中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)01-0173-15

论文引用格式:Jiang Z Y, Zhang Z X, Liu J Y and Liu R S. 2025. Multi-modality feature fusion-based wide field-of-view image generation. Journal of Image and Graphics, 30(01):0173-0187(姜智颖, 张曾翕, 刘晋源, 刘日升. 2025. 多模态数据特征融合的广角图像生成. 中国图象图形学报, 30(01):0173-0187)[DOI:10.11834/jig.240056]

多模态数据特征融合的广角图像生成

姜智颖[1],张曾翕[2],刘晋源[3*],刘日升[2]

1. 大连海事大学信息科学技术学院,大连　116026; 2. 大连理工大学软件学院、大连理工大学—立命馆大学国际信息与软件学院,大连　116081; 3. 大连理工大学机械工程学院,大连　116081

摘　要: **目的**　图像拼接通过整合不同视角的可见光数据获得广角合成图。不利的天气因素会使采集到的可见光数据退化,导致拼接效果不佳。红外传感器通过热辐射成像,在不利的条件下也能突出目标,克服环境和人为因素的影响。**方法**　考虑到红外传感器和可见光传感器的成像互补性,本文提出了一个基于多模态数据(红外和可见光数据)特征融合的图像拼接算法。首先利用红外数据准确的结构特征和可见光数据丰富的纹理细节由粗到细地进行偏移估计,并通过非参数化的直接线性变换得到变形矩阵。然后将拼接后的红外和可见光数据进行融合,丰富了场景感知信息。**结果**　本文选择包含530对可拼接多模态图像的真实数据集以及包含200对合成数据集作为测试数据,选取了3个最新的融合方法,包括RFN(residual fusion network)、ReCoNet(recurrent correction network)和DATFuse(dual attention transformer),以及7个拼接方法,包括APAP(as projective as possible)、SPW(single-perspective warps)、WPIS(wide parallax image stitching)、SLAS(seam-guided local alignment and stitching)、VFIS(view-free image stitching)、RSFI(reconstructing stitched features to images)和UDIS++(unsupervised deep image stitching)组成的21种融合—拼接策略进行了定性和定量的性能对比。在拼接性能上,本文方法实现了准确的跨视角场景对齐,平均角点误差降低了53%,避免了鬼影的出现;在多模态互补信息整合方面,本文方法能自适应兼顾红外图像的结构信息以及可见光图像的丰富纹理细节,信息熵较DATFuse-UDIS++策略提升了24.6%。**结论**　本文方法在结合了红外和可见光图像成像互补优势的基础上,通过多尺度递归估计实现了更加准确的大视角场景生成;与常规可见光图像拼接相比鲁棒性更强。

关键词: 多模态图像融合;图像拼接;卷积神经网络(CNN);红外—可见光图像;多尺度金字塔

Multi-modality feature fusion-based wide field-of-view image generation

Jiang Zhiying[1], Zhang Zengxi[2], Liu Jinyuan[3*], Liu Risheng[2]

1. College of Information Science and Technology, Dalian Maritime University, Dalian　116026, China; 2. School of Software Technology & DUT-RU International School of Information Science and Engineering, Dalian University of Technology, Dalian　116081, China; 3. School of Mechanical Engineering, Dalian University of Technology, Dalian　116081, China

Abstract: **Objective** Image stitching, a cornerstone in the field of computer vision, is dedicated to assembling a comprehensive field-of-view image by merging visible data captured from multiple vantage points within a specific scene. This

收稿日期:2024-01-29;**修回日期**:2024-04-29;**预印本日期**:2024-05-06

*通信作者:刘晋源　atlantis918@hotmail.com

基金项目:国家重点研发计划资助(2022YFA1004101);国家自然科学基金项目(62302078, U22B2052);中国博士后科学基金项目(2023M730741)

Supported by: National Key R&D Program of China(2022YFA1004101); National Natural Science Foundation of China(62302078, U22B2052); China Postdoctoral Science Foundation(2023M730741)

fusion enhances scene perception and facilitates advanced processing. The current state-of-the-art in image stitching primarily hinges on the detection of feature points within the scene, necessitating their dense and uniform distribution throughout the image. However, these approaches encounter significant challenges in outdoor environments or when applied to military equipment, where adverse weather conditions such as rain, haze, and low light can severely degrade the quality of visible images. This degradation impedes the extraction of feature points, a critical step in the stitching process. Furthermore, factors such as camouflage and occlusion can lead to data loss, disrupting the distribution of feature points and thus compromising the quality of the stitched image. These limitations often manifest as ghosting effects, undermining the effectiveness of the stitching and its robustness in practical applications. In this challenging context, infrared sensors, which detect thermal radiation to image scenes, emerge as a robust alternative. They excel in highlighting targets even under unfavorable conditions, mitigating the impact of environmental and human factors. This capability makes them highly valuable in military surveillance applications. However, a significant drawback of thermal imaging is its inability to capture the rich texture details that are abundant in visible images. These details are crucial for an accurate and comprehensive perception of the scene. **Method** This paper proposes a groundbreaking image stitching algorithm to overcome the limitations inherent in conventional visible image stitching and to extend the applicability of stitching technology across various environments. This algorithm is based on the fusion of features from multi-modality images, specifically, infrared and visible images. By exploiting the complementary characteristics of infrared and visible data, our approach integrates the precise structural features of infrared images with the rich, detailed attributes of visible images. This integration is crucial for achieving accurate homography matrix estimation for scenes viewed from multiple angles. A distinctive aspect of our method is the incorporation of a learnable feature pyramid structure. This structure is instrumental in estimating sparse offsets in a gradual, coarse-to-fine manner, thus deriving the deformation matrix through a non-parametric direct linear transformation. An innovative aspect of our approach is the fusion of stitched infrared and visible data to enrich the perceptual information of the generated scene. This fusion process entails mining deep features of the scene for contextual semantic information while also utilizing shallow features to address the deficiencies in upsampled data. This strategy aims to produce more accurate and reliable fused results. **Result** We selected a real-world dataset comprising 530 pairs of stitchable multi-modal images and a synthetic dataset containing 200 pairs of data as test datasets. It compared the qualitative and quantitative performance of 21 fusion stitching strategies, incorporating three of the latest fusion methods — namely, residual fusion network, recurrent correction network, and dual-attention transformer (DATFuse) — and seven stitching methods — namely, as projective as possible, single-perspective warps, wide parallax image stitching, seam-guided local alignment and stitching, view-free image stitching, reconstructing stitched features to images, and unsupervised deep image stitching (UDIS++). In terms of stitching performance, our method achieved accurate cross-view scene alignment, with an average corner error reduced by 53%, preventing ghosting and abnormal distortion, even outperforming existing feature-point-based stitching algorithms in challenging large-baseline scenarios. With regard to the integration of multi-modal complementary information, our method adaptively balanced the robust imaging capability of infrared images to highlight structural information and the rich texture details of visible images, resulting in an information entropy increase of 24.6% compared with the DATFuse-UDIS++ strategy, demonstrating significant advantages. **Conclusion** The proposed infrared and visible-based image stitching method marks a significant advancement in the field of computer vision. It effectively addresses the limitations of traditional image stitching methods, particularly under adverse environmental conditions. Moreover, it broadens the scope of stitching technology, making it more versatile and applicable in diverse settings. The combination of infrared and visible imagery in this algorithm could revolutionize scene perception and processing, especially in military and outdoor applications where accuracy, detail, and robustness are of utmost importance. Furthermore, the algorithm's ability to fuse different types of data opens up new avenues for research and application. It suggests potential uses in other fields such as environmental monitoring, search-and-rescue operations, and even in artistic and creative domains where novel visual representations are sought. The fusion technique employed in our algorithm not only enhances the visual quality of the stitched images but also adds a layer of information that could be vital in critical applications such as surveillance and reconnaissance. It effectively addresses the key challenges of traditional image stitching, particularly under adverse environmental conditions, and demonstrates superior performance over existing methods. This advancement not only enhances our ability to perceive and pro-

cess scenes more effectively but also paves the way for future innovations in image processing and analysis.

