

Article

Coupled Spectral–Spatial Fusion-Enabled Multi-Scale Panoramic Imaging

Ke Yin ^{1,2}, Zheng Wen ², Yuan Liao ², Shubin Liu ^{3,*} , Xiyang Zhi ^{1,*}  and Guangzhen Bao ¹

¹ School of Aeronautics, Harbin Institute of Technology (HIT), Harbin 150006, China; ke_yin@stu.hit.edu.cn (K.Y.); guangzhenbao@stu.hit.edu.cn (G.B.)

² Institute of Remote Sensing Satellite, China Academy of Space Technology, Beijing 100086, China; zwen1423@usc.edu.cn (Z.W.); yuan_liao@zuu.zju.edu.cn (Y.L.)

³ College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China

* Correspondence: shubin_liu@fzu.edu.cn (S.L.); zhixiyang@hit.edu.cn (X.Z.)

Abstract

Conventional imaging systems often suffer from a coupled limitation of narrow field of view, single modality, and insufficient effective resolution. Wide-angle imaging preserves scene context but compresses distant or small-scale targets into limited pixels, while narrow-field imaging improves details at the cost of global perception. Moreover, single-modal visible imaging is sensitive to illumination and contrast variations, whereas infrared imaging lacks fine spatial texture. To increase information at the imaging source, we propose coupled spectral–spatial fusion-enabled multi-scale panoramic imaging, a dual-field-of-view (FOV) visible–infrared framework for wide-field high-resolution perception. Two imaging units acquire paired visible and infrared images from adjacent overlapping views. For each view, a visible–infrared fusion super-resolution model integrates visible structural details with infrared radiative cues to reconstruct a high-resolution fused image. A multi-scale stitching algorithm then extracts robust features, estimates cross-view correspondences, and merges the two fused images into a large-FOV panoramic result. Outdoor experiments demonstrate that the proposed method improves local contrast, suppresses pixelation artifacts, enhances readable fine details, and expands the observable field of view, providing an effective route toward multimodal panoramic imaging.

Keywords: imaging system; optical imaging; computational imaging; super resolution

1. Introduction

Wide-field, high-resolution imaging is becoming increasingly important in remote observation, intelligent surveillance, autonomous perception, and outdoor scene understanding. In these scenarios, an imaging system is expected to capture a broad spatial context while retaining fine local details. This requirement is difficult to satisfy with conventional single-camera architectures. A wide-angle optical system can enlarge the observable scene, but distant objects are compressed into only a few pixels, making small-scale targets, text, edges, and structural details difficult to detect or interpret. Conversely, a narrow-field system improves local resolution but sacrifices global scene awareness. This field-of-view–resolution contradiction has long been recognized in computational imaging and remote sensing, and it becomes more severe when targets exhibit large variations in scale, orientation, distance, and contrast [1–3]. In aerial and outdoor imagery, small objects are especially vulnerable to information loss, and recent studies have shown that multi-



Received: 15 June 2026

Revised: 3 July 2026

Accepted: 13 July 2026

Published: 14 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

scale representation and super-resolution are key strategies for improving small-object perception [4–6].

A number of optical and computational strategies have been developed to alleviate this limitation. Multiscale and gigapixel cameras have demonstrated that large fields of view and high pixel counts can be achieved through sophisticated optical architectures and distributed sensor arrays [1]. However, these systems usually require complex hardware integration, precise calibration, and heavy image formation pipelines, which limit their flexibility in compact field deployment. Image stitching provides another route. Classical panorama generation methods use invariant local features, geometric transformation estimation, and image blending to merge overlapping images into a wider view [2,7–10]. SIFT and SURF laid the foundation for robust local feature matching under scale, rotation, and illumination changes [7,8], while automatic panorama stitching and as-projective-as-possible warping improved multi-image alignment in practical scenes [9,10]. Nevertheless, most stitching pipelines assume that the input images already contain sufficient texture, comparable resolution, and reliable overlapping information. This assumption is often violated in long-distance outdoor imaging, where repetitive building facades, weak textures, parallax, low local contrast, and sensor noise can lead to unstable matching, visible seams, or local ghosting.

Another important limitation is modality. Visible imaging provides rich textures, sharp edges, and intuitive spatial structures, but it is sensitive to illumination, haze, shadow, and background reflectance. Infrared imaging is complementary. It captures radiation-related contrast and can provide more stable responses under weak illumination or complex environmental conditions. However, infrared images generally exhibit lower texture density, weaker boundary sharpness, and lower spatial resolution than visible images. Therefore, neither visible nor infrared imaging alone can provide a sufficiently complete description of complex outdoor scenes. Infrared-visible image fusion has been widely investigated to combine the structural information of visible images and the radiative saliency of infrared images [11]. Deep-learning-based fusion methods, such as DenseFuse, FusionGAN, DDcGAN, and SwinFusion, have further improved feature extraction, cross-modal interaction, and information preservation [12–15]. These studies confirm the value of visible–infrared complementarity. Yet most of them focus on single-view image enhancement. They improve the information content of one image but do not directly address the limited field of view of the imaging system.

