同时，实现从单帧焦点切片到目标聚焦索引的高效步长预测。实验结果表明，本文方法在SLFAF数据集上不仅取得了单帧对焦预测精度的最优表现，且保持了较低的推理时间复杂度。消融实验进一步验证了各模块设计的合理性与有效性。未来工作将进一步引入梯度能量、局部方差、纹理熵等客观纹理度量，对不同纹理复杂度样本进行标准化划分，并在真实采集数据和低纹理场景下进一步验证模型的泛化能力。

参考文献(References)

Abuolaim A, Delbracio M, Kelly D, Brown M S and Milanfar P. 2021. Learning to reduce defocus blur by realistically modeling dual-pixel data//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada: IEEE: 2289-2298 [DOI: 10.1109/ICCV48922.2021.00229]

Birchfield S and Tomasi C. 2002. A pixel dissimilarity measure that is insensitive to image sampling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(4): 401-406 [DOI: 10.1109/34.677269]

Chen C Y, Hwang R C and Chen Y J. 2010. A passive auto-focus camera control system. Applied Soft Computing, 10(1): 296-303 [DOI: 10.1016/j.asoc.2009.07.007]

Chlubna T, Milet T and Zemčík P. 2021. Real-time per-pixel focusing method for light field rendering. Computational Visual Media, 7(3): 319-333 [DOI: 10.1007/s41095-021-0205-0]

Choi M, Lee H and Lee H. 2023. Exploring positional characteristics of dual-pixel data for camera autofocus//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 13158-13168 [DOI: 10.1109/ICCV51070.2023.01210]

Dai Y C, Zhang F Y, Pan L Y, Xiang M C and He M Y. 2022. Dual-pixel imaging and applications: an overview. Journal of Image and Graphics, 27(12): 3395-3414 (戴玉超, 章飞宇, 潘利源, 项末初, 何明一. 2022. 全像素双核成像技术及应用研究综述. 中国图象图形学报, 27(12): 3395-3414) [DOI: 10.11834/jig.210984]

Garg R, Wadhwa N, Ansari S and Barron J T. 2019. Learning single camera depth estimation using dual-pixels//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea: IEEE: 7628-7637 [DOI: 10.1109/ICCV.2019.00772]

Guo X, Nie R, Cao J, Zhou D and Qian W. 2018. Fully convolutional network-based multifocus image fusion. Neural Computation, 30(7): 1775-1800 [DOI: 10.1162/neco_a_01098]

Han J W, Kim J H, Lee H T and Ko S J. 2011. A novel training based auto-focus for mobile-phone cameras. IEEE Transactions on Consumer Electronics, 57(1): 232-238 [DOI: 10.1109/TCE.2011.5735507]

Heng H J, Ye H B, Zhou M and Huang R. 2021. Coarse-to-fine multi-scale defocus blur detection. Journal of Image and Graphics, 26(3): 581-593 (衡红军, 叶何斌, 周末, 黄睿. 2021. 由粗到精的多尺度散焦模糊检测. 中国图象图形学报, 26(3): 581-593) [DOI: 10.11834/jig.200126]

Herrmann C, Bowen R S, Wadhwa N, Garg R, He Q, Barron J T, et al. 2020. Learning to autofocus//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 2230-2239 [DOI: 10.1109/CVPR42600.2020.00230]

Ho C J, Chan C C and Chen H. 2020. Af-net: A convolutional neural network approach to phase detection autofocus. IEEE Transactions on Image Processing, 29: 6386-6395 [DOI: 10.1109/TIP.2019.2947349]

Krotkov E. 1988. Focusing. International Journal of Computer Vision, 1(3): 223-237 [DOI: 10.1007/BF00127822]

Lee S, Choi M, Lee N and Lee H. 2025. Stable autofocus with focal consistency loss//2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Tucson, AZ, USA: IEEE: 640-649 [DOI: 10.1109/WACV61041.2025.00072]

