# MultiPoint: Cross-spectral registration of thermal and optical aerial imagery

**Florian Achermann**[†], **Andrey Kolobov**[‡], **Debadeepta Dey**[‡], **Timo Hinzmann**[†], **Jen Jen Chung**[†], **Roland Siegwart**[†], **Nicholas Lawrance**[†]
[†]Autonomous Systems Lab, ETH Zürich; [‡]Microsoft Research, USA
{acfloria, hitimo, chungj, rsiegwart, lawrancn}@ethz.ch
{akolobov, dedey}@microsoft.com

**Abstract:** While optical cameras are ubiquitous in robotics, some robots can sense the world in several sections of the electromagnetic spectrum simultaneously, which can extend their capabilities in fundamental ways. For instance, many fixed-wing UAVs carry both optical and thermal imaging cameras, potentially allowing them to detect temperature difference-induced atmospheric updrafts, map their locations, and adjust their flight path accordingly to increase their time aloft. A key step for unlocking the potential offered by multi-spectral data is generating consistent, multi-spectral maps of the environment. In this work, we introduce MultiPoint, a novel data-driven method for generating interest points and associated descriptors for registering optical and thermal image pairs without knowledge of the relative camera viewpoints. Existing pixel-based alignment methods are accurate but too slow to work in near-real time, while feature-based methods such as SuperPoint are fast but produce poor-quality cross-spectral matches due to interest point instability in thermal images. MultiPoint capitalizes on the strengths of both approaches. An *offline* mutual information-based procedure is used to align cross-spectral image pairs from a training set, which are then processed by our generalized multi-spectral homographic adaptation stage to generate highly repeatable interest points that are invariant across viewpoint changes in both spectra. These are used to train a MultiPoint deep neural network by exposing this model to both same-spectrum and cross-spectral image pairs. This model is then deployed for fast and accurate *online* interest point detection. We show that MultiPoint outperforms existing techniques for feature-based image alignment using a dataset of real-world thermal-optical imagery captured by a UAV during flights in different conditions and release this dataset, the first of its kind.

**Keywords:** Multi-spectral robot vision; Cross-spectral image registration

## 1 Introduction

Robots have long had access to multi-spectral sensing capabilities that allow them to observe the world well beyond the limitations of human vision. Demonstrated examples include the use of sensing in visible and near-infrared spectra for monitoring crop health [1], ground penetrating radar for landmine detection [2], and others. Since many fixed-wing UAVs carry both optical and thermal infrared (TIR, a.k.a. long-wave infrared) cameras, this potentially allows them to detect regions of warm rising air called *thermals* around them, as illustrated in Fig. 1. Thermals arise due to parts of the ground surface absorbing more solar radiation than surrounding areas and heating up the air immediately above. Birds [3] and small UAVs [4, 5, 6] can gain potential energy and extend flight duration thanks to a thermal *if they happen to* fly through one. Detecting and mapping thermals *remotely* would enable UAVs to deliberatively plan their flight paths so as to maximize their time aloft and extend their range. While thermals themselves are invisible in both optical and TIR spectra, their generating regions on

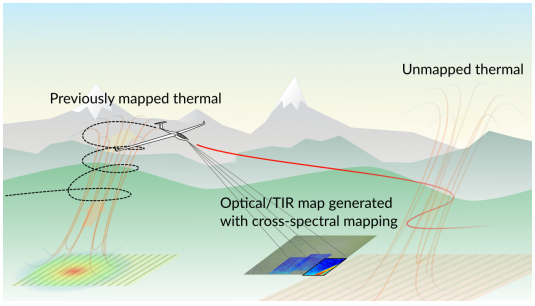


Figure 1: Remote thermal detection and mapping with TIR and visible-spectrum cameras.

4th Conference on Robot Learning (CoRL 2020), Cambridge MA, USA.

the ground can be sensed by a thermal camera on a small UAV. Crucially, generating a TIR map in-flight using standard mapping pipelines is thwarted by the poor performance of existing feature matching techniques [7, 8] on cross-spectral data [9]. To enable use cases such as this, we need a *fast, robust, and accurate* cross-spectral image registration method that would match features in thermal images with their optical counterparts and thereby allow us to leverage existing tools for optical image-based mapping to align and stitch together the corresponding TIR data.

**Proposed approach.** We present MultiPoint, a method that enables cross-spectral image registration by detecting and describing interest points common to both visible- and thermal-spectrum images. MultiPoint is fast enough to run on standard UAV hardware during flight, and makes only mild assumptions about the optical and thermal camera installation. The cameras can have different fields of view, may introduce different amounts of warping to the images, and their shutters need not be perfectly synchronized; all MultiPoint requires is that the thermal and optical images have a reasonable overlap.

MultiPoint operates via a three-stage pipeline by (1) using a base detector to label interest points, (2) robustifying the detector via homographic adaptation, i.e., generating multiple warped versions of an image to recognize only those interest points that can be consistently identified from different viewpoints, and (3) training a network that jointly identifies interest points and generates associated descriptors. Compared to SuperPoint [10], a similarly structured method for visible-spectrum features, MultiPoint differs in three critical aspects that make it successful at cross-spectral registration. In stage (1), for identifying interest points in both optical and thermal images, MultiPoint uses SURF features rather than SuperPoint's MagicPoint base detector, as the latter performs poorly at identifying consistent features in unstructured images like those of agricultural and forested terrain. In stage (2), MultiPoint generalizes homographic adaptation [10] to the multi-spectral setting by warping optical-thermal image pairs to produce an interest point detector that is consistent not only across viewpoint variations but also across spectra. In stage (3), MultiPoint trains a joint interest point detector and descriptor model using all three types of image pairs (optical-optical, thermal-thermal, and optical-thermal), with the pairs matched offline using a mutual information (MI)-maximization approach. In empirical evaluation on real-world data, we show that, thanks to these techniques, MultiPoint significantly outperforms existing approaches at cross-spectral feature detection, description, and image registration.

**Related work.** Image alignment techniques can be roughly split into pixel-based and feature-based approaches (*a more detailed treatment of related work is in the Supplement*). Pixel-based approaches perform pixel-to-pixel comparisons between images and solve for the optimal alignment, e.g., by maximizing MI. These methods work well in medical applications for aligning CT, PET and MRI scans [11], and recent improvements have further increased their accuracy in multi-spectral settings [12]. However, their computational cost is on the order of seconds for pairwise image alignments, while we are targeting real-time applications that need to operate at approximately $5\,\mathrm{Hz}$. Phase correlation methods are faster, but perform poorly when attempting to align images from multiple spectra [13, 14, 15].

Feature-based methods align images by first identifying distinct matching regions (features) and then performing alignment based only on those features. Unfortunately, classic visual features such as SIFT [7] and SURF [8] struggle when faced with multi-spectral image alignment [9] due to the non-linear pixel intensity variations that exist between images from different spectra. To address this, several methods including log-Gabor histogram descriptor (LGHD) use feature descriptors based on region information rather than pixel information [16, 17, 18]. This improves performance over standard descriptors, but the average precision values are still only around 0.24. Furthermore, evaluating over regions loses many of the computational benefits of feature-based methods, with LGHD requiring on the order of seconds to perform alignment. Even with efficiency improvements such as multi-spectral feature descriptor (MFD) [19], these methods still do not run at real-time speeds.

