光学学报

聚焦堆栈中空间几何结构的深度估计

罗天琦，邓小娟，刘畅，邱钧*

北京信息科技大学应用数学研究所，北京 100101

摘要 利用聚焦堆栈估计场景深度是计算成像领域中的重要技术手段。提出三维自适应加权全变分计算框架，用于解决场景中弱纹理区域和遮挡区域深度线索丢失导致深度估计不准确的问题。相比传统二维引导滤波方法，所提三维优化框架不仅考虑聚焦堆栈和聚焦测度中共同蕴藏的场景几何结构，避免在深度图中错误地引入场景物理信息，还充分考虑聚焦堆栈和聚焦测度沿图像序列方向的结构特点，实现更高程度的数据保真。模拟数据和实际数据实验结果表明，所提方法能够有效提升聚焦堆栈估计深度的精度。

关键词 图像处理；聚焦堆栈；聚焦测度；深度估计；结构一致性

中图分类号 TP391 **文献标志码** A **DOI:** 10.3788/AOS230645

1 引 言

深度估计可以从常规成像系统获取的二维图像中感知和重建三维场景，是计算机视觉领域的重要研究方向[1]，已广泛应用于自动驾驶[2]、工业测量[3]和虚拟现实[4]等领域。根据成像系统传感器数量的不同，可以将深度估计方法分为两类：基于多目图像和基于单目图像的深度估计[5-8]。由聚焦堆栈估计深度是一种以聚焦程度为深度线索的被动式单目视觉技术，具有成像设备体积小、计算成本低等优点[9-10]。这类方法对图像聚焦程度的度量严重依赖场景的纹理信息。光照不足、场景平滑或遮挡等区域的聚焦程度难以被准确度量，导致对这些区域的深度信息估计不准确。针对这一问题，本文提出基于三维自适应加权全变分模型的聚焦测度修正方法，所提方法利用聚焦堆栈和聚焦测度中几何结构的关联关系，对场景深度的边界位置进行有效定位，提高光滑和遮挡区域深度估计精度的同时，有效保留深度的结构信息。所提方法的思路将有助于从不同角度给优化深度的方法提供借鉴和帮助。

2 相关工作

根据聚焦堆栈估计深度的步骤，可以将现有方法分为3类：设计更理想的聚焦测度算子，以准确度量像素点的聚焦程度；优化聚焦测度体数据，修正深度线索丢失或测度算子不理想带来的聚焦测度误差；利用全聚焦图像对初始深度图进行引导滤波。

已有文献提出了很多种算法来度量整幅图像或像素点的聚焦程度，比如，基于图像一阶导数的梯度算子[11-12]、基于图像二阶导数的拉普拉斯算子[10,13]、基于图像高频信息的小波变换或离散余弦变换算子[14-16]、基于图像信息熵的统计特性算子[17]、融合图像特征点密度的组合算子[18-19]等。以上不同类型聚焦测度算子的设计旨在从图像不同变换域度量图像的清晰程度，降低图像对比度、饱和度、噪声水平及局部窗口大小等因素对聚焦程度的影响。

聚焦测度算子的设计很难融入场景本身的先验信息，如场景颜色和深度信息的分片光滑属性。针对该问题，Moeller等[20]首次利用聚焦测度值随深度变化的光滑性，建立关于深度变量的全变分优化模型，提升深度估计的精度。该工作打开了利用优化方法解决聚焦堆栈估计深度的思路。后续工作从不同角度挖掘场景和数据的先验信息，建立更具鲁棒性的优化模型。如Ali等[15,21-22]将聚焦测度体数据作为优化变量，建立了一系列的三维优化模型，以得到更能准确反映聚焦程度的测度值。

相比修正三维聚焦测度值[23]，优化初始估计的二维深度图更加简单、高效。这类方法主要利用场景全聚焦图中的结构信息对深度图进行引导滤波[24-25]。全聚焦图和深度图反映了场景不同属性的信息，直接用颜色信息引导几何结构，会在深度图中引入伪影。针对这类不同属性图像之间的引导滤波问题，Liu等[26]

收稿日期：2023-03-08；**修回日期**：2023-04-22；**录用日期**：2023-05-06；**网络首发日期**：2023-06-28

基金项目：国家自然科学基金(61931003，62101056，62171044)、北京市自然科学基金(4222004)、北京信息科技大学'勤信人才'培育计划

通信作者：*qiujun@bistu.edu.cn

提出了基于非凸优化模型的保边界去噪方法，动态更新参考图像和目标图像，以保留二者共同的结构信息。Guo 等[27]将参考图像和目标图像之间的互信息定义为引导结构，对目标图像进行引导滤波（MuGIF）。以上这类互引导滤波方法在提升初始深度图的精度上效果明显，但只考虑二维图像之间的结构引导，没有充分利用聚焦堆栈和聚焦测度数据之间更高维度的关联关系。

受上述二维互引导滤波方法的启发，本文考虑三维聚焦堆栈和聚焦测度体数据之间的结构关系，构造三维结构一致性算子，建立三维自适应加权全变分模型，修正三维聚焦测度体数据，提高初始估计深度的精度。主要贡献如下：将二维互引导滤波的机理泛化到三维引导滤波中，便于考虑更原始的高维数据中蕴含的场景几何先验信息；将场景深度的光滑性和颜色信息及聚焦测度值随深度变化的光滑性建模成三个方向的全变分（TV）正则项，便于利用成熟的分裂 Bregman 算法框架推导出稳定、高效的数值求解方法。

3　基于结构一致性加权的三维 TV 优化方法

基于 TV 极小化的优化模型最早由 Rudin 等[28]于 1992 年提出，该方法在二维图像去噪、图像修补及图像重建方面都取得了不错的效果，广泛应用于计算机视觉、医学影像和图像处理等领域[29]。传统二维 TV 优化模型在去噪的同时具有一定的保边界能力，但是，当噪声梯度幅值大于边界梯度幅值时，该模型将面临权衡去噪和保持边缘细节的难题。基于引导滤波的方法利用参考图像的边界信息对目标图像进行去噪，可以有效避免权衡去噪与保边界的难题。就有了加权 TV 的优化模型，但此模型对二维图像进行引导滤波时可以引入的优化信息较少，因此尝试将其扩展到三维图像领域。由于三维数据具有更丰富的信息量，加权三维 TV 正则化模型可以更好地平衡去噪和保边界能力。

图 1 是聚焦测度剖线图。图 1(a)中五角星点的光照不足位置沿着 Z 轴聚焦测度最大值对应的层数并不是图 1(b)准确深度对应的层数，其聚焦测度沿层的方向没有规律，不能准确反映聚焦程度的变化。但它周围的点聚焦测度有规律，如图 1(c)和图 1(d)所示，两条线的峰值都在 13/14 层的位置，能准确反映聚焦程度的变化，因此可以利用周围的点对当前点进行修正，周围像素点的图像序列维度的信息可以更好地修正深度估计的结果。其中 FM 为聚焦测度值。

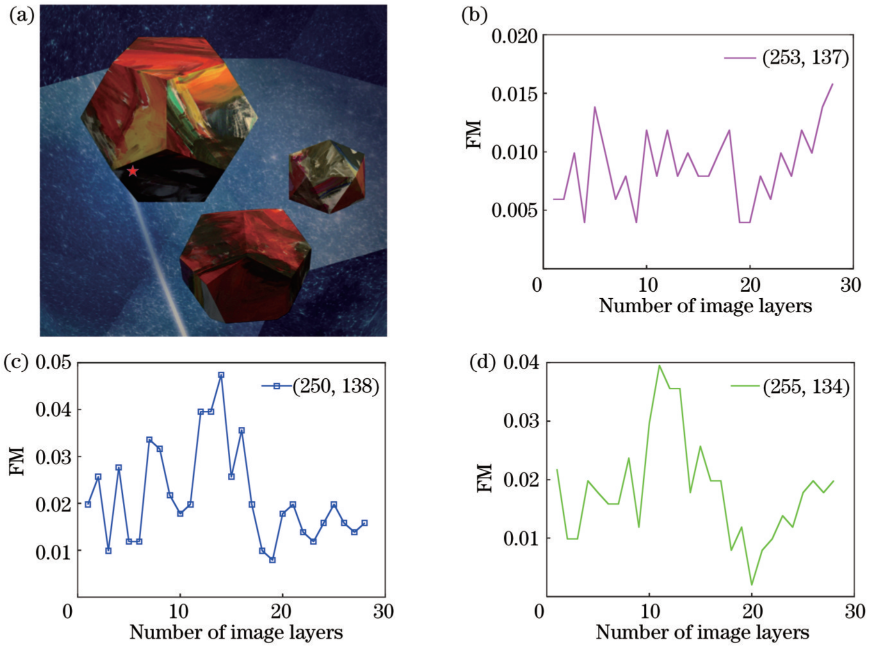


图 1　聚焦测度剖线图。(a)场景；(b)初始 FM；(c)(d)标记点周围像素 FM

Fig. 1　Focus measure profile. (a) Scene; (b) initial FM; (c)(d) pixel FM around the mark point

