Thermal and Visible Image Registration Using Deep Homography

Benoit Debaque, Hughes Perreault, Jean-Philippe Mercier, Marc-Antoine Drouin, Rares David, Benedicte Chatelais, Nicolas Duclos-Hindie and Simon Roy

June 10, 2022

Thermal and Visible Image Registration Using Deep Homography

B. Debaque[1], H. Perreault[1], J-P. Mercier[1], M-A. Drouin[2], R. David[1], B. Chatelais[1], N. Duclos-Hindié[1], and S. Roy[3]

[1]Thales Group, Thales Digital Solutions, Québec, Canada, benoit.debaque@ca.thalesgroup.com
[2]National Research Council of Canada, Ottawa, Canada, Marc-Antoine.Drouin@nrc-cnrc.gc.ca
[3]DRDC Valcartier Research Centre, Québec, Canada, Simon.Roy@drdc-rddc.gc.ca

***Abstract*—Fusing thermal and visible images is a recurring challenge in computer vision, especially when the images of the two modalities are not well registered. This registration problem is traditionally solved by matching descriptors and depends on the richness and discriminating power of the representation. Ensuring that detected features are dense and uniformly distributed is not necessarily guaranteed. More recently, machine learning methods addressed the issue of visible to visible matching, but few address the multi-modality setting. In this paper, we propose to address the special case of thermal-visible image registration with small baseline parallax correction. Our deep homography model is evaluated on an open thermal and visible dataset with two training settings, unsupervised and supervised. Results demonstrate the feasibility of the approach, and performances comparison to state-of-the-art models is evaluated.**

I. Introduction

The goal of multimodal image fusion is to improve image rendering to expand one's visual ability under varying illumination conditions and support the decision making process. This paper addresses multimodal image fusion where images come from both infrared (IR) and visible cameras (VI). The main challenge when performing IR/VI fusion is about extracting informative features from the visible and the thermal sources and effectively combining the two to enhance the content of the resulting composite image [24], [25], [36].

In the most general setting, the two cameras are not visually aligned, i.e., the two central perspectives do not emerge from the same projection centre causing perspective parallax. Without loss of generalization, horizontal parallax is usually observed while vertical offsets are also present, and are directly related to object distances or object depths with respect to the camera's image plane [18].

Many registration methods have been developed to solve this alignment problem [7], [17], [23], [35], [40], [65]. Here, we focus on deep learning approaches to compute the residual geometrical correction where features detection and descriptions are learned from the input images. Deep homography network is exploited [64] and adapted to multi-modality imagery, with special attention to IR/VI (photoconductive and photodiode devices) fusion.

II. Problem statement

A perspective camera can reasonably approximate the image formation with distortions [55]. When the IR and VI camera do not share the same principal points, the world or object points are related to the image pair by a direct mapping, which results in recovering the depth information of the scene by triangulation. A parallax encodes this depth information. It is responsible for the misalignment of features in fused images that sometimes create a ghosting effect. Some high constraint designs that are not discussed here avoid parallax by using a coaxial configurations [41], [53].

When world points lay on a planar surface, the relation between corresponding points from the IR and VI images can be described using homography, which is rarely the case. A special case though is met when all the points lay beyond the image resolution and are approximated by a plane at infinity. In this special case, the homography between the two views can be recovered up to a rotation (and camera internal parameters) [18].

A. Parallax challenge

For objects with complex geometry close to the cameras, parallax is observed and non-constant. Homographies are good initial approximates to align local image regions [18]. However, other mapping are possible depending on the geometric local relation, i.e., pure rotation (d.o.f. = 1), pure translation (d.o.f. = 2), similarity (d.o.f. = 4), affinity (d.o.f. = 6) and finally 2D homography or projectivity (d.o.f. = 8). Reducing the degree of freedom (d.o.f.) of geometric relations permits constraining the problem to a few unknown parameters, guaranteeing better convergence qualities. That been said, those simplified geometric models with appealing convergence property cannot fully capture the richness of the geometry present in some scenes.

B. Multimodal challenge

Thermal imagery exhibits fundamental differences compared to visible spectrum cameras. The detected signal is essentially the emitted signal from the scene and employs a different detector technology that measures the electric fluctuation of a material in the presence of thermal radiation. Another fundamental aspect with respect to parallax correction is the low texture information content present in a thermal scene, which reduces the possible number of matches between visible and thermal. There is also a non-negligible potential delocalization of features that may look similar in the two modalities; however, thermal diffusion does not follow the law of material reflections.

Most handcrafted approaches follow a detect-then-describe process. While the description of image patches around the feature brings the discriminative content and the local feature location repeatability to ensure a correct mapping, they typically perform poorly under extreme appearance changes (day, night, weak textures [14]). The lack of repeatability is the primary reason for the drop in performance; a small image region only brings low-level information. In this case, a describe-and-detect approach would be better, as the detection stage is postponed at the same stage of finding good descriptors.