Key words: multi-modality image fusion; image stitching; convolutional neural network (CNN); infrared and visible images; multi-scale pyramid

0 引 言

由于视角范围有限，单幅图像无法显示完整的场景信息。为了解决这个问题，可以通过对不同视角的图像进行拼接来获得更宽视野的合成图像。图像拼接的实际应用场景十分广泛，如无人机航拍（冷佳旭 等，2023；石争浩 等，2023）、自动驾驶（李熙莹等，2023）。图像拼接是进一步进行图像理解的基础步骤，拼接效果的好坏直接影响接下来的工作，所以一个好的图像拼接算法非常重要。

传统图像拼接方法大致可以分为4个步骤：特征点检测、特征点匹配、图像配准以及图像融合。在上述过程中，图像配准是影响拼接性能的关键性步骤。图像配准通过估计一个 3 × 3 的矩阵来表示从目标图像域到参考图像域的形变模型。然而，在实际场景中，不同视角下的目标通常处于不同的深度层面，仅仅使用单一的全局单应性估计结果进行拼接通常会出现重影的问题。为了克服这一问题，已有的算法（Lou 和 Gevers，2014；Gao 等，2011）通过学习位置自适应地变换，例如一幅图像可以被分成多个小区域，每个小区域采用一个变换模型，这样重叠区域能在一定程度上实现更有效的配准结果。另外一类传统的拼接方法是基于拼接缝驱动的（Zhang 和 Liu，2014；Liao 等，2023）。这些方法通过约束拼接缝的误差最小化来去除伪影的影响。

虽然上述方法均在一定程度上克服了重影的问题，但是其极大依赖于特征点的均匀分布以及特征点的质量。然而在户外环境应用时，不利的天气因素时常会干扰高质量数据的获取，例如在雨天、雾天时采集到的可见光数据通常会出现遮挡，远距离的场景信息出现模糊甚至丢失；在夜晚黑暗条件下，由于光照不足，物体的反射光有限，所获取图像大部分区域会呈现模糊甚至是黑色。因此针对恶劣条件下采集数据的特征点匹配十分困难，获取的特征点分布也是分布不均，极大地限制了拼接算法的表现。因此，基于传统的拼接算法在实际应用时鲁棒性较差。

近年来，深度学习在计算机视觉领域各个任务中均有出色的表现，在图像拼接任务中也出现了一些基于深度学习的方法（Shen 等，2019；Nie 等，2021，2023）。这些方法通常通过深层的卷积神经网络来直接实现单应性估计，并且只适用于特定视角下的图像拼接任务，因其要求不同视角图像间基线范围较小，泛化能力有限。

考虑到上述传统图像拼接算法与基于深度学习的拼接算法的不足，本文提出了一个基于多模态数据的拼接方法。首先，红外传感器通过感知环境中的热辐射进行场景成像，不受遮挡，光照等诸多因素的影响，避免了常规可见光成像方式易受干扰的特性。但也正因为红外传感器的这一成像方式，其对目标中的纹理细节感知能力较差。众所周知，知道纹理细节作为特征信息在环境感知中是十分重要的，可见光图像通过正常的光线反射可以获得丰富的纹理信息。因此，针对红外数据和可见光数据的互补性，本文在进行配准时兼顾两种模态数据的优势，采用多尺度特征金字塔结构，由粗到细地进行多模态数据的变形参数估计，并且利用非参数化的直接线性变换实现变形矩阵估计。为了充分利用不同模态数据的特征信息，本文方法在得到了目标图像和参考图像的变形参数后利用重构模块同时实现多视角场景的拼接以及多模态数据的融合。在该重构模块中，通过挖掘场景的深层特征，提供上下文语意信息，并利用浅层特征改善上采样的信息不足，从而实现更加精准可靠的融合拼接结果。

本文的主要贡献如下：1）提出一个基于多模态数据的广角图像生成算法，该方法兼顾红外数据和可见光数据的特征优势，克服了自然条件下不利因素对常规拼接算法的影响，并且在重构过程中结合多模态数据的特征融合，不仅使拼接结果更加可靠而且提升了场景信息量，更有利于场景感知。2）多模态数据通过多尺度特征金字塔回归全局相关损失，非参数化地计算变形矩阵，使得拼接结果更加精准可靠；此外，重构模块中上下文语意感知能有效实现信息补偿。3）在真实数据和合成数据上定性、定量的结果表明，本文算法相比于已有传统的以及基于深度学习的拼接算法拼接效果更加准确，并且鲁

棒性和泛化能力更强。

1 相关工作

1.1 基于传统方法的图像拼接

基于传统拼接方法在本质上易出现重影的问题，许多克服伪影的方法相继提出。Gao等人(2011)提出针对背景和前景目标的双单应性估计方法，该方法在对齐的过程中使用单应性以实现非线性扭曲。一旦图像被几何拼接，它们将被进一步处理减少拼接缝区域的伪影干扰。Lin等人(2011)提出一种平滑变形场，通过重叠区域变形并进行外推的形式减少视差在拼接过程中带来的影响，同时在一定程度上满足了对不同时间域下运动物体的容错要求。Zaragoza等人(2013)研究了当数据不完全满足投影模型的基础假设时的投影估计并提出APAP(as projective as possible)方法。该方法允许局部非投影偏差来解释对不满足成像假设条件的问题，通过一种移动直接线性变换的新型估计技术实现无缝拼接。然而，APAP算法假设在相邻区域的扭曲变化很小。在实际应用中，相邻区域的深度会发生显著变化，这导致在图像边界附近仍会显示视差伪影。Chuang等人(2014)提出一种新的参数扭曲算法，其合理结合了投影变换和相似变换，在投影的基础上针对非重叠区域图像进行合理外推，在满足良好对齐精度的同时，保留原始图像的透视信息。Lin等人(2015)在满足图像局部变换的同时，也考虑到变形场的平滑性，减少拼接图像的弯曲和伪影，改善了整体的拼接效果。Lee和Sim(2020)提出一个变形残差向量来区分不同深度平面的匹配特征。该算法对于视差较大的图像，通过使用相应的估计单应性扭曲不同的平面，可以实现更精确的对齐。

与此同时，另一类传统拼接方法专门针对拼接区域进行研究，近年来也取得了不错的性能。在Gao等人(2013)的工作中，提出了单应性的接缝损失来测量扭曲目标图像和参考图像之间的不连续性并且选择具有最小切缝损失的单应性来实现最佳拼接。该方法不是根据特征对应的最佳拟合来估计几何变换，而是根据接缝切割的视觉质量来评估变换的优劣。实验结果表明这种新的图像拼接策略通常可以产生比现有方法更好的感知结果，尤其是对于具有挑战性的场景。Zhang和Liu(2014)引入了内容保持扭曲(content preserving warp，CPW)来对齐重叠区域以进行小的局部调整，同时使用单应性来保持全局图像结构。具体来讲，该方法采用了一种混合对齐模型，实现了单应性和维持内容不变的扭曲，不仅得到了较为准确的视差估计，而且避免了局部失真的问题；此外，该方法还开发了一种接缝查找方法，通过考虑几何对齐和图像内容，仅从大致对齐的图像中就可以估计出合理的接缝。然后使用得到的单应性预对齐输入图像，并进一步使用内容保留扭曲来局部细化对齐。最终使用标准的接缝切割算法和多波段混合算法将对齐的图像组合在一起。与对齐重叠区域的像素不同，Lin等人(2016)使用估计的接缝来指导优化局部对齐的过程，以便在每次迭代中提高接缝质量。此外，引入了一种新的结构保持变形方法，以在变形过程中保留显著的曲线和线条结构。这些策略极大提高了该方法在处理大视差的各种具有挑战性的图像方面的有效性。Jia等人(2021)探索全局共线结构并将其引入目标函数，以引导平衡图像变形所需的特征，从而可以在减轻失真的同时保留局部和全局结构。Liao等人(2023)使用针对接缝中低质量的像素，将对齐图像中的突出图像块分开，并通过SIFT流提取修改后的密集对应关系来对图像进行局部对齐。

1.2 基于深度学习的图像拼接

基于卷积神经网络(convolutional neural network，CNN)在各计算机视觉任务上的优异性能，越来越多的研究人员尝试将CNN应用于图像拼接。DeTone等人(2016)首次将单应性估计运用于卷积神经网络，即通过估计一个自由度为8的单应性矩阵将不同视角的图像扭曲至同一个空间域。但由于线性变换存在局限性，该算法只能适用于视差较小的小基线场景，在面对图像深度差异较大的情况时，难以实现具有良好感知信息的空间域转换。另外，单应变换无法有效解决接缝处的连续性问题，因此在很多实际场景中拼接效果往往不尽如人意。在Hoang等人(2020)和Shi等人(2020)的算法中，CNN用于提取特征点，而在Lai等人(2019)和Shen等人(2019)提出的方法中，CNN用于拼接具有固定视角的图像。但是这些方法要么只利用网络实现部分目标，要么只能用于拼接固定视角的图像，在实际应用仍然有巨大的局限性。Nie等人(2021)提出了满足任意视角的深度图像拼接方法，填补了深度学习在

