

Article

Evaluating Deep Learning Local Features for RGB-Thermal Image Matching and 3D InfraRed Thermography

Luca Morelli ¹, Neil Sutherland ², Francesco Ioli ³, Alfonso Vitti ⁴, Stuart Marsh ², Jon Mills ⁵, Paul Bryan ² and Fabio Remondino ^{1,*}

¹ 3D Optical Metrology Unit, Bruno Kessler Foundation, 38121 Trento, Italy; lmorelli@fbk.eu

² Nottingham Geospatial Institute, University of Nottingham, Nottingham NG7 2TU, UK; neil.sutherland@nottingham.ac.uk (N.S.); stuart.marsh@nottingham.ac.uk (S.M.); paul.bryan151@btinternet.com (P.B.)

³ Department of Civil and Environmental Engineering, Politecnico di Milano, 20133 Milano, Italy; francesco.ioli@polimi.it

⁴ Department of Civil, Environmental and Mechanical Engineering, University of Trento, 38121 Trento, Italy; alfonso.vitti@unitn.it

⁵ School of Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, UK; jon.mills@newcastle.ac.uk

* Correspondence: remondino@fbk.eu

Abstract

InfraRed Thermography (IRT), a non-invasive, non-contact, and non-destructive testing (NDT) technique, has become an established tool in the assessment of a building's behavior and energy performance. However, the inherent low spatial resolution of thermal infrared (TIR) cameras has led recent work to fuse thermographic and geometric data to generate accurate 3D representations of buildings encapsulating temperature information. Whilst existing data fusion methods have relied on sensors in fixed relative orientation (RO), the co-registration of independent TIR and RGB blocks using ground control points (GCPs), or the reprojection of TIR images onto additional geometric or parametric models, approaches that directly match multi-modal images remain limited. In principle, if multi-modal tie points were available, it would be possible to directly align the RGB block with the TIR block; however, such matching is extremely challenging due to the substantial differences in radiometric properties. The main contribution of this paper is to demonstrate the applicability of off-the-shelf deep learning-based image matching algorithms, originally trained on mono-modal datasets, to multi-modal matching tasks for InfraRed Thermography 3D-Data Fusion (IRT-3DDF). We conduct a comparative evaluation of the principal algorithms developed in recent years, with particular emphasis on 3D accuracy and computational efficiency, under the hypothesis that, owing to the inherently local nature of the problem they address, these algorithms can generalize from a mono-modal training domain to a multi-modal application domain. The results are benchmarked against existing hand-crafted open-source multi-modal reference methods. Importantly, the proposed method is fully-automatic, obviating the need for sensor pre-calibration, manual co-registration, or associated positioning information. Results demonstrate that DL-based image matching, using pre-trained neural networks outside of their expected training domain, provides a viable approach for IRT-3DDF capable of co-registering blocks of multi-modal images across varying scales, settings, sensors, and subjects. Our results indicate accuracy in 3D is up to seven times better than multi-modal hand-crafted algorithms, while hand-crafted mono-modal methods fail to co-register images in their entirety.



Academic Editors: Lidia M. Ortega Alvarado and María I. Ramos Galan

Received: 10 June 2026

Revised: 13 July 2026

Accepted: 20 July 2026

Published: 29 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.

Keywords: multi-modal image matching; InfraRed Thermography (IRT); 3D-data fusion; deep learning; SuperPoint; RIFT

1. Introduction

InfraRed Thermography (IRT), a non-invasive, non-contact, and non-destructive testing (NDT) technique, is a recognized instrument in the evaluation of a building's energy efficiency, thermal comfort, and structural integrity [1]. Producing rapid, real-time, and easily-interpretable imagery representative of the infrared radiation emitted by objects above absolute zero, IRT enables both qualitative and quantitative temperature data to be visualized in 2D thermal images [2]. However, whilst IRT has demonstrated its competency as a stand-alone NDT technique, a growing body of research has emerged looking to exploit novel photogrammetric computer vision and surveying engineering techniques to fuse complementary 2D and 3D datasets of contemporary [3] and historic [4] buildings. InfraRed Thermography 3D-Data Fusion (IRT-3DDF), an emerging field exploiting the growing ubiquity of TIR cameras and photogrammetric solutions, has spearheaded the generation of geometrically accurate 3D models encapsulating quantitative thermal information [5]. IRT-3DDF has enabled the segmentation of thermal leakages [6]; the evaluation of thermal comfort and energy efficiency [7]; and the real-time simulation of temperature changes [8], paving the way for future applications in building diagnostics and conservation.

However, thermal infrared (TIR) cameras, capturing radiation in the 7–14 μm spectral range, provide several challenges when processed photogrammetrically. Firstly, thermal images from consumer-grade TIR cameras have inherently low resolution with respect to RGB cameras, making it difficult to extract reliable features that can aid data fusion processes [9]. Furthermore, as most TIR cameras do not provide any direct positioning data, the orientation of resultant imagery is not provided for direct incorporation into subsequent modelling methods. To date, IRT-3DDF research has relied on several broad approaches for the fusion of TIR images with additional geometries: the co-registration of independently oriented TIR and RGB image blocks using ground control points (GCPs) [10] or iterative closest point (ICP) algorithms [11]; the use of sensors in fixed relative orientation to determine TIR poses [12]; or the fusion of TIR images with additional geometric [7] or parametric [13] 3D models. However, the effectiveness of these methods is often restricted by the following: the manual collection of topographic measurements and identification of GCP projections; initial manual co-registration between the TIR and RGB blocks prior to ICP; the inability for surveys to be undertaken independently; or not benefiting from complementary observations in both TIR and RGB image blocks. An alternative approach in IRT-3DDF is the automatic detection of multi-modal tie points between TIR and RGB image blocks. This enables the simultaneous orientation of all multi-modal images within a single adjustment, eliminating the need for a posteriori registration. A key advantage of this strategy is the potential to improve orientation accuracy by increasing the redundancy of observations across the two modalities, which would otherwise be limited when relying solely on a limited number of ground control points. However, the poor performance of traditional handcrafted mono-modal feature detectors and descriptors when applied to multi-modal imagery has so far limited the practical adoption of this approach [14].

Multi-modal image matching represents the process of establishing correspondences between images acquired with different types of sensors. Whilst the process of image matching is well-established for imagery captured in the visible spectrum, with research developments from the early 1980s through to the early 2000s resulting in algorithms capable of matching features invariant to scale, rotation, and perspective [15], the use of multi-modal data presents distinctive challenges for these traditional (often termed 'hand-crafted') matching algorithms (Figure 1). Feature matching across domains (i.e., TIR-RGB) relies on detecting, describing, and matching features accounting for non-linear radial distortion (NRD) and markedly different resolutions between image sources. As a result, this has limited the use of multi-modal image matching for photogrammetric processes,

forcing IRT-3DDF research to determine manual or semi-automatic methods to fuse thermal and visible images.

However, recent deep learning (DL) methods have provided new opportunities to explore multi-modal image matching through efficient and robust algorithms capable of determining correspondences across image domains [16]. As opposed to hand-crafted image matchers, DL approaches utilize convolutional neural networks (CNNs) and transformers to train (both independently and in conjunction) feature detectors and descriptors on extensive benchmark datasets commonly used in computer vision [17]. These benchmarks primarily consist of images with varying scales, illuminations, and poses, but generally lack multi-modal imagery [18–21]. These new DL-based algorithms, although not specifically trained on multi-modal datasets, exhibit a greater capability to identify distinctive features under substantially different radiometric conditions. We hypothesize that their enhanced discriminative power, acquired through training on datasets characterized by significant geometric and illumination variations, enables the establishment of reliable correspondences across multi-modal imagery. This, in turn, makes it possible to address IRT-3DDF through a multi-modal bundle adjustment framework. Given the recent availability of several DL-based architectures pre-trained on millions of images, and their demonstrated ability to generalize to challenging matching scenarios beyond their original training domain [15,17], this paper investigates the performance of a selection of these pre-trained models without task-specific fine-tuning. The results demonstrate their effectiveness for the proposed case study, suggesting that current state-of-the-art models are approaching a level of generalizability that reduces the need for dedicated retraining. This represents not only a practical advantage but also a more computationally and environmentally sustainable approach, as retraining increasingly large models from scratch—or even extensive task-specific fine-tuning—for every new application entails substantial computational costs, energy consumption, and associated carbon emissions. By exploring developments in DL image matching and assessing the utility of pre-trained neural networks applied to multi-modal images, the generation of 3D thermal models with fully automatic workflows can become an established practice for building diagnostics and conservation practices.

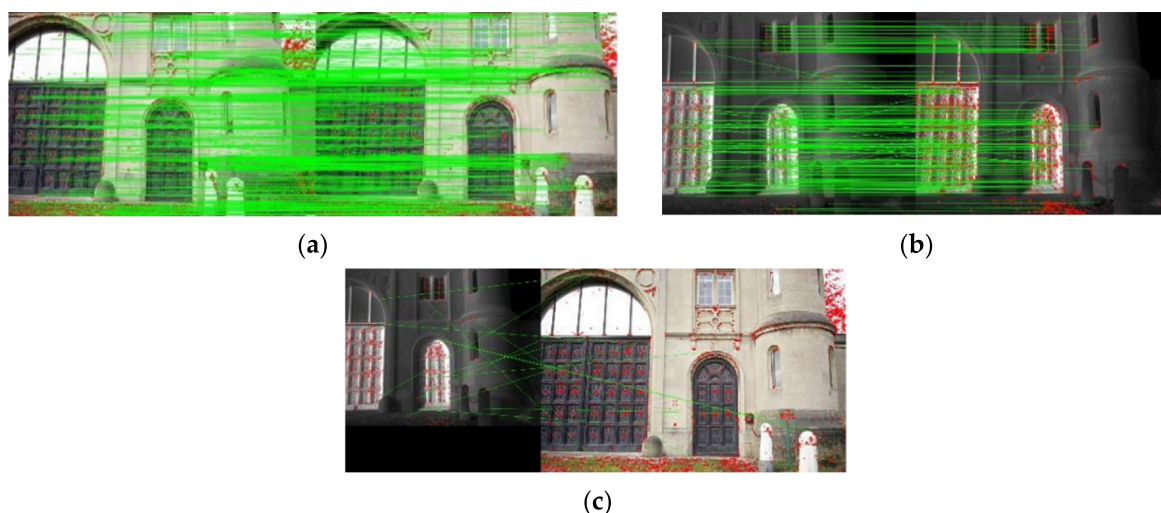


Figure 1. Comparing a hand-crafted image matching approach (SIFT feature extractor [22], nearest neighbor matching [23] and PyDEGENSAC geometric verification [24]) for (a) RGB-RGB, (b) TIR-TIR, and (c) TIR-RGB image pairs, with the number of matched features being 724, 192, and 13, respectively.

Aims

This paper hypothesizes that deep learning-based image matching algorithms trained on mono-modal datasets can generalize outside their training domain for multi-modal image matching between RGB and thermal images. This set of multi-modal tie points is then used to co-register the two multi-modal image blocks with a combined bundle block adjustment. For this purpose, off-the-shelf deep learning-based image matching algorithms with officially released pre-trained weights are considered.

Targeting thermal modeling of existing building applications, the proposed approach provides a robust and highly flexible IRT-3DDF workflow that can be readily applied across multi-modal, multi-temporal, and multi-sensor fusion scenarios without requiring fixed relative orientation between sensors, manually annotated ground control points, or simultaneous data acquisition.

The main contributions are:

- To evaluate the capability of pre-trained deep learning-based image matchers to establish reliable correspondences between TIR and RGB images, demonstrating their ability to generalize beyond the mono-modal domain on which they were trained;
- To benchmark several state-of-the-art DL-based algorithms against mono-modal and multi-modal hand-crafted image matching algorithms for TIR–RGB matching, highlighting the performance gap in terms of accuracy and computation time;
- To assess the applicability and accuracy of hand-crafted multi-modal local features for terrestrial photogrammetry, extending their evaluation beyond the remote sensing scenarios that have been the primary focus of previous studies;
- To demonstrate, through a combined bundle block adjustment, that accurate automatic co-registration of independent TIR and RGB blocks can be achieved using standard photogrammetric principles once reliable multi-modal tie points are available.
- Accuracy evaluation was performed using classical geomatics methodology based on the use of check points in the object space measured with millimeter-level topographical surveys.

This work complements the work introduced by [25] and the associated Deep-Image-Matching (DIM) library (<https://github.com/3DOM-FBK/deep-image-matching>, accessed on 15 December 2024) by extending the framework to multi-modal image matching. DIM has been significantly expanded to include additional local features and matchers with respect to previous work.

To the best of the authors' knowledge, this study is the first to provide rigorous photogrammetric validation in the 3D object space for multi-modal matching with off-the-shelf mono-modal algorithms, supported by millimeter-level check points derived from topographic survey data. This is a keypoint of the present work, as existing studies on deep learning features are often tested on small-scale images—e.g., those collected from the internet—or otherwise downsampled images, which are far from real-world applications. In addition, accuracy thresholds are less stringent than those typically required in photogrammetry, or also preferring pseudo-ground [26,27].

The structure of this paper is organized as follows: Section 2 reviews the state of the art in RGB–TIR multi-modal image block matching; Section 3 presents the proposed methodology, including the overall workflow, hardware setup, algorithms, and software implementation, as well as the datasets and evaluation metrics employed; Section 4 reports the experimental results; and Section 5 discusses the findings and their implications. Finally, Section 6 summarizes the main conclusions and outlines potential directions for future work.