Image super-resolution offers a further means of recovering spatial details from low-resolution observations. Deep super-resolution methods, including SRCNN, EDSR, SRGAN, ESRGAN, and diffusion-based SR3, have significantly advanced high-frequency reconstruction and perceptual image restoration [16–20]. However, single-modal super-resolution remains an ill-posed inverse problem. When the original image lacks sufficient information, the reconstructed textures may be over-smoothed or hallucinated. This is particularly problematic for outdoor panoramic imaging, where small signs, building textures, and distant local structures must be faithfully recovered rather than visually fabricated. A more reliable strategy is to introduce complementary physical information before reconstruction. Visible images can constrain geometry, contours, and fine textures, while infrared images can strengthen target contrast and radiative cues. Coupling visible–infrared fusion with super-resolution reconstruction therefore provides a source-level way to increase effective information before panoramic stitching.

Despite these advances, existing methods still treat field expansion, modality fusion, and resolution enhancement as largely separate tasks. Optical wide-field systems expand scene coverage but do not necessarily improve target interpretability. Image stitching enlarges the field of view but depends heavily on the quality of the input images. Infrared–

visible fusion enhances modal information but usually remains confined to a single field of view. Super-resolution improves image details but may generate unreliable textures when only one modality is available. This separation creates an urgent need for a system–algorithm co-designed framework that increases information at the source, enhances each local view before registration, and then expands the spatial coverage through robust multi-scale stitching.

Here, we propose coupled spectral–spatial fusion-enabled multi-scale panoramic imaging, a dual-FOV visible–infrared imaging framework designed to overcome the simultaneous limitations of narrow field of view, single modality, and insufficient effective resolution. The system consists of two visible–infrared imaging units that capture adjacent fields of view with an overlapping region. For each view, the visible and infrared images are first processed by a coupled fusion super-resolution model to generate a high-resolution fused representation. The visible branch preserves edges, textures, and geometric structures, while the infrared branch provides complementary radiative contrast and target saliency. The two super-resolved fused views are then registered and merged by a multi-scale panoramic stitching algorithm, which performs feature extraction, overlap perception, coarse-to-fine matching, geometric alignment, and seamless fusion. The main contributions of this work are threefold. First, we establish a dual-FOV visible–infrared imaging architecture that enhances source information in both spectral and spatial dimensions. Second, we develop a fusion super-resolution model that reconstructs high-resolution multimodal views before stitching, improving local contrast and feature reliability. Third, we introduce a multi-scale stitching strategy that converts two enhanced local views into a large-FOV panoramic image while preserving fine details. Outdoor experiments demonstrate that the proposed framework improves local clarity, enhances readable details such as distant text, suppresses mosaic-like degradation, and produces a more informative panoramic representation than raw single-modal imaging.

2. Materials and Methods

2.1. Principle

The overall workflow of the proposed system is illustrated in Figure 1. Two visible–infrared camera units are used to observe adjacent fields of view with a shared overlapping region. For each view, the visible image and its field-matched infrared counterpart are first acquired as a paired input. The visible channel provides fine spatial textures, edges and structural details, whereas the infrared channel contributes radiation-sensitive contrast and more robust target responses under illumination variations. In this way, each local view is enriched at the source level before any panoramic reconstruction is performed.

After acquisition, the visible–infrared pair from each camera unit is processed by the proposed diffusion-based super-resolution fusion model. This model reconstructs a high-resolution fused image for the left view and another high-resolution fused image for the right view. The two reconstructed views are then passed to a multi-scale stitching module, which estimates the overlap, aligns the two views and blends them into a large-field panoramic image. Rather than directly stitching low-resolution or single-modal images, the proposed pipeline first increases the information density of each local view and then expands the spatial coverage. This coupled spectral–spatial design is the key to overcoming the conventional trade-off among field of view, resolution and modality richness.