端到端图像拼接上的空白。在该任意视角的解决方案中，深度图像拼接可通过深度单应性模块、空间变换模块和深度图像细化模块来完成。通过结构到内容的拼接模块尽可能减少伪影和接缝部分的影响。随后，为了提高应用的泛化能力，Nie 等人(2023)引入局部薄板样条差值变换以实现更灵活的扭曲，并通过迭代策略实现跨分辨率应用中的自适应变形。遗憾的是，在真实环境中获取拼接标签是十分困难的，截至目前还没有用于图像拼接的真实监督数据集。

2　本文方法

基于上述分析，常规可见光图像拼接易受环境因素的影响，例如：雨、雪、雾、沙尘、弱光、强光、遮挡等。而红外传感器通过热辐射成像，不易受环境因素的影响，但因红外传感器只能获取场景中的大尺度结构信息，缺少对场景内细节的特征表示，因此不利于常规算法对场景的感知。而可见光图像由反射光成像，能够很好地呈现细节信息，与红外成像特征形成互补关系。因此为了实现任意场景、任意视角下的广角图像生成，本文提出了一个基于红外图像和可见光图像的多模态数据拼接方法。

具体来讲，首先将在不同视角获取的红外图像对(infrared image，IR)和可见光图像对(visible image，VIS)送入多尺度可学习金字塔网络，由粗到细地估计出特定模态的场景偏移。为了实现两模态数据的特征互补，体现多模态处理的先进性，本文方法对两模态预测出的潜在偏移量进行融合，并获得更加精确的单应性矩阵。其中，红外数据的结构特征可以与可见光数据的细节内容相结合。然后，对变形的图像进行全局上下文引导的图像重建，这样的做法不仅可得到任意不同视角下场景的拼接结果，而且实现了不同模态数据的信息融合，得到了信息量丰富、清晰准确无伪影的广角图像。整体网络结构如图 1 所示。整体的网络结构可以分为 3 个阶段，这 3 个阶段相辅相成，实现了由粗到细地变形参数估计。

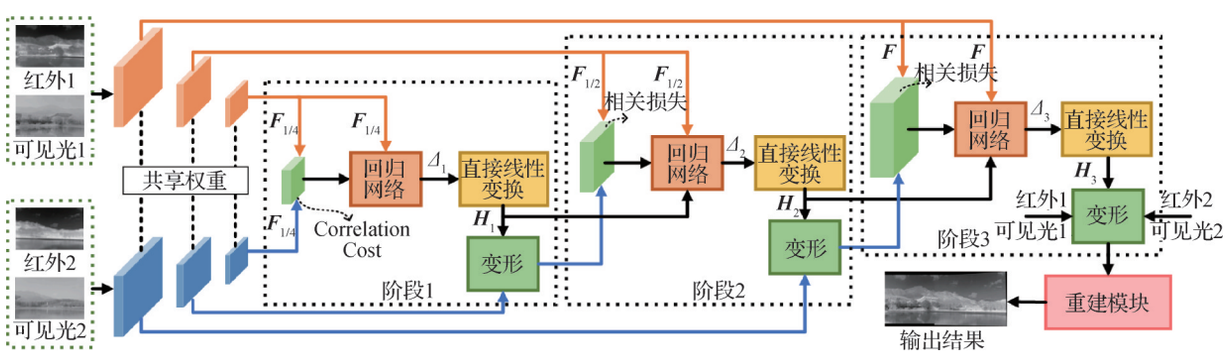


图 1　本文算法的整体流程

Fig. 1　Workflow of the proposed method

2.1　基于多尺度金字塔的单映性估计

将同一模态不同视角下的图像输入网络，通过两层卷积操作进行特征提取。然后对该浅层特征进行最大池化和两个连续卷积级联的操作并连续执行 3 次，分别得到 $\boldsymbol{F}$、$\boldsymbol{F}_{1/2}$ 和 $\boldsymbol{F}_{1/4}$ 这 3 组特征以构建多尺度特征金字塔，其中每次执行操作后获得的特征尺度均是原特征尺度的一半。本方法从大小为 16 × 16 的最小尺度特征 $\boldsymbol{F}_{1/4}$ 开始估计单应性矩阵。首先，同一尺度下的特征通过计算全局相关性，实现特征层每个像素与全局特征的相关性表示。在提取金字塔特征并计算特征相关性之后，采用一个简单的回归网络，该网络由 3 个卷积层和两个全连接层组成，来预测 8 个坐标偏移参数 Δ_1，基于这些偏移参数可以通过一个非参数化的方式确定该变形的单应性矩阵。具体为

$$\Delta_1 = R\left(\boldsymbol{F}_{\mathrm{A},\,1/4}, \boldsymbol{F}_{\mathrm{B},\,1/4}\right) \tag{1}$$

$$\boldsymbol{H}_1 = DLT\left\{R\left(\boldsymbol{F}_{\mathrm{A},\,1/4}, \boldsymbol{F}_{\mathrm{B},\,1/4}\right)\right\} \tag{2}$$

式中，$\boldsymbol{F}_{\mathrm{A},\,1/4}$，$\boldsymbol{F}_{\mathrm{B},\,1/4}$ 分别表示两个视角图像的特征，$R(\cdot,\cdot)$ 表示回归网络，$DLT\{\cdot\}$ 表示非参数化的变形计算。在计算完第 1 个尺度的单应性矩阵之后，将其作用于第 2 个尺度的特征 $\boldsymbol{F}_{\mathrm{B},\,1/2}$ 上，通过对特征 $\boldsymbol{F}_{\mathrm{B},\,1/2}$ 进行初步的扭曲以实现一个粗略变形的目的，

然后将粗略变形后的特征$\tilde{\boldsymbol{F}}_{\mathrm{B},1/2}$与$\boldsymbol{F}_{\mathrm{A},1/2}$一起进行局部相关性计算，再经过回归网络以及计算变形得到第2个尺度下的单应性矩阵$\boldsymbol{H}_2$。通过在局部特征上更加精细的相关性计算，第2阶段得到的变形矩阵对比第1阶段的结果更加精准，为了进一步提升深度学习的性能，本文算法共采用了3个阶段递归估计，即在第3阶段中，继续利用第2阶段得到的变形矩阵对第3个尺度的特征进行变形，然后仿照前两次的操作进行处理。至此，在经过3次单应性矩阵的更新后，最终得到了对单一模态数据的单应性估计。以此类推，可以用相同的做法得到另一种模态的单应性矩阵。

由于红外数据和可见光数据在成像内容上存在巨大差异，这两种模态得到的估计矩阵往往不尽相同。为了更好地利用两种数据的优势，同时规避两个模态各自存在的问题，本文方法对这两个单应性矩阵进行整合，即将两模态图像估计出的单应性矩阵一起通过一个由两层全连接操作构成的回归网络，整合得到一个融合了两种模态优势的单应性估计结果。

2.2 全局相关性计算

以1/4尺度下的特征$\boldsymbol{F}_{\mathrm{A},1/4}$，$\boldsymbol{F}_{\mathrm{B},1/4}$为例，对于两视角下特征的全局相关性，计算为

$$GC(x_1,x_2)=\frac{\left\langle \boldsymbol{F}_{\mathrm{A},1/4}(x_1),\boldsymbol{F}_{\mathrm{B},1/4}(x_2)\right\rangle}{\left|\boldsymbol{F}_{\mathrm{A},1/4}(x_1)\right|\left|\boldsymbol{F}_{\mathrm{B},1/4}(x_2)\right|},\quad x_1,x_2\in\boldsymbol{Z}^2 \tag{3}$$

式中，x_1，x_2表示特征图中的位置，$GC(x_1,x_2)$的数值范围是0到1，$\langle\cdot\rangle$表示点乘。通过计算两个视角特征的相关性，即可得到两个图像的匹配信息。当两个特征图越接近，相关性值越大。

2.3 基于多模态数据的广角图像重构

针对多尺度金字塔结构得到的单应性矩阵，本文将采集的红外数据对和可见光数据对进行变形，得到可用于拼接的预对齐数据。但是，单应性矩阵只能够在原图像具有相同深度的情况下才能够实现令人满意的变换，在实际场景中，很难满足这样的条件。同时，由于两幅输入图像可能于不同时间、不同传感器采集，相同位置信息的亮度与颜色往往会有差别，这会导致如果直接对扭曲后的图像进行简单的整合操作，在接缝区域会有明显的色差。因此，在通过单应性矩阵进行变形后，还需要对其整体进行重构处理，在保留两个视角信息的条件下，生成感知信息丰富的高质量全景图像。

