Deep Thermal Image Descriptors: A Comparative Analysis of CNNs and Vision Transformers for Feature-Based Registration

by Ahmed M. Abdelbaset*, Anastasia Topalidou**, Wei Quan*, Bogdan J. Matuszewski*

* *Institute for Engineering and Technology Innovation (InETI), Computer Vision and Machine Learning (CVML) Group, School of Engineering and Computing, University of Lancashire, Preston, United Kingdom*

** *Institute for Relational Research (INTERRELATE), School of Nursing and Midwifery, University of Lancashire, Preston, United Kingdom*

Abstract

This paper studies learned local descriptors for thermal image matching and registration, with emphasis on patch size. We compare a CNN-based and adapted Vision (ViT) Transformer descriptor across patch sizes from 8×8 to 96×96. Using a fixed-correspondence benchmark, ViT achieves stronger descriptor separability than the CNN, with 16×16 producing the lowest FPR@95TPR. A Ground Truth (GT)-labelled matching test shows that this improvement transfers to stricter match correctness, rather than simply increasing match count. In a shared-detector registration pipeline, both learned descriptors substantially outperform RIFT2, with 16×16 ViT giving the best geometric accuracy.

1. Introduction

Thermal imaging is increasingly used in non-destructive testing and evaluation (NDT/E), including the inspection of composite and aerospace structures for delamination, impact damage, disbonds, and subsurface defects [1], as well as the assessment of civil infrastructure and buildings for delamination, moisture ingress, insulation defects, cracks, and air leakage [2]. It is also used in industrial inspection [3], medical assessment [4], and other low-visibility sensing settings because it measures emitted radiation rather than reflected visible light [5]. In these applications, the ability to establish reliable correspondences between thermal images is essential for downstream registration. Registration is essential when thermal images are compared across time, fused or aligned with other modalities (e.g.visible images) for diagnosis, monitoring, localisation, or scene understanding. However, thermal imagery poses challenges that differ from visible light vision. Thermal images often contain weak gradients, low texture, sparse salient structure, and appearance changes that do not necessarily correspond to visible surface texture or object boundaries [6]. Additionally thermal image intensities are driven by surface thermal radiation, which depends on temperature, emissivity, reflected environmental radiation, and acquisition condition. As a result, assumptions that hold for visible-image local features do not automatically transfer to thermal data.

In feature-based registration, local descriptors play a critical role because they determine whether corresponding patches can be separated from distractors. Even when the detector is fixed, descriptor quality directly affects the number of false correspondences, the burden on geometric verification, and the quality of the estimated transformation. This makes descriptor evaluation especially important in thermal matching pipelines, where outliers may arise from radiometric instability, repetitive structure, or low local distinctiveness. Other correspondence frameworks, including detector-free matchers [7] and learned matching pipelines [8], address image matching at a more integrated system level by coupling feature representation, contextual reasoning, correspondence assignment, and match filtering. These approaches are important, but they are outside the scope of the present study because they do not allow the effect of the local descriptor support region to be isolated directly.

In this paper, we study learned thermal descriptors under a controlled pipeline in which descriptor quality can be evaluated directly and then related to downstream registration performance. Our focus is not on detector design, nor on end-to-end registration, but on the local patch representation used to describe thermal keypoints. Within this scope, we compare CNN- and transformer-based descriptor architectures and investigate how the choice of descriptor support size influences their relative performance in thermal patch matching and feature-based registration.

Patch size is often treated as a convenient hyperparameter inherited from practice developed for visible-light imagery. However, in thermal imagery, the choice of support size may be especially important. Small patches may preserve the most local and discriminative thermal structure, while larger patches may add context at the cost of including irrelevant background, low-contrast intensity transitions, or radiometric clutter. For transformer models, support size also changes the number of spatial tokens and therefore the form of the learned representation itself. These considerations motivate comparing CNN- and transformer-based descriptors across multiple support sizes rather than relying on a single arbitrary configuration.

To this end, we compare two descriptor families with contrasting inductive biases: a modified Convolutional Neural Network (CNN) descriptor [9] and an adapted Vision Transformer (ViT) descriptor [10]. Both models are trained and evaluated under the same controlled setting, using the same thermal datasets, synthetic transformation generation procedure, correspondence construction, descriptor distance, and evaluation metrics. We sweep patch sizes from 8 to 96 pixels and measure descriptor separability using FPR@95TPR [11] and related distance-based metrics. We also include a transform-type diagnostic analysis to examine how descriptor behaviour changes under translation, rotation, scale, shear, and full

License: https://creativecommons.org/licenses/by/4.0/deed.en

affine transformation.

A descriptor benchmark alone, however, does not fully answer the practical question of whether better descriptor scores lead to better registration. We therefore add a second stage: a shared-detector sparse registration benchmark in which the detector, matcher, and geometric verification remain fixed, while only the descriptor backend is swapped. This allows us to compare the learned descriptors against the RIFT2 descriptor [12] in a common detect–describe–match–MAGSAC [13] pipeline and to assess whether gains in descriptor separability translate into fewer outliers, higher inlier ratios, and more accurate transformation estimation.

The contributions of this paper are as follows: We present a controlled comparison of two learned local descriptor architectures for thermal imagery: a modified CNN and an adapted ViT descriptorS. We perform a systematic patch-size study over 8, 16, 32, 48, 64, and 96 pixels and show that 16×16 yields the strongest descriptor separability among the evaluated patch sizes for both architectures. We introduce a complementary transform-type diagnostic analysis that reveals how descriptor behaviour changes across different transformation components. We show, through a shared-detector registration experiment, that learned descriptors substantially outperform RIFT2 in downstream registration, while also revealing the trade-off between patch size, match set consistency, and final geometric accuracy.

The rest of the paper is organised as follows. Section 2 reviews relevant work on handcrafted and learned local descriptors in multimodal and thermal matching. Section 3 describes the datasets, descriptor architectures, training procedure, and evaluation protocols. Section 4 presents the descriptor and registration results. Section 5 discusses their implications and limitations. Section 6 concludes the paper.