**Contributions.** In summary, our contributions are as follows:

- We propose MultiPoint, a method for training a deep neural network capable of performing both interest point identification and descriptor generation for cross-spectral image registration that is fast and accurate enough for real-time image alignment and mapping.
- We present a dataset collected from a UAV with a pair of downward-facing RGB and thermal cameras over 10 flights in different conditions. This is the first dataset of this kind, to our knowledge, and includes a set of aligned cross-spectral image pairs that can be used to train a model such as the one we evaluate in the experiments.

## 2 MultiPoint – Cross-spectral interest point detection and description

MultiPoint is a learning framework for generating labeled interest points with descriptors that are consistent between images coming from visible and thermal spectra as well as from different viewpoints. A trained MultiPoint network takes as input a pair of images from the same or different spectra and overlapping fields of view (but unknown relative camera poses), and returns a list of interest points and corresponding feature descriptors from both images. The resulting interest points can be matched between images and used either to estimate the relative homography between them (image alignment), or to create a sparse map of optimized multi-spectral feature locations and camera poses in 3D space with the help of a visual mapping framework.

As previously mentioned, training this network consists of three stages and requires a dataset of aligned optical and thermal image pairs. In the rest of the paper, we refer to each pair as a *multi-spectral image pair*. Details on how we collected and processed this dataset from flight data are provided in Section 3. In this section, we outline the intuition behind each of the three stages of MultiPoint and describe in detail how each of them works. Our open-source implementation of MultiPoint and the training and testing datasets are available at `https://github.com/ethz-asl/multipoint`.

### 2.1 Stage 1: Interest point label generation

Interest points are salient features in images that can be repeatably identified in multiple images under changes in viewpoint, lighting, and, in our case, across multiple spectra. The goal of MultiPoint's first stage is to produce a crude interest point identification mechanism that subsequent stages will bootstrap from and improve upon. *In particular, we would like the sets of optical and thermal interest points produced by this step to have a significant overlap, so that we can use them to match images from different spectra.*

A state-of-the-art approach for aligning *visible-spectrum* images, SuperPoint [10], trains a base detector called MagicPoint for a similar purpose. It is trained using a synthetic dataset constructed by projecting synthetic 3D scenes containing cuboids, checkerboards, and line segments to 2D. Since the corners of such shapes are naturally stable features, they are used as ground truth targets to train an interest point detector. After training MagicPoint on the same synthetic data and refining on MS COCO 14 [20] and multi-spectral data, we observed that in the latter case the interest points produced by the resulting detector cluster around edges in an image. We suspect that the synthetic dataset does not represent the multi-spectral data well, causing the transfer from the synthetic to the real domain to fail.

Our key insight for addressing this challenge is that, counterintuitively, a *non-data-driven* detector can be more consistent at generating interest points across multimodal data distributions covering conventional (optical) and unstructured and otherwise unusual images, e.g., thermal ones. Accordingly, we propose using classical detectors like SURF or SIFT instead of MagicPoint for preliminary cross-spectral interest point identification. A caveat, however, is that, unlike MagicPoint, SURF and SIFT return the locations of individual candidate interest points instead of a heatmap. This makes them very prone to small pixel-wise errors in the detected interest points. To circumvent this problem, we first construct a binary heatmap using interest points predicted by SURF and SIFT and then smooth it using a Gaussian filter.

As shown in Fig. 2, the label sets generated by the SURF/SIFT base detectors after multi-spectral homographic adaptation do not exhibit clustering around strong edges, as opposed to those generated by the MagicPoint base detector.

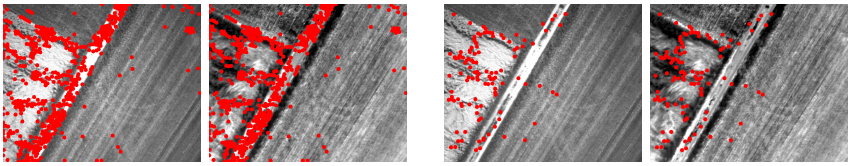


Figure 2: A comparison of different exported labels as a result of multi-spectral homographic adaption. The learned detector results in clustered interest points while the pipeline based on the SURF detector spreads interest points more uniformly.

### 2.2 Stage 2: Multi-spectral homographic adaptation

Although Stage 1 is designed to generate reasonable interest point candidates, beyond the heuristic use of SIFT/SURF to identify these candidates it takes no explicit measures to ensure that the identified points are consistent across spectra. Stage 2 improves on Stage 1's output by zeroing in on the cross-spectrally consistent subset of candidate points by employing a generalized version of a technique called *homographic adaptation* [10].

Homographic adaptation is a form of data augmentation that effectively simulates viewpoint changes by sampling and applying multiple homographies to a single image to generate a set of warped views. Interest points (identified with the detector from the previous step) can be mapped between the warped images using the known homographies, allowing the network to learn to recognize points that can be consistently identified after warping. In existing work, homographic adaptation is used to ensure feature consistency across images from a single, visible spectrum [10].

In this work, we generalize homographic adaptation to the multi-spectral domain, modifying it to boost the cross-spectral consistency of interest point detections. To do so, we apply homographic adaptation to *multi-spectral image pairs* instead of individual frames in a process depicted in Fig. 3. For a pre-aligned multi-spectral image pair, an ideal interest point detector would return the same set of locations for both images. Accordingly, we sample homographies like the original homographic adaptation does, but use each random homography to warp both images, which yields a new multi-spectral pair of aligned images with effectively a new viewpoint. Next, the detector from Stage 1 is applied to both images to obtain a heatmap pair. The intersection of the heatmap pair is determined by pixel/element-wise multiplication, represented by the $\odot$ operator. Finally, the resulting heatmap is warped back into the original frame and aggregated across the $N_h$ randomly sampled homographies. To generate "good" homographies, we follow the original homographic adaptation's approach of decomposing the transformation into a random scale, translation, in-plane rotation, and perspective distortion and sample those values from uniform distributions with pre-determined ranges.

The result is a multi-spectrum-aware scoring function $\hat{F}$ that takes as input an aligned cross-spectral image pair and returns an aggregated heatmap for generating consistent cross-spectral interest point labels for use in Stage 3:

$$\hat{F}\left(I_o; I_t; f\right) = \frac{1}{N_h} \sum_{i=1}^{N_h} H_i^{-1}\left(f\left(H_i\left(I_o\right)\right) \odot f\left(H_i\left(I_t\right)\right)\right), \tag{1}$$

where $I_o$ and $I_t$ represent the optical and thermal images respectively, $f$ is the base detector, and $H$ is a randomly sampled homography[1].

### 2.3 Stage 3: Joint detector and descriptor training

While Stages 1 and 2 could by themselves serve as an interest point detector, running them is far too slow for real-time use. Stage 3 of MultiPoint distills them into a DNN that, given a pair of *unaligned* images, can accurately identify cross-spectrally consistent interest points in them along with their descriptors in a fraction of a second.