III. State of the art

A. *Visible to visible matching*

Concerning the problem of image registration, some authors focused on the generalization of different geometric transforms [3], [27], [31], [38], [45], [63]. In contrast, others concentrate on discriminative representations [13], [22], [50], [59], or even run time [11], [42], [46], [51], [57] and compute more simple geometric relations. Some authors calculate a displacement field directly from neural networks [19], [39], [56], [58] to approximate an elastic deformation. Note that stereo matching is another problem that is essentially solved by working on the matching cost. Some methods can handle non-linear intensity mapping between images pairs [20].

Yang et al. [61] propose an incremental framework that searches for coarse-to-fine correspondence. AANet [60] replaces costly 3D convolutions with ingenuous multi-cost volumes. MobileStereoNet [48] implements two light networks based on MobileNet [21] blocks, one 2D and one 3D, significantly reducing computational expense. Finally, CFNet [49] improves the robustness and generalization capacity by fusing different low-resolution cost volumes.

B. *Thermal to visible matching*

Computing a local registration requires the association of pairs of points. In the case of IR/VI image pairs, this association is challenging [5], [9], [10], [29], [34], [62]. Although Deep-Learning (DL) approaches have proved their efficiency in many computers vision tasks, the advantages in using DL have not yet been demonstrated in the specific field of image matching and keypoint detection. First, defining a ground truth is not straightforward as selecting a “good” feature. Second, Convolutional Neural Networks (CNN) have difficulties achieving good performances in terms of repeatability [26], [33].

An influential subtask of thermal to visible matching is patch matching, multiple works design CNNs based on Siamese networks. Aguilera et al. [1] present three different CNN architectures and show that CNNs outperform the state-of-the-art methods. Q-Net [2] proposes a quadruplet network where the inputs are two matching pairs. AFD-Net [43] capitalizes on feature learning, specifically the multi-level feature differences. Two papers by Beaupre et al. [5], [6] work on the design of Siamese network for disparity estimation, first using a summation layer, and in the second work using a correlation and a concatenation layer.

Other methods estimate the matching without using patches. CMM-Net [52] presents a CNN designed to learn modality-specific information. Deshpande et al. [10] propose a triplet Siamese CNN where the three inputs are an anchor from an RGB image and a positive and a negative patch from a thermal image. Krishnan et al. [28] designed an intensity-based cross-modality image registration technique. CMTR [30] uses the vision transformer [12] network and adapts it to visible-infrared person re-identification. In SuperThermal [34], the authors propose a complete pipeline to learn a feature detector and descriptor for thermal images.

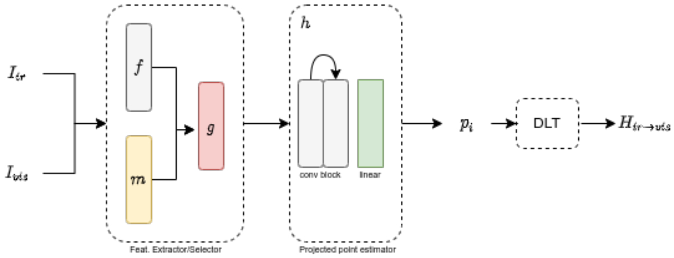


Fig. 1. Network structure

IV. Deep Homography

Correcting parallax from two views is still a problem that has not yet received a final complete solution. Employing information taken solely from passive image sensors is a challenge. Highly constrained geometrical detectors (points) are needed to guarantee a robust estimation, and correctness depends on their description richness for viewpoint invariance. Keypoints with generalization representation capabilities are usually exploited to cover as much as possible various scene content at the expense of losing their unique description.

Recent work first proposed in [11] estimate a homography to correct two visible images. Then [64] a non-supervised homography learning was proposed. The originality of the approach is that detectors and descriptors are learned from the images to obtain an optimal alignment. The current work also focuses on small baselines, i.e., viewpoints differences are relatively small.

A. *Deep Homography Estimation*

The entire network architecture contains three sections (see Figure 1); a feature extractor $f(\mathbf{x})$, a mask predictor $m(\mathbf{x})$ and a mapping $h_{ir\rightarrow vis}$, where $\mathbf{x}$ is a data vector of very high dimension $\mathbf{x} \in \Omega$ and $\{f, m, h\} \subset K$. The lower thermal resolution guided our mapping direction, as no information is loss during feature extraction in this way.

Neural networks provide approximate functions $\tilde{f}$, $\tilde{m}$ and $\tilde{h}$ from q training samples, in high dimensions, regularities are not well understood, and very often, the network architecture is found by experiments [37]. The distance of the observed values to the unknown parametrization is evaluated as an optimization procedure implying strong assumptions, i.e., K is a compact set (existence of a global optimum). This is far from being the case, as parallax offsets are not constant, and strong offsets may be significant in neighbouring regions, hence exposing multiple local minima.

Furthermore, in the vicinity of the optimum, the gradient descent requirement expressed as the Lipschitz condition [8] is a strong assumption about the slope of the manifold. Furthermore, when extracted features have naturally strong intensity dissemblance with low discriminating textural differences, the estimated set of points p do not necessarily generalize the parallax, and an over-parametrization of the manifold will diverge the search from the optimum. An under-parametrization will promote local optima or convergence will fail.