2. Related Work

2.1 Classical and multimodal local descriptors

Local feature design has a long history in computer vision. SIFT [14] established the classical detector–descriptor pipeline and remains a foundational reference point for scale- and rotation-invariant sparse matching. SURF [15] and ORB [16] later offered alternative trade-offs in speed and descriptor design. However, these methods were primarily developed for visible-band imagery and tend to degrade when radiometric relationships become nonlinear or modality-dependent.

For multimodal settings, a different family of descriptors emphasised local structural consistency rather than raw appearance. MIND [17] showed that local self-similarity can survive cross-modal intensity differences in deformable medical registration. DASC [18] extended this idea with dense adaptive self-correlation for multimodal and multispectral correspondence. In remote sensing and multimodal image registration, handcrafted structural descriptors such as CFOG [19], oriented-structure-map descriptors [20], modality-free handcrafted features [21], and PIIFD-based pipelines [22] further demonstrated that robust multimodal matching often depends on structural invariants rather than visible-image appearance cues.

Among multimodal descriptors, RIFT [23] is particularly notable because it was explicitly designed for radiation variation insensitive matching. It replaced gradient-based visible-image assumptions with phase-congruency-based feature handling and showed strong results across several multimodal scenarios. RIFT2 [12], extends RIFT with a faster rotation-invariance strategy while retaining the radiation-variation-insensitive design that makes RIFT suitable for multimodal matching. Together, these methods demonstrate how modality-aware descriptor design can improve robustness when conventional visible-image assumptions no longer hold. Consequently, RIFT2 provides a strong handcrafted baseline against which learned thermal descriptors can be evaluated.

2.2 Thermal and beyond-visible local feature evaluation

Benchmark studies have shown that local-feature behaviour in infrared imagery differs substantially from visible-band expectations. Johansson et al. [24] evaluated detector–descriptor combinations for infrared images and showed that performance trends are not identical to those observed in standard visible light benchmarks. Mouats et al. [25] analysed feature detectors and descriptors “beyond the visible” and demonstrated that non-uniformity noise, low contrast, and modality-specific imaging conditions materially alter matching performance. These studies are important because they justify a thermal-specific descriptor study rather than assuming that design choices for visible spectrum images transfer directly.

More recent work has continued to highlight the difficulty of thermal-visible or beyond-visible feature matching. Deshpande et al [26]. proposed a thermal local feature framework based on treating thermal imagery as colour-like structure, while several remote-sensing studies developed modality-aware local descriptors under geometric and radiometric discrepancies [27], [28]. These studies further emphasise the need for feature representations that remain robust under modality-dependent appearance changes.

Taken together, this literature supports three claims that motivate this study. First, local descriptors are central to registration because descriptor failure propagates directly to correspondence failure and unstable transformation estimation. Second, thermal imagery presents particular challenges for local matching and often requires structure-oriented or modality-robust representations. Third, despite increasing interest in learned descriptors and end-to-end correspondence models, there is still limited controlled evidence on learned local descriptors specifically for thermal-to-thermal matching

under a fixed keypoint-generation protocol.

2.3 Learned local descriptors and transformers

Deep local descriptors have largely developed within visible light computer vision. L2-Net [29], HardNet [30], and SOSNet [31] established compact patch descriptors trained with metric-learning losses, showing that learned features can outperform many handcrafted alternatives under controlled patch matching. At a system level, SuperPoint [32], D2-Net [33] and R2D2 [34] demonstrated that local features can also be learned jointly with detectors. These methods provide a methodological template for learned descriptor evaluation, but they were validated mainly on visible light benchmarks rather than thermal imagery.

Recent learned descriptor work has also explored whether purely local CNN processing is sufficient for robust feature description. Methods such as Soft Point-Wise Transformer [35], ContextDesc [36], MTLDesc [37], and AWDesc [38] suggest that broader spatial context, attention, or descriptor augmentation can improve distinctiveness when local evidence is weak, repetitive, or ambiguous. This is relevant to the present study because both descriptor architecture and support size determine how much local thermal structure is encoded. We therefore focus on a controlled CNN-versus-transformer descriptor comparison, rather than evaluating complete detector-free or end-to-end matching systems.

This motivates a controlled CNN-versus-transformer comparison for thermal descriptors. Within this framework, patch size becomes an important design parameter rather than a background implementation choice, allowing its influence on descriptor quality and downstream registration performance to be examined directly.

3. Methodolgy

3.1 Problem formulation

Given a source patch x_s, its corresponding transformed patch x_t, and a non-matching false patch x_f, the goal is to learn a descriptor function $f(\cdot)$ such that the source patch descriptor is closer to the true transformed patch than to the false patch in the descriptor space:

$$d(f(x_s),\, f(x_t)) < d(f(x_s),\, f(x_f))\,,$$

where $d(\cdot,\cdot)$ denotes the descriptor distance. In this paper, we focus on descriptor learning rather than detector learning. The detector was kept fixed during registration benchmarking so that any downstream performance changes can be attributed primarily to the descriptor.

3.2 Datasets

3.2.1 Training set

Training usesd infrared images from the HDRT dataset available through Hugging Face [39]. We used only the thermal infrared TIFF images at a spatial resolution of 640×480. A total of 10,000 thermal images were used for training and validation with split ration of 80/20. This dataset provided sufficient scale to train learned descriptors under synthetic geometric augmentation while keeping the modality purely thermal.

3.2.2 Test set

Held-out testing usesd the thermal subset of the Thermal Visual Paired Dataset from IEEE DataPort [40]. Only the thermal images were used in this study. The original thermal image size was 320×240, but the images were cropped to 269×216 in order to remove the FLIR logo and the embedded temperature scale bar. The resulting evaluation set contains 259 cropped thermal images, from which multiple transformed views were generated for held-out descriptor and registration evaluation.

3.3 Training keypoint generation

Training keypoints were produced offline using RIFT2 detector-only rather than the full RIFT2 matching pipeline. For each training image, the detector returns candidate keypoints and a response value derived from FAST corner response on the normalised phase-congruency magnitude map. Keypoints were sorted in descending order of response, then filtered to ensure full patch support. In the main patch-size experiments reported here, we retain the top 100 patch-valid keypoints per image. This kept the training supervision fixed across the patch-size comparison and avoids entangling descriptor quality with detector abundance.