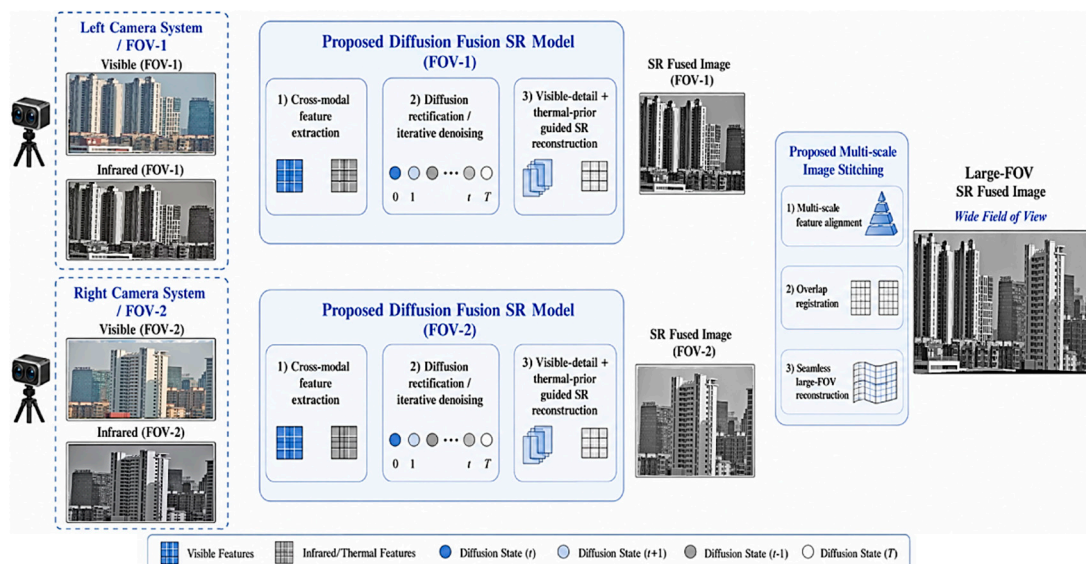


Figure 1. Principle of coupled spectral-spatial fusion-enabled multi-scale panoramic imaging.

2.2. Imaging System

The experimental prototype is shown in Figure 2. The system consists of two imaging units mounted on a stable optical platform. Their relative baseline and viewing directions are adjusted to provide two adjacent fields of view with sufficient overlap. This overlap is essential for subsequent feature matching and geometric registration. Too little overlap would lead to unstable correspondence estimation, whereas excessive overlap would reduce the effective gain in the field of view. The experimental configuration therefore balances panoramic coverage and registration reliability. The key parameters of the whole system are shown in Table 1.



Figure 2. Experimental setup of the dual-FOV visible-infrared imaging system.

Table 1. The key parameters of the system.

System	Sensor Resolution	Focal Length	Pixel Size	Overlapping Ratio
Value	2560×1920	80 mm	$2 \mu\text{m}$	20–30%

This hardware arrangement is not merely a conventional binocular setup. Each imaging unit is designed to provide visible and infrared observations from the corresponding scene region, enabling dual-modal information acquisition within each field of view. The visible image captures geometric structures and fine textures, while the infrared image supplies complementary radiative contrast. Such a configuration allows the system to enhance scene information in two dimensions: the binocular layout extends the spatial

coverage, and the visible–infrared pairing improves the information content of each local view. This provides the physical basis for the proposed coupled spectral–spatial panoramic imaging framework.

2.3. Visible–Infrared Fusion Super-Resolution Network

The visible–infrared fusion super-resolution network is presented in Figure 3. The input consists of a low-resolution visible image and a field-matched infrared image captured from the same view. The network begins with modality-specific shallow feature extraction. The visible branch focuses on high-frequency information, including edges, contours and fine textures, while the infrared branch emphasizes radiation-related contrast and target-saliency cues. These modality-specific features are then introduced into a cross-modal shallow coding and interaction module, where complementary information is exchanged in a shared latent space rather than being simply combined at the pixel level. Before visible–infrared fusion, spatial registration was performed to compensate for the differences in optical axis, image resolution, lens distortion, and sampling geometry between the visible and infrared cameras. First, intrinsic calibration and distortion correction were applied to each camera. Then, a set of corresponding points in the visible and infrared images was selected from stable structural features, such as building corners, window boundaries, and roof edges. A homography transformation was estimated using RANSAC to suppress mismatched correspondences. The infrared image was then warped into the visible-image coordinate system and resampled to the same resolution as the visible image. The registered visible–infrared pair was used as the input of the fusion super-resolution network.

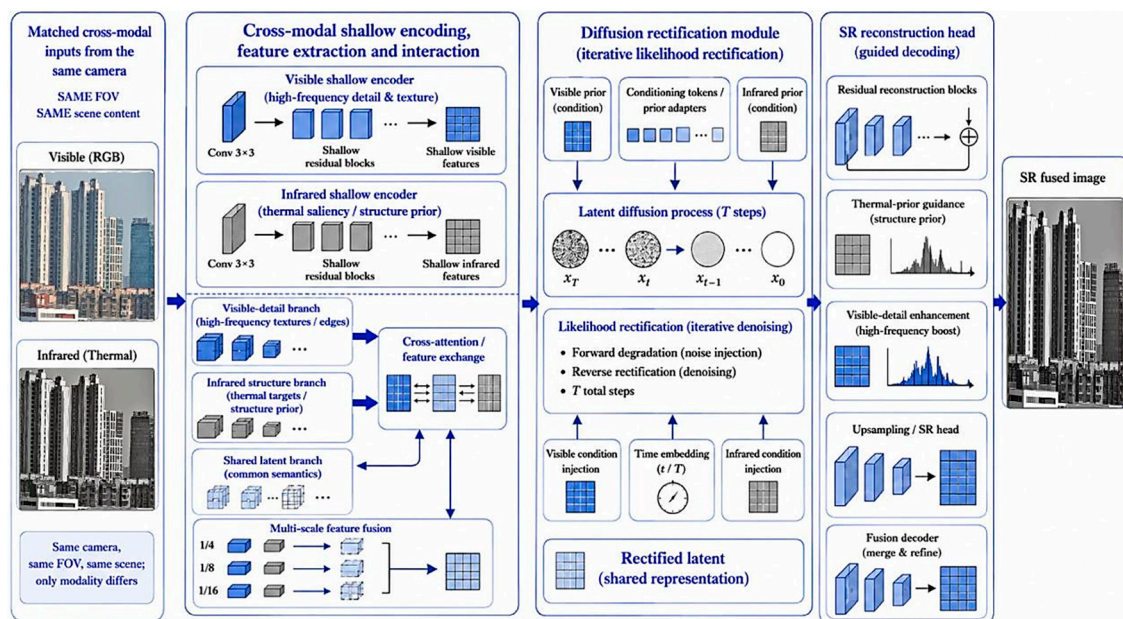


Figure 3. Coupled visible–infrared fusion super-resolution method.