3.1　结构一致性

图像的结构使用像素梯度值来衡量。当聚焦堆栈中的物理结构梯度跳跃和聚焦测度中的几何结构梯度跳跃在同一位置发生较大变化时，认为该位置的结构是一致的，可以用于信息交换。如果在同一位置上，两种属性信息内一个结构较大而另一个较小则为不一致结构，或两者都很小则是平滑结构，不一致或者平滑结构都不进行信息交换。结构一致性指两种数据结构之间的相似程度。

图 2 为聚焦堆栈与聚焦测度结构。从水平线位置的剖线图可以看出，聚焦堆栈和聚焦测度数据在 X 轴上存在几个大阶梯跳跃的位置，如图 2(d)所示。如果某个位置在两个数据结构中都存在大结构，则认为该位置是结构一致的，并且可以进行信息交换。如图 2

(f)所示，只有前两个大梯度跳跃的位置是两者共存的，因此这两个位置在X轴上具有较大的结构一致性，Y轴上同理。

三维的聚焦堆栈和聚焦测度都包含X、Y和Z三个维度。每个位置Z维度的像素值由于景深不同而不同。在从失焦到聚焦再到失焦的数据采集过程中，像素值也会随之由小变大再变小。聚焦测度体数据中每个位置的像素在图像序列方向上具有一个先变大后变小的结构，如图 3 所示。二者共有的先变大后变小的变化趋势可以视为Z维度上的结构一致性。使用三维结构一致性进行引导的优势在于，二维一致结构可以提供信息进行优化外，第三个维度的空间结构也可以为优化过程提供更多的信息，从而使优化结果更加准确。

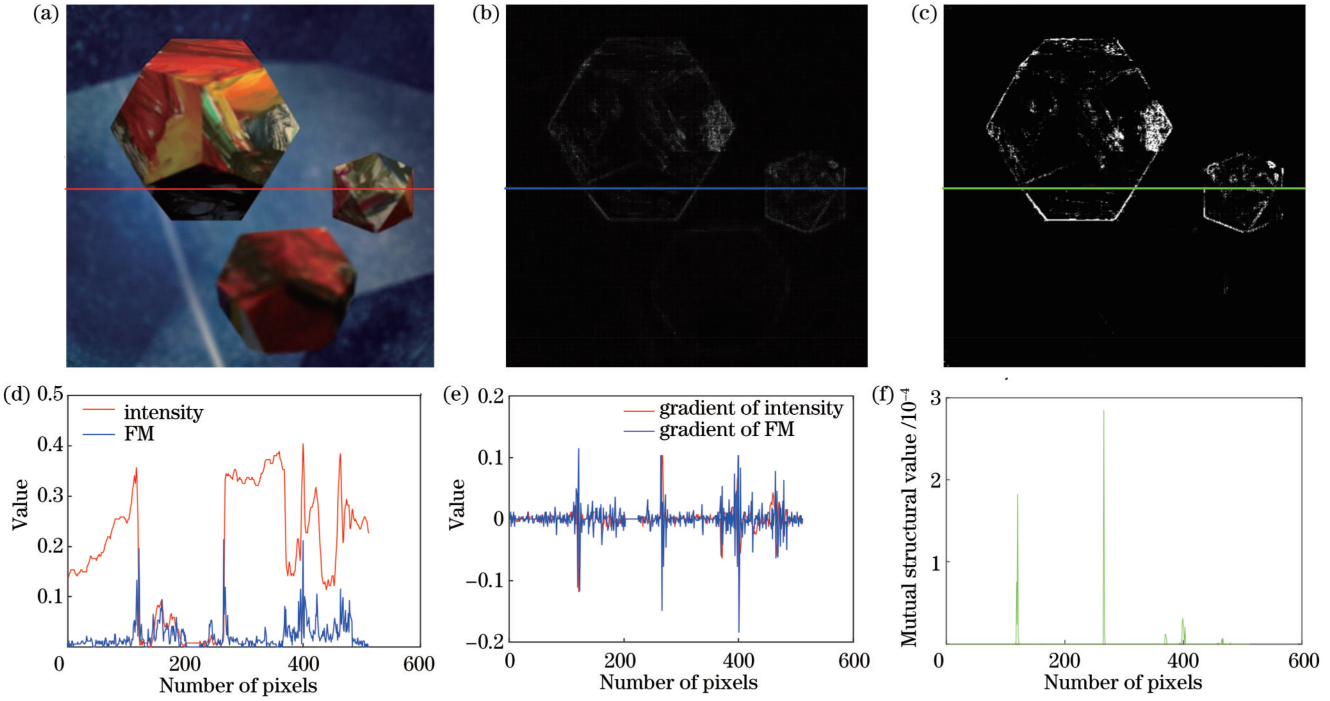


图 2 聚焦堆栈与聚焦测度结构。(a)聚焦堆栈第 13 层；(b)聚焦测度第 13 层；(c)三维结构第 13 层；(d)水平剖线；(e)剖线梯度；(f)结构值

Fig. 2 Profile of the focal stack and focus measure. (a) The 13th layer of the focal stack; (b) the 13th layer of the focus measure; (c) the 13th layer of the mutual structure; (d) horizontal profile; (e) gradient of profile; (f) structural value

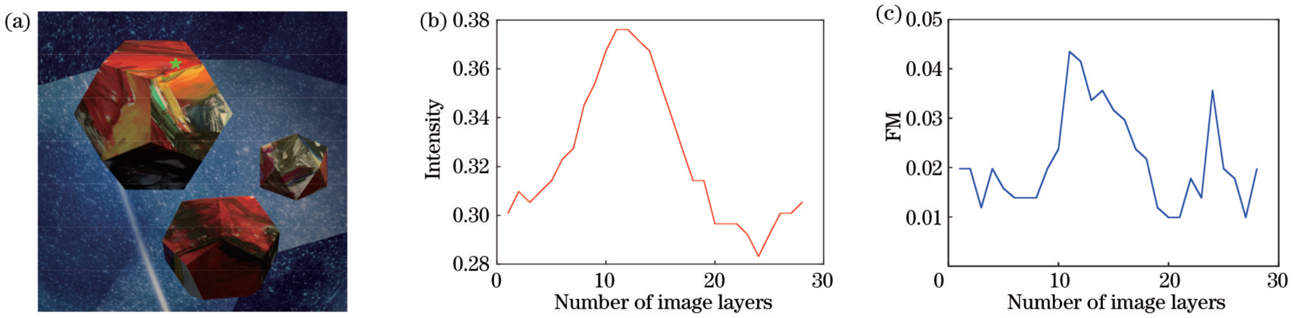


图 3 Z轴方向剖线图。(a)场景；(b)像素值；(c)聚焦测度值

Fig. 3 Profile along Z-axis direction. (a) Scene; (b) pixel; (c) FM

聚焦堆栈$\boldsymbol{I}$和聚焦测度体数据$\boldsymbol{u}$之间的三维结构一致性定义为

$$S=\left[(\nabla_x \boldsymbol{I})(\nabla_x \boldsymbol{u})\right]^2+\left[(\nabla_y \boldsymbol{I})(\nabla_y \boldsymbol{u})\right]^2+\left[(\nabla_z \boldsymbol{I})(\nabla_z \boldsymbol{u})\right]^2, \tag{1}$$

式中：∇_x、∇_y、∇_z是三个方向的梯度。当一个像素在聚焦堆栈和聚焦测度中都具有较大的梯度值时，说明该像素是一致结构的一部分，可以引入信息进行聚焦测度的优化。

3.2 基于三维结构一致性加权全变分的聚焦测度优化模型

将优化聚焦测度的过程建模成三维的加权 TV 正则，其自适应权重由结构一致性确定。将聚焦测度正则化问题表示为能量最小化问题并有目标函数，

$$E(\boldsymbol{u})=\frac{\lambda}{2}\|\boldsymbol{u}-\boldsymbol{f}\|_2^2+\omega(\|\nabla_x\boldsymbol{u}\|_1+\|\nabla_y\boldsymbol{u}\|_1+\|\nabla_z\boldsymbol{u}\|_1),\tag{2}$$