3.4 Patch-pair generation

Training samples were generated on the fly from the source-keypoint index. For a given source keypoint, the pipeline extracts a square source patch of size $P \times P$, where $P \in \{8, 16, 32, 48, 64, 96\}$. A random affine transformation,

represented as a 3×3 homography matrix, was then sampled by combining independent horizontal and vertical scaling, rotation, shear, and translation. The scale factors were sampled uniformly from $[0.8, 1.2]$, the rotation angle from $[-15°, 15°]$, the shear parameter from $[-0.1, 0.1]$, and the translations from $[-0.1w, 0.1w]$ and $[-0.1h, 0.1h]$ in the horizontal and vertical directions, respectively, where w and h are the image width and height. The source image was warped by this transformation, and the source keypoint was projected into the warped image. If the projected location supported a fully valid positive patch, the corresponding target patch was extracted; otherwise, the sample was rejected and another transformation was drawn.

This procedure ensures that the network always sees anchor–positive pairs (where anchor refers to the extracted patch in the original image and positive refers to its correct transformed patch in the warped image) with known geometric correspondence but varied local appearance induced by thermal structure and synthetic viewpoint change. Both source and target patches are represented as a vector in the range $[0, 1]$.

3.5 Descriptor architectures

3.5.1 Modified ResNet34 descriptor

The CNN(ResNet34) descriptor reuses the main residual trunk of an ImageNet-pretrained ResNet34, but it is modified for single-channel thermal patch description rather than image classification. The standard visible light images classification stem is replaced with a grayscale 3×3 convolution with stride 1 and padding 1. The early max-pooling layer is removed, and the first block of layer4 is modified so that it no longer performs additional spatial downsampling. Global average pooling and the final classification layer are also removed. Instead, the network uses a patch-size-dependent convolutional descriptor head that maps the final trunk feature map to a 128-dimensional descriptor, followed by L_2 normalisation.

The modified ResNet34 therefore consists of a stride-1 grayscale stem, the four standard residual stages of ResNet34, and a convolutional descriptor head. The residual stages retain the ResNet34 block structure of 3, 4, 6, and 3 basic blocks in layers 1–4, respectively, but the spatial reduction is limited to layer2 and layer3. For an input patch of size $P \times P$, the trunk produces a feature map of size $512 \times (P/4) \times (P/4)$. The descriptor head then reduces this feature map to $128 \times 1 \times 1$, which is flattened to a 128-dimensional unit-normalised descriptor vector.

For example, for a 16×16 input patch, the tensor shapes are:

$$B{\times}1{\times}16{\times}16 \to B{\times}64{\times}16{\times}16 \to B{\times}64{\times}16{\times}16 \to B{\times}128{\times}8{\times}8 \to B{\times}256{\times}4{\times}4 \to B{\times}512{\times}4{\times}4 \to B{\times}128{\times}1{\times}1 \to B{\times}128,$$

corresponding to the input, stem, layer1, layer2, layer3, layer4, descriptor head, and final flattened descriptor, respectively. This design preserves more local spatial structure than a standard classification ResNet, where early downsampling and global pooling would remove fine patch-level information that is important for local descriptor learning.

3.5.2 Adapted ViT-B/16 descriptor

The transformer descriptor is based on the ViT-B16 architecture, adapted for single-channel thermal patch description. The patch embedding is modified from visible light to grayscale by averaging the original visible light weights. In the standard ViT-B/16 formulation, a 16 by 16 input patch would produce only one spatial token, which is limiting for local descriptor learning because the transformer would have little internal spatial structure to model. To avoid this one-token regime, we use a 4 by 4 internal tokenisation strategy, so that a 16 by 16 input patch produces a 4 by 4 grid of 16 spatial tokens, plus the class token. For the 8 by 8 setting, the same tokenisation gives a 2 by 2 grid of four spatial tokens. Positional embeddings are interpolated to the token grid required by each input patch size. The transformer encoder processes the class token and spatial tokens, and the final class-token representation is projected using a 128-dimensional linear descriptor head followed by L2 normalisation.

3.6 Training objective

Let $\mathcal{X} = \{(x_{s,i}, x_{t,i})\}_{i=1}^{B}$ denote a mini-batch of B matching patch pairs, where $x_{s,i}$ is the source (anchor) patch and $x_{t,i}$ is its corresponding transformed (positive) patch. Their descriptors are given by

$$\mathbf{a}_i = f(x_{s,i}), \qquad \mathbf{p}_i = f(x_{t,i}),$$

where both descriptors are L_2-normalised.

Following the HardNet [30] training strategy, we compute the pairwise distance matrix

$$\mathbf{D} \in \mathbb{R}^{B \times B}, \qquad D_{ij} = d(\mathbf{a}_i, \mathbf{p}_j),$$

where the distance between two normalised descriptors is

$$d(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|_2 = \sqrt{2 - 2\mathbf{a}^\top \mathbf{b}}.$$

For the i-th matching pair, the positive distance is the diagonal entry

$$d_i^+ = D_{ii}.$$

The hardest negative distance is defined as the closest non-matching descriptor in the corresponding row or column of $\mathbf{D}$:

$$d_i^- = \min\left(\min_{j:\, j\neq i} D_{ij},\ \min_{k:\, k\neq i} D_{ki}\right).$$

Thus, d_i^- corresponds to the most confusing false match for the i-th pair within the current mini-batch.

The final margin-based triplet loss is

$$\mathcal{L} = \frac{1}{B}\sum_{i=1}^{B} \max\left(0,\ m + d_i^+ - d_i^-\right),$$

where m is the margin that defines the desired separation between the positive pair and the hardest negative pair. In other words, the loss encourages the hardest negative descriptor to be at least m farther from the anchor than the true positive descriptor. Although the original HardNet formulation used $m = 1.0$, preliminary experiments with $m \in \{0.60, 0.75, 1.0\}$ showed that $m = 0.60$ gave more stable training and lower FPR@95TPR in our thermal descriptor setting, particularly for the CNN-based model. Therefore, in our main experiments we set $m = 0.60$.