The network further incorporates a diffusion rectification module to refine the fused latent representation. In this stage, visible features can constrain geometric boundaries and structural details, while infrared features guide the reconstruction of regions with weak visible contrast or unstable illumination. The diffusion process progressively suppresses noise and modality inconsistency, reducing the risk of over-smoothed details or visually sharp but unreliable artifacts. After latent refinement, the super-resolution reconstruction head decodes and upsamples the fused features to generate a high-resolution visible–infrared image. This output is not only a visually enhanced result; it also serves as a feature-rich and resolution-enhanced input for the following panoramic stitching stage.

Specifically, the super-resolution diffusion backbone used in this work was adapted from an unconditional model trained on the official ImageNet ILSVRC2012 Training Set. The training set contains approximately 1,281,167 natural-scene RGB images from 1000 categories, with an original download size of approximately 138 GB. During training, the images were center-cropped and downsampled/resized to 256×256 pixels. Although ImageNet contains category annotations, the model used in this work is unconditional, and no category labels were used as model inputs. The size of the model checkpoint used in this work is approximately 2.21 GB. These revisions clarify the origin, training data, and usage mode of the model.

2.4. Multi-Scale Panoramic Stitching Strategy

The multi-scale stitching strategy is summarized in Figure 4. The inputs to this module are the two super-resolved fused images generated from the left and right visible-infrared views. A multi-scale feature pyramid is first constructed for each input image. Coarser scales capture global scene structures and provide a stable basis for initial alignment, whereas finer scales preserve local edges and texture details for accurate refinement. This design is particularly important for urban scenes, where repeated windows, vertical building edges and similar façade patterns can easily produce false matches if only a single feature scale is used.

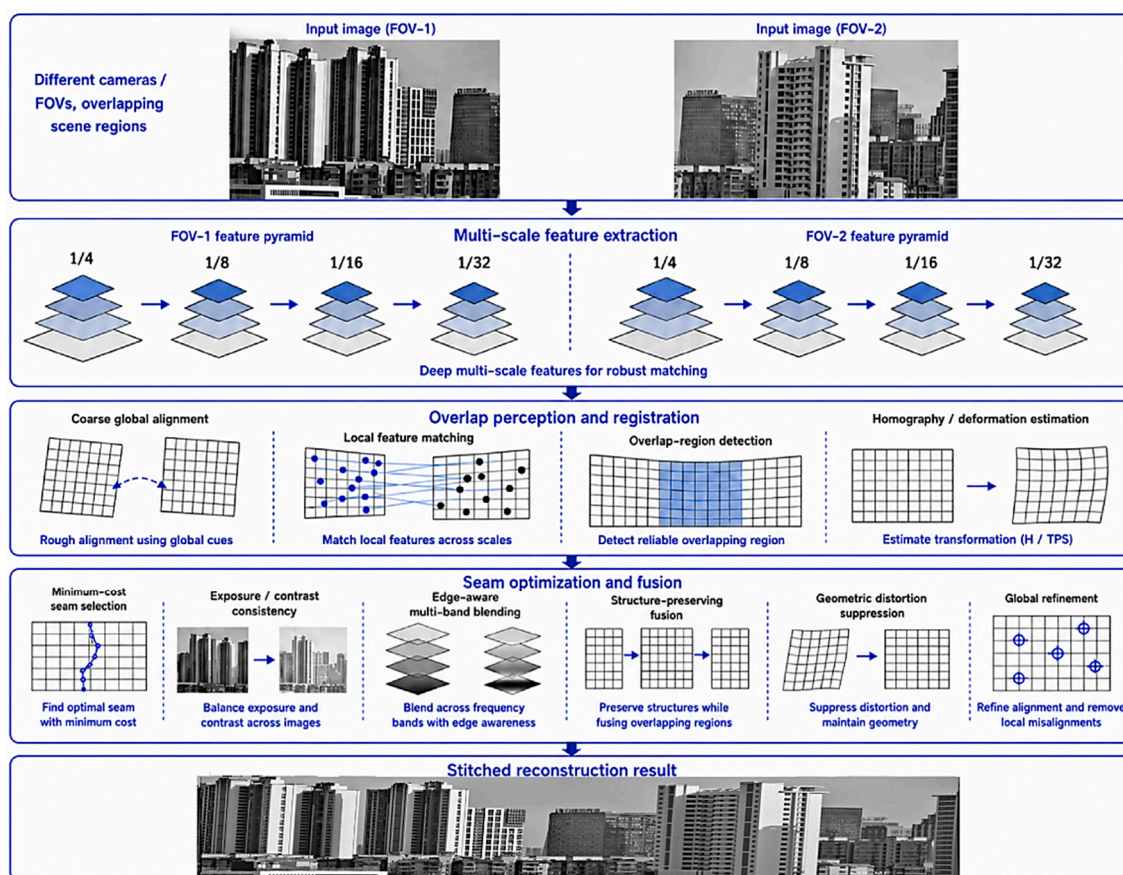


Figure 4. Multi-scale panoramic stitching workflow.