We train this model using image pairs related with a known homography labeled with interest point locations generated by the multi-spectral homographic adaptation. To get such a pair, we randomly sample either a multi- or same-spectrum pair of aligned images and warp one of the images using a random but known homography. This results in the network seeing 50% multi-spectrum pairs and 50% same-spectrum (thermal-thermal or visible-visible) pairs during training.

The loss function is composed of a detector loss – a fully-convolutional cross entropy loss separately evaluated on each image, and a descriptor loss – a hinge loss on point correspondences. The two losses are balanced using a manually determined weighting parameter. Overall, the loss function is identical to the one used to train SuperPoint[10], but we lowered the threshold parameter for the homography-induced correspondence between the $(h, w)$ cell and the $(h', w')$ cell:

$$s_{hwh'w'} = \begin{cases} 1, \text{ if } \left\| \widehat{H} p_{hw} - p_{h'w'} \right\| \leq 4 \\ 0, \text{ otherwise.} \end{cases} \tag{2}$$

[1] We use the notation introduced in DeTone et al. [10], where $H(I)$ denotes warping the entire image $I$ with $H$ and $Hx$ represents applying the homography $H$ to the interest points $x$.

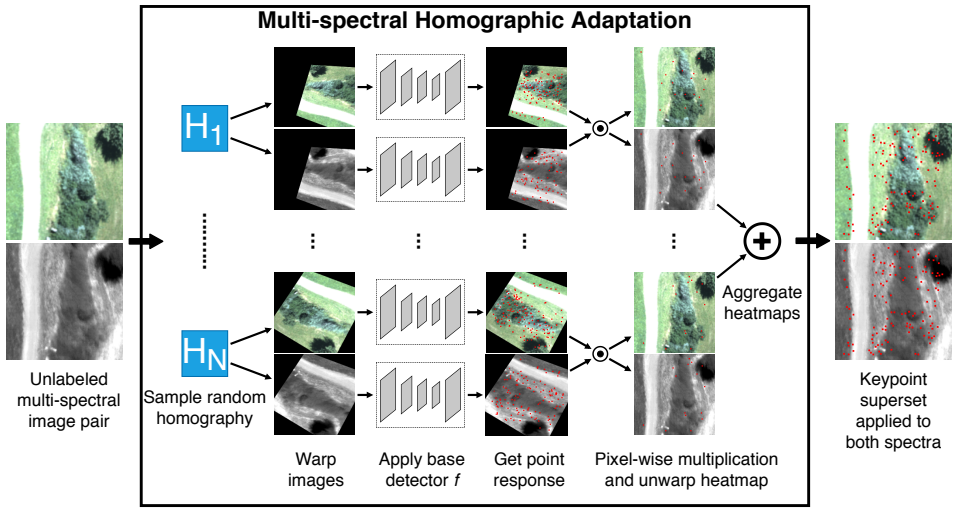


Figure 3: Multi-spectral homographic adaptation procedure for boosting the cross-spectral consistency of the interest point detector.

The center of a cell in the warped frame is represented by $p_{h'w'}$, and $\widehat{H}p_{hw}$ equals warping the cell centers from the unwarped into the warped frame. We observed a more distinctive descriptor response by decreasing this threshold from the original 8 to 4. We not only saw this effect when training MultiPoint with multi-spectral image pairs but also when retraining SuperPoint with optical-only pairs on the MS COCO 14 dataset.

## 3 Multi-spectral aerial image dataset

### 3.1 Data collection

To train the network on representative multi-spectral data, we collected data from an unmanned aerial vehicle (UAV) and compiled a dataset of aligned multi-spectral aerial images. The images were taken using a fixed-wing UAV equipped with two downward-facing cameras, one visual (UI-5261SE Rev. 4 with a $16\,\mathrm{mm}$ focal length lens) and one TIR (FLIR TAU2 $19\,\mathrm{mm}$, spectral band $7.5$–$13.5\,\mu\mathrm{m}$). The thermal camera captures images at $5\,\mathrm{Hz}$ and triggers the optical camera at the same rate, resulting in an average time offset of $63\,\mathrm{ms}$ determined by Kalibr [21], which was used to calibrate the cameras using a pinhole radial-tangential camera model. The aircraft was flown at altitudes from $80$ to $150\,\mathrm{m}$ AGL, above a mix of farmed and lightly forested terrain in Switzerland (see Fig. 4). In total, 25647 image pairs were captured in ten different flights over two days with take-off times ranging from 9 a.m. to 3 p.m., resulting in varying thermal landscapes. We observed temperature changes up to $30°\mathrm{C}$ between the early morning and afternoon flights.

### 3.2 Offline multi-spectral feature matching

To provide labeled data for training our network, we require multi-spectral image pairs with known transformations (planar homographies) between the images. However, raw images from the aerial dataset can have varying relative transformations between the images due to the trigger time offset, exposure times, and different motions within that time frame (see Fig. 5). During the data collection flights we regularly observed roll rates around $20\mathrm{deg}/\mathrm{s}$ which leads, together with the average time offset of $63\mathrm{ms}$, to a pixel error of $25.5\mathrm{px}$ in the thermal image frame ($20\mathrm{px}/\mathrm{deg}$). To calculate the 'ground-truth' homography between multi-spectral image pairs, we developed a pipeline based on the MI score. We also tested other multi-spectral matching approaches including LGHD but found that MI produced the most consistent results. At this stage, we were not concerned with running time since this dataset was generated to provide training data for the proposed MultiPoint approach.

First, we performed limited pre-processing on the raw images. RGB images were converted to greyscale and normalized. Raw thermal images are single-channel, 14-bit images where each pixel is the absolute temperature, representing a temperature range of $-40\,°\mathrm{C}$ to $160\,°\mathrm{C}$. We normalized each thermal image to the range between the 1st and 99th percentile of the raw temperature data to remove large temperature outliers and increase contrast.

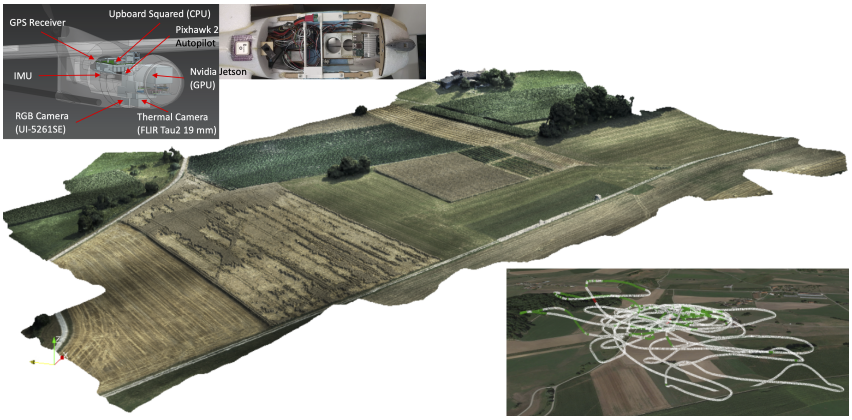


Figure 4: Mixed agricultural region where data were collected, reconstructed offline from visible-spectrum images using Pix4D (https://www.pix4d.com/). Top insets show the flight hardware. Bottom inset shows the path taken during one flight overlaid on the Google Earth (https://www.google.com/earth/) image. Green indicates regions of thermal updrafts and red indicates regions of downdrafts.