Optimisation is performed using AdamW with learning rate 10^{-4}, weight decay 0.05, and batch size 128. Training is scheduled for up to 400 epochs, with early stopping patience set to 15 epochs, so runs may terminate before the full epoch budget is reached. For each run, both the best and latest checkpoints are retained.

3.7 Descriptor evaluation protocols

3.7.1 Affine Transformation test

Held-out descriptor evaluation uses a fixed geometric correspondence index built from the cropped thermal test images. In this test images are paired with five independently sampled mixed transformations per image. This yields a fixed set of source target correspondences under combined geometric transformations and serves as the primary pooled descriptor test set.

For the i-th correspondence, let $\mathbf{a}_i$ denote the descriptor of the source patch and $\mathbf{p}_i$ the descriptor of its true matching target patch. The positive distance is defined as

$$d_i^+ = d(\mathbf{a}_i, \mathbf{p}_i).$$

To construct a hard false match, we compare $\mathbf{a}_i$ against all non-matching target descriptors from the same view and retain the closest one:

$$d_i^- = \min_{j\neq i} d(\mathbf{a}_i, \mathbf{p}_j).$$

Thus, d_i^- represents the closest negative distance for the i-th correspondence.

From these distances, we compute

$$\text{mean\_d\_pos} = \frac{1}{N}\sum_{i=1}^{N} d_i^+, \qquad \text{mean\_d\_neg} = \frac{1}{N}\sum_{i=1}^{N} d_i^-,$$

where N is the total number of evaluated correspondences. Lower mean_d_pos and higher mean_d_neg indicate better descriptor separation.

We also compute the ordering fraction

$$\text{Ordering Fraction} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[d_i^+ < d_i^-\right],$$

which measures the proportion of correspondences for which the true match is closer than the hardest false match.

Our main metric is FPR@95TPR. Let τ be the distance threshold corresponding to the 95th percentile of positive distances:

$$\tau = \text{Percentile}_{95}\left(\{d_i^+\}_{i=1}^{N}\right).$$

This threshold retains 95% of true matches. The false positive rate at this operating point is then

$$\text{FPR@95TPR} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[d_i^- \leq \tau\right].$$

Lower FPR@95TPR indicates better separation between true correspondences and hard false matches.

3.7.2 Ground Truth(GT)-labelled descriptor matching test

The descriptor evaluation described above tests patch descriptors at known corresponding locations. However, in a registration pipeline, keypoints are detected independently in the source and target images before descriptor matching. Therefore, we introduce an intermediate GT-labelled descriptor matching test between descriptor verification and full registration.

In this test, keypoints are detected independently in the source image and in the transformed target image using the same RIFT2 detector-only stage. Descriptors are then extracted at the detected keypoints using the selected descriptor backend. A tentative match refers to a candidate source–target keypoint pair whose descriptors are matched by BFMatcher and retained after applying the Lowe ratio threshold of 0.95. No robust transformation estimation is performed in this test. The known ground-truth transformation, H_{gt}, is used only to label the tentative descriptor matches as correct or incorrect. For each tentative match, the source keypoint is projected into the transformed target image using H_{gt}, and the distance between this projected location and the descriptor-selected target keypoint is measured. A match is counted as GT-correct if this distance is within a predefined pixel threshold.

We report GT-correct matches and matching precision at 1 px and 3 px. The 1-pixel threshold provides a strict measure of near-exact correspondence accuracy, while the 3-pixel threshold provides a slightly more tolerant measure that accounts for detector localisation uncertainty and interpolation effects in the transformed image. To avoid artificial border effects from the transformed target image, target keypoints whose descriptor patches overlap invalid warped regions are excluded before descriptor extraction. This ensures that the GT-labelled matching test measures descriptor matching quality rather than artefacts caused by black or invalid transformed-image borders.

3.8 Shared-detector registration test

To test whether descriptor quality translates into practical registration gains, we construct a shared-detector registration benchmark. For each held-out view, the source image is loaded and the target image is created by applying the ground-truth transformation to the source image. Keypoints are then detected independently in both images using the RIFT2 detector-only stage. To ensure a fair comparison across descriptor patch sizes, all methods are evaluated using a shared support region. Specifically, we use the `common_p96` protocol, where only keypoints that can support a full 96×96 patch are retained. This prevents smaller-patch descriptors from using border keypoints that would be unavailable to larger-patch descriptors. For the transformed target image, keypoints whose descriptor patches overlap invalid warped regions are also excluded before descriptor extraction. This avoids artificial matches from black or invalid regions introduced by the geometric transformation.

Descriptors are then computed at the retained keypoints using one of three backends: the native RIFT2 descriptor, the learned ResNet descriptor, or the learned ViT descriptor. Matches are obtained using BFMatcher with a Lowe ratio threshold of 0.95, and outliers are removed by robust transformation estimation using OpenCV's `USAC_MAGSAC` estimator.For each view, we record the number of tentative matches, inliers, outliers, and the inlier ratio, together with the mean and median inlier reprojection error relative to the ground-truth transformation and the mean transformation corner error. By keeping the detector, matcher, support protocol, and robust estimation stage fixed, this benchmark isolates the effect of the descriptor backend.

4. Results

4.1 FPR@95TPR across patch sizes

The descriptor-level test shows a clear effect of patch size on matching separability. Across all tested patch sizes, the ViT descriptor achieves lower FPR@95TPR than the CNN descriptor, indicating better separation between true correspondences and hard false matches. The best overall result is obtained at patch size 16×16, where the ViT descriptor achieves an FPR@95TPR of 0.007328. The corresponding CNN result at the same patch size is 0.061969.

For the CNN descriptor, FPR@95TPR decreases from 0.109846 at p8 to 0.061969 at p16, but then increases as the patch size becomes larger: 0.114695 at p32, 0.174440 at p48, 0.248703 at p64, and 0.319462 at p96. This indicates that the CNN benefits from moving from very small patches to p16, but larger support regions reduce descriptor discrimination.

The ViT descriptor follows a similar but consistently stronger trend. It obtains 0.028154 at p8, improves to 0.007328 at p16, and then gradually worsens as patch size increases: 0.022834 at p32, 0.025328 at p48, 0.046440 at p64, and 0.091401 at p96. Although performance decreases at larger patch sizes, the ViT descriptor remains better than the CNN descriptor at every tested patch size.