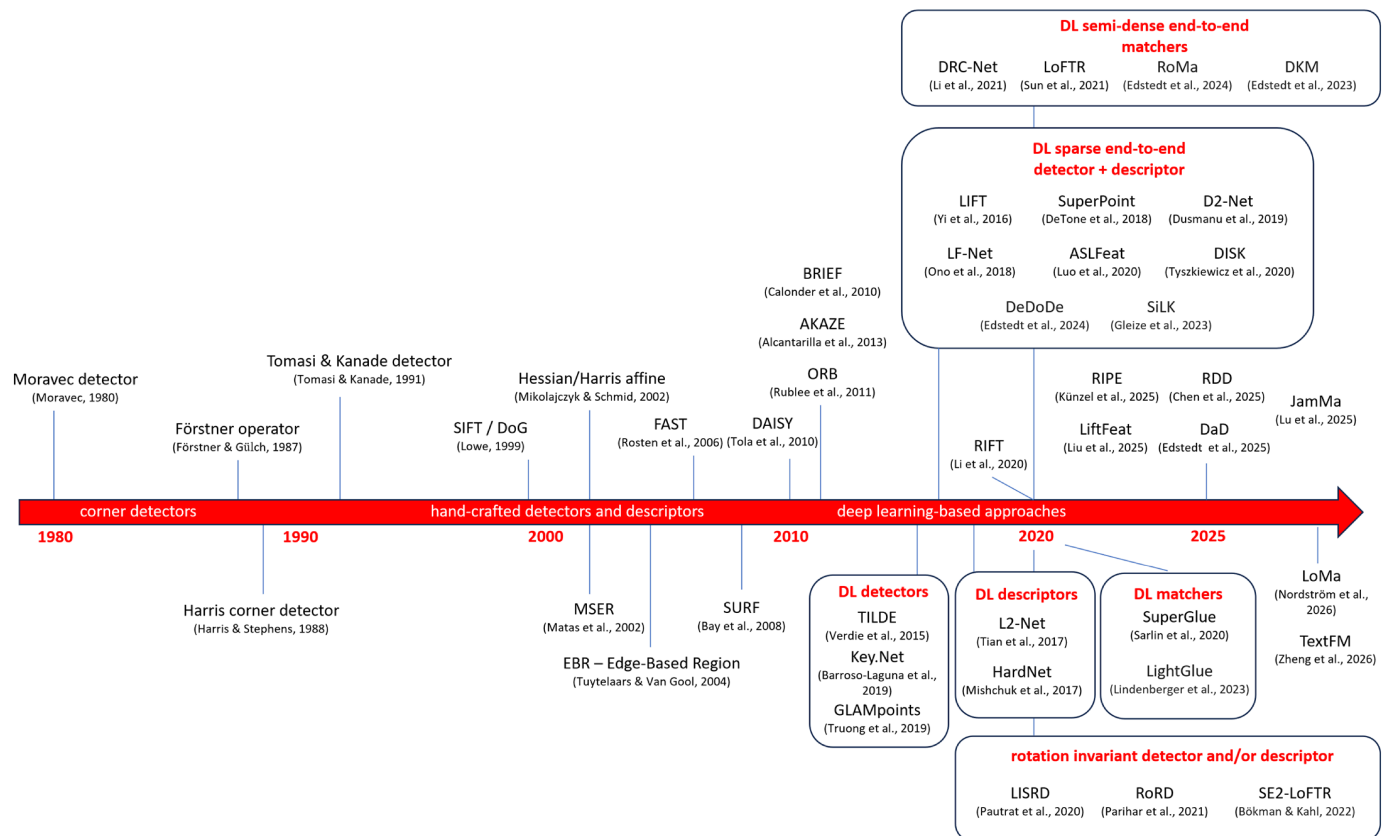


Figure 2. Timeline of image matching detectors, descriptors, and matchers from preliminary works, hand-crafted, machine learning, and deep learning local features [28–71]—adapted from [72].

2. State of the Art

Image matching is a fundamental component of photogrammetry and computer vision, with traditional hand-crafted methods providing well-established algorithms for image registration, 3D reconstruction, object detection, and localization [73]. Notably, SIFT (Scale Invariant Feature Transform), despite its limited invariance to illumination and affine transformations, marked a significant milestone as the first detector and descriptor demonstrating sufficient invariance for a broad spectrum of applications [22]. Even today, SIFT remains the benchmark feature matcher among non-DL-based approaches [26], and together with subsequent hand-crafted feature-based approaches such as SURF (Speeded-Up Robust Features) [36], ORB (Oriented FAST and rotated BRIEF) [39] and FAST (Features from Accelerated Segment Test) [40] is a robust method for achieving matches in mono-modal image datasets under limited radiometric and geometric distortions.

Nevertheless, whilst the aforementioned methods are adequate for the mono-modal matching of RGB-RGB and TIR-TIR image pairs [74], both intensity (detection) and gradient (description) information are sensitive to NRD between TIR and RGB images [75] (Figure 1). This has led to the development of several hand-crafted multi-modal image matchers specifically designed to be insensitive to NRD and applicable across multiple image modalities. Among these, RIFT (Radiation Invariant Feature Transform) [55] provides a SIFT-like multi-modal image matcher using frequency-domain information (phase congruency) and maximum index mapping to detect and describe keypoints. RIFT2 improves upon this by integrating rotation invariance using dominant index mapping in the feature description [76]. Similarly, SRIF (Scale and Rotation Invariant Feature Transform) [77] integrates a pyramid scale space strategy to project keypoints within scale space, enabling scale invariance for each image, and a local intensity binary transform (LIBT) to

provide greater structural information for feature description. Although these hand-crafted multi-modal image matchers provide suitable thermal-specific methods for IRT-3DDF [78], their ability to provide robust multi-view matches across challenging scales, poses, and baselines requires assessment. In fact, it is well established that mono-modal hand-crafted matching algorithms are unable to establish reliable correspondences between multi-modal images [55,77], whereas multi-modal hand-crafted approaches are computationally expensive [79]. Furthermore, the accuracy of multi-modal hand-crafted methods for terrestrial and close-range photogrammetry has not been thoroughly documented, as existing studies have primarily focused on remote sensing applications, where images are typically acquired from nadir viewpoints and exhibit limited depth variation. These limitations highlight the need for alternative solutions and for a rigorous validation of image matching methods in terrestrial photogrammetric scenarios, which limits the use of IRT-3DDF methods based on multi-modal tie points.

In recent years, the proliferation of DL approaches for image matching has sought to overcome the limitations of hand-crafted matching algorithms by using CNNs to build new detectors and descriptors (Figure 2). Detectors and descriptors can be trained independently, as KeyNet detector [43] and HardNet descriptor [54], or in an end-to-end manner, where SuperPoint [46] has been a key representative work, with a corner detector trained on synthetic elementary shapes and with a descriptor that shares the encoder part and is trained with homographic adaptation. Seminal works that followed are represented by: DISK [50], which formulates keypoint detection as a reinforcement learning problem to directly optimize matching performance; ALIKED [80], which introduces deformable descriptors and score maps for robust keypoint detection under severe appearance variations; and DeDoDe [51] that explores a foundation-model paradigm for local features by leveraging large-scale pretraining and dense feature learning to jointly optimize detection and description, producing highly repeatable and discriminative correspondences. In addition, feature matching, after local feature detection, can be trained in isolation using graph neural networks as a substitute for nearest neighbor matching and epipolar geometry verification, with SuperGlue [63] and LightGlue [64] presenting versatile matching solutions for RGB images. Their key feature is the ability to link descriptors while also taking into account the 2D distribution of keypoints, thereby increasing the number of matches found even in the presence of repetitive patterns. Conceptually, any combination of hand-crafted and learning-based components (i.e., detector, descriptor, matcher) is feasible for the determination of correspondences [26]. These networks are pre-trained on extensive mono-modal datasets with varying illumination, perspective changes, and wide baselines [42,43,54], which, when used for multi-modal image matching, are applied beyond their original mono-modal training domain.

Whilst the paper focuses on RGB-TIR images, multi-modal image matching embraces a much wider variety of images characterized by inherent geometric, radiometric, and spectral differences, e.g., RGB vs. LiDAR or RGB vs. SAR. Learning-based methods have presented useful tools capable of matching features across image domains [16]. In particular, pre-trained mono-modal DL-features have started to be applied to multi-modal image matching, even if 3D validation is still missing. PhotoMatch [17], a multi-modal image matching software, demonstrates the application of DL image matching methods with ‘detect-and-describe’ approaches like D2-Net [47] and R2D2 [81], both superior to hand-crafted methods in six multi-modal scenarios (inc. scanning electron microscopy, infrared, LiDAR, aerial, and magnetic resonance images). Similarly, ref. [82] presents a novel lightweight network built to exploit local and global feature information for infrared-visible image matching, achieving competitive results with RIFT, R2D2, SuperPoint + LightGlue, and LoFTR [60]. These cases demonstrate the success of ‘out-the-box’ DL feature detec-

tors and matchers in 2D scenarios (i.e., only matching TIR-*RGB* images); however, the application of DL neural networks for IRT-3DDF remains open for investigation.

Importantly, the use of multi-modal image matching for IRT-3DDF and methods looking to use oriented images for 3D reconstruction point to several critical conclusions. Firstly, hand-crafted local features, although successful mono-modally [11], are unsuitable for comprehensive multi-modal image matching [83], incapable of providing a suitable volume of verified matches for successful 3D reconstruction. Secondly, image orientation through combined bundle block adjustments represents a viable solution when correspondences are determined either manually [10], semi- [14] or fully –automatically [84]. Finally, DL multi-modal image matching approaches, increasingly robust to NRD and perspective changes, represent the logical evolution of IRT-3DDF approaches across sensors, scales, and subjects. With increasing access, availability, and applications for DL image matchers, the ability for multi-sensor correspondences to be used within a single bundle adjustment has become well-established, utilized for 2.5D TLS [85], historical [15] and day-and-night [86] images. However, the development of methods specific to TIR images appreciating the unique characteristics of infrared radiation, specific use cases, and desired applications, is required for the complete assessment of multi-modal image matching for IRT-3DDF [87].

3. Methodology

The proposed workflow introduces learning-based image matching for IRT-3DDF, demonstrating the utility of pre-trained DL feature matching algorithms for multi-modal 3D reconstruction. To this end, this paper compares the performance of several off-the-shelf hand-crafted and DL-based approaches to evaluate their versatility for simultaneous multi-modal image orientation. Importantly, although insights into the radiometric properties of TIR cameras are highlighted, the focus of this work remains on the orientation of multi-modal imagery and photogrammetric reconstruction.

Specifically, the methodology proposed for IRT-3DDF consists of directly establishing sparse correspondences between the *RGB* and TIR image blocks through a brute-force matching strategy. This approach enables the identification of *RGB*–*RGB*, TIR–TIR, and *RGB*–TIR correspondences. The availability of robust and locally accurate correspondences, at pixel or sub-pixel precision, is essential to ensure the accuracy of the final orientation of the image blocks.

The key stage of correspondence extraction is addressed through a comprehensive evaluation of existing matching algorithms belonging to two main categories: hand-crafted methods (both mono-modal and multi-modal) and learned multi-modal approaches. The main contribution of this work is to demonstrate that developing new deep learning-based matching algorithms is not necessary. Instead, the generalization capability of state-of-the-art mono-modal deep learning methods proves to be sufficiently robust and accurate for geomatics applications, particularly in terrestrial photogrammetry. A broad range of methods is systematically compared, encompassing the most significant recent advances in the field, including independently trained detector-and-matcher pipelines, end-to-end self-supervised approaches, methods optimized for computational efficiency, reinforcement learning-based techniques, foundation models, and sparse, semi-dense, and dense matching methods.

The extracted correspondences are subsequently used to orient the two image blocks through an incremental Structure-from-Motion (SfM) pipeline based on an iterative process of triangulation and resection, alternated with local and global bundle adjustments. Owing to the limited scale invariance exhibited by current deep learning-based matching methods, the *RGB* images are initially downsampled to facilitate robust matching, while the lost spatial information is subsequently reintroduced in a dedicated refinement stage.

A further distinctive contribution of this work is the acquisition and open-source release of two terrestrial photogrammetric datasets representing civil engineering buildings, both provided with millimeter-level ground truth. The accuracy of the proposed methodology, whose implementation details are described in the following sections, is assessed using different matching algorithms and quantitatively evaluated through established photogrammetric accuracy assessment procedures. The two datasets are freely-accessible and open-sourced within the Thermitage repository [<https://github.com/3DOM-FBK/Thermitage>] (accessed on 19 July 2026) [87].

3.1. Case Studies

To evaluate multi-modal image matching for IRT-3DDF, two case studies of contrasting subjects, scales, and settings were selected. Firstly, to assess the implementation of IRT-3DDF for energy efficiency and thermal comfort applications, a 21st-century contemporary dwelling in Warwickshire, UK, was employed (Figure 3a). Secondly, Lenton Lodge Gatehouse, a 19th-century Grade II* listed property, was used as a heritage case study in Nottingham, UK (Figure 3b). The structure, built by Jeffry Wyattville and largely comprising Mansfield sandstone, represents one of the finest examples of Elizabethan Revival in England.



Figure 3. (a) Dunrobin House, Warwickshire, UK; (b) Lenton Lodge Gatehouse, Nottinghamshire, UK.

3.2. Cameras

For TIR image capture, a Workswell WIRIS Pro (WWP) UAV thermal camera was configured to be used terrestrially (Workswell s.r.o, 2021). The WWP features an uncooled vanadium oxide (VOx) focal plane array (FPA) sensor with a resolution of 640×512 px, capturing infrared radiation in the $7\text{--}13.5\ \mu\text{m}$ spectral range. Although the WWP features both thermal and visible sensors fixed within the same camera housing, the latter has a limited image resolution of 1920×1080 px. Therefore, a Sony $\alpha 7\text{R II}$ (Sony Group Corporation, Tokyo, Japan) full-frame mirrorless camera with a resolution of 7952×5304 px was used to capture higher-quality RGB images. Notably, as the two cameras used for TIR and RGB acquisitions were not fixed in relative orientation, the proposed method does not assume a fixed baseline between the two cameras, enabling images to be captured asynchronously. Table 1 summarizes the key geometric and radiometric properties of both the WWP and Sony $\alpha 7\text{R II}$.