考虑到红外数据和可见光数据刻画信息的不同，在重构阶段，本方法同时实现红外数据和可见光数据的融合，使得最终的拼接结果不仅具有更宽广的视野，同时还能够包含两个模态的互补信息。为了实现上述目标，本文首先引入Unet(Ronneberger等，2015)网络，通过挖掘场景的深层特征，提供上下文语意信息，并利用浅层特征进一步改善上采样信息不足的问题；此外，在编码和解码部分对应层的跳跃链接过程中引入注意力机制(Hu等，2018)，通过自监督的方式实现准确的拼接。然后将红外数据的拼接结果和可见光数据的结果分别输入两个密集连接块(Huang等，2017)进行编码，不同模态的编码特征通过加权操作得到融合互补后的特征，再通过一个密集连接块进行解码，重构出融合后的拼接结果，详细的网络结构图如图2所示。至此，通过输入的不同模态、不同视角的图像数据，本文方法得到了一个结合了多模态数据优势的广角图像。

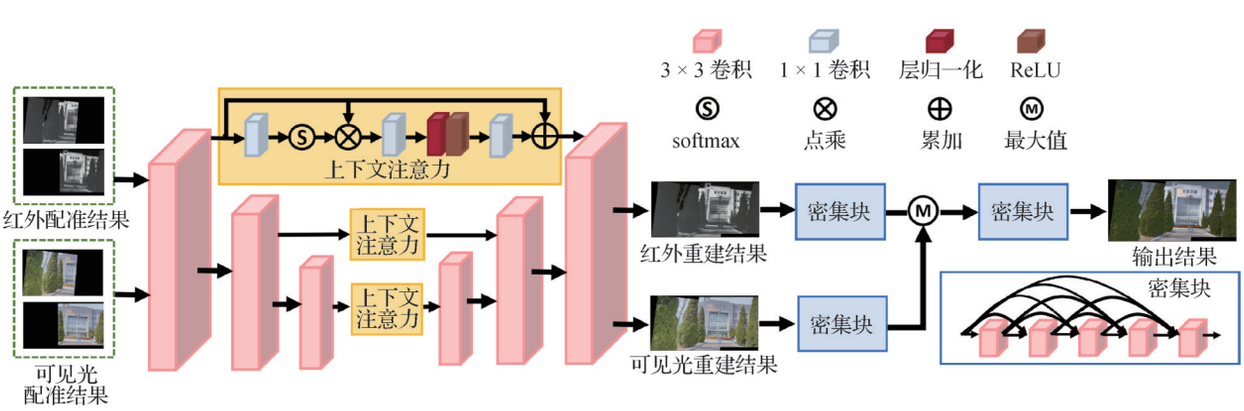


图2 图像重构模块的详细结构

Fig. 2 Detailed architecture of the image reconstruction module

2.4 损失函数

对于多尺度单应性估计模块，本文采用无监督的训练方式，通过针对不同视角下变形后的两幅图像的共有重叠区域进行范式约束，以保证变形对齐后的两幅图像在共有重叠区域上的内容保持一致，间接保证了变形估计的精确程度。在这里本文方法分别针对3个尺度得到的单应性参数进行计算，因此该模块的损失函数可以表示为

$$L_{\mathrm{H}} = \lambda_1 \left| \varepsilon(\boldsymbol{I}_{\mathrm{A}}, \boldsymbol{I}_{\mathrm{B}}, \boldsymbol{H}_1) \right|_1 + \lambda_2 \left| \varepsilon(\boldsymbol{I}_{\mathrm{A}}, \boldsymbol{I}_{\mathrm{B}}, \boldsymbol{H}_2) \right|_1 + \lambda_3 \left| \varepsilon(\boldsymbol{I}_{\mathrm{A}}, \boldsymbol{I}_{\mathrm{B}}, \boldsymbol{H}_3) \right|_1 \tag{4}$$

式中，$\boldsymbol{I}_{\mathrm{A}}, \boldsymbol{I}_{\mathrm{B}}$表示同一模态不同视角下的两幅图像，$\varepsilon$表示选取这两幅图像共有区域的操作，具体实现时可以表示为

$$\varepsilon(\boldsymbol{I}_{\mathrm{A}}, \boldsymbol{I}_{\mathrm{B}}, \boldsymbol{H}_1) = \boldsymbol{I}_{\mathrm{A}} \times W(\boldsymbol{E}, \boldsymbol{H}_1) - W(\boldsymbol{I}_{\mathrm{B}}, \boldsymbol{H}_1) \tag{5}$$

式中，$W(\cdot,\cdot)$表示不改变窗口大小的扭曲操作，$\boldsymbol{E}$为与$\boldsymbol{I}_{\mathrm{B}}$大小相同的全1掩码。在扭曲的操作中，超过其窗口以外的图像信息将会被隐藏。在取共有区域的操作中，参考图像$\boldsymbol{I}_{\mathrm{A}}$不需要进行额外扭曲。式(4)通过预测出的单应性矩阵对目标视角下的图像变形，得到与基准视角相配准的结果。λ_1、λ_2和λ_3表示3个尺度下对应的权重系数。

对于重构模块中的拼接部分，本文采用有监督的方式，这里将拼接结果分成两部分进行参考，即拼接缝区域和非拼接缝区域。对于拼接缝区域要保证其连续且光滑，而非拼接缝区域则要保证其与真值接近。因此本文首先计算得到4个掩码$\{\boldsymbol{M}_{\mathrm{S1}}, \boldsymbol{M}_{\mathrm{S2}}\}$，$\{\boldsymbol{M}_{\mathrm{C1}}, \boldsymbol{M}_{\mathrm{C2}}\}$来分别表示拼接缝区域和非拼接缝区域，具体为

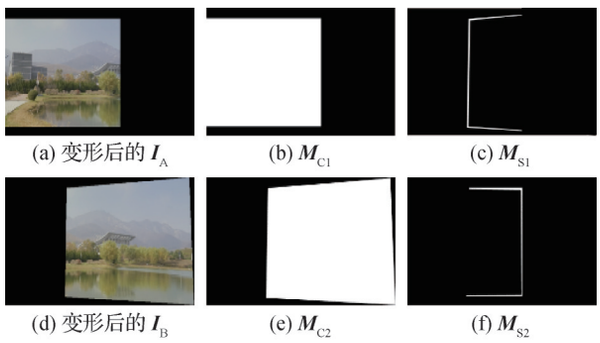
$$\begin{cases} \boldsymbol{M}_{\mathrm{C1}} = A(\boldsymbol{I}_{\mathrm{A}}, \boldsymbol{H}) \\ \boldsymbol{M}_{\mathrm{C2}} = A(\boldsymbol{I}_{\mathrm{B}}, \boldsymbol{H}) \end{cases} \tag{6}$$

$$\begin{cases} \nabla \boldsymbol{M}_{\mathrm{C1}} = \left| \boldsymbol{M}_{\mathrm{C1},i,j} - \boldsymbol{M}_{\mathrm{C1},i-1,j} \right| + \left| \boldsymbol{M}_{\mathrm{C1},i,j} - \boldsymbol{M}_{\mathrm{C1},i,j-1} \right| \\ \nabla \boldsymbol{M}_{\mathrm{C2}} = \left| \boldsymbol{M}_{\mathrm{C2},i,j} - \boldsymbol{M}_{\mathrm{C2},i-1,j} \right| + \left| \boldsymbol{M}_{\mathrm{C2},i,j} - \boldsymbol{M}_{\mathrm{C2},i,j-1} \right| \end{cases} \tag{7}$$

$$\begin{cases} \boldsymbol{M}_{\mathrm{S1}} = C(S(S(S(\nabla \boldsymbol{M}_{\mathrm{C2}})))) \odot \boldsymbol{M}_{\mathrm{C1}} \\ \boldsymbol{M}_{\mathrm{S2}} = C(S(S(S(\nabla \boldsymbol{M}_{\mathrm{C1}})))) \odot \boldsymbol{M}_{\mathrm{C2}} \end{cases} \tag{8}$$

式中，i,j表示笛卡尔坐标系的坐标，$A(\cdot,\cdot)$表示自适应窗口大小的图像扭曲。$S(\cdot)$表示SOBEL算子。$C(\cdot)$将得到的掩码的进行二值化。图3提供了掩码的可视化结果示例。

对于拼接缝区域，通过L1范式对其进行约束，具体为



(a) 变形后的$\boldsymbol{I}_{\mathrm{A}}$　(b) $\boldsymbol{M}_{\mathrm{C1}}$　(c) $\boldsymbol{M}_{\mathrm{S1}}$

(d) 变形后的$\boldsymbol{I}_{\mathrm{B}}$　(e) $\boldsymbol{M}_{\mathrm{C2}}$　(f) $\boldsymbol{M}_{\mathrm{S2}}$