This trend is also reflected in the positive and negative descriptor distances. For the CNN descriptor, the separation between `mean_d_neg` and `mean_d_pos` is largest at p16, with a gap of 0.546807. For the ViT descriptor, the largest gap is also obtained at p16, with a gap of 0.692107. This supports the FPR@95TPR result and shows that p16 provides the best balance between local detail and sufficient contextual support for this thermal descriptor task.

Overall, these results suggest that smaller local support regions are more effective for thermal patch description in

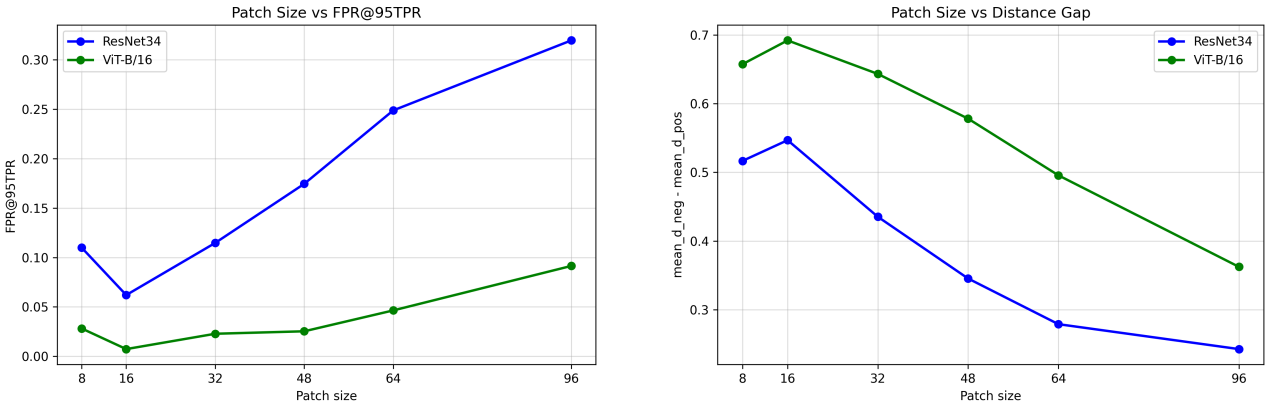

Fig. 1. *Descriptor-level performance across patch sizes under the affine transformation test. Left: FPR@95TPR, where lower values indicate better separation between true and false matches. Right: descriptor distance gap, computed as* $\texttt{mean\_d\_neg} - \texttt{mean\_d\_pos}$*, where larger values indicate stronger separation between matching and non-matching patches. The ViT descriptor consistently outperforms the CNN descriptor, with the best performance observed at p16.*

this setting. Very small patches such as p8 contain useful local structure but are less stable than p16, while larger patches appear to introduce additional thermal appearance variation that reduces descriptor separability. The p16 ViT descriptor therefore provides the strongest descriptor-level performance and is expected to provide a cleaner set of tentative matches in the downstream matching and registration tests.

4.2 Distance-distribution analysis

The distance-distribution analysis provides additional insight into why descriptor performance changes with patch size. Figure 2 shows the evolution of `mean_d_pos` and `mean_d_neg` across patch sizes and compares the positive and negative distance distributions at p16 and p96.

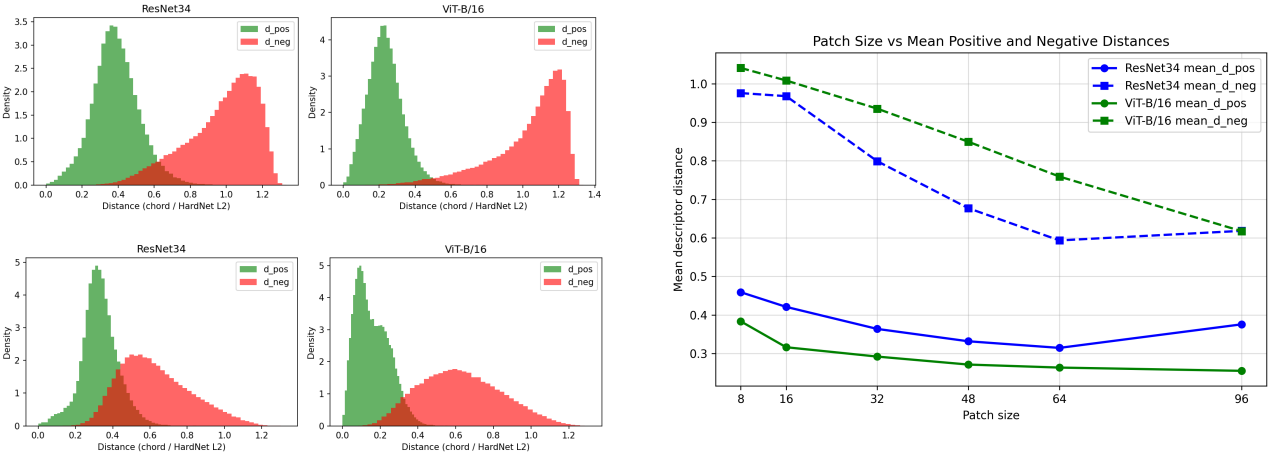

Fig. 2. *Distance-distribution analysis under the affine transformation test. The upper-left panel shows the positive and hard-negative descriptor-distance distributions at p16, while the lower-left panel shows the same distributions at p96. The increased overlap at p96 indicates weaker separation between true and false matches at larger patch sizes. The right panel summarises the mean positive and negative descriptor distances across patch sizes, showing that the ViT descriptor maintains a larger separation than the CNN descriptor, with the strongest separation observed at p16*

A clear pattern emerges. For both descriptors, increasing patch size generally reduces the negative distance, meaning that hard non-matching patches become more similar in descriptor space. This is the main reason why FPR@95TPR worsens as patch size becomes larger. For the CNN descriptor, `mean_d_neg` decreases from 0.975724 at p8 to 0.618244