式中：第一项是数据保真度项，第二项是正则化项，λ控制这两个项的相对重要性；$\boldsymbol{f}$为初始聚焦堆栈体数据，$\boldsymbol{u}$为优化后的聚焦堆栈体数据；权重是结构一致性定义的指数函数，$\omega(S)=\exp(-\beta\cdot S)$，$\beta$为可调节参数。聚焦堆栈和聚焦测度中的像素结构越相似，它们所起的作用就越大。该能量泛函使用Split Bregman方法进行求解[30-37]。首先引入辅助变量$\boldsymbol{d}_x=\nabla_x\boldsymbol{u}$、$\boldsymbol{d}_y=\nabla_y\boldsymbol{u}$、$\boldsymbol{d}_z=\nabla_z\boldsymbol{u}$，和松弛变量$\boldsymbol{b}_x$、$\boldsymbol{b}_y$、$\boldsymbol{b}_z$。在离散情况下，将Split Bregman算法应用到三维TV模型中，则有

$$\begin{cases}(\boldsymbol{u}^{(k+1)},\boldsymbol{d}_x^{(k+1)},\boldsymbol{d}_y^{(k+1)},\boldsymbol{d}_z^{(k+1)})=\underset{\boldsymbol{u},\boldsymbol{d}_x,\boldsymbol{d}_y,\boldsymbol{d}_z}{\arg\min}\,T(\boldsymbol{u},\boldsymbol{d}_x,\boldsymbol{d}_y,\boldsymbol{d}_z)\\ \boldsymbol{b}_x^{(k+1)}=\boldsymbol{b}_x^{(k)}+\nabla_x\boldsymbol{u}^{(k+1)}+\boldsymbol{d}_x^{(k+1)}\\ \boldsymbol{b}_y^{(k+1)}=\boldsymbol{b}_y^{(k)}+\nabla_y\boldsymbol{u}^{(k+1)}+\boldsymbol{d}_y^{(k+1)}\\ \boldsymbol{b}_z^{(k+1)}=\boldsymbol{b}_z^{(k)}+\nabla_z\boldsymbol{u}^{(k+1)}+\boldsymbol{d}_z^{(k+1)}\end{cases},\tag{3}$$

其中，$T(\boldsymbol{u},\boldsymbol{d}_x,\boldsymbol{d}_y,\boldsymbol{d}_z)=\frac{\lambda}{2}\|\boldsymbol{u}-\boldsymbol{f}\|_2^2+\omega\|\boldsymbol{d}_x\|_1+\omega\|\boldsymbol{d}_y\|_1+\omega\|\boldsymbol{d}_z\|_1+\frac{\boldsymbol{u}}{2}(\|\boldsymbol{d}_x-\nabla_x\boldsymbol{u}\|_2^2+\|\boldsymbol{d}_y-\nabla_y\boldsymbol{u}\|_2^2+\|\boldsymbol{d}_z-\nabla_z\boldsymbol{u}\|_2^2)$。则式(2)求解可转换为

$$\min_{\boldsymbol{u}}\frac{\lambda}{2\omega}\|\boldsymbol{u}-\boldsymbol{f}\|_2^2+\|\boldsymbol{d}_x\|_1+\|\boldsymbol{d}_y\|_1+\|\boldsymbol{d}_z\|_1+\frac{\boldsymbol{u}}{2\omega}(g_1+g_2+g_3),\tag{4}$$

式中：$g_1=\|\boldsymbol{d}_x-\nabla_x\boldsymbol{u}-\boldsymbol{b}_x^{(k)}\|_2^2$，$g_2=\|\boldsymbol{d}_y-\nabla_y\boldsymbol{u}-\boldsymbol{b}_y^{(k)}\|_2^2$，$g_3=\|\boldsymbol{d}_z-\nabla_z\boldsymbol{u}-\boldsymbol{b}_z^{(k)}\|_2^2$。其中，$\boldsymbol{b}_x^{(k)}$、$\boldsymbol{b}_y^{(k)}$、$\boldsymbol{b}_z^{(k)}$的大小与分裂迭代次数$k$有很强的相关性。使用迭代极小化的方式来求解这个极小值问题时就需要解决它的子问题：

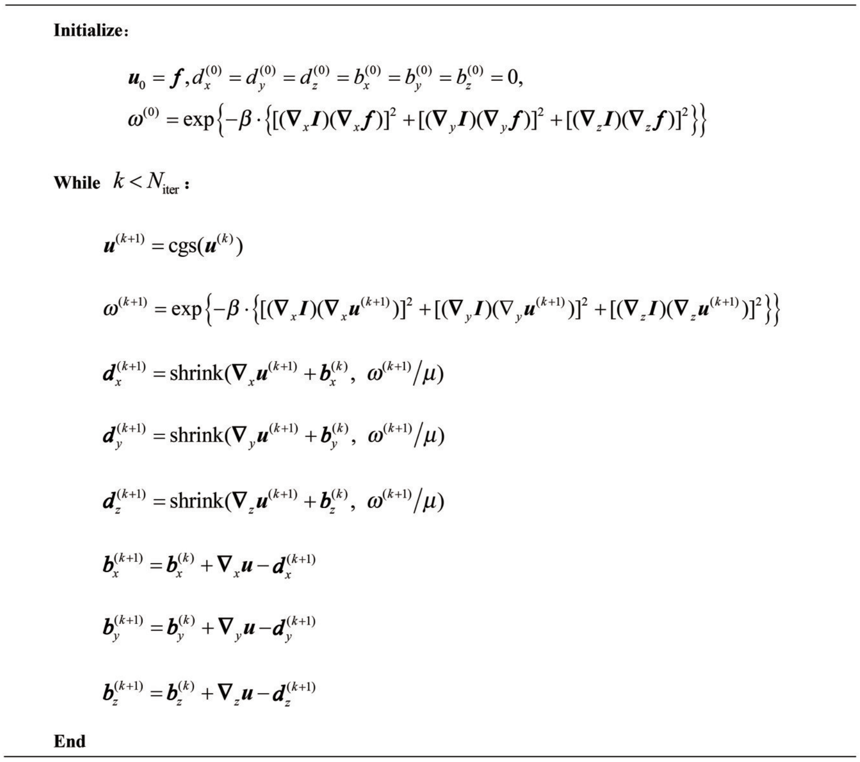
$$\boldsymbol{u}^{(k+1)}=\min_{\boldsymbol{u}}\frac{\lambda}{2}\|\boldsymbol{u}-\boldsymbol{f}\|_2^2+\frac{\boldsymbol{u}}{2\omega}(g_1+g_2+g_3),\tag{5}$$

对于这个子问题，使用共轭梯度法来解决，所以使用Split Bregman方法求解能量泛函极小值问题的步骤如图4所示，其中$\mathrm{shrink}(\alpha,\theta)=\begin{cases}\alpha-\theta, & \alpha>\theta\\ 0, & -\theta\leqslant\alpha\leqslant\theta\\ \alpha+\theta, & \alpha<-\theta\end{cases}$，cgs(·)为共轭梯度法。