Table 1. Camera specifications for the Workswell WIRIS Pro (WWP) and Sony $\alpha 7\text{R II}$ (RGB).

Workswell WIRIS Pro (WWP)	
Resolution	640×512 px
Sensor Size (FPA)	10.88×8.71 mm
Pixel Size	$17.00\ \mu\text{m}$

Table 1. *Cont.*

Workswell WIRIS Pro (WWP)	
Nominal Focal Length	13.00 mm
IR Spectral Range	7.5–13.5 μm
Temp. Sensitivity	0.05 $^{\circ}\text{C}$
Temp. Accuracy	± 2 $^{\circ}\text{C}$
Sony $\alpha 7\text{RII}$ (RGB)	
Resolution	7952 $\times$ 5304 px
Sensor Size (CMOS)	35.90 $\times$ 24.00 mm
Pixel Size	4.50 μm
Nominal Focal Length	35.00 mm

3.3. Image Acquisition and Pre-Processing

The underlying assumption of the proposed method, given its reliance on multi-modal matches, is the presence of high contrast and low noise in both the TIR and RGB datasets for successful matching to be achieved. In both cases, the aim is to capture local features in the images that are repeatable, distinctive, and unique [88]. To mitigate the low resolution of the TIR images, image strips were captured with 90% horizontal and 70% vertical overlap (Figure 4). For Dunrobin House, a fixed TIR ground sampling distance (GSD) of 9.8 mm was used for all images, whilst for Lenton Lodge Gatehouse, successive strips at different distances were captured, with GSD ranging between 9.8 mm and 26.2 mm.

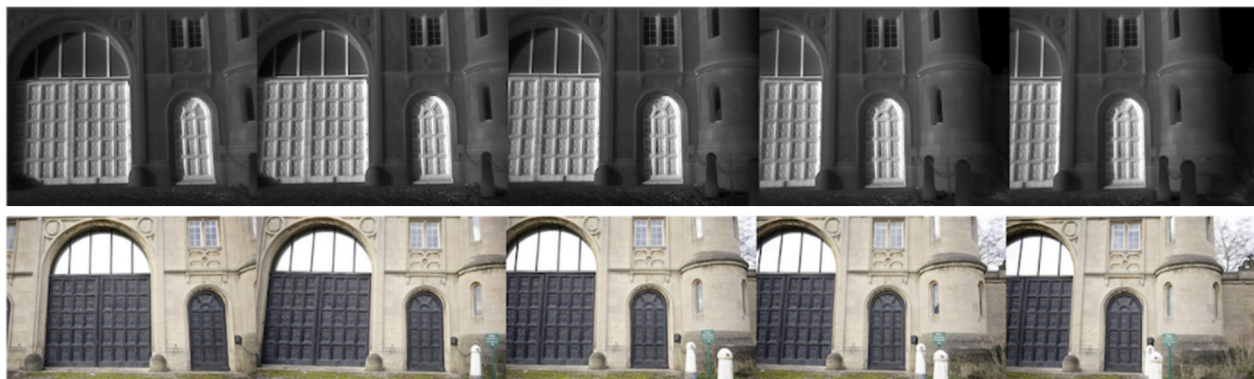


Figure 4. Example image strips for TIR (**top**) and RGB (**bottom**) images of Lenton Lodge Gatehouse.

To increase the number of detectable features between TIR and RGB images prior to image matching, the WWP's 16-bit TIR images were imported into FIJI (v2.16.0) (FIJI Is Just ImageJ v2.16.0) and enhanced with an UnSharp Masking filter [89]. This procedure significantly increased the number of matches extracted by both hand-crafted and DL-based matching algorithms [16,90]. To facilitate effective multi-modal matching, RGB images were downsampled from their original resolution (7952 $\times$ 5304 px) to 1/8th of their original size (994 $\times$ 663 px), with 1/8th resolution providing the most matches across initial investigations compared to $\frac{1}{2}$, $\frac{1}{4}$, and $\frac{1}{8}$ of the original resolution. Therefore, the GSD of the RGB images is required to be comparable to that of the TIR images (see Table 2). Notably, none of the tested DL approaches found keypoints by directly matching TIR and full-resolution RGB images before resampling, showing a limited scale invariance of the DL local features tested [91].

For both structures, a network of GCPs and check points (CPs) was acquired using custom thermal-specific survey markers to validate the accuracy of the proposed pipeline. For each marker, 3D coordinates were measured using a Leica TS10 total station (Precision:

angular: 0.1 mgon; range: 0 m–500 m = 2 mm + 2 ppm). To guarantee the best accuracy in GCP/CP positioning, projections of the targets in RGB images were annotated on the full-resolution images, subsequently downsampled to match the $\frac{1}{8}$ resolution of the images processed in Section 3.5. Details of both surveys, including the acquired camera networks, GCP/CP network, and survey statistics, can be seen in Figure 5 and Table 2.

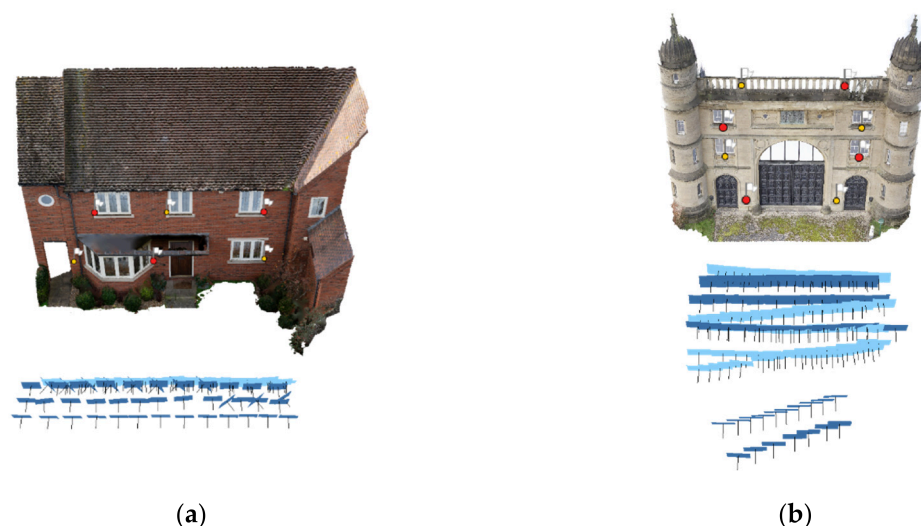


Figure 5. TIR (light blue) and RGB (blue) camera networks for (a) Dunrobin House and (b) Lenton Lodge Gatehouse, including the allocation of GCPs (red) and CPs (yellow) across both building facades.

Table 2. Data acquisition statistics for both TIR and RGB data sets.

Statistics	Case Study	
	Dunrobin House	Lenton Lodge Gatehouse
No. TIR Images	45	90
TIR res. (px)	640 × 512	640 × 512
Min. TIR GSD (mm)	9.8	9.8
Max. TIR GSD (mm)	9.8	26.2
No. Full RGB Images	50	55
Full RGB res. (px)	7952 × 5304	7952 × 5304
Full RGB Min. GSD (mm)	1.0	1.0
Full RGB Max. GSD (mm)	1.3	2.6
Downsampled RGB res. (px)	994 × 663	994 × 663
Downsampled Min. RGB GSD (mm)	8.0	8.0
Downsampled Max. RGB GSD (mm)	10.4	20.8
No. Image Pairs (Bruteforce)	4465	10,440
No. GCPs	3	4
No. CPs	3	4

3.4. Image Matching

3.4.1. Learning-Based Image Matching: Deep-Image-Matching

Deep-Image-Matching (DIM), a flexible, open-source Python (v3.11.9) library designed for robust multi-view image matching, was used and extended for the implementation of the multi-modal image matching [25]. DIM provides front-end access to a wide range of state-of-the-art computer vision algorithms that determine correspondences across unordered image sets. Subsequently, this provides the essential foundation for bundle adjustments and accurate photogrammetric reconstruction. DIM employs a modular

workflow designed for effective multi-view matching, including pipeline selection, image pair selection (matching strategy), and image pre-processing (Figure 6). Importantly, DIM does not perform either the bundle adjustment or 3D scene reconstruction itself, but exports matches in formats compatible with various photogrammetric software. Whilst applications of DIM for RGB images have been previously presented [85,92], this is the first investigation on multi-modal images.

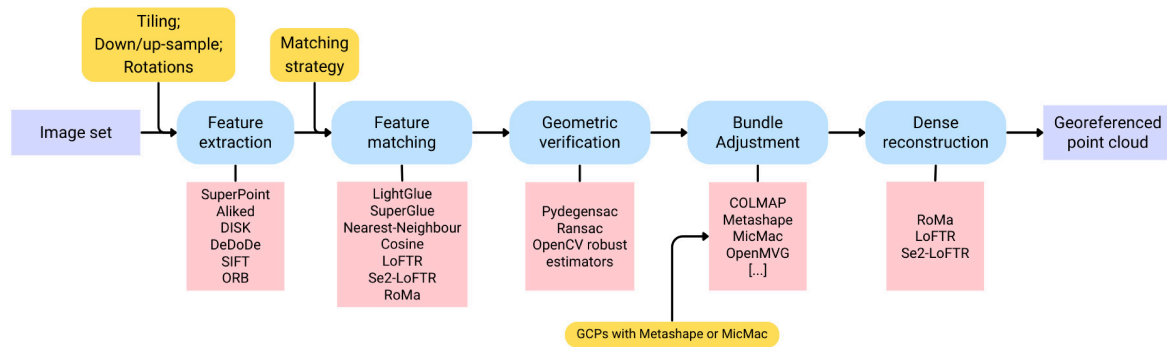


Figure 6. Deep-Image-Matching (DIM) modules and workflow.

As DIM supports various combinations of local feature extraction and matching algorithms, various hand-crafted and DL-based models were tested to compare their respective performances. The following DL feature extractors and matchers were included from DIM: (1) SuperPoint feature extractor [46] combined with SuperGlue matcher [63]; (2) SuperPoint, ALIKED [80] and DISK [50] feature extractors combined with LightGlue matcher [64]; and (3) DeDoDe extractor [51], KeyNet detector [43] and a HardNet descriptor [54] combined with the nearest neighbor (NN) matcher. In addition, the comparison was extended to include the use of LoFTR [60], a semi-dense method with a featureless approach. The family of semi-dense methods has expanded in recent years, but since, as will be seen below, this method is computationally expensive and less accurate than the sparse local features, it was decided to limit the analysis to this one, as a representative of the entire family.

Let an unordered set of images acquired from multiple modalities be

$$I = \{I_1, I_2, \dots, I_N\}, I_i : \Omega_i \subset \mathbb{R}^2 \rightarrow \mathbb{R}^C \quad (1)$$

where C depends on the modality, i.e., $C = 3$ for the RGB and $C = 1$ for the TIR images. For each image I_i , a sparse feature extractor ϕ produces a set of keypoints and descriptors

$$\phi_{sparse}(I_i) = \left\{ (x_{ik}, d_{ik}) \mid x_{ik} \in \Omega_i, d_{ik} \in \mathbb{R}^D \right\}_{k=1}^{K_i} \quad (2)$$

with $K_i \leq 8000$, and where x_{ik} are 2D keypoint coordinates, d_{ik} are learned or hand-crafted descriptors, and ϕ is one of the local features in Sections 3.4.1 and 3.4.2. The number of keypoints K_i directly affects camera pose accuracy. Therefore, 8000 features per image have been chosen since it is a standard in the field (given the order of magnitude of the image size used in the experiments) [26]. Pairwise matching for an image pair (I_i, I_j) where for brute-force matching pairs are defined as $P = \{(i, j) \mid 1 \leq i < j \leq N\}$, computes tentative correspondences as

$$M_{ij} = \Psi(\phi_{sparse}(I_i), \phi_{sparse}(I_j)) = \left\{ (x_{ik}, x_{jl}) \mid (k, l) \in \mathcal{A}_{ij} \right\} \quad (3)$$

where $\mathcal{A}_{ij}$ denotes the set of matched descriptor indices by a learning-based matcher or nearest neighbor (NN) operator Ψ .

The DIM ‘bruteforce’ matching strategy was used to exhaustively match all images in the combined TIR-RGB image blocks, resulting in 4465 and 10,440 pairs for Dunrobin House and Lenton Lodge Gatehouse, respectively (Table 2). For each DIM pipeline, a maximum of 8000 keypoints K_i was extracted per image, a quantity commonly used in local features and matcher accuracy assessment for similar image sizes [26], and an important parameter to be kept fixed across different matching algorithms, since it directly affects camera pose accuracy. Additionally, DIM settings remained consistent across each pipeline, including: no tiling, highest quality image sampling, and DEGENSAC geometric verification [93].

Each correspondence set is therefore geometrically verified by estimating the fundamental matrix F_{ij} as $x_{jl}^T F_{ij} x_{ik} = 0$ and retaining only inliers satisfying $d_{epi}(x_{ik}, x_{jl}, F_{ij}) < \tau$, $\tau = 4$ px. Since COLMAP [94] is used for incremental SfM, the geometric filtering parameter for matches in DIM was set to COLMAP’s default value. This yields the inlier set $\widehat{M}_{ij} \subset M_{ij}$. Subsequently, the mono-modal and multi-modal tie points generated from DIM were exported into a SQLite database compatible with COLMAP.