Table 1. *GT-labelled descriptor matching results before transformation estimation stage.*

Descriptor	Patch	Tentative matches	GT-correct @1px	Precision @1px	GT-correct @3px	Precision @3px
CNN	p8	537.7	205.2	0.3755	306.2	0.5623
ViT	p8	518.7	224.0	0.4248	329.6	0.6273
CNN	p16	526.6	221.3	0.4125	392.3	0.7340
ViT	p16	525.6	230.2	0.4314	407.6	0.7656
CNN	p32	518.3	216.6	0.4101	397.3	0.7563
ViT	p32	537.8	228.3	0.4182	413.5	0.7596
CNN	p48	513.2	203.9	0.3884	377.8	0.7241
ViT	p48	542.0	227.9	0.4145	411.3	0.7494
CNN	p64	505.9	197.1	0.3791	363.8	0.7057
ViT	p64	546.1	226.1	0.4081	410.7	0.7431
CNN	p96	494.8	196.4	0.3874	361.4	0.7189
ViT	p96	551.2	223.6	0.3991	405.7	0.7262

at p96. For the ViT descriptor, it decreases from 1.041019 at p8 to 0.617356 at p96. This shrinkage in negative distance reduces the separation between true and false matches.

The behaviour of the positive distance is more subtle. For the ViT descriptor, `mean_d_pos` decreases steadily from 0.383575 at p8 to 0.255022 at p96, indicating that true correspondences remain relatively close even for larger patches. However, this does not translate into better overall discrimination because the negative distances collapse at the same time. In other words, although the descriptor still brings matching patches closer together, it also brings hard negatives closer, which increases overlap between the two distributions. For the CNN descriptor, `mean_d_pos` decreases up to p64 and then rises again at p96, indicating less stable behaviour across scales.

The p16 setting gives the best overall trade-off. At p16, the gap between `mean_d_neg` and `mean_d_pos` is largest for both architectures: 0.546807 for the CNN descriptor and 0.692107 for the ViT descriptor. This is consistent with the lowest FPR@95TPR values observed at the same patch size. The density plots make this especially clear. At p16, the positive and negative distributions are well separated, with relatively limited overlap. At p96, the overlap is visibly larger for both models, especially for the CNN descriptor, confirming that larger patches reduce descriptor discriminability. The ViT descriptor remains stronger than the CNN descriptor across all tested patch sizes. In the density plots, the ViT consistently places positive matches at lower distances and negative matches at higher distances, resulting in a cleaner separation. This shows that the advantage of the ViT is not limited to a single summary metric such as FPR@95TPR, but is also visible directly in the full distance distributions.

4.3 GT-labelled descriptor matching results

The GT-labelled matching results at Table 1 show that ViT produces more strict GT-correct matches than the CNN across all patch sizes. At the 1 px threshold, ViT achieves higher correct-match counts than the CNN at every patch size. For example, at p16, ViT produces 230.2 GT-correct matches compared with 221.3 for the CNN. The same trend continues at p32, p48, p64, and p96, showing that the ViT descriptor gives more accurate tentative correspondences after independent keypoint detection and descriptor matching.

The precision values also broadly support the descriptor-level FPR@95TPR trend. For ViT, precision is strongest at p16, with Precision@1px of 0.4314 and Precision@3px of 0.7656. After p16, both precision values gradually decrease as patch size increases, reaching 0.3991 at 1 px and 0.7262 at 3 px for p96. This agrees with the descriptor benchmark, where p16 also produced the lowest FPR@95TPR and the largest positive-negative distance separation.

For the CNN descriptor, the trend is similar but less perfectly monotonic. Precision@1px improves from p8 to p16 and then generally decreases at larger patch sizes. Precision@3px is highest at p32, but the overall pattern still shows that very large patches do not improve matching quality. This suggests that larger support regions may introduce additional thermal appearance variation, making the descriptor less discriminative.

Overall, the GT-labelled matching test confirms that the descriptor-level advantage of ViT transfers to the independent matching stage. The p16 ViT descriptor gives the best balance between strict correspondence accuracy and matching precision, supporting its selection as the strongest descriptor configuration before the final registration test.

Overall, this intermediate test confirms that the lower FPR@95TPR values of the ViT descriptor are not only visible in fixed-patch descriptor verification, but also lead to cleaner tentative matches after independent keypoint detection and descriptor matching. This provides the missing link between the descriptor-level benchmark and the final registration benchmark.

4.4 registration results

The registration test evaluates whether the descriptor improvements observed in the previous tests translate into more accurate final transformation estimation. The results in Table 2 show a clear advantage for the learned descriptors

Table 2. *Masked registration results using independent RIFT2 keypoint detection, descriptor matching, Lowe ratio filtering, and MAGSAC-based transformation estimation.*

Descriptor	Patch	Tentative matches	Inliers	Inlier ratio	Mean reproj. error	Mean transformation corner error
RIFT2	–	170.0	121.0	0.6208	7.8756	92.2938
CNN	8	537.7	322.9	0.5940	1.0682	2.4641
ViT	8	518.7	341.8	0.6511	0.9867	2.1096
CNN	16	526.6	431.4	0.8088	1.3414	2.3472
ViT	16	525.6	434.8	0.8178	1.2431	2.0534
CNN	32	518.3	461.5	0.8819	1.5450	3.4203
ViT	32	537.8	467.8	0.8613	1.4419	2.8969
CNN	48	513.2	449.9	0.8674	1.6494	4.5286
ViT	48	542.0	469.7	0.8577	1.4721	3.5211
CNN	64	505.9	437.3	0.8545	1.6970	5.2200
ViT	64	546.1	478.8	0.8692	1.5526	4.1609
CNN	96	494.8	432.3	0.8653	1.6750	5.1780
ViT	96	551.2	476.3	0.8560	1.5856	4.7039

over the native RIFT2 descriptor. RIFT2 produces only 170.0 tentative matches and 121.0 inliers on average, with a mean transformation corner error of 92.2938 px. In contrast, the learned descriptor models reduce the mean transformation corner error to between 2.0534 px and 5.2200 px.