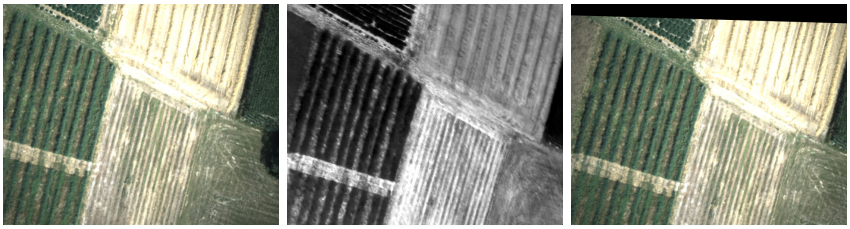

Figure 5: *Left*: Optical image warped with the homography determined from the camera calibration. *Middle*: TIR image. *Right*: cropped and aligned optical image using the optimized homography found with MI.

Next, we used an optimizer to solve for the relative homography between images using the MI score as the cost function [22]. In our setup, the thermal field of view (FoV) is a subset of the optical FoV, so we solved for the homography from the optical to the thermal frame maximising the overlap. This yielded a resolution of $512 \times 640$ for each image in each aligned image pair. A (fixed) initial estimate of the relative homography was calculated from the known camera calibration and used as a starting point for the optimizer. Our method used the Nelder-Mead solver [23] from the Python SciPy package [24]. Computing the MI score requires binning the data and then computing a 2D-histogram. We chose twice the number of bins for the thermal compared to the visible spectrum to match the resolution of the pixel values. We performed alignment with three different visible ($b_v$) and thermal ($b_t$) bin number pairs in parallel $(b_v, b_t) \in \{(32, 64), (100, 200), (256, 512)\}$. The best match was selected using the MI score. We found using these settings produced the most consistent results as image quality varied with lighting, exposure, and camera orientation. Each optimization took approximately $20\,\mathrm{s}$ on a single CPU thread.

The image pairs from the MI method were filtered for poor matches in a two-step process. First, bad matches were automatically rejected using hand-tuned thresholds for the changes in the MI-score and homography. Second, all remaining image pairs were manually checked for visually acceptable alignment with the optimized homography estimate. Finally, 13731 out of 25647 image pairs were accepted (53.54%). See Fig. 5 for an example of an accepted image pair pre- and post-alignment.

## 4 Results

MultiPoint is intended to provide interest points and corresponding descriptors for multi-spectral image pairs with different unknown viewpoints. Although the ultimate goal is to incorporate Muti-Point into a feature-based mapping framework for reconstructing dense aligned optical and thermal maps from aerial imagery, here we focus on traditional interest point metrics in order to demonstrate the core capabilities of MultiPoint versus existing detection and description methods. We show the following properties of MultiPoint on multi-spectral image pairs with different viewpoints:

1. Interest points are detected in the same locations (repeatable).

2. Descriptors provide distinct and correct matches (descriptor precision and matching scores).
3. Using standard feature-based image alignment with MultiPoint interest points results in accurate estimates of the relative homography between images (homography estimation).

### 4.1 Experimental setup

**Interest point label generation.** We used the implementations of the SURF and SIFT detectors provided by OpenCV. We evaluated different label sets, where we varied the Hessian threshold, $C_1$, for SURF and the number of retained features for SIFT. Our multi-spectral dataset has a large variety of textures where SURF often struggled to detect any interest point locations for a given detection threshold. In cases where the detector would return less than 50 interest points, we reran the detector with a lower threshold $C_2$ to mitigate this issue and still generate some interest point labels. The SIFT detector did not require this two-step process and was directly used as the base detector, as it would always return the $n$ best interest points according to the local contrast. The final set of labels to train MultiPoint were generated using the SURF base detector ($C_1 = 1500$, $C_2 = 300$). We set the filter size of the Gaussian filter $K_{gb}$ to 3 and used $N_h = 100$ random homographies for the multi-spectral homographic adaptation.

**MultiPoint training.** The training of MultiPoint was done using PyTorch [25]. We used the Adam solver with default parameters to train the model for 3000 epochs with a batch size of 32 and a learning rate of 0.001. We used photometric augmentation, specifically applying motion blur, illumination changes, contrast changes, and additive shades, plus random Gaussian and speckle noise. Additionally, as an augmentation method, as well as allowing for a batch size of 32, we randomly sampled 240x320 patches out of the full resolution images (512x640).

We partitioned the multi-spectral dataset from Section 3 into training and test sets across flights. We used a total of seven flights (9340 image pairs) for training and three flights (4391 image pairs) as test data to assess the performance of the models. All flights are over similar and potentially overlapping regions, but on different days and under different lighting conditions.

### 4.2 MultiPoint detector and descriptor performance

We assessed the detector and descriptor performance of all models under comparison on the test set that contains previously unseen multi-spectral image pairs. We compared MultiPoint, SURF, SIFT, LGHD, and SuperPoint, for which we used the weights released by MagicLeap[2]. We used the default OpenCV implementations for SURF and SIFT and a custom Python implementation of LGHD. For each image pair, at the resolution of $512 \times 640$, we computed descriptors and interest points and then evaluated with the same set of metrics as in DeTone et al. [10].

Repeatability – the ratio of interest points detected in both images (where interest points are detected at the same reprojected location in both images within a pixel threshold $\delta = 4$) to the total number of detections [26] – is a measure of interest point stability. We also report the average number of interest point detections ($N_{kp}$). Detecting too many points slows down the subsequent matching steps, and $N_{kp}$ tends to be positively correlated with repeatability since more total points are matched within the threshold but these may not be correct matches. The nearest-neighbour mean average precision (NN mAP) is the area under the precision-recall curve, and is a measure of the discriminating power of the descriptors. Matching score (M. Score) is the ratio of true positive matches over the total number of matches (match precision) and thus measures the combined performance of the interest point detections and descriptors.

Homography estimation is used to determine if the correct matches would be sufficient for accurate image alignment. An alignment is considered 'correct' if all four reprojected corners of the warped image using the estimated homography lie within a pixel threshold ($\epsilon$) of the corners of the true homography projection. We evaluated homography estimation using three different thresholds: $\epsilon = 2, 5$, and 10 pixels. We used OpenCV implementations for brute-force ($\ell_2$-norm) matching (`cv2.BFMatcher`) and to estimate the homography between matched interest points that minimizes back-projection error (`cv2.findHomography`).