The original DIM implementation has been extended to support additional local feature extractors and matchers, specifically DeDoDe, SRIF, LoFTR, and RoMa [61]. In particular, LoFTR and RoMa have been adapted for multi-view image matching, as their original implementations are designed for two-view matching and produce pair-dependent keypoint indexing. To address this limitation, keypoints are clustered based on their 2D image coordinates to ensure consistent indexing across multiple image pairs.

Therefore, while for sparse local features $x_{ik} \in \Omega_i$, for semi-dense methods, local features depend on the specific image pair for which they have been computed, and can be written as

$$C_{ij} = \phi_{semi-dense}(I_i, I_j) = \left\{ \left(x_{im}^{(ij)}, x_{jm}^{(ij)} \right) \right\}_{m=1}^{M_{ij}} \quad (4)$$

Consistent indexing across multiple views is obtained by clustering keypoints in image space. For each image I_i , define the union of all matched points:

$$x_{ik} = \mathcal{C} \left(\bigcup_{j \neq i} \{x_{im}^{(ij)}\}_{m=1}^{M_{ij}} \right) \quad (5)$$

through the operator $\mathcal{C}$ that groups nearby points within a distance of 1 px. A bin sized 1 pixel was selected to preserve the semi-dense behavior of the matcher. Although a smaller bin size would have yielded sub-pixel localization accuracy, it would also have significantly reduced the number of extracted tie points. As demonstrated in the following sections, this trade-off does not adversely affect the overall orientation accuracy as shown by SuperPoint, which predicts keypoints at pixel rather than sub-pixel precision, yet it consistently achieves the highest levels of accuracy (see experiment section). Following preliminary experiments, RoMa was excluded from the analysis due to its high computational cost and the excessive number of dense tie points, which hindered SfM processing.

3.4.2. Hand-Crafted Image Matching: RIFT2, SRIF, SIFT

To provide a benchmark for comparison with the DL matching algorithms, two hand-crafted matchers specifically designed for multi-modal images were included: RIFT2 and SRIF. RIFT2 was configured in MATLAB (R2024b) to generate local features and geometrically verified matches (4 px), including: (1) the number of convolution orientations of the log-Gabor filter ($N_O = 6$); (2) the number of scales at which the log-Gabor filter is generated ($N_S = 4$); and (3) the patch size of the feature descriptor ($J = 96$). Similarly, for SRIF, configuration settings were implemented for the patch size ($J = 128$), grid size ($N_{grid} = 8$) and histogram bins ($N_{bin} = 4$) with geometric verification set at 4 px. Finally, as a hand-crafted method for mono-modal matching, SIFT + NN matching was included, since is still considered the benchmark among traditional matching algorithms [26]. For

each hand-crafted approach, all features and matches for subsequent image orientation were exported in a database compatible with COLMAP.

3.5. Combined TIR-RGB Bundle Block Adjustment

All geometrically verified multi-modal matches were imported into COLMAP for the Combined TIR-RGB Bundle Block Adjustments (CBBA). As WWP's FPA sensor is assumed to follow the Brown model [95], initial nominal values for the interior orientation ($f, c_X, c_Y, k_1, k_2, p_1, p_2$) of both sensors were provided for self-calibration. Images were incrementally oriented using a resection-intersection approach to generate sparse 3D point clouds for each matching procedure. Notably, this process co-registers both image blocks simultaneously through shared observations of local features using both mono-modal (i.e., TIR-TIR, RGB-RGB) and multi-modal (i.e., TIR-RGB) image pairs. As a result, 2D image coordinates, 3D tie points, individual camera poses, and associated intrinsics were generated. Finally, camera networks for each of the multi-modal image matching approaches were exported using the Bundler format [96] into Agisoft Metashape (Pro v:2.2.0) [97] for referencing using GCPs, metric evaluation on marked CPs, and the constrained bundle adjustments (see Section 3.6). Thus, the multi-modal bundle adjustment (BA) takes its classical form, where the optimization jointly refines the camera parameters and the 3D point coordinates using multi-modal observations. Specifically, $P_i = [R_i | t_i]$ denotes the extrinsic parameters (camera pose) of the i -th image, composed of the rotation matrix R_i and translation vector t_i . K_i represents the intrinsic calibration matrix of the corresponding camera, while X^m is the three-dimensional coordinate of the m -th scene point in the world reference frame. The objective function minimizes the reprojection error over all available multi-modal observations:

$$E = \sum_m \sum_{i \in V^m} \left\| x_i^{(m)} - \pi(P_i K_i X^m) \right\|^2 \quad (6)$$

where $\pi(\cdot)$ is the camera projection model, V^m is the set of images observing the 3D point m , and observations include RGB-RGB, TIR-TIR, and RGB-TIR. As visible in the software structure reported in Figure 6, incremental SfM can be run directly in the DIM library, which uses pycolmap as a backend for SfM.

3.6. Constrained Bundle Adjustment

As previously stated, due to the limited scale invariance of DL local features, the current workflow does not determine matches using the full-resolution RGB images; instead, it determines multi-modal correspondence from the downsampled images ($1/8$ th res.). Moreover, the downsampled RGB images are insufficient for the generation of an accurate 3D geometry in eventual multi-view-stereo procedures. Therefore, to implement a data fusion method that maximizes the potential of both datasets, RGB camera poses were estimated using the original full-resolution RGB images in a separate bundle adjustment. For each existing multi-modal camera network, these full-resolution RGB poses were introduced as known reference and constrained in an additional bundle adjustment. This refined the poses of the TIR-RGB blocks, allowing each to benefit from the information of the full-resolution RGB images. Therefore, let $P_i^{HR} = (R_i^{HR}, t_i^{HR})$ be the camera poses independently estimated from the full-resolution RGB images. The multi-modal bundle is subsequently refined by minimizing:

$$E = \sum_m \sum_{i \in V^m} \left\| x_i^{(m)} - \pi(P_i K_i X^m) \right\|^2 + \lambda_t \sum_{i \in RGB} \left\| t_i - t_i^{HR} \right\|^2 + \lambda_r \sum_{i \in RGB} d_R(R_i, R_i^{HR})^2 \quad (7)$$

where $d_R(\cdot)$ denotes the rotation distance, and the weighting terms are derived from the a priori standard deviation ($\sigma_{xyz} = 1$ mm; $\sigma_{\omega \phi \kappa} = 0.1$ deg). These values should be

compared with the GSD of the datasets, and their purpose is to force the poses into the position derived from the independently oriented RGB block. A further reduction in the a priori precision parameters does not result in a significant change in the pose following the multi-modal BA.

3.7. Evaluation

As highlighted by [26], accuracy assessments for image matching and SfM applications should occur downstream of the entire reconstruction pipeline, not solely relying on image metrics as a means of evaluation. Therefore, to determine the efficacy of multi-modal image matching for IRT-3DDF, assessments were undertaken using several metrics. To assess the performance of multi-modal image matching, the number of triangulated 3D tie points, Mean Reprojection Error (MRE), and inlier match ratio (the ratio between successfully triangulated tie points and the number of raw matches) were utilized [15]. For the assessment of the derived camera poses in 3D space, the Root Mean Squared Error (RMSE) on CPs was employed, split by each image modality: $RMSE_{TIR}$ and $RMSE_{RGB}$. Finally, a comparison between the ‘Original’ combined TIR- RGB bundle block adjustments (Section 3.5) and the ‘Constrained’ bundle adjustments (Section 3.6) is provided to determine the suitability of re-introducing the full-resolution RGB images.

The Mean Reprojection Error (MRE) is computed from the reprojection residuals:

$$MRE = \frac{1}{N} \sum_m \sum_{i \in V^m} \left\| \mathbf{x}_i^{(m)} - \pi(\mathbf{P}_i \mathbf{K}_i \mathbf{X}^m) \right\| \quad (8)$$

where N is the total number of image observations. The inlier match ratio is defined as

$$\rho = \frac{N_{triangulated}}{N_{matches}} \quad (9)$$

where $N_{matches}$ is the number of verified image correspondences and $N_{triangulated}$ is the number of successfully triangulated 3D tie points. While RMSE on check points is calculated as

$$RMSE = \sqrt{\frac{1}{M} \sum_{m=1}^M [(X_m - \hat{X}_m)^2 + (Y_m - \hat{Y}_m)^2 + (Z_m - \hat{Z}_m)^2]} \quad (10)$$

where M is the number of check points. The metric is evaluated separately for the TIR- and RGB -derived camera networks: $RMSE_{TIR}$, $RMSE_{RGB}$.

4. Experimental Results

The following section outlines the results of the proposed IRT-3DDF workflow, assessing the performance of different DL-based and hand-crafted algorithms for multi-modal image matching. This evaluation is undertaken using both image- and pose-specific metrics, with image metrics used to characterize the reliability of multi-modal tie points and camera pose metrics used to determine reconstruction accuracy in 3D space. The results of the CBBAs are summarized in Tables 3 and 4, with the 2D matched features and 3D tie points of each matching algorithm shown in Figures 7 & 8 and 9 & 10, respectively. In addition, the results for the constrained bundle adjustments (Section 3.6), reintroducing the full-resolution RGB images, are presented in Tables 5 and 6, where downsampled RGB images (‘Original’) are compared to those obtained with full-resolution RGB images (‘Constrained’). Finally, the constrained results are compared with an additional IRT-3DDF method to assess the performance of the proposed workflow against a common IRT-3DDF approach, where independently generated TIR and RGB point clouds were co-registered using GCPs.

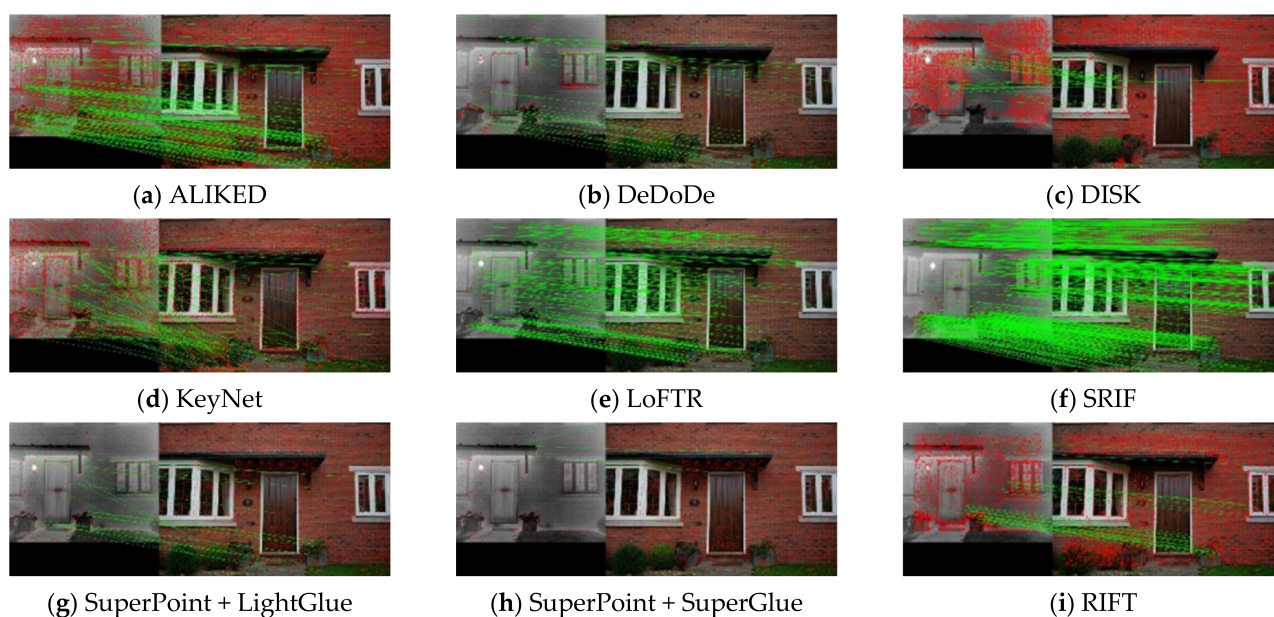


Figure 7. Multi-modal image matching results for a TIR-RGB pair of Dunrobin House using both DL and hand-crafted matchers.

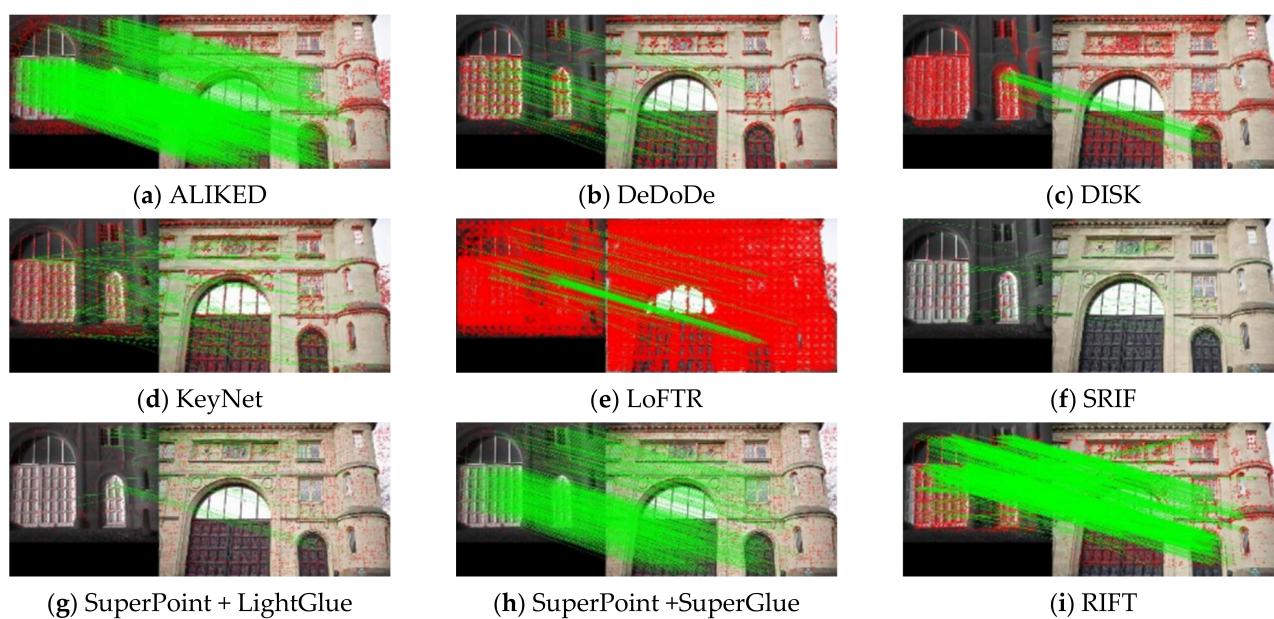


Figure 8. Multi-modal image matching results for a TIR-RGB pair of Lenton Lodge Gatehouse using both DL and hand-crafted matchers.