The ViT descriptor gives lower transformation corner error than the CNN descriptor at every tested patch size. At p8, the error decreases from 2.4641 px for CNN to 2.1096 px for ViT. At p16, the error decreases from 2.3472 px for CNN to 2.0534 px for ViT. The same pattern continues at p32, p48, p64, and p96. This shows that the descriptor-level advantage of ViT is not limited to the FPR@95TPR test or the GT-labelled matching test, but also appears in the final registration stage.

The best overall registration result is obtained by the ViT descriptor at p16, with a mean transformation corner error of 2.0534 px. This is consistent with the earlier descriptor-level result, where p16 ViT achieved the lowest FPR@95TPR, and with the GT-labelled matching result, where p16 ViT achieved the strongest strict matching precision. Therefore, p16 appears to provide the best balance between local detail and sufficient contextual support for this thermal registration task.

However, the results also show that the link between descriptor separability and final registration accuracy is not perfectly linear. For example, some larger ViT patches produce more inliers than p16, but they do not produce lower transformation corner error. ViT p64 gives 478.8 inliers on average, compared with 434.8 for ViT p16, but its mean transformation corner error is worse, increasing from 2.0534 px to 4.1609 px. Similarly, CNN p32 has the highest CNN inlier ratio at 0.8819, but its transformation corner error is higher than CNN p16. This indicates that the number of inliers alone is not enough to explain registration accuracy.

The MAGSAC and GT agreement values provide further insight into the registration behaviour. MAGSAC is the robust estimator used after descriptor matching to select matches that are geometrically consistent with the estimated transformation. For ViT p16, 433.6 MAGSAC inliers are GT-correct on average, while only 1.2 MAGSAC inliers are GT-wrong. This means that very few incorrect matches are accepted as inliers. Therefore, the remaining transformation error is unlikely to be caused mainly by gross false matches. Instead, it is more likely affected by keypoint localisation accuracy, the spatial distribution of the inlier matches, and the geometric conditioning of the estimated transformation.

Overall, the registration results support the trend observed in the earlier tests. Lower FPR@95TPR generally leads to cleaner GT-labelled matching, and this translates into better final registration accuracy. This is most clearly shown by the p16 ViT descriptor, which achieves the lowest FPR@95TPR, strong GT-labelled matching performance, and the lowest transformation corner error. However, FPR@95TPR should still be interpreted as a descriptor-level indicator rather than a complete predictor of final registration error, because the final estimate also depends on the geometry and localisation of the matched keypoints.

5. Discussion

Our results demonstrate that patch size is a critical design variable for learned thermal descriptors. The best descriptor-level performance was obtained at 16×16, with performance generally degrading as the support region increased beyond this size. This suggests that useful correspondence information in the tested thermal images is concentrated in relatively local structures. Thermal images often contain smoother radiometric patterns, weaker texture, and less stable contextual information than visible-light images; therefore, larger support regions may introduce redundant or unstable thermal context rather than additional discriminative information.

The comparison between the CNN and ViT descriptors shows that the ViT representation is better suited to this thermal descriptor task. Across patch sizes, ViT produced lower FPR@95TPR, stronger GT-labelled matching performance, and lower final transformation corner error than CNN. This suggests that ViT forms a more discriminative representation of local thermal structure. However, the best ViT performance still occurs at p16, showing that attention-based modelling does not automatically benefit from larger support regions; the input patch must remain sufficiently local and

stable.

The GT-labelled matching test provides an important link between descriptor verification and registration. While FPR@95TPR measures descriptor separability under controlled fixed correspondences, the GT-labelled test confirms that the ViT advantage persists after independent keypoint detection, descriptor matching, and Lowe ratio filtering. This supports the conclusion that the lower FPR@95TPR translates into cleaner tentative matches, not only better fixed-patch verification.

The registration results further support this trend, but also show that descriptor evaluation alone is not a complete predictor of final geometric accuracy. The p16 ViT descriptor achieved the best final transformation accuracy, consistent with its strong FPR@95TPR and GT-labelled matching results. However, larger ViT patches sometimes produced more inliers without reducing transformation corner error. This indicates that final registration also depends on keypoint localisation precision, the spatial distribution of accepted matches, and the conditioning of the transformation estimation problem.

The weak performance of the native RIFT2 descriptor highlights the value of learning task-specific thermal descriptors. Since the detector was fixed across all methods, the improvement is mainly explained by the descriptor backend rather than by detecting different keypoints. Replacing the descriptor while keeping the detector, matcher, and robust estimator fixed substantially improved the match set and final transformation estimate.

Several limitations should be considered when interpreting these findings. First, the evaluation isolates descriptor effects and does not address detector–descriptor co-design. Second, the experiments focus on thermal-to-thermal matching using synthetically transformed target images, which provides controlled ground truth but does not fully represent real independently acquired image pairs. Third, the study does not yet extend to thermal-visible matching, where modality differences may change the optimal patch size and descriptor architecture. Future work should evaluate the descriptors on real multi-view thermal data, extend the framework to cross-modal registration, and investigate detector adaptation.

6. Conclusion

This paper presented a controlled study of learned thermal local descriptors focusing on patch size and downstream registration performance. Using modified CNN and ViT descriptors, we evaluated patch sizes from 8 to 96 pixels under fixed thermal matching and registration protocols. The results showed that 16×16 patches provided the strongest descriptor separability for both architectures, while ViT consistently outperformed CNN in the descriptor benchmark. In the registration benchmark, both learned descriptors substantially outperformed RIFT2. Although larger patches improved match cleanliness and inlier ratios, the best geometric accuracy was achieved with 16×16 patches. These findings suggest that support-region choice is a critical design variable in learned thermal descriptors and should not be treated as a fixed implementation detail. Future work will investigate detector–descriptor co-design and broader cross-modal evaluation.

References

[1] Bing Wang, Shuncong Zhong, Tung-Lik Lee, Kevin S Fancey, and Jiawei Mi. Non-destructive testing and evaluation of composite materials/structures: A state-of-the-art review. *Advances in mechanical engineering*, 12(4):1687814020913761, 2020.

[2] Ko Tomita and Michael Yit Lin Chew. A review of infrared thermography for delamination detection on infrastructures and buildings. *Sensors*, 22(2):423, 2022.