Li W, Wang G, Hu X and Yang H. 2018. Scene-adaptive image acquisition for focus stacking//Proceedings of the 25th IEEE International Conference on Image Processing (ICIP). Athens: IEEE: 1887-1891 [DOI: 10.1109/ICIP.2018.8451455]

Liu Y, Chen X, Peng H and Wang Z. 2017. Multi-focus image fusion with a deep convolutional neural network. Information Fusion, 36: 191-207 [DOI: 10.1016/j.inffus.2016.12.001]

Liu Y, Ou L, Fu Q, Amata H, Heidrich W and Peng Y. 2025. Learned Binocular-Encoding Optics for RGBD Imaging Using Joint Stereo and Focus Cues//Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. Nashville, TN, USA: IEEE: 15833-15842 [DOI: 10.1109/CVPR52734.2025.01476]

Nimisha T M, Kumar Singh A and Rajagopalan A N. 2017. Blur-invariant deep learning for blind-deblurring//Proceedings of the IEEE International Conference on Computer Vision: 4752-4760 [DOI: 10.1109/ICCV.2017.509]

Ray S. 2002. Applied photographic optics. London, UK: Routledge [DOI: 10.4324/9780080499253]

Ren Z, Xu Z and Lam E Y. 2018. Learning-based nonparametric autofocusing for digital holography. Optica, 5(4): 337-344 [DOI: 10.1364/OPTICA.5.000337]

Schuler C J, Hirsch M, Harmeling S and Schölkopf B. 2015. Learning to deblur. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(7): 1439-1451 [DOI: 10.1109/TPAMI.2015.2481418]

Stein F. 2004. Efficient computation of optical flow using the census transform//Joint Pattern Recognition Symposium. Berlin, Heidelberg: Springer Berlin Heidelberg: 79-86 [DOI: 10.1007/978-3-540-28649-3_10]

Szeliski R. 2022. Computer vision: algorithms and applications. Cham, Switzerland: Springer Nature [DOI: 10.1007/978-3-030-34372-9]

Thelen A, Frey S, Hirsch S and Hering P. 2008. Improvements in shape-from-focus for holographic reconstructions with regard to focus operators, neighborhood-size, and height value interpolation. IEEE Transactions on Image Processing, 18(1): 151-157 [DOI: 10.1109/TIP.2008.2007049]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention is all you need. Advances in Neural Informa-

tion Processing Systems，30：5998-6008［DOI：10.48550/ arXiv.1706.03762］

Wadhwa N，Garg R，Jacobs D E，Feldman B E，Kanazawa N，Carroll R，et al. 2018. Synthetic depth-of-field with a single-camera mobile phone. ACM Transactions on Graphics（ToG），37（4）：1-13［DOI：10.1145/3197517.3201329］

Wang C，Huang Q，Cheng M，Ma Z and Brady D J. 2021. Deep learning for camera autofocus. IEEE Transactions on Computational Imaging，7：258-271［DOI：10.1109/TCI.2021.3059497］

Xu G，Wang X，Ding X and Yang X. 2023. Iterative geometry encoding volume for stereo matching//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，BC，Canada：IEEE：21919-21928［DOI：10.48550/arXiv.2303.06615］

Zabih R and Woodfill J. 1994. Non-parametric local transforms for computing visual correspondence//European Conference on Computer Vision. Berlin，Heidelberg：Springer Berlin Heidelberg：151-158［DOI：10.1007/BFb0028345］

Zhu S，Li C，Jiang Y，Wei L，Kan N，Zheng Z，et al. 2025. Stabilizing and Accelerating Autofocus with Expert Trajectory Regularized Deep Reinforcement Learning//Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. Nashville，TN，USA：IEEE：26440-26450［DOI：10.1109/CVPR52734.2025.02462］

作者简介

周广彬，男，助教，博士研究生，研究方向为深度学习、图像恢复和图像深度估计。E-mail：alexgbzhou@snut.edu.cn

肖照林，通信作者，男，教授，主要研究方向为计算摄影学。E-mail：xiaozhaolin@xaut.edu.cn

杨延伟，男，硕士研究生，主要研究方向为图像恢复。E-mail：2926921221@qq.com