Table 3. Combined TIR-RGB bundle block adjustment for Dunrobin House.

CBBA	Local Feature	Matcher	No. 3D Points	Inlier Match Ratio	MRE [px]	RMSE _{TIR} [mm]	RMSE _{RGB} [mm]
COLMAP	ALIKED	LightGlue	64,296	0.83	1.0	14.1	1.9
	DeDoDe	Nearest Neighbor	9524	0.75	0.7	8.8	3.1
	DISK	LightGlue	47,964	0.75	1.2	11.1	3.2
	SuperPoint	LightGlue	17,295	0.84	0.3	8.4	1.7
	SuperPoint	SuperGlue	17,040	0.84	0.3	5.7	1.7
	LoFTR		282,504	0.72	0.7	13.5	1.6
	RIFT2		46,800	0.22	0.9	7.7	13.3

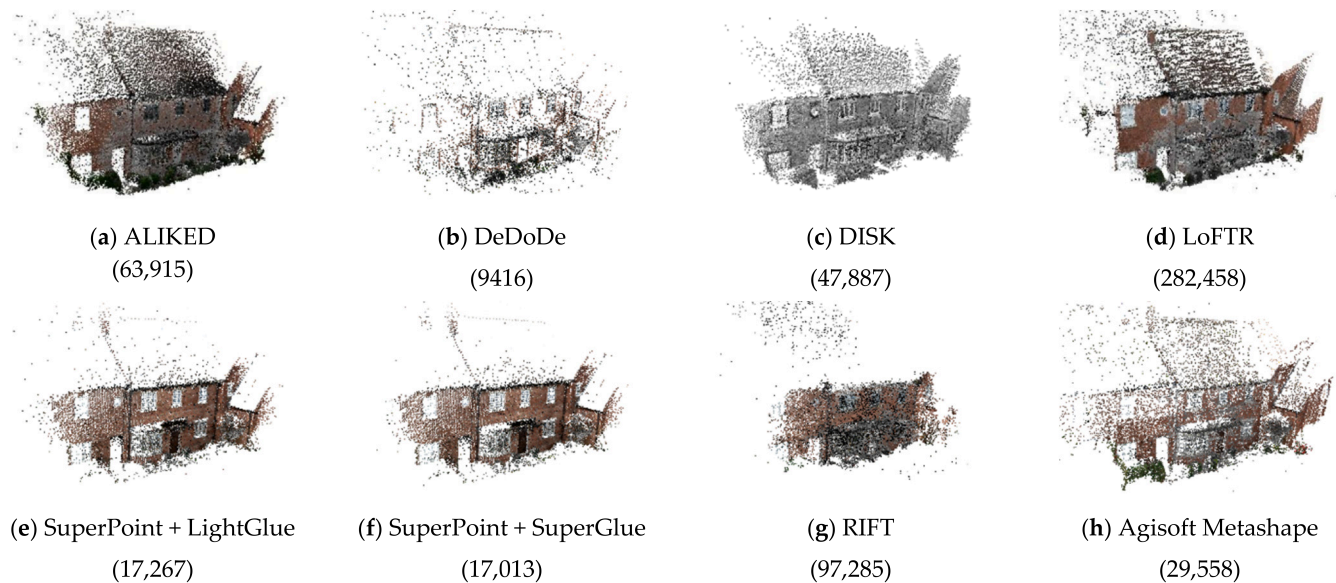


Figure 9. Multi-modal image matching results for Dunrobin House's façade ($A \approx 55 \text{ m}^2$), featuring the number of tie points generated.

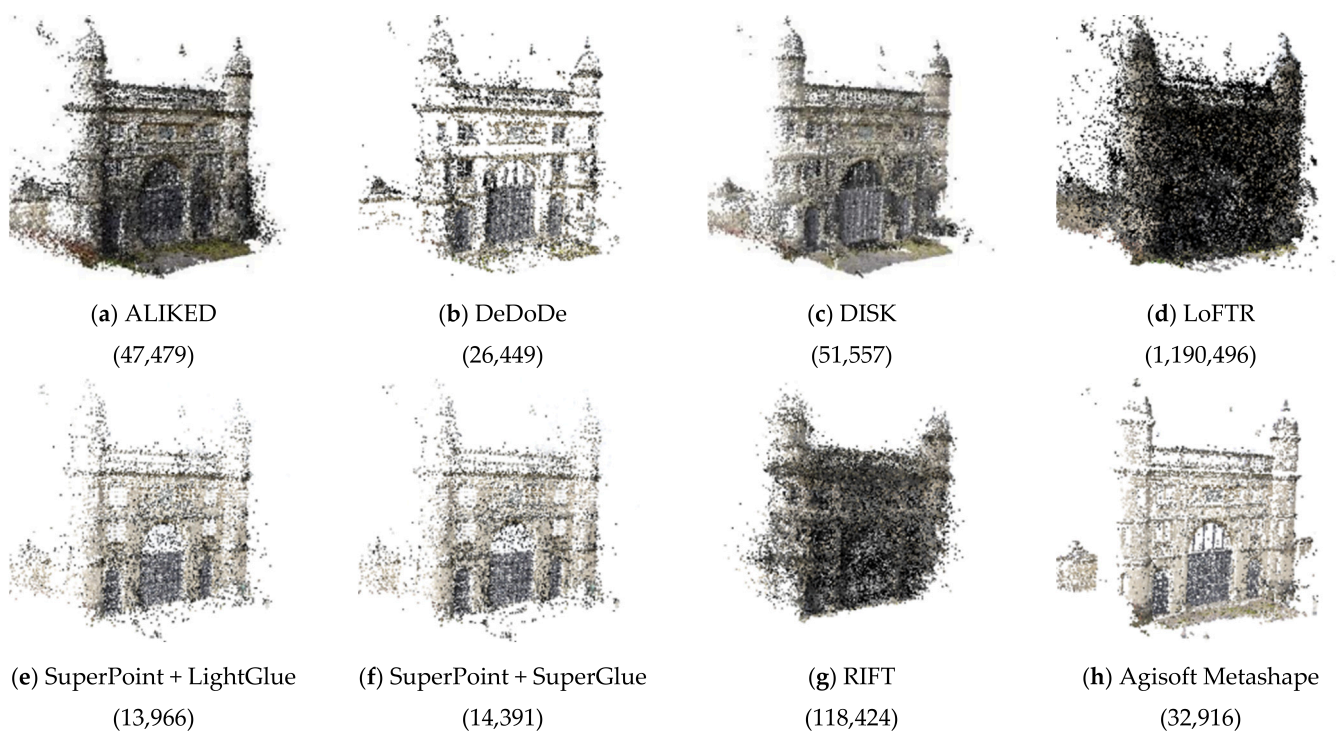


Figure 10. Multi-modal image matching results for Lenton Lodge Gatehouse's façade ($A \approx 272 \text{ m}^2$), featuring the number of tie points generated.

Firstly, it must be noted that not all approaches were able to orient the TIR-RGB blocks for both datasets, which means they were excluded from evaluation. This included KeyNet + HardNet, SRIF, and SIFT, which failed due to insufficient matches across modalities (see Figures 7 and 8). KeyNet + HardNet and SIFT + NN failed to produce any quantity of verified matches between multi-modal images that would enable orientation of the TIR-RGB blocks. Surprisingly, although SRIF seems to extract a good number of matches on Dunrobin House, they are not triangulated successfully in COLMAP, thus preventing the resection of TIR images in the RGB block. Conversely, matches extracted for Lenton Lodge Gatehouse were very limited.

Table 4. Combined TIR-RGB bundle block adjustment for Lenton Lodge Gatehouse.

CBBA	Local Feature	Matcher	No. 3D Points	Inlier Match Ratio	MRE [px]	RMSE _{TIR} [mm]	RMSE _{RGB} [mm]
COLMAP	ALIKED	LightGlue	47,882	0.91	0.4	10.8	6.3
	DeDoDe	Nearest Neighbor	27,175	0.74	0.4	21.4	4.7
	DISK	LightGlue	51,852	0.92	0.5	15.5	7.2
	SuperPoint	LightGlue	13,973	0.84	0.6	15.2	10.0
	SuperPoint	SuperGlue	14,409	0.84	0.6	17.0	9.9
	LoFTR		1,194,197	0.21	0.6	13.1	34.2
	RIFT2		118,463	0.50	0.8	22.3	36.6

Table 5. Comparison of Original and Constrained TIR-RGB bundle block adjustment for Dunrobin House.

Local Features	Matcher	Original		Constrained	
		RMSE _{TIR} [mm]	RMSE _{RGB} [mm]	RMSE _{TIR} [mm]	RMSE _{RGB} [mm]
ALIKED	LightGlue	14.1	1.9	12.5	1.7
DeDoDe	Nearest Neighb.	8.8	3.1	9.8	1.9
DISK	LightGlue	11.1	3.2	10.5	1.4
SuperPoint	LightGlue	8.4	1.7	11.3	1.4
SuperPoint	SuperGlue	5.7	1.7	8.6	1.2
LoFTR		13.5	1.5	16.8	3.7
RIFT2		7.7	13.3	9.8	3.5
Agisoft Metashape (Downsamp. Res.)		12.9	1.3	-	-
Agisoft Metashape (Full Res.)		-	-	12.9	1.2

Table 6. Comparison of Original and Constrained TIR-RGB bundle block adjustment for Lenton Lodge Gatehouse.

Local Features	Matcher	Original		Constrained	
		RMSE _{TIR} [mm]	RMSE _{RGB} [mm]	RMSE _{TIR} [mm]	RMSE _{RGB} [mm]
ALIKED	LightGlue	10.8	6.3	11.6	3.7
DeDoDe	Nearest Neighb.	21.4	4.7	19.1	2.7
DISK	LightGlue	15.5	7.2	10.3	2.0
SuperPoint	LightGlue	15.2	10.0	13.7	2.6
SuperPoint	SuperGlue	17.0	9.9	15.2	2.3
LoFTR		13.1	34.2	11.6	8.3
RIFT2		36.6	22.3	29.5	19.3
Agisoft Metashape (Downsamp. Res.)		15.3	2.0	-	-
Agisoft Metashape (Full Res.)		-	-	15.3	2.0

4.1. Combined Bundle Block Adjustments

For Dunrobin House, the results of the combined bundle block adjustments and multi-modal image matchers show favorable results, both in terms of image and camera pose measures. All DL matchers achieve an inlier match ratio > 72%, meaning that

~7/10 matched 2D features result in 3D tie points. Across the DL-based approaches, $RMSE_{RGB}$ ranges from 1.6 to 3.2 mm, which is an optimal result considering the GSD of the downsampled RGB images (8.0–10.4 mm). This reiterates the suitability of DL matchers for traditional RGB imagery, despite their low resolution or the presence of other modalities as part of the CBBA [17,85,98]. Expectedly, $RMSE_{TIR}$, with a GSD for TIR images of 9.8 mm, is an order of magnitude higher, ranging from 5.7 to 14.1 mm. LoFTR, as a semi-dense matcher, expectedly produced more tie points than any other approach ($\sim 4\times$) and an adequate $RMSE_{TIR}$ of 13 mm (Figure 9d). The best of the DL approaches, SuperPoint with both SuperGlue and LightGlue, produces similar results in terms of MRE (0.3 pix) and RMSE. As LightGlue is an evolution of SuperGlue, designed to improve computational efficiency, the only noticeable difference between them is the execution time (Table 7). The results from Dunrobin House demonstrate that with a favorable camera network and appropriate GSD, DL feature matching algorithms can handle multi-modal image matching for close-range building facades (Figure 9).

Table 7. Comparison of matching time per image pair using exhaustive (‘bruteforce’) matching with DIM and RIFT2.

CBBA	Detector + Descriptor	Matcher	Matching Time [s/pair]
COLMAP	ALIKED	LightGlue	0.12
	DeDoDe	Nearest Neighb.	0.09
	DISK	LightGlue	0.70
	SuperPoint	LoFTR	1.80
		SuperGlue	2.05
		LightGlue	0.25
	RIFT2		58.05

RIFT2, the only reference hand-crafted algorithm to successfully orient the TIR-RGB blocks, presents markedly different results from the DL-based matching approaches. For Dunrobin House, whilst RIFT2 exhibits an $RMSE_{TIR}$ in line with the DL-based approaches, reporting a RMSE of 7.7 mm and a MRE of 0.9 pix, only 22% of the 2D matches are used as effective 3D tie points, suggesting that the geometrically verified matches are not successful within traditional SfM workflows. In addition, its $RMSE_{RGB}$ is 10x higher than that of other methods, an image block likely richer in features with respect to the TIR block. Interestingly, RIFT2 required specific conditions for initialization to successfully co-orient the multi-modal block, including: (1) an additional geometric verification with epipolar geometry and error threshold for inlier matches (4.0 pix to 1.5 pix); and (2) the RGB block was initialized first during the combined TIR-RGB bundle block adjustment, mitigating its complete automation.