图3　接缝掩码和非解封掩码示例

Fig. 3　Illustration of seam mask and content mask

((a) transformed $\boldsymbol{I}_{\mathrm{A}}$; (b) $\boldsymbol{M}_{\mathrm{C1}}$; (c) $\boldsymbol{M}_{\mathrm{S1}}$; (d) transformed $\boldsymbol{I}_{\mathrm{B}}$; (e) $\boldsymbol{M}_{\mathrm{C2}}$; (f) $\boldsymbol{M}_{\mathrm{S2}}$)

$$L_{\mathrm{S}} = \left| \boldsymbol{I}_{\mathrm{A}} \times \boldsymbol{M}_{\mathrm{S1}} - \boldsymbol{I}_{\mathrm{S}} \times \boldsymbol{M}_{\mathrm{S1}} \right|_1 + \left| \boldsymbol{I}_{\mathrm{B}} \times \boldsymbol{M}_{\mathrm{S2}} - \boldsymbol{I}_{\mathrm{S}} \times \boldsymbol{M}_{\mathrm{S2}} \right|_1 \tag{9}$$

式中，$\boldsymbol{I}_{\mathrm{S}}$表示拼接后的图像。利用两个视角场景中的拼接缝区域的掩膜分别与重构图像和真值图像进行点乘，以减少非拼接缝内容对拼接缝内容的影响和干扰。而对于非拼接缝区域，为了使得重建后图像的特征可以与通过单应性矩阵扭曲后的目标图像特征尽可能相似，本文采用VGG(Visual Geometry Group)网络(Simonyan和Zisserman，2015)进行深层的特征提取，然后通过对高层特征进行约束来保持感知信息上的一致，从而缓解因输入图像深度不一所带来的拼接信息差异问题，即

$$L_{\mathrm{C}} = \left| VGG(\boldsymbol{I}_{\mathrm{A}}) - VGG(\boldsymbol{I}_{\mathrm{S}} \times \boldsymbol{M}_{\mathrm{C1}}) \right|_2 + \left| VGG(\boldsymbol{I}_{\mathrm{B}}) - VGG(\boldsymbol{I}_{\mathrm{S}} \times \boldsymbol{M}_{\mathrm{C2}}) \right|_2 \tag{10}$$

而对于重构中的融合模块，本文方法额外训练了一个自监督的编解码结构，编码由两个密集连接块构成，解码部分由一个密集连接块实现。在测试阶段，两种模态数据将分别通过编码器进行编码，然后将编码后的特征进行相应加权后直接进行解码，即可得到对红外数据和可见光数据的融合结果。该编解码网络的训练通过对可见光图像的自我学习实现，其目标函数表示为

$$L_{\mathrm{F}} = \left| N(\boldsymbol{I}) - \boldsymbol{I} \right|_1 + (1 - SSIM(N(\boldsymbol{I}), \boldsymbol{I})) \tag{11}$$

式中，$N(\cdot)$表示编解码操作，$SSIM(\cdot,\cdot)$表示融合图像相似程度的评估指标，当两幅输入图像越接近时其值越大，因此为了便于损失函数的计算，采用了$1 - SSIM(N(\boldsymbol{I}), \boldsymbol{I})$的形式。

3 实 验

本文在合成数据集和真实数据上与目前已有的性能优异的图像拼接算法进行实验对比，基于传统特征点检测的方法包括APAP（Zaragoza等，2013）、SPW（single-perspective warps）（Liao和Li，2020）、WPIS（wide parallax image stitching）（Jia等，2021）和SLAS（seam-guided local alignment and stitching）（Liao等，2023）；基于深度学习的方法包括VFIS（view-free image stitching）（Nie等，2020），RSFI（reconstructing stitched features to images）（Nie等，2021）和UDIS++（unsupervised deep image stitching）（Nie等，2023）。

考虑到本文算法在重构阶段引入了红外数据和可见光数据的融合模块，为了实现更公平的对比，本文考虑在执行常规图像拼接算法的基础上进一步通过红外和可见光图像融合实现对等的性能评估。由于红外图像和可见光图像在成像内容上差异明显，不同模态数据得到的单应性矩阵存在不同，从而导致对红外图像对和可见光图像对进行拼接处理会出现拼接结果形状异构的情况。对这样的拼接结果进行融合，非但不能实现互补信息的整合，反而会出现严重的鬼影干扰。考虑到图像融合是逐像素地进行整合，不会出现场景形状的变化，因此实验对比时，本文在执行其他拼接算法前，预先执行一次融合算法，这里采用了发表在顶级期刊或会议上融合表现稳定且融合效果优异的方法，包括RFN（residual fusion network）（Li等，2021）、ReCoNet（recurrent correction network）（Huang等，2022）和DATFuse（dual attention transformer）（Tang等，2023），使得参加对比的拼接算法是针对融合后的数据进行的。

3.1 数据集介绍

对于多模态图像拼接任务，本文在真实条件下基于实时同步红外—可见光双摄相机采集了一个连续拍摄可用于拼接任务的真实数据集。该数据库包含530对配准好的红外和可见光图像，涵盖12种真实场景，包括自然风光、人文建筑、野外环境及室内环境。场景复杂多变，连续拍摄间隔随机，具有丰富拼接条件的多样性。此外，为了更客观地评价算法性能，本文基于Nie等人（2021）提出的拼接图像合成方法对已有的红外可见光数据集RoadScene（Xu等，2020）进行处理，由于该数据集是非连续拍摄得到的，每个场景中仅一对多模态数据，因此从该场景中任意位置取尺寸为128×128像素的图像块作为参考图像，将该图像的4个顶点位置随机移动，并将该任意四边形场景强制变形为128×128像素的正方形以得到目标图像块。基于RoadScene数据集共生成了10 000对训练数据和200对测试数据，并且严格保证训练和测试集不存在交集。

3.2 实验设定

本文方法的实验代码由Tensorflow框架实现，在RTX3090 GPU上进行模型的训练。首先将输入图像输入到多模态单应估计网络，得到单应性矩阵后将输入图像进行初步的对齐。随后将对齐的图像输入到重构网络，以得到拼接缝区域过渡平滑的融合拼接图像。针对单应性估计网络的预训练，一批图像的数量设置为4，训练轮数设置为50，图像的输入尺寸设置为128×128像素。初始学习率设置为1E-4。在训练过程中，损失函数的加权系数λ_1、λ_2和λ_3分别设置为16、14和1。针对特征重构网络，一批图像的数量设置为8，训练轮数设置为20，图像的输入尺寸设置为128×256像素，初始学习率设置为1E-3，L_S、L_C和L_F的加权系数分别设置为1、1E-5和1。在分别对单应性估计网络和特征重构网络进行预训练后，需要对两个模块进行联合训练，一批图像的数量被设置为1，训练轮数设置为30，初始学习率设置为2E-5，L_H、L_S、L_C和L_F的加权系数分别设置为1、1、1E-5和1。在所有网络的训练中，均采用Adam优化器进行训练，学习率的衰减系数设置为0.96。

3.3 可视结果对比

图4展示的是与传统图像拼接算法的结果对比。IR表示输入的红外数据，VIS表示输入的可见光数据。为了避免融合算法对拼接性能的影响，本文从3种融合算法得到的结果中选取性能最好的结果进行后续拼接操作。通过观察可以发现，在第1个例子上，APAP和SPW算法的拼接结果表明其对于共有区域和拼接缝区域的处理不够平滑，生成的广角图像中边界明显，产生了明显的色差。影响了拼接图像的可视效果。WPIS和SLAS对于目标视角的变形估计出现了偏差，场景扭曲异常，在破坏了原始场景直线框架的同时，在重叠区域也产生了明显误差。而对于第2个例子中APAP匹配失败，直接导致拼接图像丢失大量信息。SPW、WPIS和SLAS在

其重叠区域产生了误差，使图像的场景产生了畸变，并出现了伪影，破坏了视觉效果。而本文算法在两个例子中的结果均估计准确。避免了伪影的干扰，清晰完整地保留了图像的原始信息。

图 5 展示的是本文算法与深度学习算法的结果对比。在两个例子中，VFIS、RSFI 和 UDIS++算法都没有很好地利用两个视角场景的内容信息，对于多视角场景的配准估计均出现明显问题，导致最终拼接结果损失了大量的信息。而本文算法得到了较好的拼接结果，在共有区域的过渡十分平滑，重建出的最终场景清晰准确，能够为后续算法对环境的感知提供有力保障。

图 4　与传统拼接算法的可视结果对比

Fig. 4　Visual comparison with the conventional stitching methods

((a) IR; (b) VIS; (c) APAP; (d) SPW; (e) WPIS; (f) SLAS; (g) ours)

图 5　与深度学习拼接算法的可视结果对比

Fig. 5　Visual comparison with the deep learning based stitching methods

((a) IR; (b) VIS; (c) VFIS; (d) RSFI; (e) UDIS++; (f) ours)