4 实验分析与讨论

4.1 实验设置

实验使用了模拟图像序列数据集[38]和真实图像序列数据集[39-40]。这些图像序列在表1中进行了描述。



```
Initialize:
    $\boldsymbol{u}_0=\boldsymbol{f},\boldsymbol{d}_x^{(0)}=\boldsymbol{d}_y^{(0)}=\boldsymbol{d}_z^{(0)}=\boldsymbol{b}_x^{(0)}=\boldsymbol{b}_y^{(0)}=\boldsymbol{b}_z^{(0)}=0,$
    $\omega^{(0)}=\exp\{-\beta\cdot\{[(\nabla_x\boldsymbol{I})(\nabla_x\boldsymbol{f})]^2+[(\nabla_y\boldsymbol{I})(\nabla_y\boldsymbol{f})]^2+[(\nabla_z\boldsymbol{I})(\nabla_z\boldsymbol{f})]^2\}\}$
While $k<N_{\mathrm{iter}}$:
    $\boldsymbol{u}^{(k+1)}=\mathrm{cgs}(\boldsymbol{u}^{(k)})$
    $\omega^{(k+1)}=\exp\{-\beta\cdot\{[(\nabla_x\boldsymbol{I})(\nabla_x\boldsymbol{u}^{(k+1)})]^2+[(\nabla_y\boldsymbol{I})(\nabla_y\boldsymbol{u}^{(k+1)})]^2+[(\nabla_z\boldsymbol{I})(\nabla_z\boldsymbol{u}^{(k+1)})]^2\}\}$
    $\boldsymbol{d}_x^{(k+1)}=\mathrm{shrink}(\nabla_x\boldsymbol{u}^{(k+1)}+\boldsymbol{b}_x^{(k)},\ \omega^{(k+1)}/\mu)$
    $\boldsymbol{d}_y^{(k+1)}=\mathrm{shrink}(\nabla_y\boldsymbol{u}^{(k+1)}+\boldsymbol{b}_y^{(k)},\ \omega^{(k+1)}/\mu)$
    $\boldsymbol{d}_z^{(k+1)}=\mathrm{shrink}(\nabla_z\boldsymbol{u}^{(k+1)}+\boldsymbol{b}_z^{(k)},\ \omega^{(k+1)}/\mu)$
    $\boldsymbol{b}_x^{(k+1)}=\boldsymbol{b}_x^{(k)}+\nabla_x\boldsymbol{u}-\boldsymbol{d}_x^{(k+1)}$
    $\boldsymbol{b}_y^{(k+1)}=\boldsymbol{b}_y^{(k)}+\nabla_y\boldsymbol{u}-\boldsymbol{d}_y^{(k+1)}$
    $\boldsymbol{b}_z^{(k+1)}=\boldsymbol{b}_z^{(k)}+\nabla_z\boldsymbol{u}-\boldsymbol{d}_z^{(k+1)}$
End
```

图4 Split Bregman方法流程

Fig. 4 Flow chart of Split Bregman method

模拟图像序列来自HCI 4D光场数据，对于每一个图像序列场景，使用工具箱[41]生成了28张不同深度层的图像。由于模拟图像序列存在真实深度图（GT），因此可以对估计深度图与真实深度图进行定量比较，计算了多种类型的定量指标，包括均方根误差（RMSE）[42]、相关系数（Corr）[43]和结构相似性（SSIM）[44]。

RMSE是预测值与真实值偏差的定量指标，表达式为

$$Q_{\mathrm{RMSE}}=\sqrt{\frac{1}{N}\sum_{x}\left[D(x)-d(x)\right]^{2}}，\tag{6}$$

式中：D和d分别表示GT和估计的深度图；N表示图中的像素总数。

Corr是用来反映变量之间相关密切程度的统计指标，表达式为

$$Q_{\mathrm{Corr}}=\frac{\sum_{x}\left[D(x)-\overline{D}\right]\left[d(x)-\overline{d}\right]}{\sqrt{\sum_{x}\left[D(x)-\overline{D}\right]^{2}}\sqrt{\sum_{x}\left[d(x)-\overline{d}\right]^{2}}}，\tag{7}$$

式中：$\overline{D}$和$\overline{d}$分别表示GT和估计的深度图均值。

SSIM是一种衡量两幅图像结构相似度的指标。从图像组成的角度将结构信息定义为亮度、对比度和结构三个不同因素的组合。用均值作为亮度的估计，标准差作为对比度的估计，协方差作为结构相似程度的度量。

$$Q_{\mathrm{SSIM}}=\frac{(2\overline{D}\,\overline{d}+c_{1})(2\sigma_{Dd}+c_{2})}{(\overline{D}^{2}+\overline{d}^{2}+c_{1})(\sigma_{D}^{2}+\sigma_{d}^{2}+c_{2})}，\tag{8}$$

式中：σ_{D}^{2}和σ_{d}^{2}分别表示GT和估计的深度图的方差；σ_{Dd}为它们的协方差；$c_{1}=(k_{1}L)^{2}$，$c_{2}=(k_{2}L)^{2}$，$k_{1}=0.01$，$k_{2}=0.03$，L是像素值的动态范围。

4.2　模型参数选取

模型包含4个参数，即迭代次数、保真项系数、一般参数、权重系数，分别用N_{iter}、λ、μ和α表示。这些参数的取值变化会影响聚焦测度优化的质量，这些影响可以从图5所示的深度图中观察到。其中图5（a）是迭代次数N_{iter}分别为3、8、13的实验结果，可以看出较小

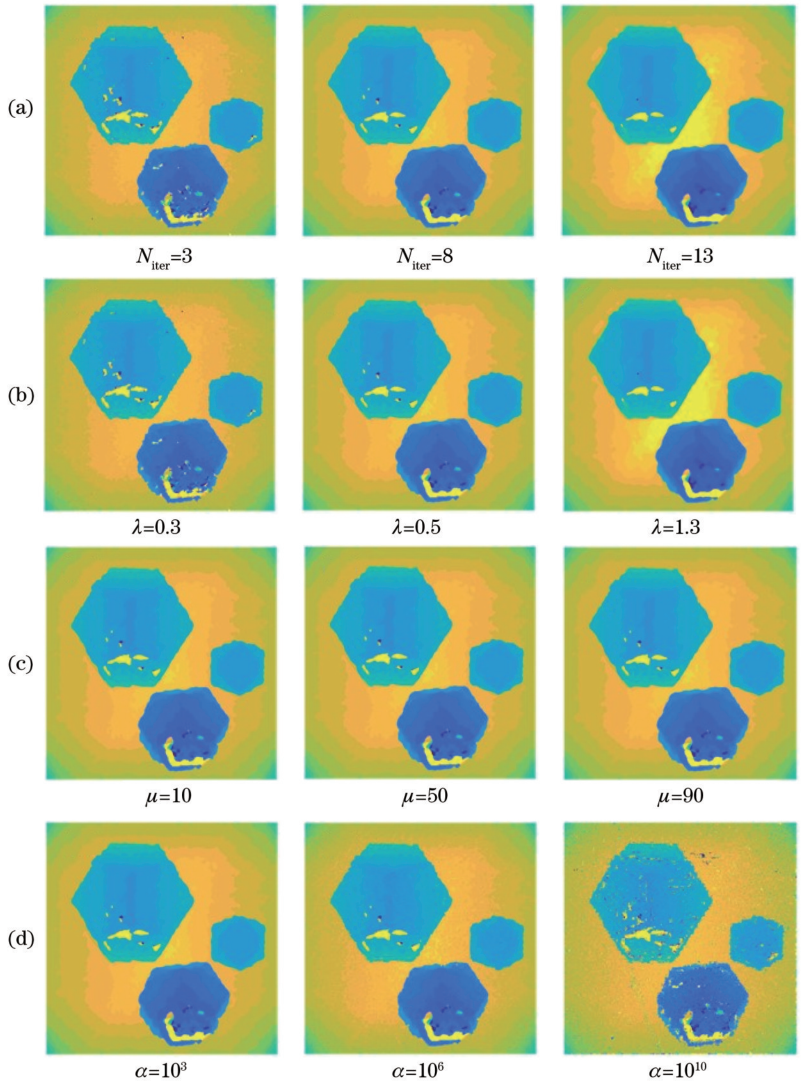


图5　参数N_{iter}、λ、μ和α的影响，默认值设定为$\mu=50$、$\lambda=0.8$、$\alpha=10^{4}$、$N_{\mathrm{iter}}=8$

Fig. 5　Effect of parameters N_{iter}, λ, μ, and α, the default values have been set to $\mu=50$, $\lambda=0.8$, $\alpha=10^{4}$, $N_{\mathrm{iter}}=8$

表1 实验数据描述

Table 1 Description of experimental dataset

Dataset	Scene	Number of layers	Dimension
Synthetic	HCI	28	512 × 512
Real	Motion	14	518 × 518
	Balls	25	320 × 320
	Kitchen	12	518 × 518
	Keyboard	32	360 × 360
	Home1-4	49	320 × 320
	Home2-3	49	320 × 320
	Microkitchen1_44	49	320 × 320
	Pillowroom1_8	49	320 × 320

的迭代次数使得图像不够平滑，而较大的迭代次数使得场景的背景无法保留该有的渐变效果。图5(b)是保真项系数λ分别为0.3、0.5、1.3的实验结果，该系数越小，正则化项起到的作用越大，图片越平滑。图5(c)是参数μ分别为10、50、90的实验结果，该系数为求解过程中引入的一个中间系数，对优化的结果不产生显著的影响。图5(d)是权重系数α分别为10^3、10^6、10^{10}的实验结果，较小的权重系数不足以调节聚焦测度中过大的差异，因此在降低深度图的噪声焦点测量值(异常值)过程中观察到很少的改善。然而由于结构一致并不是完全正确的，较大的权重系数值会使优化过程中引入异常值，选择合适的量级进行优化是十分重要的。