Additionally, the results of the CBBAs for Lenton Lodge Gatehouse show the success of the proposed IRT-3DDF workflow (Figure 10). To reiterate, Lenton Lodge Gatehouse’s camera network captured images over increasing distances, with resulting GSDs ranging from 9.8 to 26.2 mm and 8.0 to 20.8 mm for both TIR and downsampled RGB images, respectively. For all DL approaches, combined TIR-RGB bundle block adjustments achieved $RMSE_{RGB}$ between 4.7 and 34.2 mm, and $RMSE_{TIR} \sim <20$ mm. ALIKED, DISK, and SuperPoint + LightGlue performed favorably, with an $RMSE_{TIR}$ of 10.8, 15.5, and 15.2 mm, respectively. Interestingly, ALIKED and DISK produce comparatively more tie points than other DL approaches, a pattern observable across both datasets and supported by their remarkably high match ratio (91% and 92% for Lenton Lodge Gatehouse). For Lenton Lodge Gatehouse, LoFTR exhibits a behavior distinct from the previous dataset, producing a particularly noisy semi-dense point cloud and high $RMSE_{RGB}$ (34.2 mm). This drop in

performance is also reflected in its very low inlier match ratio (21%). However, LoFTR, being a semi-dense matcher, triangulates a substantially higher number of 3D tie points compared to the other methods (Figure 10d).

The comparably poorer results of RIFT2 are mirrored in Lenton Lodge Gatehouse, achieving the worst $RMSE_{RGB}$ (36.6 mm) and $RMSE_{TIR}$ (22.3 mm). Upon inspection, RIFT2 presents a markedly uneven tie point distribution on RGB pairs, which, when combined with the need for initialization using the RGB images for a successful CBBA, may have led to a poor block orientation. Overall, TIR- RGB image orientation results are preferable for the DL-based methods, both in terms of multi-modal image matching performance and 3D reconstruction reliability. However, the performance of data fusion methods must be assessed through the constrained bundle adjustment processes, determining their effectiveness by utilizing all the value present with each modality.

4.2. Constrained Bundle Block Adjustments

For Dunrobin House, the process of reintroducing the detailed information from full-resolution RGB images has a negligible effect on the $RMSE_{TIR}$. Whilst improving the RGB block could theoretically improve the orientation of the TIR block, it is likely that the maximum achievable accuracy with the multi-modal features has been reached at the respective GSD. However, significant improvements in the $RMSE_{RGB}$ are observed, decreasing from an average of 3.8 mm for the downsampled RGB images to 2.1 mm with the full-resolution RGB images. This brings results in line with the reference data fusion results from Agisoft Metashape processed at full resolution, where the two image blocks (RGB and TIR) are independently oriented and co-registered via GCPs. Importantly, these results ensure that the maximum geometric accuracy is maintained for the RGB block, on which the construction of a 3D 'base geometry' can be based. LoFTR is the only method that does not improve with this procedure, likely due to the quantity of semi-dense points and influence of the TIR block, increasing the MRE to ~ 8 pix. RIFT2 also shows substantial improvement, although its $RMSE_{RGB}$ remains higher than that of the other methods. Finally, all the results of the feature + matcher DL-based algorithms achieve a better $RMSE_{TIR}$ than that of the reference co-registration method, suggesting the superiority of the proposed data fusion method for IRT-3DDF.

Similarly, the results from Lenton Lodge Gatehouse largely confirm the trends observed for Dunrobin House, with all $RMSE_{RGB}$ results improving when constrained on the full-resolution RGB images. Although the $RMSE_{RGB}$ does not always reach the 2 mm achieved by Agisoft Metashape, it remains within a range of 2.0–3.7 mm for all feature + matcher DL-based approaches. Additionally, all approaches improve their $RMSE_{TIR}$ (apart from ALIKED), demonstrating the benefit that can be gained (possibly at greater GSDs) when introducing the full-resolution RGB images. For LoFTR, which shows similarly noisy results for Dunrobin House, the $RMSE_{RGB}$ improves from 34 mm to 8 mm; however, 8 mm is still considerably higher compared to the other methods. Finally, RIFT2 shows minimal change between the $RMSE_{TIR}$ and $RMSE_{RGB}$ values when constrained on the full-resolution RGB poses. This marks a significant outlier in the results across all constrained BA processes, which is likely a result of poor matches as part of both original and constrained CBBA. Overall, the results from Lenton Lodge Gatehouse show promise for the proposed data fusion method, with the implemented constrained bundle block adjustments allowing results to be refined after initial, low-resolution, multi-modal orientation.

5. Discussion

The results of the combined bundle block adjustments in the previous section, despite varying levels of accuracy, demonstrate the success of DL-based methods to simultaneously

orient TIR and RGB image blocks without any user intervention (e.g., GCPs), marking a significant junction for IRT-3DDF. This validates the continued exploration of ‘image-to-image’ data fusion methods, a facet of IRT-3DDF (and by extension other 3D data fusion protocols) which has been previously under-researched and hampered by the limitations of hand-crafted algorithms [15]. The results in Tables 3 and 4 demonstrate the following: (1) the success of DL-based feature + matchers to translate multi-modal matched keypoints into viable 3D tie points (all sparse DL features with an inlier match ratio > 74%); (2) the ability for RGB orientation to be highly accurate even with the influence of multi-modal correspondence; and (3) the diversity of multi-modal features (e.g., linear, corner, blob, or regions) being detected irrespective of NRD differences [19]. Critically, this success is achieved using neural networks outside of their intended training domain and using models pre-trained on vast quantities of traditional RGB images, not tailored towards multi-modal datasets.

To this end, the computational time required for such processes points supports the feasibility of DL approaches at varying scales. Table 7 reports the average time for matching an image pair executed on a NVIDIA GeForce RTX 3050 Laptop GPU. The fastest matcher is DeDoDe + NN, completing the task in 0.09 s/pair, followed by ALIKED + LightGlue at 0.12 s/pair. For SuperPoint, switching from SuperGlue to LightGlue significantly reduces the matching time from 2.05 s/pair to 0.25 s/pair. However, the necessity for RGB images to be downsampled for successful matching with TIR images must be noted as a current limitation, highlighting a limited scale invariance of the DL-based methods yet to be comprehensively solved. This limitation could potentially be addressed by automating feature detection across an image pyramid at multiple scales, similar to the approach employed by SIFT.

Furthermore, the shortcomings of some of the multi-modal matchers during and after the combined bundle block adjustments point to several key considerations for future implementation. Firstly, it is noteworthy that KeyNet + HardNet, SRIF, and SIFT were unable to co-orient both image blocks during the CBBAs, and they failed to handle the significant radiometric variations between images, and were only able to orient individual modalities depending on the initialization of the bundle adjustment (i.e., starting with either TIR-TIR or RGB-RGB matches). Upon further inspection of the exported multi-modal matches, the quantity and quality of verified matches for these approaches were insufficient for effective image orientation. Additionally, LoFTR, the sole semi-dense matcher, although a viable option for multi-modal image matching, produced $RMSE_{RGB}$ values three to four times higher than other DL-based methods, which were heavily dependent on the subject and available texture (as seen from the drop in inlier match ratio between Dunrobin House and Lenton Lodge Gatehouse). Future work should weigh the benefit of utilizing feature + matcher approaches as opposed to a semi-dense matcher for certain multi-modal applications.

Finally, the poor performance of RIFT2 and SRIF, both state-of-the-art hand-crafted multi-modal matchers, suggests the need for multi-modal matchers tailored specifically towards photogrammetric 3D reconstruction. Firstly, only available as a MATLAB implementation, RIFT2 lacks any built-in parallelization, resulting in the matching time being unreasonably long and unsuitable for typically large photogrammetric datasets [82]. Secondly, primarily used for aerial and satellite datasets (and only used for pairing operations within IRT-3DDF) [78], additional tests and tuning are required to assess RIFT2 (and similar hand-crafted multi-modal matchers) as a dedicated tool for multi-view 3D reconstruction. In summary, continued testing of DL-based approaches on new multi-modal datasets, development of multi-modal-specific evaluation criteria, and comparisons with established

hand-crafted matchers are required to cement the success of multi-modal image matching for IRT-3DDF [19,82,99].

Evaluating the presented IRT-3DDF workflow, the method successfully achieved the intended goal: the automatic co-orientation of TIR and RGB blocks utilizing multi-modal matches. Importantly, this workflow is executed without the need for sensor pre-calibration, fixed relative orientation between sensors, manually annotated features, GCPs, or ICP for co-registration, simultaneous data capture, or auxiliary positioning and referencing data. Notably, the proposed method is fully automatic, which, when compared with the reference data fusion method using Agisoft Metashape, required 200 and 479 manually annotated 2D marker coordinates for Dunrobin House and Lenton Lodge Gatehouse, respectively. Although some targets are utilized as CPs for metric evaluation, our method completely removes the need for GCPs as features for co-registration, a time-consuming procedure that can be challenging in cases where targets must be placed on building facades with appropriate access and installation. Finally, the proposed IRT-3DDF workflow enables asynchronous acquisition of TIR and RGB datasets, particularly advantageous in settings where optimal conditions for data collection (e.g., lighting, ambient temperature, and desired applications) do not coincide. For IRT-3DDF, this provides flexibility for additional TIR surveys to be integrated, enabling the comparison of TIR surveys over time or under different experimental conditions.

With the reintroduction of the full-resolution RGB imagery in the outlined constrained bundle adjustment process, sparse DL-based algorithms generally reach similar accuracy to the reference method (Agisoft Metashape), with minimal variations, either positive or negative, in $RMSE_{TIR}$ and $RMSE_{RGB}$. This final constrained bundle adjustment significantly improves the RMSE for the RGB blocks (especially for Lenton Lodge Gatehouse over varying GSDs), as demonstrated in Tables 5 and 6. Importantly for IRT-3DDF, with the possibility of full-resolution images being utilized in the generation of an accurate 3D base geometry for subsequent thermal texturing, the reintroduction of full-resolution RGB poses is warranted for the evolution of the proposed workflow through to complete thermal building models. To this end, given the co-registered RGB and thermal image blocks, the full-resolution RGB image block is used to generate a dense point cloud that serves as the basis for mesh reconstruction, while the thermal image is employed to project and map the corresponding thermal information (Figures 11 and 12). Critically, the importance of the proposed workflow is in the orientation of the TIR image block, which, when derived to a degree of accuracy, can then be used to texture a wealth of 3D base geometries (inc. photogrammetric- [100], laser scan- [101] or parametric building models [102]).



Figure 11. (a) RGB and (b) TIR textured 3D polygonal models for Dunrobin House.

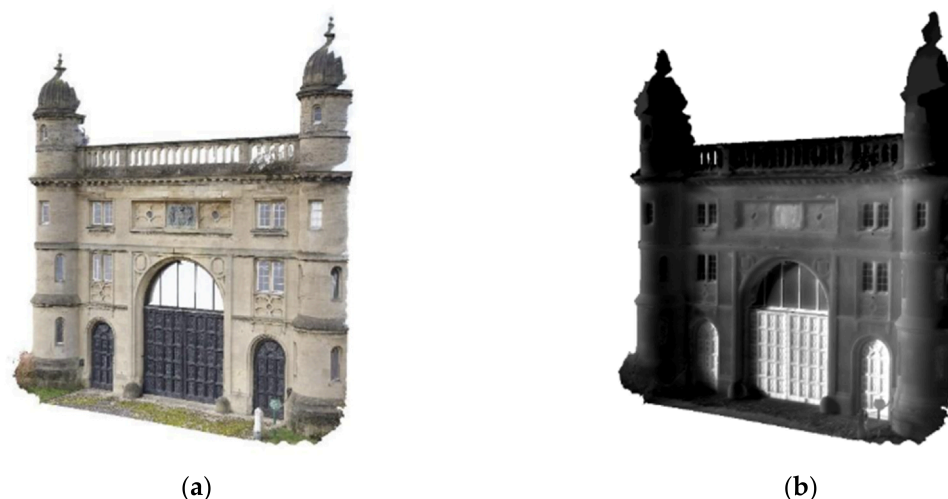


Figure 12. (a) RGB and (b) TIR textured 3D polygonal models for Lenton Lodge Gatehouse.

6. Conclusions

The generation of 3D models incorporating temperature information can have a significant impact on the assessment of a building's performance, expanding traditional two-dimensional thermal imaging techniques into novel products and purposes. This paper contributes to the growing field of InfraRed Thermography 3D-Data Fusion, demonstrating the application of pre-trained neural networks for multi-modal image matching and 3D thermal modeling. Importantly, these neural networks, trained on vast datasets of RGB imagery with significant radiometric and geometric variations, show that they can generalize outside their training domain to accommodate multi-modal image matching. Exploiting a novel, flexible, open-source image matching package, this work shows the successful use of multi-modal image matching for the fully automatic co-registration of TIR and RGB imagery in a combined bundle block adjustment, a previously under-utilized approach for IRT-3DDF. When applied to building facades and terrestrial image acquisition, DL-based image matchers demonstrate superior solutions when compared to established hand-crafted methods, demonstrating higher geometric accuracy and computational efficiency whilst managing the significant radiometric differences between TIR and RGB images. More specifically, the results demonstrate that off-the-shelf, pre-trained, mono-modal DL-based matching algorithms constitute a valid alternative to hand-crafted approaches. While mono-modal hand-crafted methods fail completely, the proposed DL-based methods achieve performance improvements of up to almost seven times over the best-performing multi-modal hand-crafted algorithm. From a computational standpoint, the available DL-based methods are parallelized to run on GPUs and perform two/three orders of magnitude better in terms of matching time.