The performance metrics of the different models can be seen in Tab. 1. MultiPoint clearly outperforms all baseline methods, boosting the descriptor metrics by a factor of 4. This results in superior homography estimation abilities. Since LGHD is not viewpoint-invariant, it performs considerably

[2] https://github.com/magicleap/SuperPointPretrainedNetwork

| | Detector Metrics | | Descriptor Metrics | | Homography Estimation | | |
|---|---|---|---|---|---|---|---|
| | Repeatability | $N_{kp}$ | NN mAP | M. Score | $\epsilon = 2$ | $\epsilon = 5$ | $\epsilon = 10$ |
| SURF | 0.195 | 523 | 0.002 | 0.012 | 0.003 | 0.013 | 0.022 |
| SIFT | 0.264 | 624 | 0.036 | 0.041 | 0.027 | 0.139 | 0.203 |
| LGHD | 0.162 | 599 | 0.002 | 0.005 | 0.005 | 0.035 | 0.062 |
| SuperPoint | 0.163 | 255 | 0.071 | 0.032 | 0.015 | 0.089 | 0.148 |
| MultiPoint | **0.281** | 424 | **0.271** | **0.134** | **0.125** | **0.507** | **0.667** |

Table 1: Detector and descriptor metrics on the test set with random homographies applied. MultiPoint outperforms all baseline methods due to a significantly increased descriptor performance.

worse on estimating the homography the larger the warp between the images. A similar trend can be noted for MagicLeap's SuperPoint model. This may be related to the magnitude of warps sampled during their training process versus the ones used for MultiPoint. This reveals a general drawback of learned descriptors, which tend to be less effective if test data is not represented well by training data. However, a preliminary evaluation of MultiPoint on the MS COCO 14 test set shows similar performance to SuperPoint, indicating that MultiPoint generalises well across different images and landscapes. A qualitative comparison of the different methods is shown in Fig. 6.

### 4.3 Timing

We evaluated the running time of the pipeline on flight-grade hardware such as NVIDIA Xavier NX. On the test hardware, a mixed-precision forward pass of MultiPoint for a single image pair requires, on average, $112\,\mathrm{ms}$. Non-maximum suppression requires $26\,\mathrm{ms}$ and interpolation for interest points takes $10\,\mathrm{ms}$. The total running time of the detector and descriptor pipeline is $148\,\mathrm{ms}$ ($6.75\,\mathrm{Hz}$), meeting our target performance of $5\,\mathrm{Hz}$.

## 5 Conclusion

We presented MultiPoint, a DNN and a method of training it capable of predicting interest point locations and descriptors for cross-spectral image registration. Additionally, we introduced a novel dataset for training MultiPoint that consists of aligned multi-spectral image pairs collected from a UAV, as well as a pipeline for creating consistent cross-spectral interest point labels.

In future work we plan to extend MultiPoint by: 1) exploring alternative model structures, especially smaller networks, to further improve prediction performance or inference times, and 2) incorporating MultiPoint in an online mapping framework to create consistent multi-spectral maps[27, 28].

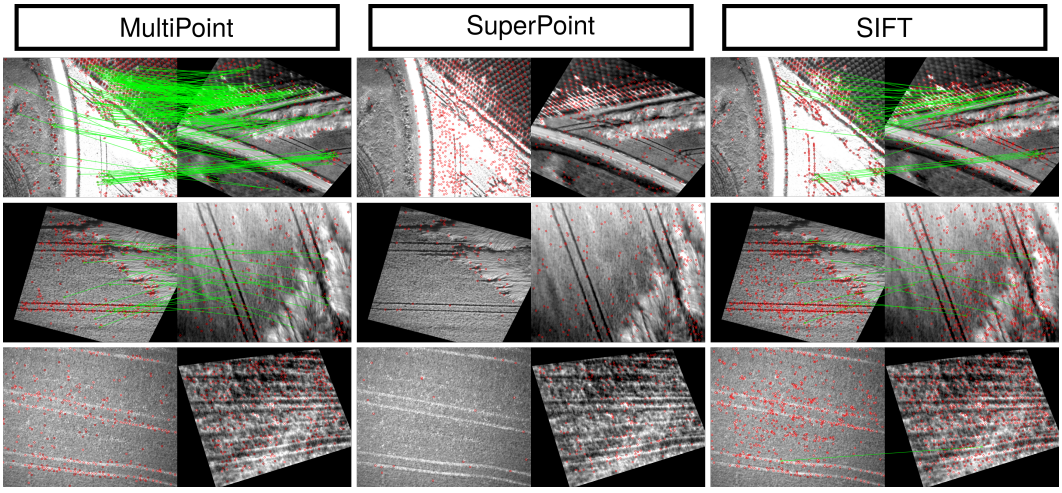


Figure 6: Qualitative results on the multi-spectral dataset. Correct matches, with a threshold of 4 pixels, are highlighted in green. Both SIFT and MultiPoint can generate correct matches if there is enough texture present in the image pair, but MultiPoint does so at a higher rate. SuperPoint struggles in all three examples because of the large warp between the images, and all three methods break down when cross-spectral differences are too dramatic *(bottom row)*.

**Acknowledgments**

This research was funded by the Microsoft Swiss Joint Research Center. We would also like to thank Thomas Mantel for his assistance in the assembly of the sensor pod used for data collection and Felix Graule for his inputs on alternative matching techniques.

## References

[1] I. Sa, Z. Chen, M. Popović, R. Khanna, F. Liebisch, J. Nieto, and R. Siegwart. weednet: Dense semantic weed classification using multispectral images and MAV for smart farming. *IEEE Robotics and Automation Letters*, 3(1):588–595, 2017.

[2] M. Schartel, R. Burr, W. Mayer, N. Docci, and C. Waldschmidt. UAV-based ground penetrating synthetic aperture radar. In *2018 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*, pages 1–4. IEEE, 2018.

[3] Z. Akos, M. Nagy, S. Leven, and T. Vicsek. Thermal soaring flight of birds and unmanned aerial vehicles. *Bioinspiration & Biomimetics*, 5(4), 2010.

[4] M. Allen. Autonomous soaring for improved endurance of a small uninhabitated air vehicle. In *43rd AIAA Aerospace Sciences Meeting and Exhibit*, page 1025, 2005.

[5] P. Oettershagen, A. Melzer, T. Mantel, K. Rudin, T. Stastny, B. Wawrzacz, T. Hinzmann, S. Leutenegger, K. Alexis, and R. Siegwart. Design of small hand-launched solar-powered uavs: From concept study to a multi-day world endurance record flight. volume 34, pages 1352–1377, 2017. URL http://dx.doi.org/10.1002/rob.21717.

[6] I. Guilliard, R. Rogahn, J. Piavis, and A. Kolobov. Autonomous thermalling as a partially observable markov decision process. In *RSS*, 2018.

[7] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157. IEEE, 1999.

[8] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[9] J. Cronje and J. De Villiers. A comparison of image features for registering LWIR and visual images. In *23rd Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, pages 1–8, 2012.

[10] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 337–349, 2018.

[11] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16(2):187–198, 1997.

[12] X. Shen, L. Xu, Q. Zhang, and J. Jia. Multi-modal and multi-spectral registration for natural images. In *European Conference on Computer Vision*, pages 309–324. Springer, 2014.

[13] C. D. Kuglin and D. C. Hines. The phase correlation image alignment method. In *Proceedings of the 1975 International Conference on Cybernetics and Society*, pages 163–165, 1975.

[14] E. De Castro and C. Morandi. Registration of translated and rotated images using finite Fourier transforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(5): 700–703, 1987.

[15] R. Szeliski. Image alignment and stitching: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 2(1):1–104, 2007.