[3] FLIR Systems. *FLIR Thermal Imaging Guidebook for Industrial Applications*, 2022. Accessed: Nov. 2022. [Online]. Available: https://www.flirmedia.com/MMC/THG/Brochures/T820264/T820264_EN.pdf.

[4] Anastasia Topalidou, Nazmin Ali, Slobodan Sekulic, and Soo Downe. Thermal imaging applications in neonatal care: A scoping review. *BMC Pregnancy and Childbirth*, 19(1):381, 2019.

[5] AN Wilson, Khushi Anil Gupta, Balu Harshavardan Koduru, Abhinav Kumar, Ajit Jha, and Linga Reddy Cenkeramaddi. Recent advances in thermal imaging and its applications using machine learning: A review. *IEEE Sensors Journal*, 23(4):3395–3407, 2023.

[6] Yawen Lu and Guoyu Lu. Superthermal: Matching thermal as visible through thermal feature exploration. *IEEE Robotics and Automation Letters*, 6(2):2690–2697, 2021.

[7] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021.

[8] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020.

[9] Muhammad Shafiq and Zhaoquan Gu. Deep residual learning for image recognition: A survey. *Applied sciences*, 12(18):8972, 2022.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] Juan Terven, Diana-Margarita Cordova-Esparza, Julio-Alejandro Romero-González, Alfonso Ramírez-Pedraza, and Edgar A Chavez-Urbiola. A comprehensive survey of loss functions and metrics in deep learning. *Artificial Intelligence Review*, 58(7):195, 2025.

[12] Jiayuan Li, Y. Xu, and Mingyao Ai. RIFT2: Speeding-up RIFT with a new rotation-invariance technique. *arXiv preprint arXiv:2303.00319*, 2023.

[13] Daniel Barath, Jiri Matas, and Jana Noskova. Magsac: Marginalizing sample consensus. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10197–10205, 2019.

[14] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[15] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[16] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2564–2571, 2011.

[17] Mattias P. Heinrich, Mark Jenkinson, M. Bhushan, T. Matin, Fergus V. Gleeson, Michael Brady, and Julia A. Schnabel. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical Image Analysis*, 16(7):1423–1435, 2012.

[18] Seungryong Kim, Dongbo Min, Bumsub Ham, Seunghwan Ryu, and Kwanghoon Sohn. DASC: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2103–2112, 2015.

[19] Yuanxin Ye, Li Shen, and Chenglu Zhu. Fast and robust matching for multimodal remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11):9058–9070, 2019.

[20] T. Ma, J. Shen, and X. Gao. A local feature descriptor based on oriented structure maps with guided filtering for multispectral remote sensing image matching. *Remote Sensing*, 11(8):951, 2019.

[21] S. Cui, Y. Shen, and J. Ma. Modality-free feature detector and descriptor for multimodal remote sensing image registration. *Remote Sensing*, 12(18):2937, 2020.

[22] X. Chen, L. Liu, J. Zhang, and W. Shao. Registration of multimodal images with edge features and scale invariant PIIFD. *Infrared Physics & Technology*, 111:103549, 2020.

[23] Jiayuan Li, Qingwu Hu, and Mingyao Ai. RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Transactions on Image Processing*, 29:3296–3310, 2020.

[24] Johan Johansson, Martin Solli, and Atsuto Maki. An evaluation of local feature detectors and descriptors for infrared images. In *European Conference on Computer Vision*, pages 711–723. Springer, 2016.

[25] Tarek Mouats, Nabil Aouf, David Nam, and Stephen Vidas. Performance evaluation of feature detectors and descriptors beyond the visible. *Journal of Intelligent & Robotic Systems*, 92(1):33–63, 2018.

[26] Bhavesh Deshpande, Sourabh Hanamsheth, Yawen Lu, and Guoyu Lu. Matching as color images: Thermal image local feature detection and description. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1905–1909. IEEE, 2021.

[27] Yunyun Dong, Weili Jiao, Tengfei Long, Lanfa Liu, Guojin He, Chengjuan Gong, and Yantao Guo. Local deep descriptor for remote sensing image feature matching. *Remote Sensing*, 11(4):430, 2019.

[28] Song Cui, Miaozhong Xu, Ailong Ma, and Yanfei Zhong. Modality-free feature detector and descriptor for multimodal remote sensing image registration. *Remote Sensing*, 12(18):2937, 2020.

[29] Yurun Tian, Bin Fan, and Fuchao Wu. L2-Net: Deep learning of discriminative patch descriptor in euclidean space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[30] Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenović, and Jiří Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems*, 2017.

[31] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. SOSNet: Second order similarity regularization for local descriptor learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[32] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

[33] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A trainable CNN for joint detection and description of local features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[34] Jerome Revaud, Philippe Weinzaepfel, Cesar de Souza, and Martin Humenberger. R2D2: Repeatable and reliable detector and descriptor. In *Advances in Neural Information Processing Systems*, 2019.

[35] Zihao Wang, Xueyi Li, and Zhen Li. Local representation is not enough: Soft point-wise transformer for descriptor and detector of local features. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 1150–1156, 2021.

[36] Zixin Luo, Tianwei Shen, Lei Zhou, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. Contextdesc: Local descriptor augmentation with cross-modality context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2527–2536, 2019.

[37] Changwei Wang, Rongtao Xu, Yuyang Zhang, Shibiao Xu, Weiliang Meng, Bin Fan, and Xiaopeng Zhang. Mtldesc: Looking wider to describe better. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2388–2396, 2022.

[38] Changwei Wang, Rongtao Xu, Ke Lv, Shibiao Xu, Weiliang Meng, Yuyang Zhang, Bin Fan, and Xiaopeng Zhang. Attention weighted local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10632–10649, 2023.

[39] Jingchao Peng, Thomas Bashford-Rogers, Francesco Banterle, Haitao Zhao, and Kurt Debattista. Hdrt: A large-scale dataset for infrared-guided hdr imaging. *Information Fusion*, 120:103109, 2025.

[40] Suranjan Goswami, Nand Kumar Yadav, and Satish Kumar Singh. Thermal visual paired dataset, 2020.