The proposed methodology, co-orienting TIR images with additional reference RGB images, demonstrates the possibility for asynchronous and repeated surveys to be undertaken, allowing images to be generated at optimal times for each modality and capturing specific conditions for varying applications. This workflow is achieved without a reliance on sensors in fixed relative orientation, manually annotated GCPs, or supplementary positioning information, drastically increasing efficiency and applicability. The future developments of this work will look to expand the data fusion procedures by demonstrating the versatility of the proposed workflow across new sensors, subjects, and scales (e.g., aerial and UAV images, industrial objects, etc.), generating complete 3D thermal models that can inform building diagnostics and conservation practices. Furthermore, the significant limitations of scale invariance across deep-learning multi-modal image matchers, demonstrated here by

the necessity to downsample RGB images for effective matching, should become a priority in the development of future models and training datasets.

Author Contributions: All authors contributed to the study’s conception and design. N.S. conducted the data collection. N.S. and L.M. performed experiments and data analysis. L.M. and F.I. developed the DIM Python library and prepared the supporting materials. N.S. and L.M. wrote the first draft of the manuscript, and all authors provided feedback on previous versions. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the UK Research and Innovation (UKRI) Engineering and Physical Sciences Research Council (EPSRC) as part of the Geospatial Systems Centre for Doctoral Training (CDT), grant number: EP/S023577/1; Sensor equipment used in the research was funded by UKCRIC—UK Collaboratorium for Research in Infrastructure & Cities: Newcastle Laboratories (EP/R010102/1).

Data Availability Statement: The datasets used in the experiments, including the images and ground truth annotations, are available at <https://github.com/3DOM-FBK/Thermitage> (accessed on 19 July 2026).

Acknowledgments: We would like to thank the University of Nottingham Estates Team for granting access and providing assistance for the control surveys of Lenton Lodge Gatehouse. We would also like to thank Jon Bedford and Historic England’s Geospatial Survey Team for lending the Sony $\alpha 7R II$ used in both experiments.

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

References

1. Fox, M.; Coley, D.; Goodhew, S.; De Wilde, P. Thermography Methodologies for Detecting Energy Related Building Defects. *Renew. Sustain. Energy Rev.* **2014**, *40*, 296–310. [\[CrossRef\]](#)
2. Usamentiaga, R.; Venegas, P.; Guerediaga, J.; Vega, L.; Mollada, J.; Bulnes, F. Infrared Thermography for Temperature Measurement and Non-Destructive Testing. *Sensors* **2014**, *14*, 12305–12348. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Ramón, A.; Adán, A.; Javier Castilla, F. Thermal Point Clouds of Buildings: A Review. *Energy Build.* **2022**, *274*, 112425. [\[CrossRef\]](#)
4. Adán, A.; Quintana, B.; García Aguilar, J.; Pérez, V.; Castilla, F.J. Towards the Use of 3D Thermal Models in Constructions. *Sustainability* **2020**, *12*, 8521. [\[CrossRef\]](#)
5. Sutherland, N.; Marsh, S.; Priestnall, G.; Bryan, P.; Mills, J. InfraRed Thermography and 3D-Data Fusion for Architectural Heritage: A Scoping Review. *Remote Sens.* **2023**, *15*, 2422. [\[CrossRef\]](#)
6. Motayyeb, S.; Samadzedegan, F.; Dadrass Javan, F.; Hosseinpour, H. Fusion of UAV-Based Infrared and Visible Images for Thermal Leakage Map Generation of Building Facades. *Heliyon* **2023**, *9*, e14551. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Adán, A.; Pérez, V.; Vivancos, J.-L.; Aparicio-Fernández, C.; Prieto, S.A. Proposing 3D Thermal Technology for Heritage Building Energy Monitoring. *Remote Sens.* **2021**, *13*, 1537. [\[CrossRef\]](#)
8. Natephra, W.; Motamedi, A.; Yabuki, N.; Fukuda, T. Integrating 4D Thermal Information with BIM for Building Envelope Thermal Performance Analysis and Thermal Comfort Evaluation in Naturally Ventilated Environments. *Build. Environ.* **2017**, *124*, 194–208. [\[CrossRef\]](#)
9. Ham, Y.; Golparvar-Fard, M. An Automated Vision-Based Method for Rapid 3D Energy Performance Modeling of Existing Buildings Using Thermal and Digital Imagery. *Adv. Eng. Inform.* **2013**, *27*, 395–409. [\[CrossRef\]](#)
10. Scaioni, M.; Rosina, E.; L’Erario, A.; Diaz-Vilariño, L. Integration of Infrared Thermography and Photogrammetric Surveying of Built Landscape. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *XLII-5/W1*, 153–160. [\[CrossRef\]](#)
11. Maset, E.; Fusiello, A.; Crosilla, F.; Toldo, R.; Zorretto, D. Photogrammetric 3D Building Reconstruction from Thermal Images. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *IV-2/W3*, 25–32. [\[CrossRef\]](#)
12. Dlesk, A.; Vach, K.; Pavelka, K. Transformations in the Photogrammetric Co-Processing of Thermal Infrared Images and RGB Images. *Sensors* **2021**, *21*, 5061. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Hoegner, L.; Stilla, U. Automatic 3D Reconstruction and Texture Extraction for 3D Building Models from Thermal Infrared Image Sequences. In Proceedings of the QIRT 2016—13th International Conference on Quantitative InfraRed Thermography, Gdańsk, Poland, 4–8 July 2016.
14. Previtali, M.; Barazzetti, L.; Redaelli, V.; Scaioni, M.; Rosina, E. Rigorous Procedure for Mapping Thermal Infrared Images on Three-Dimensional Models of Building Façades. *J. Appl. Remote Sens.* **2013**, *7*, 073503. [\[CrossRef\]](#)

15. Morelli, L.; Bellavia, F.; Menna, F.; Remondino, F. Photogrammetry Now and Then—From Hand-Crafted to Deep-Learning Tie Points—. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, XLVIII-2/W1-2022, 163–170. [[CrossRef](#)]
16. Velesaca, H.O.; Bastidas, G.; Rouhani, M.; Sappa, A.D. Multimodal Image Registration Techniques: A Comprehensive Survey. *Multimed. Tools Appl.* **2024**, 83, 63919–63947. [[CrossRef](#)]
17. Ruiz De Oña, E.; Barbero-García, I.; González-Aguilera, D.; Remondino, F.; Rodríguez-Gonzálvez, P.; Hernández-López, D. PhotoMatch: An Open-Source Tool for Multi-View and Multi-Modal Feature-Based Image Matching. *Appl. Sci.* **2023**, 13, 5467. [[CrossRef](#)]
18. Yang, M.; Fan, F.; Huang, J.; Ma, Y.; Mei, X.; Cai, Z.; Ma, J. Multimodal Image Matching Based on Cross-Modality Completion Pre-Training. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence; International Joint Conferences on Artificial Intelligence Organization: Montreal, QC, Canada, 2025*; pp. 2206–2214.
19. Jiang, X.; Ma, J.; Xiao, G.; Shao, Z.; Guo, X. A Review of Multimodal Image Matching: Methods and Applications. *Inf. Fusion* **2021**, 73, 22–71. [[CrossRef](#)]
20. Wu, P.; Yao, Y.; Zhang, W.; Wei, D.; Wan, Y.; Li, Y.; Zhang, Y. MapGlue: Multimodal Remote Sensing Image Matching. *arXiv* **2025**, arXiv:2503.16185.
21. Han, S.; Liu, X.; Dong, J.; Liu, H. Remote Sensing Multimodal Image Matching Based on Structure Feature and Learnable Matching Network. *Appl. Sci.* **2023**, 13, 7701. [[CrossRef](#)]
22. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, 60, 91–110. [[CrossRef](#)]
23. Riba, E.; Mishkin, D.; Ponsa, D.; Rublee, E.; Bradski, G. Kornia: An Open Source Differentiable Computer Vision Library for PyTorch. In *Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV); IEEE: Snowmass, CO, USA, 2020*; pp. 3663–3672.
24. Mishkin, D.; Matas, J.; Perdoch, M. MODS: Fast and Robust Method for Two-View Matching. *Comput. Vis. Image Underst.* **2015**, 141, 81–93. [[CrossRef](#)]
25. Morelli, L.; Ioli, F.; Maiwald, F.; Mazzacca, G.; Menna, F.; Remondino, F. Deep-Image-Matching: A Toolbox for Multiview Image Matching of Complex Scenarios. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2024**, XLVIII-2/W4-2024, 309–316. [[CrossRef](#)]
26. Jin, Y.; Mishkin, D.; Mishchuk, A.; Matas, J.; Fua, P.; Yi, K.M.; Trulls, E. Image Matching Across Wide Baselines: From Paper to Practice. *Int. J. Comput. Vis.* **2021**, 129, 517–547. [[CrossRef](#)]
27. Berton, G.; Goletto, G.; Trivigno, G.; Stoken, A.; Caputo, B.; Masone, C. EarthMatch: Iterative Coregistration for Fine-Grained Localization of Astronaut Photography. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); IEEE: Seattle, WA, USA, 2024*; pp. 4264–4274.
28. Moravec, H.P. *Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover*; Stanford University: Stanford, CA, USA, 1980.
29. Förstner, W.; Gülch, E. A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features. In *Proceedings of the ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data, Interlaken, Switzerland, 2–4 June 1987; Volume 6*, pp. 281–305.
30. Harris, C.; Stephens, M. A Combined Corner and Edge Detector. In *Proceedings of the Alvey Vision Conference 1988; Alvey Vision Club: Manchester, UK, 1988*; pp. 147–152.
31. Tomasi, C.; Kanade, T. Detection and Tracking of Point. *Int. J. Comput. Vis.* **1991**, 9, 3.
32. Lowe, D.G. Object Recognition from Local Scale-Invariant Features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision; IEEE: Kerkyra, Greece, 1999; Volume 2*, pp. 1150–1157.
33. Mikolajczyk, K.; Schmid, C. An Affine Invariant Interest Point Detector. In *Computer Vision—ECCV 2002; Heyden, A., Sparr, G., Nielsen, M., Johansen, P., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2002; Volume 2350*, pp. 128–142, ISBN 978-3-540-43745-1.
34. Matas, J.; Chum, O.; Urban, M.; Pajdla, T. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In *Proceedings of the British Machine Vision Conference 2002; British Machine Vision Association: Cardiff, UK, 2002*; pp. 384–393.
35. Tuytelaars, T.; Van Gool, L. Matching Widely Separated Views Based on Affine Invariant Regions. *Int. J. Comput. Vis.* **2004**, 59, 61–85. [[CrossRef](#)]
36. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.* **2008**, 110, 346–359. [[CrossRef](#)]
37. Rosten, E.; Drummond, T. Machine Learning for High-Speed Corner Detection. In *Computer Vision—ECCV 2006; Leonardis, A., Bischof, H., Pinz, A., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2006; Volume 3951*, pp. 430–443, ISBN 978-3-540-33832-1.
38. Tola, E.; Lepetit, V.; Fua, P. DAISY: An Efficient Dense Descriptor Applied to Wide-Baseline Stereo. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, 32, 815–830. [[CrossRef](#)] [[PubMed](#)]
39. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An Efficient Alternative to SIFT or SURF. In *Proceedings of the 2011 International Conference on Computer Vision; IEEE: Barcelona, Spain, 2011*; pp. 2564–2571.