After feature extraction, the algorithm performs overlap perception and coarse-to-fine registration. The true overlapping region is identified first, so that unreliable matches from non-overlapping areas can be excluded. A coarse alignment then estimates the global geometric relationship between the two views, followed by local matching and deformation correction to compensate for viewpoint differences and residual parallax.

Once the two views are aligned, seam optimization, exposure consistency adjustment and edge-preserving fusion are applied to reduce ghosting, intensity discontinuity and blurred boundaries. The final output is a large-field super-resolution panoramic image that preserves the local detail enhanced by visible–infrared fusion while extending the observable scene range through multi-view spatial reconstruction.

3. Results

Figure 5 demonstrates the complete imaging and reconstruction process in a real outdoor scene. In Figure 5a, the first imaging unit records three results from the same field of view: the original visible image, the corresponding infrared image, and the fused image generated by the proposed visible–infrared super-resolution fusion model. The visible image preserves rich architectural textures and edge structures, whereas the infrared image provides complementary radiative contrast. After fusion, the reconstructed image exhibits enhanced structural clarity and improved local contrast, indicating that the model can effectively integrate texture information from the visible channel and robust contrast information from the infrared channel.

Figure 5b shows the results acquired by the second imaging unit. Similar to the first view, the three images from left to right correspond to the visible image, the infrared image, and the fused output. Although the two imaging units observe different spatial regions, their fields of view partially overlap, providing the basis for subsequent panoramic stitching. The fused image from the second view also shows clearer building contours and stronger detail responses than the single-modal inputs, which is essential for improving the reliability of feature extraction and cross-view registration. The feature extraction and matching results of the proposed multi-scale stitching method are presented in Figure 5c. The two fused images obtained from the two imaging units are used as the inputs for stitching. Instead of matching raw visible or infrared images directly, the algorithm extracts feature points from the super-resolved fused images, where structural details and local contrast have already been enhanced. This produces more reliable correspondences in the overlapping region and reduces the risk of mismatches caused by repetitive building patterns, weak textures, or modality-dependent contrast variations. The matching result confirms that the proposed fusion-before-stitching strategy provides a stronger feature basis for geometric alignment.

The final panoramic reconstruction is shown in Figure 5d. The two fused views are geometrically aligned and seamlessly stitched into a wider-field image. The yellow box marks a local region of interest, which is enlarged on the right together with the corresponding region from the original image. To further evaluate local detail preservation, the intensity variation along the dashed line is visualized. Compared with the original image, the proposed result shows a clear increase in local contrast and sharper intensity transitions. As a result, the Chinese characters in the selected region become much more legible in the reconstructed image, whereas they are difficult to distinguish in the original image. This comparison demonstrates that the proposed framework not only expands the field of view, but also improves the effective resolution and interpretability of fine local details. Figure 5e presents a direct comparison with the classical SIFT-based image stitching method. The SIFT-based baseline directly extracts and matches features from the original input images and then performs homography estimation and image blending. However, the stitched result exhibits obvious stitching seams and fusion discontinuities in the overlapping region. Local misalignment can also be observed around building edges and repeated façade structures, indicating insufficient stitching accuracy. In contrast, the proposed method performs visible–infrared fusion super-resolution before panoramic stitching. The enhanced local contrast and reconstructed structural details provide more

reliable feature correspondences for geometric registration, thereby reducing seam artifacts and improving the geometric continuity of the final panoramic image. This result verifies the effectiveness of the proposed fusion-before-stitching strategy.