[16] C. Aguilera, F. Barrera, A. D. Sappa, and R. Toledo. A novel SIFT-like-based approach for FIR-VS images registration. In *11th International Conference on Quantitative InfraRed Thermography*, pages 1–9, 2012.

[17] C. Aguilera, F. Barrera, F. Lumbreras, A. D. Sappa, and R. Toledo. Multispectral image feature points. *Sensors*, 12(9):12661–12672, 2012.

[18] C. A. Aguilera, A. D. Sappa, and R. Toledo. LGHD: A feature descriptor for matching across non-linear intensity variations. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 178–181. IEEE, 2015.

[19] C. F. Nunes and F. L. Pádua. A local feature descriptor based on log-gabor filters for keypoint matching in multispectral images. *IEEE Geoscience and Remote Sensing Letters*, 14(10): 1850–1854, 2017.

[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[21] J. Maye, P. Furgale, and R. Siegwart. Self-supervised calibration for robotic systems. *IEEE Intelligent Vehicles Symposium, Proceedings*, 06 2013. doi:10.1109/IVS.2013.6629513.

[22] P. Viola and W. M. Wells III. Alignment by maximization of mutual information. *International journal of computer vision*, 24(2):137–154, 1997.

[23] J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 01 1965. ISSN 0010-4620. doi:10.1093/comjnl/7.4.308. URL `https://doi.org/10.1093/comjnl/7.4.308`.

[24] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[25] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, Inc., 2019.

[26] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of computer vision*, 37(2):151–172, 2000.

[27] T. Schneider, M. T. Dymczyk, M. Fehr, K. Egger, S. Lynen, I. Gilitschenski, and R. Siegwart. maplab: An open framework for research in visual-inertial mapping and localization. *IEEE Robotics and Automation Letters*, 2018. doi:10.1109/LRA.2018.2800113.

[28] T. Hinzmann, J. L. Schönberger, M. Pollefeys, and R. Siegwart. Mapping on the fly: real-time 3d dense reconstruction, digital surface map and incremental orthomosaic generation for unmanned aerial vehicles. In *Field and Service Robotics*, pages 383–396. Springer, 2018.

# Appendices

## A Related Work

The goal of cross-spectral image alignment is to find the spatial transformation that relates two images of the same subject (overlapping fields of view) taken in different spectra usually with different cameras. This can be particularly challenging because images can appear quite different in different spectra, with some features having inverted responses, or not being visible in both images. Image alignment techniques can be roughly split into pixel-based approaches and feature-based approaches. Pixel-based methods use a pixel-to-pixel metric to evaluate the difference between images combined with an optimization scheme to solve for the optimal alignment. Such methods for performing multi-spectral image alignment include the maximization of MI, which has been shown to work well in medical applications for cross-spectral alignment, in examples such as CT, PET and MRI scans [11]. More recently, Shen et al. [12] proposed a new matching cost to better handle the dramatic structure inconsistencies and gradient variations observed in multi-spectral and multi-modal images, showing alignment accuracy improvements over MI and other state-of-the-art methods. However, the difficulty with these approaches is their computational cost, which achieves pairwise image alignments in the order of seconds, while we are targeting real-time applications that need to operate in the range of $5\,\mathrm{Hz}$. On the other hand, phase correlation methods using the frequency domain representation of images provide the computational speedups needed for our targeted application. However, despite working well for visual image alignment, these methods perform poorly when attempting to align multi-spectral images [13, 14, 15].

Feature-based methods offer efficient and rapid matching by first detecting distinct regions (features) in images and then performing alignment based only on those features. Unfortunately, classic visual features such as SIFT [7] and SURF [8] struggle when faced with multi-spectral image alignment [9]. These feature descriptors operate over pixel gradients, thus the observed performance degradation is attributed to the non-linear pixel intensity variations that exist between multi-spectral images. To address this, Aguilera et al. proposed several feature descriptors based on region information rather than pixel information [16, 17, 18]. For example, the LGHD method is based on the distribution of high frequency components in a region around a point of interest. Results for matching TIR to visual images show definite improvement over other standard descriptors; however, the overall performance is still far from satisfactory, with average precision values around 0.24. Furthermore, evaluating over regions loses many of the computational benefits of feature-based methods with LGHD still taking on the order of seconds to perform alignment. Although the similar MFD proposed by Nunes and Pádua [19] was able to halve the LGHD computation time, these methods still do not provide the real-time image alignment solutions we seek that would enable online multi-spectral aerial mapping.

## B MultiPoint Hyperparameter Study

We evaluated how hyperparameter and model choices in the MultiPoint training pipeline affect the performance of the trained model. First, we show the performance of alternative interest point label methods and their parameters compared to the MagicPoint detector trained with synthetic shape data. Second, we investigate how showing the network different combinations of cross-spectral or single-spectrum image pairs during training effects final performance. Finally, we show variations in the MultiPoint architecture. The results for these variations are summarised in Tab. 4, and each model change is described below.

**Interest Point Labels.** We generated 10 different interest point label sets using the SURF or SIFT base detector, varying the detector parameters $C_1$ and $C_2$, as well as the filter kernel size $K_{gb}$. An overview of the parameter choices and the resulting label statistics, where $N_{kp}$ represents the number of interest points per image, can be found in Tab. 2.

**Image Pair Composition.** We studied the effect of always showing MultiPoint cross-spectral image pairs or randomly sampling cross-spectral and same-spectrum image pairs ('Randomized Pairs') during Stage 3 (joint detector and descriptor training). The latter resulted in the model seeing 50% cross-spectral pairs and 50% same-spectrum (thermal-thermal or visible-visible) pairs during training.

| Label | $C_1$ | $C_2$ | $K_{gb}$ | $N_{kp,Min}$ | $N_{kp,Mean}$ | $N_{kp,Max}$ |
|---|---|---|---|---|---|---|
| SURF1 | 400 | 100 | 5 | 16 | 623 | 3852 |
| SURF2 | 1500 | 300 | 3 | 0 | 133 | 2548 |
| SURF3 | 800 | 200 | 3 | 0 | 215 | 2715 |
| SURF4 | 100 | 50 | 0 | 0 | 75 | 749 |
| SURF5 | 50 | 50 | 3 | 31 | 916 | 3092 |
| SIFT1 | 1000 | - | 3 | 0 | 160 | 482 |
| SIFT2 | 2000 | - | 3 | 0 | 359 | 1021 |
| SIFT3 | 4000 | - | 3 | 0 | 695 | 2094 |
| SIFT4 | 4000 | - | 5 | 0 | 947 | 4109 |
| SIFT5 | 4000 | - | 0 | 0 | 122 | 1027 |

Table 2: Detector settings and interest point statistics for the different labels used to train MultiPoint. $N_{kp,Mean}$ denotes the average number of interest points per image on the dataset and $N_{kp,Min}$/$N_{kp,Max}$ represent the minimum/maximum respectively.

**Descriptor size.** We explored a range of descriptor sizes to understand the trade-off between complexity and performance. Allowing larger descriptor sizes has the potential to provide a richer descriptor space at the cost of requiring additional model parameters.