4.3 一致性结构对深度优化的性能分析

通过图6可以观察到，单纯利用聚焦堆栈的纹理信息对聚焦测度进行引导时会产生引导错误的问题。聚焦堆栈所包含的颜色信息和空间信息与聚焦测度体数据所包含的聚焦程度信息是不同属性的信息，因此使用定义的相互结构作为正则化权重，可以有效地对二者共存的一致结构进行保留，对二者存在的不一致结构进行平滑，从而使得聚焦测度体数据的优化过程更加准确。如图6第2行所示，在下方点位置的弱纹理区域，仅使用颜色信息引导的结果没有得到明显改善，

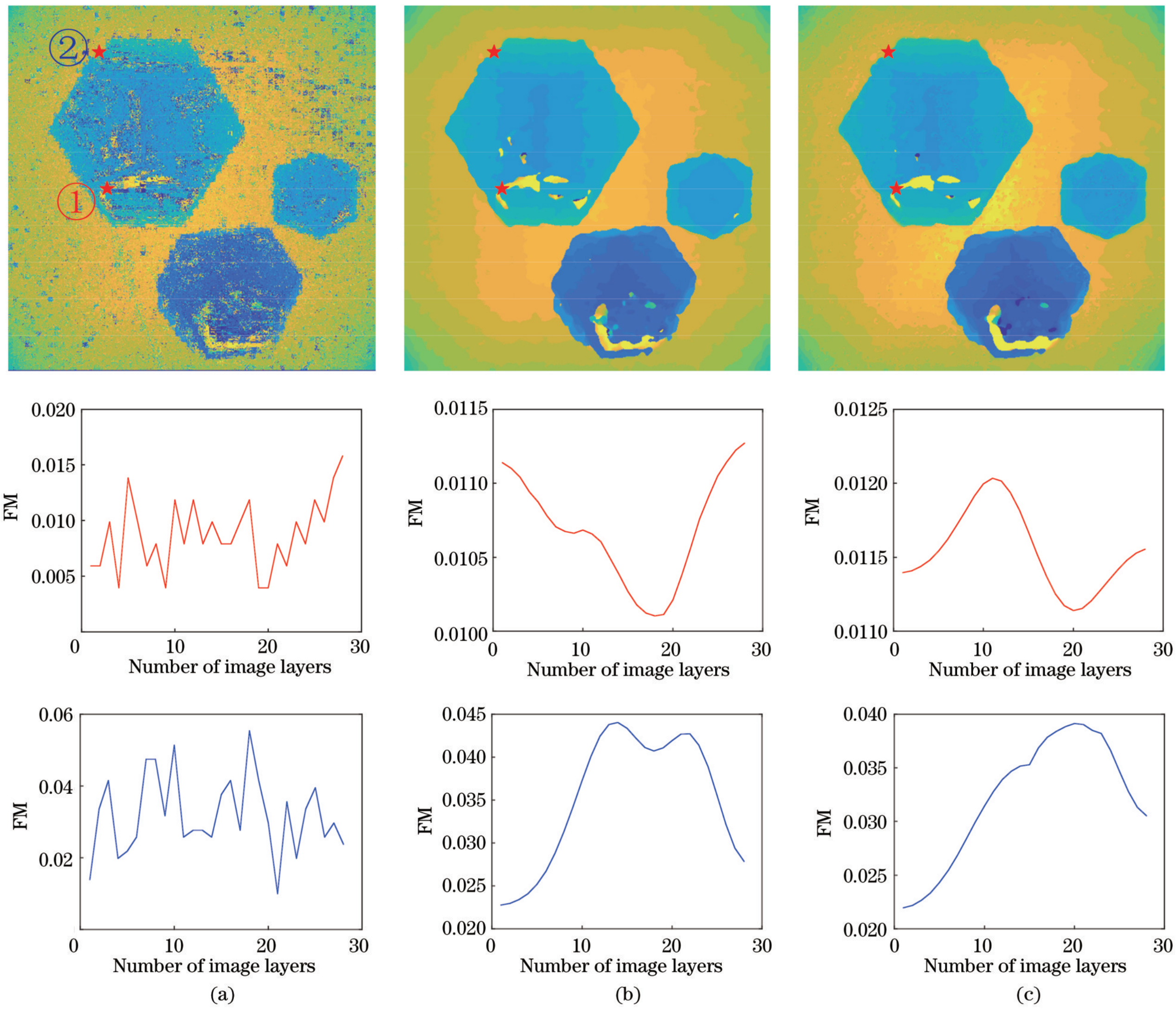


图6 优化效果对比。(a)初始深度；(b)robust focus volume(RFV)优化；(c)所提优化方法

Fig. 6 Comparison of optimization result. (a) Initial depth; (b) RFV optimization; (c) proposed optimization method

而使用三维的结构一致性不仅引入了第三个维度信息，还使二者共存的一致结构得到保留。同样对于上方点位置（边缘区域），这种优化不准确的边缘问题也有进一步的改善。

4.4　不同方法对比实验

图 7 展示了 HCI 数据集[38]中 7 个模拟场景的实验结果。所提方法与 2021 年发表的 RFV 和 2019 年发表的 MuGIF 进行对比。MuGIF 是一种全聚焦图与深度图通过一致结构相互引导的滤波方法，所提方法是利用三维聚焦堆栈与聚焦测度的一致结构引导聚焦测度的滤波方法。MuGIF 虽然在一定程度上对初始深度图起到了平滑并保留边缘的效果，但对二维图像进行后处理时所能引入的信息量较少，在引导过程中无法准确判断哪些位置需要保留哪些位置进行平滑。RFV 在对三维的聚焦测度进行优化的过程提供了更多的信息，但单纯使用聚焦堆栈进行引导的过程会引入部分纹理信息，即没有深度突变的位置也被保留了下来。