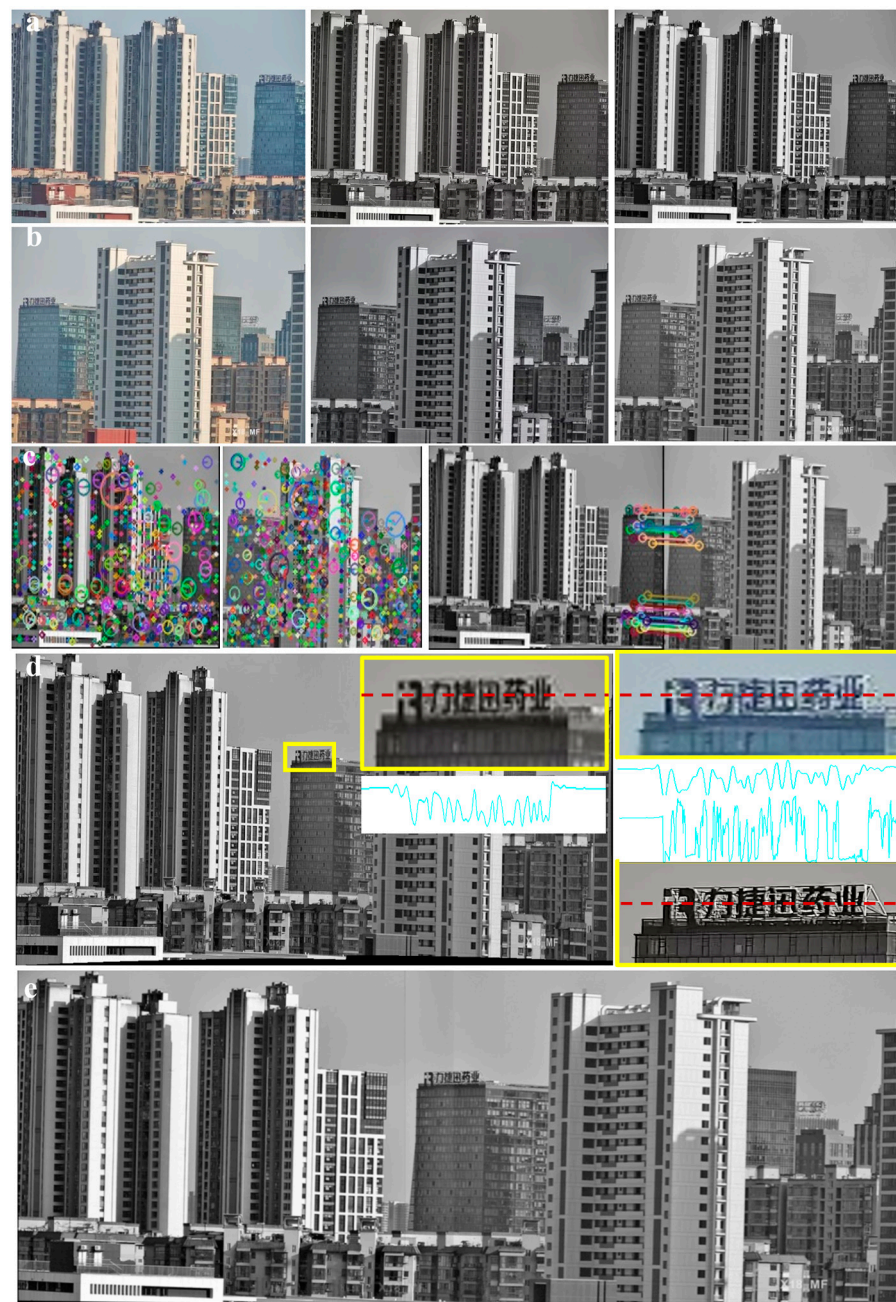


Figure 5. Experimental example of the proposed panoramic fusion framework in outdoor scene 1. (a) Results from FOV-1: original visible image, raw infrared image, and fused super-resolved result. (b) Results from FOV-2: original visible image, raw infrared image, and fused super-resolved result. (c) Feature detection and matching results obtained from the fusion-enhanced images. (d) Final panoramic reconstruction and local comparison of the original visible image, raw infrared image, and proposed reconstructed result, with the corresponding gray-value profile along the selected dashed line. (e) Comparison between classical SIFT-based stitching and the proposed method.

Figure 6 further evaluates the proposed framework in a more challenging outdoor scene. In Figure 6a, the first imaging unit captures three images from the same field of view: the visible image, the corresponding infrared image, and the fused result obtained by the proposed visible–infrared super-resolution fusion model. The visible image provides scene

structures and spatial textures, while the infrared image introduces complementary radiative contrast. After fusion, the reconstructed image shows improved structural definition and clearer local responses, indicating that the model effectively combines the spatial detail of the visible channel with the contrast information of the infrared channel.