**Multiple encoder heads.** We evaluated the effect of modifying the network structure proposed by SuperPoint by using multiple encoding heads, one per spectrum, instead of a single encoder. The goal of the multi-headed architecture was to determine if having independent encoder heads for each spectrum would improve performance, by allowing the network to specialize by spectrum. The network was structured to have two encoder heads that share a common structure but independent parameters, and the relevant encoder head would be selected by the input image spectrum. The interest point and descriptor decoders were still shared (see Fig. 7).

**Discussion.** An overview of the evaluated model configurations is shown in Tab. 3. The MultiPoint1_X variants are a special case where we only varied the interest point label method. Performance metrics of the different MultiPoint variants are presented in Tab. 4.

When comparing the different MultiPoint1 models we observed that the interest point labels used for training have a large influence on the overall performance of the model. Choosing the right label leads to 23.9% more accepted homographies with a threshold $\epsilon = 5$ from the worst, MultiPoint1_SIFT4, to the best model, MultiPoint1_SURF1. However, even MultiPoint1_SIFT4 significantly outperforms the best baseline method in the homography estimation task with similar descriptor metrics. For subsequent experiments (and the final MultiPoint model used in the paper) we decided to use the SURF2 label set because it resulted in the highest descriptor metrics while still performing well on the homography estimation task.

We observed that lowering the descriptor size to 64 (MultiPoint2) slightly boosted performance versus the standard SuperPoint descriptor size of 256. Further reduction to 32 (MultiPoint3) led to a minor performance drop leading us to keep descriptor size 64. The model with two encoder heads (MultiPoint5) performed on par with using only a single encoder for both spectra (MultiPoint2). We conclude that a single encoder is expressive enough to detect and describe cross-spectral features.

The model trained with randomized image pairs (MultiPoint2) outperforms the model trained with only cross-spectral pairs (MultiPoint4). We suspect that seeing same-spectrum pairs during training improves the intra-spectral performance which subsequently influences the cross-spectral performance. Training the model with randomized pairs might also help in the case where we still have minor alignment errors in the training image pairs.

Similar to the findings of SuperPoint on the COCO dataset[10] we observed that the repeatability is not a strong indicator of the overall performance of a model. In fact, the two models with the highest repeatability score, MultiPoint1_SIFT4 and MultiPoint1_SIFT5, have the lowest scores in the descriptor metrics and homography estimation. When comparing the repeatability score and the number of interest points ($N_{kp}$) for every model for every image pair in the test set, we observed a strong positive correlation ($r = 0.791$). This leads us to the conclusion that the models still struggle to find repeatable interest points in both spectra for the image pairs in the cross-spectral dataset (aerial images of a mixed agricultural region).

The final model selected as MultiPoint in the paper was the MultiPoint2 variant, with SURF2 interest point labeling, single encoder head, descriptor size 64 and trained with randomized input pairs.

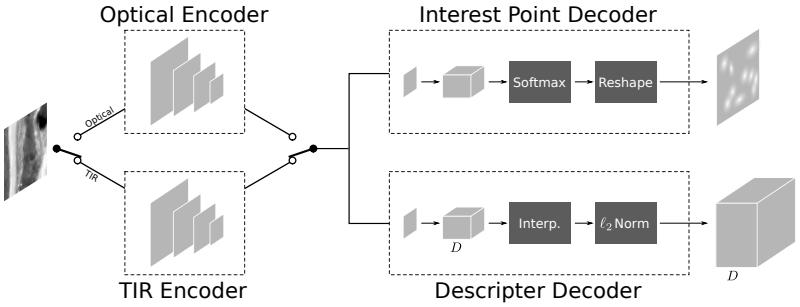


Figure 7: MultiPoint multiple encoder head architecture. Two different encoders are learned, and the image passes through either the optical or TIR encoder head depending on the image spectrum.

| Model | Labels | Multiple Heads | Randomized Pairs | Descriptor Size |
|---|---|---|---|---|
| MultiPoint1_X | X | No | Yes | 256 |
| MultiPoint2 | SURF2 | No | Yes | 64 |
| MultiPoint3 | SURF2 | No | Yes | 32 |
| MultiPoint4 | SURF2 | No | No | 64 |
| MultiPoint5 | SURF2 | Yes | Yes | 64 |

Table 3: Configuration for the different MultiPoint models. The MultiPoint1_X have the same model parameter but are trained with different interest point labels.

| | Detector Metrics | | Descriptor Metrics | | Homography Estimation | | |
|---|---|---|---|---|---|---|---|
| | Repeatability | $N_{kp}$ | NN mAP | M. Score | $\epsilon = 2$ | $\epsilon = 5$ | $\epsilon = 10$ |
| MultiPoint1_SURF1 | 0.412 | 1254 | 0.152 | 0.105 | **0.149** | **0.525** | **0.671** |
| MultiPoint1_SURF2 | 0.286 | 440 | 0.245 | 0.128 | 0.136 | 0.510 | 0.668 |
| MultiPoint1_SURF3 | 0.334 | 713 | 0.204 | 0.116 | 0.147 | 0.524 | 0.670 |
| MultiPoint1_SURF4 | 0.159 | 275 | 0.208 | 0.108 | 0.126 | 0.399 | 0.522 |
| MultiPoint1_SURF5 | 0.515 | 2197 | 0.083 | 0.081 | 0.132 | 0.467 | 0.610 |
| MultiPoint1_SIFT1 | 0.340 | 750 | 0.137 | 0.083 | 0.144 | 0.477 | 0.625 |
| MultiPoint1_SIFT2 | 0.421 | 1346 | 0.062 | 0.058 | 0.102 | 0.391 | 0.537 |
| MultiPoint1_SIFT3 | 0.563 | 2823 | 0.027 | 0.045 | 0.072 | 0.322 | 0.475 |
| MultiPoint1_SIFT4 | **0.571** | 2974 | 0.026 | 0.041 | 0.061 | 0.286 | 0.442 |
| MultiPoint1_SIFT5 | 0.311 | 1281 | 0.073 | 0.059 | 0.083 | 0.326 | 0.482 |
| **MultiPoint2** | 0.281 | 424 | 0.271 | 0.134 | 0.0125 | 0.507 | 0.667 |
| MultiPoint3 | 0.293 | 451 | 0.187 | 0.111 | 0.140 | 0.487 | 0.629 |
| MultiPoint4 | 0.271 | 473 | 0.132 | 0.088 | 0.114 | 0.427 | 0.565 |
| MultiPoint5 | 0.286 | 391 | **0.280** | **0.136** | 0.138 | 0.518 | 0.670 |

Table 4: Detector and descriptor metrics on the test set with view-point changes for the different MultiPoint variants. MultiPoint2 was selected as the best model variant for high performance across all metrics with relatively few interest points.

| | Detector Metrics | | Descriptor Metrics | | Homography Estimation | | |
|---|---|---|---|---|---|---|---|
| | Repeatability | $N_{kp}$ | NN mAP | M. Score | $\epsilon = 2$ | $\epsilon = 5$ | $\epsilon = 10$ |
| SURF | 0.248 | 630 | 0.006 | 0.022 | 0.028 | 0.088 | 0.124 |
| SIFT | 0.275 | 740 | 0.060 | 0.049 | 0.058 | 0.220 | 0.286 |
| LGHD | 0.151 | 735 | 0.088 | 0.088 | 0.122 | 0.576 | 0.752 |
| SuperPoint | 0.190 | 280 | 0.327 | 0.106 | 0.083 | 0.371 | 0.608 |
| MultiPoint2 | **0.303** | 446 | **0.462** | **0.187** | **0.257** | **0.673** | **0.793** |

Table 5: Detector and descriptor metrics on the test set without viewpoint changes. In this set of experiments LGHD and SuperPoint produce better results but are still outperformed by MultiPoint2.