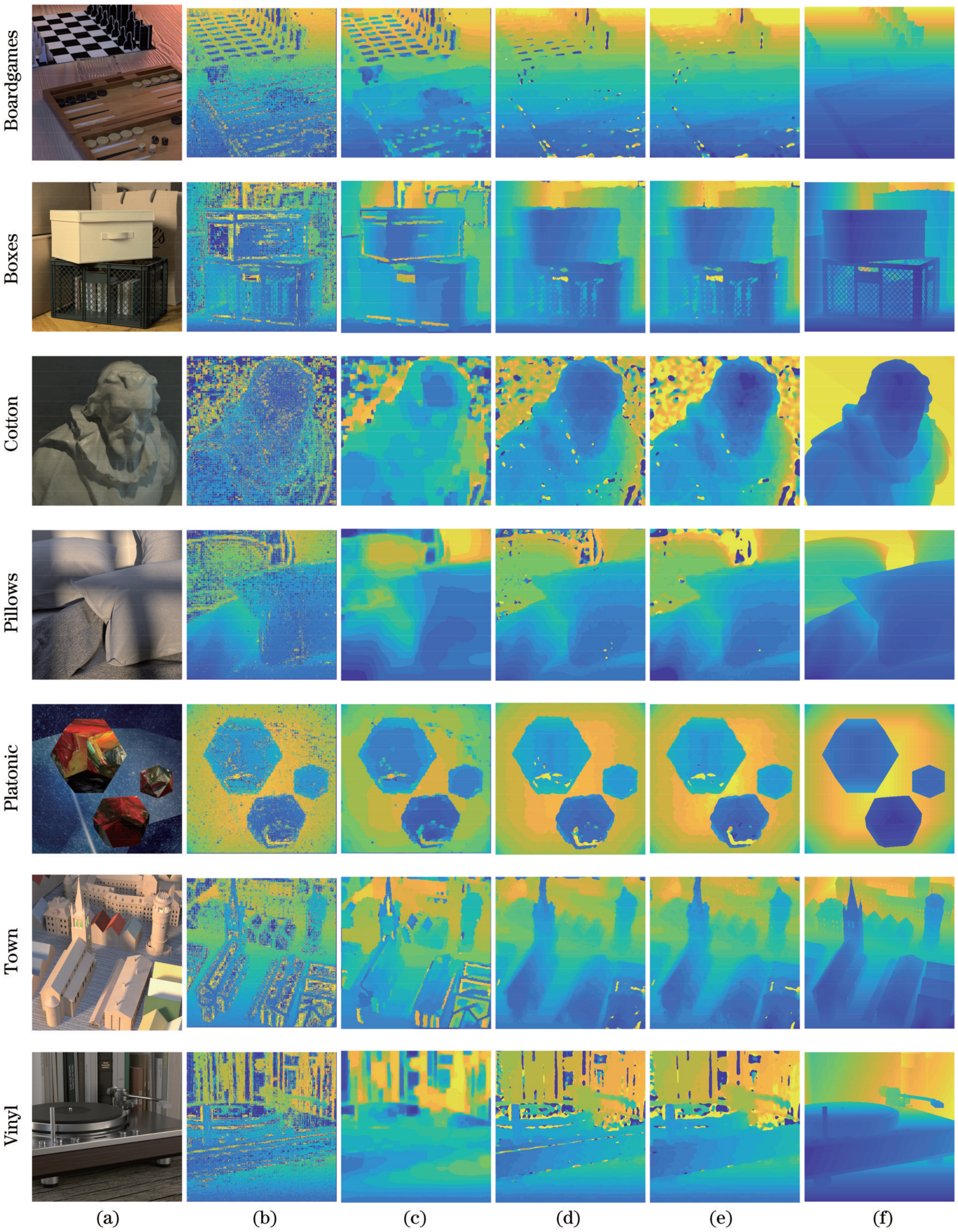


图 7　HCI 数据集部分场景的实验结果。(a)场景；(b)初始深度；(c) MuGIF；(d) RFV；(e) Ours；(f) GT

Fig. 7　Experimental results on part scenes of HCI dataset. (a) Scene; (b) initial depth; (c) MuGIF; (d) RFV; (e) Ours; (f) GT

根据图 7 展示的结果，可以发现在 Boardgames 场景中，棋盘格区域的水平方向并没有深度变化，但是存在纹理突变，而竖直方向深度是渐变的，但是被重建成了深度跳跃的巨变。由于聚焦堆栈引导优化的过程将这些突变判定为边缘保留了下来，优化后的结果存在错误信息。利用三维的一致结构对聚焦测度进行引导的方法可以有效地处理这些单纯的纹理突变而非深度突变引起的错误估计问题，并提供了与真实深度值最接近的深度。在 Platonic 场景中，可以看出所提方法不仅更准确地保留了边缘，还能更好地填充空洞并保留背景渐变。使用聚焦堆栈与聚焦测度的结构一致性，不仅提供了多一个维度的信息，还可以有效避免纹理带来的引导错误的问题。综上所述，与 MuGIF 相比，所提方法可以引入更多的信息量，并且与 RFV 相比，引入的信息更为准确。

实验结果的定量指标包括 RMSE、Corr 和 SSIM，这些数据在表 2 中呈现。从表 2 可以看出，所提方法在大多数场景下都获得了良好的指标。另外，从图 7 深度重建结果和表 2 定量结果可以看出，相比其他方法，所提方法在聚焦测度优化方面表现更加优异。

表 2　HCI 数据集部分场景的定量指标

Table 2　Quantitative index on part scenes of HCI dataset

Scene	RMSE			Corr			SSIM		
	MuGIF	RFV	Ours	MuGIF	RFV	Ours	MuGIF	RFV	Ours
Antinous	0.9981	0.7779	**0.7628**	0.3507	0.5174	**0.5253**	0.2640	0.2881	**0.3640**
Boardgames	0.3036	**0.2306**	0.2575	0.7852	0.8902	**0.9327**	0.6183	**0.6820**	0.6183
Boxes	**0.0943**	0.1577	0.1737	**0.6640**	0.4302	0.3627	0.5746	0.7017	**0.7034**
Cotton	0.6051	0.3849	**0.3672**	0.3421	0.7213	**0.7907**	0.7051	0.5301	**0.6057**
Dino	0.2499	0.2010	**0.2002**	0.8628	0.8857	**0.9274**	0.6749	0.6014	**0.6783**
Dishes	2.3292	2.0300	**1.8100**	0.6836	0.7036	**0.7687**	**0.7126**	0.4253	0.4626
Greek	0.8296	0.7202	**0.7166**	0.4506	0.5329	**0.5428**	**0.3882**	0.2796	0.2761
Medieval2	0.6085	**0.5880**	0.6286	0.8896	0.8789	**0.8917**	0.6033	0.5628	**0.6709**
Museum	0.5417	0.4584	**0.4057**	0.7386	0.7890	**0.8432**	0.5827	0.5265	**0.5922**
Pillows	0.2563	0.1893	**0.1744**	0.6718	0.7819	**0.8571**	0.8314	0.8049	**0.8375**
Platonic	1.2301	1.2540	**1.2267**	0.8754	0.8752	**0.8766**	0.4702	0.4709	**0.4375**
Sideboard	0.9688	**0.6437**	0.6744	0.7906	0.8926	**0.8936**	0.3920	0.4388	**0.4760**
Table	0.3741	0.2569	**0.2521**	0.5684	0.7714	**0.8176**	0.5599	0.5794	**0.6046**
Tomb	3.5769	1.0681	**1.0512**	0.9805	0.9922	**0.9931**	0.3973	0.3652	**0.3979**
Town	2.7154	1.5993	**1.4801**	0.4825	0.9001	**0.9220**	0.2512	0.2395	**0.2755**
Vinyl	0.1993	0.1828	**0.1800**	0.5931	0.6539	**0.7203**	0.7517	0.7072	**0.7517**

对真实图像序列[39-40]也应用了不同的深度图优化方法进行对比，并在图 8 中呈现了实验结果。观察图 8 可以发现：尽管 MuGIF 方法具有很好的平滑效果，但在边缘位置的精度相对较低，例如，在厨房场景中，后侧瓶子的边缘未能被准确检测出来，导致其与背景无法清晰分离；相比之下，所提方法和 RFV 方法在某些场景中没有使用后处理，效果也很平滑，相比初始深度图，所提方法仍然能够较好地保留深度跳跃边缘，使背景与物体的边缘清晰可见，并且弱纹理区域相对平滑。虽然 MuGIF 和 RFV 方法都取得了明显的提升，但少量信息引导或直接颜色信息引导会导致纹理信息在深度图中被保留，从而产生深度没有突变但颜色突变的错误估计。例如，在 pillowroom1_8 场景中，MuGIF 和 RFV 保留了抱枕的纹理，相比之下，所提方法利用三维的一致结构对聚焦测度进行引导，避免了纹理突变引起的错误估计问题，得到了与真实深度值最接近的深度图。从真实数据的实验结果可以看出，所提方法作用于有深度变化的大物体场景时效果会更好，目前还没有达到很好的泛化效果。接下来的工作将继续改进优化模型，使所提方法在不同场景都可以取得较好的重建结果。

5　结　　论

聚焦堆栈包含场景的物理颜色信息，聚焦测度则包含场景的纹理和几何结构信息。根据两种数据在场景几何结构方面的关联关系，提出结构一致性的概念，以更好地定位场景深度边界。结构一致性加权的三维全变分模型有效提高了模型的保边界能力，同时避免了场景颜色信息被错误地引入深度图的问题。所提方法能够有效解决聚焦堆栈估计深度中弱纹理区域和遮挡区域深度线索丢失问题。基于 TV 正则化的优化模型的数值求解方法较为成熟，但选取正则化参数时需尽量避免斑块效应，后续工作将深入研究正则项函数的设计，并挖掘数据端与深度相关的更多有效信息。