3.4　定量结果对比

为了更加客观地评估各个拼接算法的性能，针对真实数据库，本文借助 4 个图像质量评估指标，包括熵（entropy，EN）（Roberts 等，2008）、空间频率（spatial frequency，SF）（Eskicioglu 和 Fisher，1995），平均梯度（average gradient，AG）（Cui 等，2015）和无参考图像空间评价指标（blind/referenceless image spatial quality evaluator，BRISQUE）（Mittal 等，2012），并且 4 个指标的数据均与图像质量呈正相关。

实验结果如表 1 所示。在 3 种融合方法和 7 种拼接算法的共 21 种组合中，本文算法在 EN、SF 以及 BRISQUE 指标上均取得了最高值，而在 AG 上以微弱劣势低于 SPW 方法。综合来看，与传统拼接和基于深度学习的拼接算法相比，本文算法优势较为明显，鲁棒性较好。

3.5　消融实验分析

3.5.1　关于多模态单应性估计的有效性分析

本文分别估计出了红外图像和可见光图像的单应性矩阵，随后，对两模态的矩阵进行整合得到统一的单应矩阵。其中，一个统一的单应性矩阵是否能比单模态的矩阵取得更好的效果是值得探讨的。这里分别将红外模态和可见光模态所预测出的单应性矩阵以及结合了红外光信息和可见光信息的单应性矩阵进行比较。本文采用 RoadScene 合成数据库进

表1　多种融合—拼接算法在真实数据库上的定量结果对比

Table 1　Quantitative comparison against different fusion-stitching strategies

	方法	EN	SF	AG	BRISQUE
RFN	APAP	6.466	9.282	2.672	0.491
	SPW	6.476	9.231	2.804	0.482
	WPIS	6.452	9.380	2.805	0.486
	VFIS	6.786	9.193	3.312	0.493
	RSFI	6.704	8.347	2.528	0.487
	SLAS	6.702	9.046	3.303	0.492
	UDIS++	6.637	10.009	2.590	0.490
ReCoNet	APAP	5.635	10.068	3.217	0.489
	SPW	5.846	9.916	3.251	0.485
	WPIS	5.721	9.822	3.372	0.492
	VFIS	6.12	10.155	3.358	0.494
	RSFI	6.172	9.997	3.279	0.488
	SLAS	6.189	10.120	3.362	0.491
	UDIS++	6.059	9.920	3.279	0.488
DATFuse	APAP	5.636	10.108	3.342	0.492
	SPW	5.813	11.179	**3.553**	0.491
	WPIS	5.719	10.152	3.332	0.490
	VFIS	5.894	9.994	3.301	0.492
	RSFI	5.933	9.653	2.667	0.489
	SLAS	5.972	10.029	3.359	0.493
	UDIS++	5.609	9.669	3.057	0.491
	本文	**6.990**	**11.186**	3.373	**0.495**

注：加粗字体表示各列最优结果。

行实验对比，可视化结果如图6所示，采用平均角点误差（average corner error，ACE）作为评估指标，该值越低，预测的单应矩阵越准确。图中白色框表示单应性矩阵真值的变形趋势，蓝色框表示通过算法预测出的单应性矩阵的变形趋势，白框与蓝框的接近程度表示其预测出的单应性矩阵的精确程度。

定量估计如表2所示。不难发现，在所提供的两个例子中，结合了红外模态和可见光模态信息的单应性矩阵无论是在可视结果上还是在平均角点误差的指标对比中都取得了最好的效果。

3.5.2　关于多模态潜在单应性信息融合策略的有效性分析

本文基于多尺度金字塔的单应性估计模块分别对红外数据和可见光数据的单应性进行估计。在得到了两个潜在变形估计后，如何将其有效结合，使其既充分保留红外数据的结构信息又不丢失可见光特征的细节信息是值得探究的。为此，本文分别对这两个单独得到的单应性矩阵进行L1正则化、取最大值和取平均值操作，并在合成数据集上进行对比。定量结果如表3所示，可视化结果如图7所示，IR表示红外数据，VIS表示可见光数据。通过对比可以发现，当采用上述3种操作时，输出的拼接图像均会

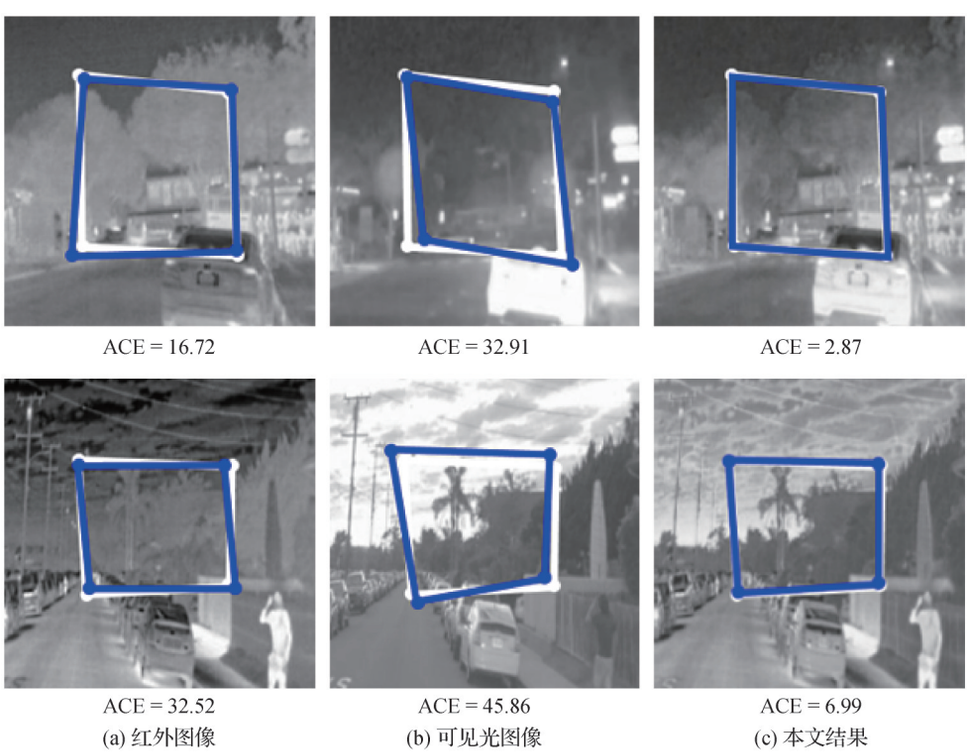


图6　针对多模态单应性估计模块的可视化分析

Fig. 6　Visual analysis on multi-modality homography estimation module((a) infrared images; (b) visible images; (c) ours)

表 2　平均角点误差定量结果对比

Table 2　Quantitative comparison in average corner error

方法	平均角点误差
红外图像的单应性矩阵	11.06
可见光图像的单应性矩阵	9.73
本文	**4.57**

注：加粗字体表示最优结果。

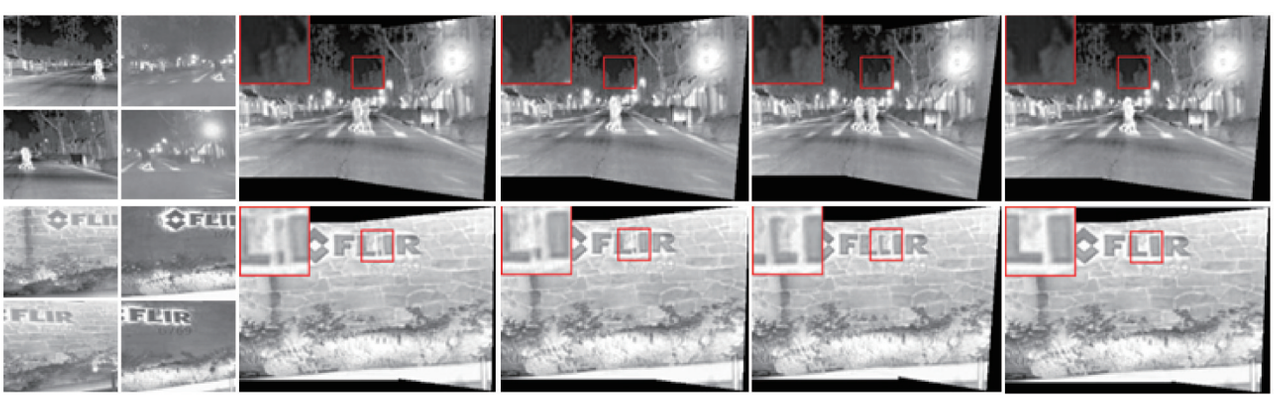
出现伪影、信息缺失等问题，而采用回归网络得到的拼接结果实现了良好的可视化效果。通过评价指标对比可知，本文方法包含了最丰富的信息。因此，本文采用回归的处理手段来融合不同模态得到的单应性估计是合理的。