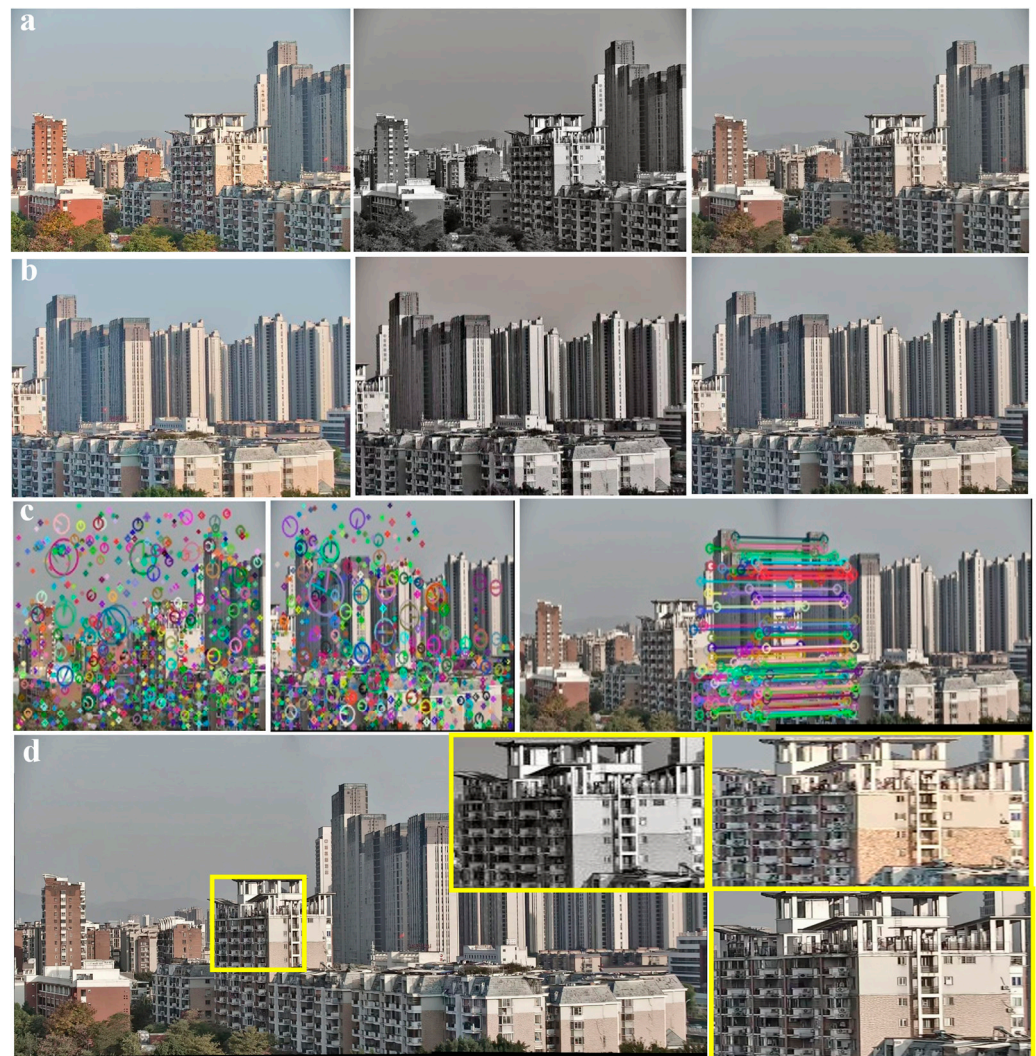


Figure 6. Experimental example of the proposed panoramic fusion framework in outdoor scene 2. (a) Results from FOV-1: original visible image, raw infrared image, and fused super-resolved result. (b) Results from FOV-2: original visible image, raw infrared image, and fused super-resolved result. (c) Feature detection and matching results obtained from the fusion-enhanced images. (d) Final panoramic reconstruction and local comparison of the original visible image, raw infrared image, and proposed reconstructed result.

Figure 6b presents the results from the second imaging unit. From left to right, the images correspond to the visible observation, the infrared observation, and the fused output. Although this view covers a different spatial region from that in Figure 6a, the two views share an overlapping area that enables subsequent panoramic reconstruction. The fused image exhibits sharper scene structures and more distinguishable local features than either single-modal input, providing a more reliable basis for feature extraction and cross-view registration.

The feature detection and matching results are shown in Figure 6c. The two fused images generated from the two imaging units are used as the inputs to the proposed multi-scale stitching algorithm. By extracting features from the fusion-enhanced images rather than from the original single-modal images, the algorithm obtains more stable keypoints

and more reliable correspondences in the overlapping region. This is particularly important in this scene, where repeated architectural patterns, weak local texture, and scale variations can easily lead to mismatches in conventional stitching pipelines.

The final fused and stitched panoramic result is presented in Figure 6d. The two super-resolved fused views are aligned and merged into a wider-field image with good geometric continuity. The yellow box marks the selected region of interest, which is enlarged on the right and compared with the corresponding region from the original image. The comparison clearly shows that the proposed method substantially improves image clarity. In the original image, the enlarged region exhibits obvious mosaic-like pixelation, making fine structures difficult to interpret. In contrast, the proposed result suppresses these pixelated artifacts and preserves smoother, more continuous structural details. This demonstrates that the proposed framework not only expands the field of view, but also improves the effective visual resolution of local regions in the final panoramic image.

For local contrast evaluation, the gray-value profiles along the selected dashed lines in Figures 5d and 6d were extracted and analyzed. The normalized contrast and mean gray gradient were calculated to quantify the gray-level separation and local edge-transition strength. In Scene 1, the normalized contrast increased from 0.536 in the original image to 0.939 in the proposed result, while the mean gray gradient increased from 9.62 to 28.54. In Scene 2, the normalized contrast increased from 0.700 to 0.741, and the mean gray gradient increased from 14.70 to 16.94. These results indicate that the proposed method enhances local contrast and improves fine-detail distinguishability.

For stitching robustness evaluation, we counted the detected keypoints and valid matched feature pairs in the overlapping regions. In Scene 1, approximately 588 keypoints were detected, and 32 valid matched feature pairs were obtained. In Scene 2, approximately 786 keypoints were detected, and 88 valid matched feature pairs were obtained. These feature correspondences provide a reliable basis for geometric transformation estimation and panoramic stitching. The detailed description is shown in Table 2.

Table 2. Quantitative analysis of local contrast and feature matching.