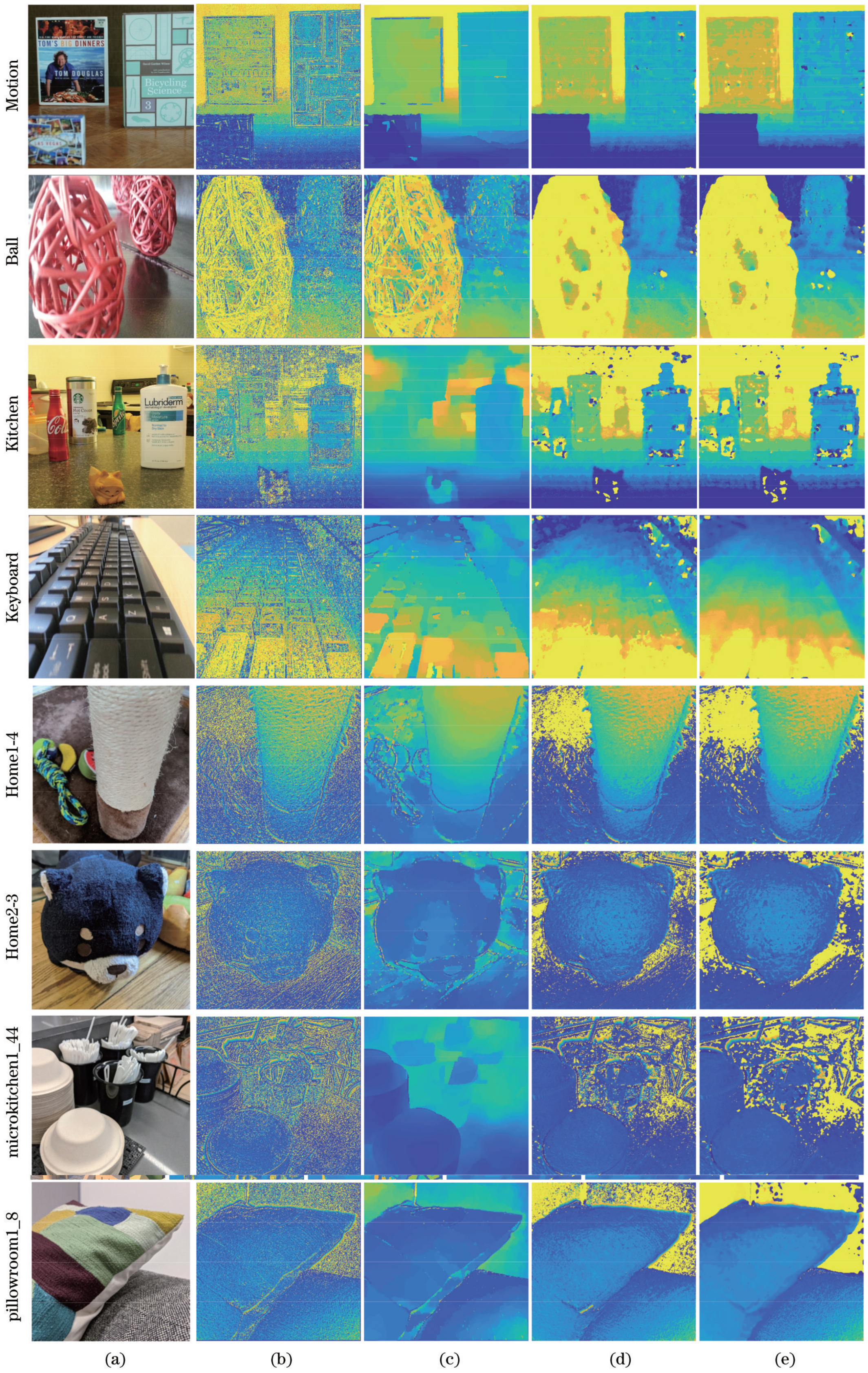


图 8　真实数据集的实验结果。(a)场景;(b)初始深度;(c) MuGIF;(d) RFV;(e) Ours

Fig. 8　Experimental results on real dataset. (a) Scene; (b) initial depth; (c) MuGIF; (d) RFV; (e) Ours

参 考 文 献

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

[2] Yan T, Hu Z G, Qian Y H, et al. 3D shape reconstruction from multifocus image fusion using a multidirectional modified Laplacian operator[J]. Pattern Recognition, 2020, 98: 107065.

[3] Alicona. Optical measurement solutions in use[EB/OL]. [2023-02-01]. https://www.alicona.com/applications/.

[4] Surh J, Jeon H G, Park Y, et al. Noise robust depth from focus using a ring difference filter[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2444-2453.

[5] 沙浩, 刘越, 王涌天, 等. 基于二维图像和三维几何约束神经网络的单目室内深度估计方法[J]. 光学学报, 2022, 42(19): 1911001.
Sha H, Liu Y, Wang Y T, et al. An indoor depth estimation method based on two-dimensional image and three-dimensional geometric constraint neural network[J]. Acta Optica Sinica, 2022, 42(19): 1911001.

[6] 何泽阳, 邓慧萍, 向森, 等. 融合一致性与差异性约束的光场深度估计[J]. 红外与激光工程, 2021, 50(11): 20210021.
He Z Y, Deng H P, Xiang S, et al. Light field depth estimation of fusing consistency and difference constraints[J]. Infrared and Laser Engineering, 2021, 50(11): 20210021.

[7] 李靖怡, 侯国家, 张孝嘉, 等. 基于场景深度估计和背景分割的水下图像复原[J]. 激光与光电子学进展, 2023, 60(2): 0210010.
Li J Y, Hou G J, Zhang X J, et al. Underwater image restoration based on scene depth estimation and background segmentation[J]. Laser & Optoelectronics Progress, 2023, 60(2): 0210010.

[8] 殷永凯, 于锴, 于春展, 等. 几何光场三维成像综述[J]. 中国激光, 2021, 48(12): 1209001.
Yin Y K, Yu K, Yu C Z, et al. 3D imaging using geometric light field: a review[J]. Chinese Journal of Lasers, 2021, 48(12): 1209001.

[9] Salokhiddinov S, Lee S. Deep spatial-focal network for depth from focus[J]. Journal of Imaging Science and Technology, 2021, 65(4): 040501.

[10] Nayar S K, Nakagawa Y. Shape from focus[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994, 16(8): 824-831.

[11] Huang W, Jing Z L. Evaluation of focus measures in multi-focus image fusion[J]. Pattern Recognition Letters, 2007, 28(4): 493-500.

[12] Fu B Y, He R Z, Yuan Y L, et al. Shape from focus using gradient of focus measure curve[J]. Optics and Lasers in Engineering, 2023, 160: 107320.

[13] Jeon H G, Surh J, Im S, et al. Ring difference filter for fast and noise robust depth from focus[J]. IEEE Transactions on Image Processing, 2019, 29: 1045-1060.

[14] Yang G, Nelson B J. Wavelet-based autofocusing and unsupervised segmentation of microscopic images[C]//Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 27-31, 2003, Las Vegas, NV, USA. New York: IEEE Press, 2004: 2143-2148.

[15] Ali U, Mahmood M T. Combining depth maps through 3D weighted least squares in shape from focus[C]//2019 International Conference on Electronics, Information, and Communication (ICEIC), January 22-25, 2019, Auckland, New Zealand. New York: IEEE Press, 2019.

[16] Lee S Y, Yoo J T, Kumar Y, et al. Reduced energy-ratio measure for robust autofocusing in digital camera[J]. IEEE Signal Processing Letters, 2009, 16(2): 133-136.

[17] Malik A S, Choi T S. A novel algorithm for estimation of depth map using image focus for 3D shape recovery in the presence of noise[J]. Pattern Recognition, 2008, 41(7): 2200-2225.

[18] Ahmed Z, Shahzad A, Ali U. Enhancement of depth map through weighted combination of guided image filters in shape-from-focus[C]//2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2), May 24-26, 2022, Rawalpindi, Pakistan. New York: IEEE Press, 2022.

[19] Ceruso S, Bonaque-González S, Oliva-García R, et al. Relative multiscale deep depth from focus[J]. Signal Processing: Image Communication, 2021, 99: 116417.

[20] Moeller M, Benning M, Schönlieb C, et al. Variational depth from focus reconstruction[J]. IEEE Transactions on Image Processing, 2015, 24(12): 5369-5378.

[21] Ali U, Mahmood M T. Energy minimization for image focus volume in shape from focus[J]. Pattern Recognition, 2022, 126: 108559.

[22] Ali U, Mahmood M T. Robust focus volume regularization in shape from focus[J]. IEEE Transactions on Image Processing, 2021, 30: 7215-7227.

[23] Yang F T, Huang X L, Zhou Z H. Deep depth from focus with differential focus volume[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 12632-12641.

[24] Ali U, Lee I H, Mahmood M T. Guided image filtering in shape-from-focus: a comparative analysis[J]. Pattern Recognition, 2021, 111: 107670.

[25] He K M, Sun J, Tang X O. Guided image filtering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(6): 1397-1409.

[26] Liu W, Chen X G, Shen C H, et al. Robust guided image filtering[EB/OL]. (2017-03-28)[2023-02-01]. https://arxiv.org/abs/1703.09379.