## C Additional Experiment without Viewpoint Changes

We conducted a set of experiments with test data consisting only of pairs of pre-aligned images (with no additional homographic warping applied) to assess the performance of generating cross-spectral matches without viewpoint change. This more closely simulates the performance of more traditional alignment approaches such as LGHD, where cross-spectral images are likely to have only mild translations. We compared the MultiPoint2 model to the baseline methods. The results, shown in Tab. 5, show that SuperPoint and LGHD perform significantly better compared to the experiment with viewpoint changes (Tab. 1) but are still outperformed by MultiPoint.

## D Additional Experiment on MS COCO 14

We wanted to evaluate how well the MultiPoint model generalises to previously unseen data. To do so we did compare the performane of MultiPoint, trained only using the multispectral dataset, to the SuperPoint on the MS COCO 14 dataset. We conducted two sets of experiments where we varied the magnitude of the viewpoint changes. The results are shown in Tab. 6. While for smaller viewpoint changes SuperPoint outperforms MultiPoint we observe the opposite for the second experiment. We conclude that MultiPoint still performs reasonably well on this set of previously unobserved images. However, both models do not generalise that well to a change in distribution of the viewpoint changes. We infer that it is more important during training to match the correct distribution in the viewpoint changes than having as similar images as possible.

## E Additional Qualitative Examples

In Fig. 8 we show additional qualitative examples for the cross-spectral matching of MultiPoint, SuperPoint, and SIFT.

| Small Viewpoint Changes | | | | | |
|---|---|---|---|---|---|
| | Detector Metrics | | Homography Estimation | | |
| | NN mAP | M. Score | $\epsilon = 2$ | $\epsilon = 5$ | $\epsilon = 10$ |
| SuperPoint | **0.941** | **0.801** | **0.972** | **0.993** | **0.997** |
| MultiPoint | 0.818 | 0.465 | 0.832 | 0.948 | 0.974 |

| Large Viewpoint Changes | | | | | |
|---|---|---|---|---|---|
| | Detector Metrics | | Homography Estimation | | |
| | NN mAP | M. Score | $\epsilon = 2$ | $\epsilon = 5$ | $\epsilon = 10$ |
| SuperPoint | 0.410 | 0.343 | 0.506 | 0.597 | 0.626 |
| MultiPoint | **0.587** | **0.465** | **0.612** | **0.870** | **0.939** |

Table 6: Descriptor metrics on the MS COCO 14 test set with small and large viewpoint changes. In this set of experiments it depends on the magnitude of the viewpoint changes which model performs best.

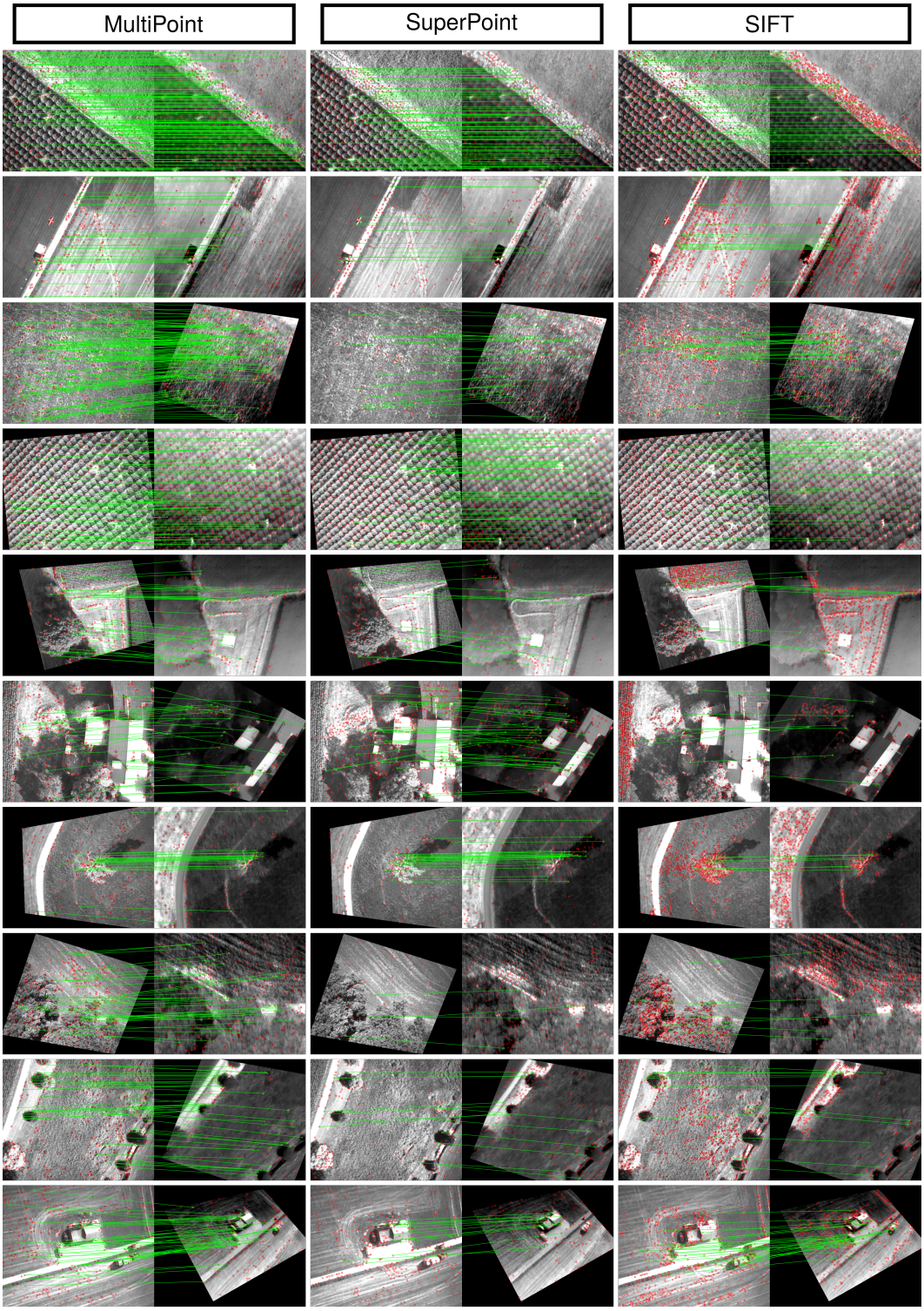


Figure 8: Additional qualitative feature matching results on the multi-spectral dataset. As in Fig. 6, correct matches with a threshold of four pixels are shown highlighted in green. The first two rows show examples with no viewpoint change (as in Appendix C).