40. Alcantarilla, P.; Nuevo, J.; Bartoli, A. Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces. In *Proceedings of the British Machine Vision Conference 2013*; British Machine Vision Association: Bristol, UK, 2013; pp. 1–11.
41. Calonder, M.; Lepetit, V.; Strecha, C.; Fua, P. BRIEF: Binary Robust Independent Elementary Features. In *Computer Vision—ECCV 2010; Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6314, pp. 778–792, ISBN 978-3-642-15560-4.
42. Verdie, Y.; Yi, K.; Fua, P.; Lepetit, V. TILDE: A Temporally Invariant Learned Detector. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Boston, MA, USA, 2015; pp. 5279–5288.
43. Barroso-Laguna, A.; Riba, E.; Ponsa, D.; Mikolajczyk, K. Key.Net: Keypoint Detection by Handcrafted and Learned CNN Filters. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Seoul, Republic of Korea, 2019; pp. 5835–5843.
44. Truong, P.; Apostolopoulos, S.; Mosinska, A.; Stucky, S.; Ciller, C.; Zanet, S.D. GLAMpoints: Greedily Learned Accurate Match Points. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Seoul, Republic of Korea, 2019; pp. 10731–10740.
45. Yi, K.M.; Trulls, E.; Lepetit, V.; Fua, P. LIFT: Learned Invariant Feature Transform. In *Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2016; Volume 9910, pp. 467–483, ISBN 978-3-319-46465-7.
46. DeTone, D.; Malisiewicz, T.; Rabinovich, A. SuperPoint: Self-Supervised Interest Point Detection and Description. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: Piscataway, NJ, USA, 2018; pp. 337–33712.
47. Dusmanu, M.; Rocco, I.; Pajdla, T.; Pollefeys, M.; Sivic, J.; Torii, A.; Sattler, T. D2-Net: A Trainable CNN for Joint Description and Detection of Local Features. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Long Beach, CA, USA, 2019; pp. 8084–8093.
48. Ono, Y.; Trulls, E.; Fua, P.; Yi, K.M. LF-Net: Learning Local Features from Images. In *Proceedings of the Advances in Neural Information Processing Systems 31*, Montréal, QC, Canada, 3–8 December 2018; Volume 31.
49. Luo, Z.; Zhou, L.; Bai, X.; Chen, H.; Zhang, J.; Yao, Y.; Li, S.; Fang, T.; Quan, L. ASLFeat: Learning Local Features of Accurate Shape and Localization. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Seattle, WA, USA, 2020; pp. 6588–6597.
50. Tyszkiewicz, M.J.; Fua, P.; Trulls, E. DISK: Learning Local Features with Policy Gradient. In *Advances in Neural Information Processing Systems 33, Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Online, 6–12 December 2020; Curran Associates, Inc.: Red Hook, NY, USA, 2020.
51. Edstedt, J.; Bökman, G.; Wadenbäck, M.; Felsberg, M. DeDoDe: Detect, Don't Describe—Describe, Don't Detect for Local Feature Matching. In *Proceedings of the 2024 International Conference on 3D Vision (3DV)*; IEEE: Davos, Switzerland, 2024; pp. 148–157.
52. Gleize, P.; Wang, W.; Feiszli, M. SiLK: Simple Learned Keypoints. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Paris, France, 2023; pp. 22442–22451.
53. Tian, Y.; Fan, B.; Wu, F. L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Honolulu, HI, USA, 2017; pp. 6128–6136.
54. Mishchuk, A.; Mishkin, D.; Radenović, F.; Matas, J. Working Hard to Know Your Neighbor's Margins: Local Descriptor Learning Loss. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017 Proceedings)*; NIPS Foundation: San Diego, CA, USA, 2017.
55. Li, J.; Hu, Q.; Ai, M. RIFT: Multi-Modal Image Matching Based on Radiation-Variation Insensitive Feature Transform. *IEEE Trans. Image Process.* **2020**, *29*, 3296–3310. [[CrossRef](#)] [[PubMed](#)]
56. Pautrat, R.; Larsson, V.; Oswald, M.R.; Pollefeys, M. Online Invariance Selection for Local Feature Descriptors. In *Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2020; Volume 12347, pp. 707–724, ISBN 978-3-030-58535-8.
57. Parihar, U.S.; Gujarathi, A.; Mehta, K.; Tourani, S.; Garg, S.; Milford, M.; Krishna, K.M. RoRD: Rotation-Robust Descriptors and Orthographic Views for Local Feature Matching. In *Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: Prague, Czech Republic, 2021; pp. 1593–1600.
58. Bökman, G.; Kahl, F. A Case for Using Rotation Invariant Features in State of the Art Feature Matchers. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: New Orleans, LA, USA, 2022; pp. 5106–5115.
59. Li, L.; Cai, Y.; Zhou, Q. An Efficient Approach for DRC Hotspot Prediction with Convolutional Neural Network. In *Proceedings of the 2021 IEEE International Symposium on Circuits and Systems (ISCAS)*; IEEE: Daegu, Republic of Korea, 2021; pp. 1–5.
60. Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; Zhou, X. LoFTR: Detector-Free Local Feature Matching with Transformers. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Nashville, TN, USA, 2021; pp. 8918–8927.

61. Edstedt, J.; Sun, Q.; Bökman, G.; Wadenbäck, M.; Felsberg, M. RoMa: Robust Dense Feature Matching. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Seattle, WA, USA, 2024; pp. 19790–19800.
62. Edstedt, J.; Athanasiadis, I.; Wadenbäck, M.; Felsberg, M. DKM: Dense Kernelized Feature Matching for Geometry Estimation. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Vancouver, BC, Canada, 2023; pp. 17765–17775.
63. Sarlin, P.-E.; DeTone, D.; Malisiewicz, T.; Rabinovich, A. SuperGlue: Learning Feature Matching with Graph Neural Networks. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Seattle, WA, USA, 2020; pp. 4937–4946.
64. Lindenberger, P.; Sarlin, P.-E.; Pollefeys, M. LightGlue: Local Feature Matching at Light Speed. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Paris, France, 2023; pp. 17581–17592.
65. Künzel, J.; Hilsmann, A.; Eisert, P. RIPE: Reinforcement Learning on Unlabeled Image Pairs for Robust Keypoint Extraction. In *Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Honolulu, HI, USA, 2025; pp. 4868–4877.
66. Chen, G.; Fu, T.; Chen, H.; Teng, W.; Xiao, H.; Zhao, Y. RDD: Robust Feature Detector and Descriptor Using Deformable Transformer. In *Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Nashville, TN, USA, 2025; pp. 6394–6403.
67. Liu, Y.; Lai, W.; Zhao, Z.; Xiong, Y.; Zhu, J.; Cheng, J.; Xu, Y. LiftFeat: 3D Geometry-Aware Local Feature Matching. In *Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: Atlanta, GA, USA, 2025; pp. 11714–11720.
68. Edstedt, J.; Bökman, G.; Wadenbäck, M.; Felsberg, M. DaD: Distilled Reinforcement Learning for Diverse Keypoint Detection. *arXiv* **2025**, arXiv:2503.07347.
69. Nordström, D.; Edstedt, J.; Bökman, G.; Astermark, J.; Heyden, A.; Larsson, V.; Wadenbäck, M.; Felsberg, M.; Kahl, F. LoMa: Local Feature Matching Revisited. *arXiv* **2026**, arXiv:2604.04931.
70. Zheng, Z.; Feng, J.; Savaliya, N.; Yeh, Z.-H.; Lang, B.; Chuah, M.C. TextFM: Robust Semi-Dense Feature Matching with Language Guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Denver, CO, USA, 2026; pp. 16635–16644.
71. Lu, X.; Du, S. JamMa: Ultra-Lightweight Local Feature Matching with Joint Mamba. In *Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Nashville, TN, USA, 2025; pp. 14934–14943.
72. Remondino, F.; Morelli, L.; Stathopoulou, E.; Elhashash, M.; Qin, R. Aerial Triangulation with Learning-Based Tie Points. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, XLIII-B2-2022, 77–84. [\[CrossRef\]](#)
73. Remondino, F.; Spera, M.G.; Nocerino, E.; Menna, F.; Nex, F. State of the Art in High Density Image Matching. *Photogramm. Rec.* **2014**, 29, 144–166. [\[CrossRef\]](#)
74. González-Aguilera, D.; Rodríguez-Gonzálvez, P.; Armesto, J.; Lagüela, S. Novel Approach to 3D Thermography and Energy Efficiency Evaluation. *Energy Build.* **2012**, 54, 436–443. [\[CrossRef\]](#)
75. Saleem, S.; Bais, A.; Sablatnig, R.; Ahmad, A.; Naseer, N. Feature Points for Multisensor Images. *Comput. Electr. Eng.* **2017**, 62, 511–523. [\[CrossRef\]](#)
76. Li, J.; Shi, P.; Hu, Q.; Zhang, Y. RIFT2: Speeding-up RIFT with A New Rotation-Invariance Technique. *arXiv* **2023**, arXiv:2303.00319.
77. Li, J.; Hu, Q.; Zhang, Y. Multimodal Image Matching: A Scale-Invariant Algorithm and an Open Dataset. *ISPRS J. Photogramm. Remote Sens.* **2023**, 204, 77–88. [\[CrossRef\]](#)
78. Lin, D.; Jarzabek-Rychard, M.; Tong, X.; Maas, H.-G. Fusion of Thermal Imagery with Point Clouds for Building Façade Thermal Attribute Mapping. *ISPRS J. Photogramm. Remote Sens.* **2019**, 151, 162–175. [\[CrossRef\]](#)
79. Guan, Z.; Zheng, J.; Wang, S. Hessian Matrix-Based Improved RIFT Algorithm for Infrared-Visible Image Registration. In *Proceedings of the 2025 IEEE 5th International Conference on Electronic Technology, Communication and Information (ICETCI)*; IEEE: Changchun, China, 2025; pp. 929–934.
80. Zhao, X.; Wu, X.; Chen, W.; Chen, P.C.Y.; Xu, Q.; Li, Z. ALIKED: A Lighter Keypoint and Descriptor Extraction Network via Deformable Transformation. *IEEE Trans. Instrum. Meas.* **2023**, 72, 1–16. [\[CrossRef\]](#)
81. Revaud, J.; Weinzaepfel, P.; De Souza, C.; Humenberger, M. R2d2: Reliable and Repeatable Detector and Descriptor. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*; ACM: Vancouver, BC, Canada, 2019.
82. Chang, R.; Yang, C.; Zhang, H.; Xie, H.; Zhou, C.; Pan, A.; Yang, Y. Speed Matters, a Robust Infrared and Visible Image Matching Method at Real-Time Speed. *J. Real-Time Image Proc.* **2024**, 21, 15. [\[CrossRef\]](#)
83. Hoegner, L.; Abmayr, T.; Tosic, D.; Turzer, S.; Stilla, U. Fusion of TLS and RGB Point Clouds with TIR Images for Indoor Mobile Mapping. In *Proceedings of the 14th Quantitative InfraRed Thermography Conference (QIRT 2018)*, Berlin, Germany, 25–29 June 2018.
84. Elias, M.; Weitkamp, A.; Eltner, A. Multi-Modal Image Matching to Colorize a SLAM Based Point Cloud with Arbitrary Data from a Thermal Camera. *ISPRS Open J. Photogramm. Remote Sens.* **2023**, 9, 100041. [\[CrossRef\]](#)

85. Morelli, L.; Perda, G.; Ioli, F.; Trybała, P.; Sterpin, A.; Rigon, S.; Sutherland, N.; Medici, M.; Remondino, F.; Vitti, A. Co-Registering Laser Scanning Point Clouds and Photogrammetric Images with Deep Learning Multi-Modal Matching. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2024**, XLVIII-2/W8-2024, 335–342. [[CrossRef](#)]
86. Morelli, L.; Perda, G.; Malek, S.; Remondino, F.; Legat, K. Night and Day Aerial Photogrammetry. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2025**, XLVIII-1/W4-2025, 81–88. [[CrossRef](#)]
87. Sutherland, N.; Morelli, L.; Mills, J.; Bryan, P.; Marsh, S.; Remondino, F. Multi-Modal and Multi-Sensor Photogrammetric Data Fusion Exploiting a New Repository for Infrared Thermography Datasets. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2025**, XLVIII-1/W6-2025, 235–242. [[CrossRef](#)]
88. Zhang, D.; Liu, Y.; Zhao, Y.; Liang, J.; Sun, B.; Chu, S. Algorithm Research on Detail and Contrast Enhancement of High Dynamic Infrared Images. *Appl. Sci.* **2023**, *13*, 12649. [[CrossRef](#)]
89. Schindelin, J.; Arganda-Carreras, I.; Frise, E.; Kaynig, V.; Longair, M.; Pietzsch, T.; Preibisch, S.; Rueden, C.; Saalfeld, S.; Schmid, B.; et al. Fiji: An Open-Source Platform for Biological-Image Analysis. *Nat. Methods* **2012**, *9*, 676–682. [[CrossRef](#)] [[PubMed](#)]
90. Akcay, O.; Erenoglu, R.C.; Erenoglu, O. Correction and Densification of UAS-based Photogrammetric Thermal Point Cloud. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, XLI-B3, 163–166. [[CrossRef](#)]
91. Marelli, D.; Morelli, L.; Farella, E.M.; Bianco, S.; Ciocca, G.; Remondino, F. ENRICH: Multi-purposE Dataset for beNchmaRking in Computer Vision and pHotogrammetry. *ISPRS J. Photogramm. Remote Sens.* **2023**, *198*, 84–98. [[CrossRef](#)]
92. Song, S.; Morelli, L.; Wu, X.; Qin, R.; Albanwan, H.; Remondino, F. Evaluating Learning-Based Tie Point Matching for Geometric Processing of Off-Track Satellite Stereo. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2024**, XLVIII-2–2024, 393–400. [[CrossRef](#)]
93. Chum, O.; Werner, T.; Matas, J. Two-View Geometry Estimation Unaffected by a Dominant Plane. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*; IEEE: San Diego, CA, USA, 2005; Volume 1, pp. 772–779.
94. Schonberger, J.L.; Frahm, J.-M. Structure-from-Motion Revisited. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Las Vegas, NV, USA, 2016; pp. 4104–4113.
95. Brown, D.C. Close-Range Camera Calibration. *Photogramm. Eng.* **1971**, *37*, 855–866.
96. Snavely, N.; Seitz, S.M.; Szeliski, R. Photo Tourism: Exploring Photo Collections in 3D. *ACM Trans. Graph.* **2006**, *25*, 835–846. [[CrossRef](#)]
97. Agisoft Metashape Professional 2024. Available online: <https://www.agisoft.com/> (accessed on 19 July 2026).
98. Ma, J.; Jiang, X.; Fan, A.; Jiang, J.; Yan, J. Image Matching from Handcrafted to Deep Features: A Survey. *Int. J. Comput. Vis.* **2021**, *129*, 23–79. [[CrossRef](#)]
99. Bellavia, F.; Morelli, L.; Menna, F.; Remondino, F. Image Orientation with a Hybrid Pipeline Robust to Rotations and Wide-Baselines. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, XLVI-2/W1-2022, 73–80. [[CrossRef](#)]
100. Zhu, J.; Xu, Y.; Ye, Z.; Hoegner, L.; Stilla, U. Fusion of Urban 3D Point Clouds with Thermal Attributes Using MLS Data and TIR Image Sequences. *Infrared Phys. Technol.* **2021**, *113*, 103622. [[CrossRef](#)]
101. Biswanath, M.K.; Hoegner, L.; Stilla, U. Thermal Mapping from Point Clouds to 3D Building Model Facades. *Remote Sens.* **2023**, *15*, 4830. [[CrossRef](#)]
102. Hoegner, L.; Kumke, H.; Schwarz, A.; Meng, L.; Stilla, U. Strategies for Texturing Building Models with Low Resolution Infrared Image Sequences. In *Proceedings of the MMT '07—Mobile Mapping Technology*, Padua, Italy, 28–31 May 2007; Volume 36.

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.