[27] Guo X J, Li Y, Ma J Y. Mutually guided image filtering[C]//Proceedings of the 25th ACM International Conference on Multimedia, October 23-27, 2017, Mountain View, California, USA. New York: ACM Press, 2017: 1283-1290.

[28] Rudin L I, Osher S, Fatemi E. Nonlinear total variation based noise removal algorithms[J]. Physica D: Nonlinear Phenomena, 1992, 60(1/2/3/4): 259-268.

[29] Zamir S W, Arora A, Khan S, et al. Multi-stage progressive image restoration[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 14816-14826.

[30] Goldstein T, Osher S. The split Bregman method for L1-regularized problems[J]. SIAM Journal on Imaging Sciences, 2009, 2(2): 323-343.

[31] Bregman L M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming[J]. USSR Computational Mathematics and Mathematical Physics, 1967, 7(3): 200-217.

[32] Cai J F, Osher S, Shen Z W. Linearized Bregman iterations for compressed sensing[J]. Mathematics of Computation, 2009, 78(267): 1515-1536.

[33] Candes E J, Romberg J K. Signal recovery from random projections[J]. Proceedings of SPIE, 2005, 5674: 76-86.

[34] Goldfarb D, Yin W T. Parametric maximum flow algorithms for fast total variation minimization[J]. SIAM Journal on Scientific Computing, 2009, 31(5): 3712-3743.

[35] Hale E T, Yin W, Zhang Y. A fixed-point continuation method for l1-regularized minimization with applications to compressed sensing[EB/OL]. [2023-02-01]. https://www.cmor-faculty.rice.edu/~yzhang/reports/tr0707.pdf.

[36] Lustig M, Lee J H, Donoho D L, et al. Faster imaging with randomly perturbed, under-sampled spirals and L_1 reconstruction[M]//Cauley S F, Abuhashem A O, Bilgic B, et al. Proceedings of ISMRM 2013, 2013: 685.

[37] Li Y Y, Santosa F. An affine scaling algorithm for minimizing total variation in image enhancement[R]. Ithaca: Cornell University, 1994.

[38] Honauer K, Johannsen O, Kondermann D, et al. A dataset and evaluation methodology for depth estimation on 4D light fields [M]//Lai S H, Lepetit V, Nishino K, et al. Computer vision -

ACCV 2016. Lecture notes in computer science. Cham: Springer, 2017, 10113: 19-34.

[39] Suwajanakorn S, Hernandez C, Seitz S M. Depth from focus with your mobile phone[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 3497-3506.

[40] Herrmann C, Bowen R S, Wadhwa N, et al. Learning to autofocus[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 2227-2236.

[41] Dansereau D. Light Field Toolbox v0.4. [EB/OL]. [2023-02-01]. https://www.mathworks.com/matlabcentral/fileexchange/49683-light-field-%toolbox-v0-4/.

[42] Chai T, Draxler R. Root mean square error (RMSE) or mean absolute error (MAE)[J]. Geoscientific Model Development Discussions, 2014, 7(1): 1525-1534.

[43] Meng X L, Rosenthal R, Rubin D B. Comparing correlated correlation coefficients[J]. Psychological Bulletin, 1992, 111(1): 172-175.

[44] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

Depth Estimation Based on Spatial Geometry in Focal Stacks

Luo Tianqi, Deng Xiaojuan, Liu Chang, Qiu Jun*

Institute of Applied Mathematics, Beijing Information Science and Technology University, Beijing 100101, *China*

Abstract

Objective Depth estimation is an important research topic in the field of computer vision, which is used to perceive and reconstruct three-dimensional (3D) scenes using two-dimensional (2D) images. Estimating depth based on a focal stack is a passive method that uses the degree of focus as a depth clue. This method has advantages including small imaging equipment size and low computational cost. However, this method relies heavily on the measurement of image focus, which is considerably affected by the texture information related to a scene. Measuring the degree of focus accurately in regions with poor lighting, smooth textures, or occlusions is difficult, leading to inaccurate depth estimation in these areas. Previous studies have proposed various optimization algorithms to increase the accuracy of depth estimation. These algorithms can generally be classified into three categories: designing satisfactory focus-measurement operators, optimizing the focus-measurement volume data to correct errors, and using all-in-focus images for guided filtering of the initial depth map. However, numerous factors, including scene texture, contrast, illumination, and window size, can affect the performance of focus-measurement operators, resulting in erroneous estimates in initial focus-measure volume data, resulting in inaccurate depth estimation. Effectiveness of the methods that optimize an initial depth map heavily depends on the accuracy of the initial depth map. Because the initial depth values may be estimated incorrectly owing to insufficient illumination, introducing considerably valid information to improve depth estimation through postprocessing is difficult. Therefore, intermediate optimization methods are ideal for improving the accuracy of a depth map. To solve the problem of inaccurate depth clues in regions showing weak texture and occlusion, this study proposes a novel method based on 3D adaptive-weighted total variation (TV) to optimize focus-measure volume data.

Methods The proposed method consists of two key parts: 1) defining a structure consistency operator based on the prior geometric information related to different dimensions between the focal stack and focus-measure volume data, which is used to locate the depth boundary and area with high reliable depth clues to increase the accuracy of depth optimization; 2) incorporating the prior geometric information related to the scene hidden in the 3D focal stack and focus-measure volume data into the 3D TV regularization model. The structure of the image is measured using pixel-gradient values. Gradient jumps in the focal stack reflect changes in physical structure, while those in the focus-measure volume data reflect changes in focus level. When the physical structure and focus level exhibit considerable variations at the same position, the structure is consistent and corresponds to an area with reliable depth change. By measuring the structural consistency between the focal stack and focus measure, we can determine the positions exhibiting reliable depth clues and guide the optimization process related to the focus-measure data highly accurately. The traditional 2D TV optimization model has some edge-preserving ability while performing denoising. However, when the noise-gradient amplitude exceeds the edge-gradient amplitude, this model faces a dilemma between balancing denoising and preserving edge details. Based on the guided filtering method, the edge information of a reference image is used to denoise the target image, effectively resolving the dilemma. This leads to a weighted TV optimization model; however, when applying guided filtering to 2D images, the optimization information that can be introduced is limited. Therefore, we attempt to extend this method to a 3D image field. A weighted 3D TV regularization model can balance denoising and edge-preserving abilities high effectively owing to the rich information in 3D data. Herein, the process of optimizing the focus-measure data is modeled as a 3D weighted TV regularization method, and the adaptive weight is determined based on structural consistency.

Results and Discussions First, an analysis is conducted on the selection of model parameters. We observe that adjusting these parameters can considerably impact the performance of the proposed algorithm, thereby optimizing the accuracy of depth estimation. Second, herein, a detailed analysis is conducted on the impact of structural consistency during the optimization process and the problems that may arise because of focusing solely on texture information for optimization. A comparative analysis is also performed with the introduction of 3D structural consistency. Finally, the proposed algorithm is tested on simulated and real image sequence datasets and the results are compared with those from two other methods: mutually guided image filtering (MuGIF) and robust focus volume (RFV). The proposed method computes 3D structural consistency, which is an additional dimension of information, as opposed to the MuGIF method, which uses consistent structural guidance filtering on inputs from all-in-focus and depth maps. The RFV method uses focal stacks to guide focus measure for optimizing depth estimation in 3D. Compared with the RFV method, the proposed method considers the property issues related to focal stack and focus measure and uses their consistent structure to guide optimization. Furthermore, three evaluation metrics are used to analyze and validate the three algorithms with respect to simulated data. The experimental results demonstrate that the proposed method exhibits better performance than the other methods, providing more accurate information for correcting the focusing measure process through 3D structural consistency. The proposed method not only preserves edge information but also preserves texture information with high accuracy and reduces errors in depth estimation.

Conclusions Focal stack contains physical color information of a scene, while focus measure contains textural and geometric structure information of the scene. In this study, we propose a method for measuring the structural consistency between the two to effectively locate the depth discontinuities. A structural consistency weighted TV model enhances the ability of the model to preserve edge information while avoiding the introduction of color information into a depth map. Thus, effectively addressing the problem of loss of depth clues related to focal-stack depth estimation in regions with weak texture and occlusion and increasing the accuracy of depth reconstruction. The computation of the L1 model of TV is relatively easy; however, this computation suffers from local distortion. Using highly advanced regularization terms may further improve the reconstruction effect. Future research needs to consider ways of incorporating increased data and investigating methods to improve the regularization term during the optimization process.

Key words image processing; focal stack; focus measure; depth estimation; mutual structure