Scene	Detected Keypoints	Valid Matched Feature Pairs	Normalized Contrast Original	Normalized Contrast Proposed	Mean Gray Gradient Proposed	Mean Gray Gradient Original
Scene 1	~588	~32	0.536	0.939	28.54	9.62
Scene 2	~786	~88	0.700	0.741	16.94	14.70

4. Discussion

In this work, we proposed coupled spectral–spatial fusion-enabled multi-scale panoramic imaging, a dual-FOV visible–infrared framework for wide-field and high-resolution scene reconstruction. The system first acquires paired visible and infrared images from two adjacent views, then reconstructs high-resolution fused images through a visible–infrared super-resolution model, and finally merges them into a large-FOV panorama using a multi-scale stitching algorithm. Outdoor experiments demonstrate that the proposed method improves local contrast, enhances the readability of fine details such as distant text, suppresses mosaic-like degradation, and preserves geometric continuity after stitching. These results indicate that coupling spectral-domain fusion with spatial-domain panoramic reconstruction provides an effective route toward multimodal, wide-field, and detail-preserving computational imaging. Although the current validation mainly focuses on outdoor building scenes, the two tested scenes differ in illumination, thermal distribution, viewing distance, texture density, and structural complexity. The consistent improvement in both visual quality and quantitative indicators suggests that the proposed method has

good potential generalization ability for outdoor multimodal panoramic imaging. More diverse scenes, such as roads, vegetation, night-time scenes, and low-visibility environments, will be further investigated in future work.

5. Conclusions

The proposed framework shows that panoramic imaging should not be regarded merely as a geometric stitching task. In conventional systems, the final panorama is largely limited by the quality of the original inputs: single-modal images lack complementary information, low-resolution views lose small-scale details, and weak local contrast often leads to unstable feature matching. By introducing visible–infrared fusion and super-resolution reconstruction before stitching, each local view is first enhanced in both spectral content and effective spatial resolution. The visible modality provides structural textures and sharp boundaries, while the infrared modality supplies radiative contrast and more robust responses under complex illumination. This fusion-before-stitching strategy improves feature distinguishability in the overlapping regions, thereby benefiting multi-scale registration and reducing artifacts caused by repetitive building patterns, weak textures, or local pixelation.

Author Contributions: S.L., Z.W. and K.Y. were responsible for the whole experiment; K.Y. and S.L. wrote the manuscript; Y.L. carried out the structure design; Y.L. designed the optical system; X.Z. and G.B. conducted the project; S.L. and K.Y. were responsible for the image collection. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Science Fund for Distinguished Young Scholars, grant number 61925112, Natural Science Foundation of Fujian Province under Grant No. 2025J01464 and in part by the Start-up Fund for Scientific Research of Fuzhou University (Grant No. 511577), State key laboratory open fund of space information system and integrated application.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

FOV Field of view

References

1. Brady, D.J.; Gehm, M.E.; Stack, R.A.; Marks, D.L.; Kittle, D.S.; Golish, D.R.; Vera, E.M.; Feller, S.D. Multiscale gigapixel photography. *Nature* **2012**, *486*, 386–389. [[CrossRef](#)] [[PubMed](#)]
2. Szeliski, R. Image alignment and stitching: A tutorial. *Found. Trends Comput. Graph. Vis.* **2006**, *2*, 1–104.
3. Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
4. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–27 July 2017.
5. Chen, G.; Wang, Y.; Wang, H.; Peng, D. A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal. *IEEE Trans. Syst. Man Cybern. Syst.* **2022**, *52*, 936–953. [[CrossRef](#)]
6. Ding, J.; Xue, N.; Xia, G.-S.; Bai, X.; Yang, W.; Yang, M.Y.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7778–7796. [[CrossRef](#)] [[PubMed](#)]
7. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]

8. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-up robust features. *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. [[CrossRef](#)]
9. Brown, M.; Lowe, D.G. Automatic panoramic image stitching using invariant features. *Int. J. Comput. Vis.* **2007**, *74*, 59–73.
10. Zaragoza, J.; Chin, T.-J.; Tran, Q.-H.; Brown, M.S.; Suter, D. As-projective-as-possible image stitching with moving DLT. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013.
11. Ma, J.; Ma, Y.; Li, C. Infrared and visible image fusion methods and applications: A survey. *Inf. Fusion* **2019**, *45*, 153–178. [[CrossRef](#)]
12. Li, H.; Wu, X.-J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2019**, *28*, 2614–2623. [[CrossRef](#)]
13. Ma, J.; Yu, W.; Liang, P.; Li, C.; Jiang, J. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion* **2019**, *48*, 11–26. [[CrossRef](#)]
14. Ma, J.; Xu, H.; Jiang, J.; Mei, X.; Zhang, X.-P. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4980–4995. [[CrossRef](#)]
15. Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; Ma, Y. SwinFusion: Cross-domain long-range learning for general image fusion via Swin Transformer. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 1200–1217. [[CrossRef](#)]
16. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [[CrossRef](#)] [[PubMed](#)]
17. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017.
18. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–27 July 2017.
19. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Loy, C.C.; Qiao, Y.; Tang, X. ESRGAN: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision Workshops, Munich, Germany, 8–14 September 2018.
20. Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D.J.; Norouzi, M. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4713–4726. [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.