表 3　单应性信息融合策略在 Roadscene 合成数据库上定量结果对比

Table 3　Quantitative comparison of the homography integration on Roadscene synthetic dataset

方法	EN	SF	AG	BRISQUE
L1	6.934	17.532	5.751	0.491
Avg	6.950	17.531	5.726	0.49
Max	6.884	17.384	5.662	0.492
本文	**6.968**	**17.770**	**5.853**	**0.495**

注：加粗字体表示各列最优结果。



(a) IR　(b) VIS　(c) L1　(d) Avg　(e) Max　(f) 本文

图 7　针对多模态潜在单应性信息融合策略的可视化分析

Fig. 7　Visual analysis on multi-modality potential homography integration ((a) IR; (b) VIS; (c) L1; (d) Avg; (e) Max; (f) ours)

3.5.3　关于注意力重建模块的有效性分析

本文中基于注意力机制的重构模块可以对红外光和可见光图像进行重构并得到信息更丰富、视野更广阔的全景图像，针对网络模块是否能在图像平滑的前提下保留多模态数据的信息也是值得讨论分析的。本文采用当今最具有代表性的主干模块，包括 Dense-121（Huang 等，2017）、ResNet-18（He 等，2016）以及 VGG-11（Simonyan 和 Zisserman，2015），将其整体替换为注意力机制重构模块，并在训练后进行测试对比，定量结果如表 4 所示，可视结果如图 8 所示。通过对比不难发现，采用注意力机制重构模块在引入感知信息辅助网络有选择性地对信息进行融合重构，因此本文提出的重构模块在视觉效果和多项评价指标中均取得了最优效果。

3.5.4　关于注意力重建模块关键损失函数的有效性分析

本文重建模块中使用到了 3 个关键的损失函数。分别是针对拼接缝区域的 L1 范式约束、针对输入图像深层特征的感知约束以及结构相似度（structural similarity index，SSIM）约束。为了探究 3 个损失函数在训练中起到的作用，在合成数据集上分别比较了这 3 个关键损失函数的总体性能，定量结果如表 5 所示，w/o 表示不使用该损失函数进行训练。可以看到，3 种损失函数的使用都对网络的训练起

表 4　重建网络的定量结果对比

Table 4　Quantitative comparison of the reconstruction module

方法	EN	SF	AG	BRISQUE
Dense-121	6.843	14.542	4.626	0.483
Res-18	6.883	15.008	4.915	0.484
VGG-11	6.815	14.352	4.756	0.483
本文	**6.968**	**17.770**	**5.853**	**0.495**

注：加粗字体表示各列最优结果。

(a) IR (b) VIS (c) Dense-121 (d) Res-18 (e) VGG-11 (f) 本文

图8 针对重建网络有效性的可视化分析

Fig. 8 Visual analysis of the effectivenesson the reconstruction module

((a) IR; (b) VIS; (c) Dense-121; (d) Res-18; (e) VGG-11; (f) ours)

表5 重建网络损失函数消融分析的定量结果对比

Table 5 Quantitative comparison of the ablation study on loss function of reconstruction module

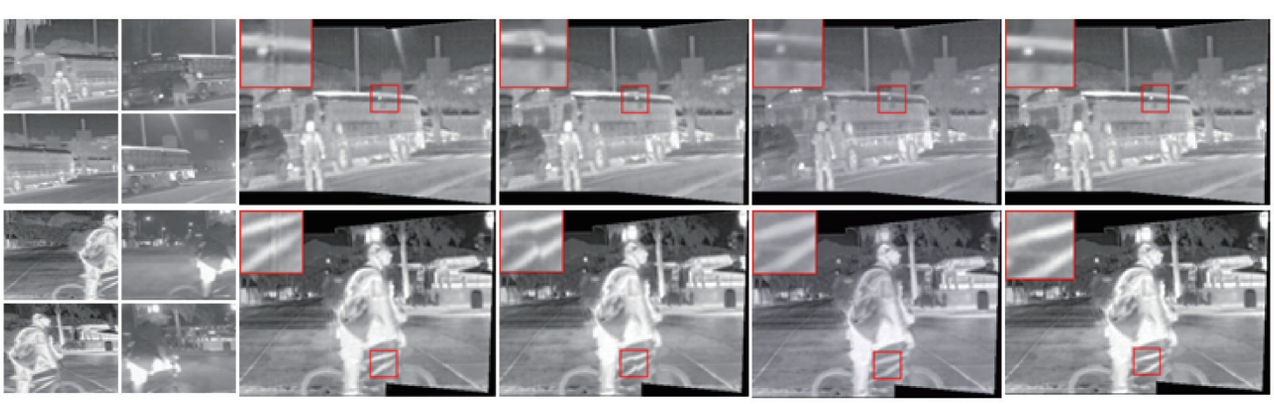
方法	EN	SF	AG	BRISQUE
w/o 接缝损失	6.916	14.811	5.688	0.491
w/o 感知损失	6.804	15.057	5.412	0.492
w/o 结构损失	6.924	16.263	5.438	0.493
本文	**6.968**	**17.770**	**5.853**	**0.495**

注:加粗字体表示各列最优结果。

到了正向作用。可视化结果对比如图9所示。当对图像的接缝处进行约束时,图像在拼接缝区域的过渡会更加平滑。当对其深层特征进行感知约束时,图像在重叠区域上的伪影会被有效地消除。而当对图像的结构特征进行约束时,图像的纹理与结构信息会更加丰富。综合来看,实验结果证实了3个损失函数的有效性。

3.6 挑战性场景性能展示

为了进一步验证本文算法的泛化能力和鲁棒性,针对真实场景中大基线场景上的性能进行了评估,即多个视角下的场景重叠区域面积较小,并且可



(a) IR (b) VIS (c) w/o接缝损失 (d) w/o感知损失 (e) w/o结构损失 (f) 本文

图9 损失函数的消融分析

Fig. 9 Ablation analysis on the loss function

((a) IR; (b) VIS; (c) w/o seam loss; (d) w/o perception loss; (e) w/o structure loss; (f) ours)

表6 在挑战性场景上的定量结果

Table 6 Quantitative results on challenging case

方法	EN	SF	AG	BRISQUE
本文	5.764	16.023	4.798	0.486

供参考的信息量较少的场景。定量结果如表6所示。图10展示的两个例子是在大基线场景下的拼接结果。可以发现，虽然该场景中的重叠区域不足20%，但是本文算法仍然能够实现较为准确的单应性估计，在重叠区域实现了平滑地衔接和过渡，并且最终的拼接结果能够很好地还原真实场景。

(a) IR 1　(b) IR 2　(c) VIS 1　(d) VIS 2　(e) 本文

图10　在挑战性场景中的性能展示

Fig. 10　Visual results on challenging cases ((a) IR 1; (b) IR 2; (c) VIS 1; (d) VIS 2; (e) ours)

4　结　论

本文提出一种基于多模态数据特征融合的广角图像生成方法。该方法结合了红外图像结构信息丰富和可见光图像纹理细节特征丰富的优势，克服了环境条件对常规可见光图像拼接算法的限制。其中，多模态数据通过多尺度特征金字塔回归全局相关损失，非参数化计算变形矩阵，使得拼接结果更加精准可靠；此外，重构模块中上下文语意感知能有效实现信息补偿。通过在合成数据和真实数据上的定性和定量实验对比表明，采用本文方法得到的拼接性能更加稳定，相较于其他基于融合—拼接策略得到的结果在处理重叠区域时更加准确，场景还原更加真实，避免了重影和异常形变的出现。

虽然本文方法在处理全局刚性变换场景的对齐时效果显著，但由于真实世界中场景复杂，场景内部深度差异明显，从而导致不同视角采集的数据之间不能仅通过一个全局单应性矩阵实现精确对齐。此外，场景中的移动目标在拼接结果中通常会呈现移动模糊和扭曲失真的情况，影响后续算法对场景内容的感知。在未来的工作中，将从以下两个方面针对这些问题进行改进：1)构建全局刚性和局部非刚性变换相结合的跨视角对齐网络，通过全局粗对齐与局部精细调整相结合的策略更适用于复杂场景下的对齐。2)构建静态场景和动态场景识别网络，针对静态区域可采用常规拼接思路，而针对动态区域，则需要独立设计拼接算法；进一步提升拼接网络的自适应能力和鲁棒性。

参考文献(References)

Chang C H, Sato Y and Chuang Y Y. 2014. Shape-preserving half-projective warps for image stitching//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 3254-3261 [DOI: 10.1109/CVPR.2014.422]

Cui G M, Feng H J, Xu Z H, Li Q and Chen Y T. 2015. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. Optics Communications, 341: 199-209 [DOI: 10.1016/j.optcom.2014.12.032]

DeTone D, Malisiewicz T and Rabinovich A. 2016. Deep image homography estimation [EB/OL]. [2023-12-27]. https://arxiv.org/pdf/1606.03798.pdf

Eskicioglu A M and Fisher P S. 1995. Image quality measures and their performance. IEEE Transactions on Communications, 43 (12): 2959-2965 [DOI: 10.1109/26.477498]

Gao J H, Kim S J and Brown M S. 2011. Constructing image panoramas using dual-homography warping//Proceedings of the CVPR 2011. Colorado Springs, USA: IEEE: 49-56 [DOI: 10.1109/CVPR.2011.5995433]

Gao J H, Li Y, Chin T J and Brown M S. 2013. Seam-driven image stitching//Eurographics (Short Papers). Girona, Spain: The Eurographics Association: 45-48

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hoang V D, Tran D P, Nhu N G, Pham T A and Pham V H. 2020. Deep

feature extraction for panoramic image stitching//Proceedings of the 12th Asian Conference on Intelligent Information and Database Systems:. Phuket, Thailand: Springer: 141-151 [DOI: 10.1007/978-3-030-42058-1_12]

Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]

Huang G, Liu Z, Van Der Maaten L and Weinberger K Q. 2017. Densely connected convolutional network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2261-2269 [DOI: 10.1109/CVPR.2017.243]

Huang Z B, Liu J Y, Fan X, Liu R S, Zhong W and Luo Z X. 2022. ReCoNet: recurrent correction network for fast and efficient multi-modality image fusion//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 539-555 [DOI: 10.1007/978-3-031-19797-0_31]

Jia Q, Li Z J, Fan X, Zhao H T, Teng S Y, Ye X C and Latecki L J. 2021. Leveraging line-point consistence to preserve structures for wide parallax image Stitching//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 12181-12190 [DOI: 10.1109/CVPR46437.2021.01201]

Lai W S, Gallo O, Gu J W, Sun D Q, Yang M H and Kautz J. 2019. Video stitching for linear camera arrays [EB/OL]. [2023-12-27]. https://arxiv.org/pdf/1907.13622.pdf

Lee K Y and Sim J Y. 2020. Warping residual based image stitching for large parallax//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8198-8206 [DOI: 10.1109/CVPR42600.2020.00822]

Leng J X, Mo M J C, Zhou Y H, Ye Y M, Gao C Q and Gao X B. 2023. Recent advances in drone-view object detection. Journal of Image And Graphics, 28(9): 2563-2586 (冷佳旭, 莫梦竟成, 周应华, 叶永明, 高陈强, 高新波. 2023. 无人机视角下的目标检测研究进展. 中国图象图形学报, 28(9): 2563-2586) [DOI: 10.11834/Jig.220836]

Li H, Wu X J and Kittler J. 2021. RFN-nest: an end-to-end residual fusion network for infrared and visible images. Information Fusion, 73: 72-86 [DOI: 10.1016/j.inffus.2021.02.023]

Li X Y, Ye Z H, Wei S K, Chen Z, Chen X T, Tian Y H, Dang J W, Fu S J and Zhao Y. 2023. 3D object detection for autonomous driving from image: a survey——benchmarks, constraints and error analysis. Journal of Image And Graphics, 28(6): 1709-1740 (李熙莹, 叶芝桧, 韦世奎, 陈泽, 陈小彤, 田永鸿, 党建武, 付树军, 赵耀. 2023. 基于图像的自动驾驶3D目标检测综述——基准、制约因素和误差分析. 中国图象图形学报, 28(6): 1709-1740) [DOI: 10.11834/Jig.230036]

Liao T L and Li N. 2020. Single-perspective warps in natural image stitching. IEEE Transactions on Image Processing, 29: 724-735 [DOI: 10.1109/TIP.2019.2934344]

Liao T L, Zhao C Y, Li L and Cao H L. 2023. Seam-guided local alignment and stitching for large parallax images [EB/OL]. [2023-12-27]. https://arxiv.org/pdf/2311.18564.pdf

Lin C C, Pankanti S U, Ramamurthy K N and Aravkin A Y. 2015. Adaptive as-natural-as-possible image stitching//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1155-1163 [DOI: 10.1109/CVPR.2015.7298719]

Lin K M, Jiang N J, Cheong L F, Do M and Lu J B. 2016. SEAGULL: seam-guided local alignment for parallax-tolerant image stitching//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 370-385 [DOI: 10.1007/978-3-319-46487-9_23]

Lin W Y, Liu S Y, Matsushita Y, Ng T T and Cheong L F. 2011. Smoothly varying affine stitching//Proceedings of the CVPR 2011. Colorado Springs, USA: IEEE: 345-352 [DOI: 10.1109/CVPR.2011.5995314]

Lou Z Y and Gevers T. 2014. Image alignment by piecewise planar region matching. IEEE Transactions on Multimedia, 16(7): 2052-2061 [DOI: 10.1109/TMM.2014.2346476]

Mittal A, Moorthy A K and Bovik A C. 2012. No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing, 21(12): 4695-4708 [DOI: 10.1109/TIP.2012.2214050]

Nie L, Lin C Y, Liao K, Liu M Q and Zhao Y. 2020. A view-free image stitching network based on global homography. Journal of Visual Communication and Image Representation, 73: #102950 [DOI: 10.1016/j.jvcir.2020.102950]

Nie L, Lin C Y, Liao K, Liu S C and Zhao Y. 2021. Unsupervised deep image stitching: reconstructing stitched features to images. IEEE Transactions on Image Processing, 30: 6184-6197 [DOI: 10.1109/TIP.2021.3092828]

Nie L, Lin C Y, Liao K, Liu S C and Zhao Y. 2023. Parallax-tolerant unsupervised deep image stitching [EB/OL]. [2023-12-27]. https://arxiv.org/pdf/2302.08207.pdf

Roberts J W, Van Aardt J A and Ahmed F B. 2008. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. Journal of Applied Remote Sensing, 2(1): #023522 [DOI: 10.1117/1.2945910]

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Shen C W, Ji X Y and Miao C L. 2019. Real-time image stitching with convolutional neural networks//Proceedings of 2019 IEEE International Conference on Real-time Computing and Robotics. Irkutsk, Russia: IEEE: 192-197 [DOI: 10.1109/RCAR47638.2019.

9044010]

Shi Z F, Li H, Cao Q J, Ren H Z and Fan B Y. 2020. An image mosaic method based on convolutional neural network semantic features extraction. Journal of Signal Processing Systems, 92(4): 435-444 [DOI: 10.1007/s11265-019-01477-2]

Shi Z H, Wu C W, Li C J, You Z Z, Wang Q and Ma C C. 2023. Object detection techniques based on deep learning for aerial remote sensing images: a survey. Journal of Image And Graphics, 28(9): 2616-2643 (石争浩, 仵晨伟, 李成建, 尤珍臻, 王泉, 马城城. 2023. 航空遥感图像深度学习目标检测技术研究进展. 中国图象图形学报, 28(9): 2616-2643) [DOI: 10.11834/jig.221085]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2023-12-27]. https://arxiv.org/pdf/1409.1556.pdf

Tang W, He F Z, Liu Y, Duan Y S and Si T Z. 2023. DATFuse: infrared and visible image fusion via dual attention transformer. IEEE Transactions on Circuits and Systems for Video Technology, 33(7): 3159-3172 [DOI: 10.1109/TCSVT.2023.3234340]

Xu H, Ma J Y, Le Z L, Jiang J J and Guo X J. 2020. FusionDN: a unified densely connected network for image fusion//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 12484-12491 [DOI: 10.1609/aaai.v34i07.6936]

Zaragoza J, Chin T J, Brown M S and Suter D. 2013. As-projective-as-possible image stitching with moving DLT//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE: 2339-2346 [DOI: 10.1109/CVPR.2013.303]

Zhang F and Liu F. 2014. Parallax-tolerant image stitching//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 3262-3269 [DOI: 10.1109/CVPR.2014.423]

作者简介

姜智颖，女，讲师，主要研究方向为计算机视觉。
E-mail: zyjiang0630@gmail.com
刘晋源，通信作者，男，助理研究员，主要研究方向为计算机视觉和多模态图像处理。E-mail: atlantis918@hotmail.com
张曾翕，男，硕士研究生，主要研究方向为计算机视觉。
E-mail: zengxizhang@mail.dlut.edu.cn
刘日升，男，教授，主要研究方向为计算机视觉和最优化方法。E-mail: rsliu@dlut.edu.cn