B. Network Architecture

The network structure takes two grayscale images I_{ir} and I_{vis} and produce a set of points p_i where $i = 4$. This set of points compute a mapping $h_{ir \to vis}$, where $|h_{ir \to vis}| = \{4, 8\}$. Thus, the minimum number of points is always insured to compute a mapping, constraining the network to discard spurious matches. The network architecture is taken from [64], and modified to accommodate multimodal inputs.

TABLE I
FEATURE EXTRACTOR f

Layer	1	2	3
Type	*conv*	*conv*	*conv*
Kernel	3	3	3
Stride	1	1	1
Channel	4	8	1

TABLE II
MASK PREDICTOR m

Layer	1	2	3	4	5
Type	*conv*	*conv*	*conv*	*conv*	*conv*
Kernel	3	3	3	3	3
Stride	1	1	1	1	1
Channel	4	8	16	32	1

TABLE III
POINT ESTIMATOR h

Layer	1	2	3	4	5
Type	*conv*	*pool*	*conv*	*pool*	*fc*
Kernel	7	3	3	-	-
Stride	2	2	1	1	-
Channel	64	-	1024	-	8

1) Feature Extractor: The feature extractor is a small subnetwork composed of five identical blocks. Each block is built with one fully convolutional layer of kernel size three by three, followed by one batch normalization layer and a ReLU (see Table I) and [64].

2) Mask Prediction: The mask prediction sub-network is similar to the feature extractor, see Table II. The mask prediction layers serve as an attention map and an outliers' removal operator by learning to weight the extracted features accordingly. The weighted feature map is expressed as:

$$g(x) = m(x) \odot f(x) \tag{1}$$

3) Interest Points Calculation and Mapping Computation: Once the weighted feature maps are calculated, a final network h that follows a ResNet backbone (see Table III) computes the coordinates of the projected point from the initial four corner points coordinates defining the ROI of the input mask (eight coefficients in total). It contains two layers of strided convolutions followed by a global average pooling layer. The homography is found by solving a direct linear transform (DLT) [55].

C. Unsupervised Learning

The network can learn without explicit knowledge, by minimizing the loss, features are automatically selected and extracted. The alignment of the two images serves as a constraint to specify features of interest.

1) Loss Expression: Our loss expression is the same as in [64].

$$\min_{f,m,h} L_n(I'_{ir}, I_{vi}) + L_n(I'_{vi}, I_{ir}) - \lambda L + \mu||\mathcal{H} - \mathcal{I}|| \tag{2}$$

where I'_{ir} and I'_{vi} are the projected feature masks of the visible and thermal image respectively and I_{ir} and I_{vi} the original images. $\mathcal{H}$ is the product of homographies to insure their symmetry when compared to identity matrix $\mathcal{I}$. The parameters are set to $\lambda = 2.0$ and $\mu = 0.01$. L is a loss that maximize discriminative features and L_n is the loss between the warped I' and I (see [64]).

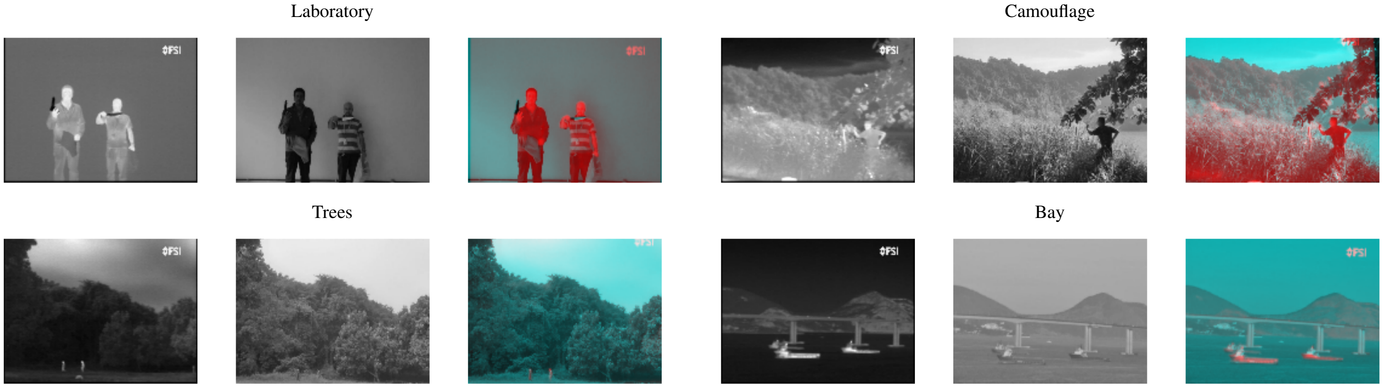
D. Supervised Learning

We selected homologous points for each sequence to evaluate our results, allowing us to compute a ground-truth homography. These annotations are used in the loss in the supervised learning setting. Since the network's output is four points with four corresponding offsets, our approach consists of calculating the homography associated with them and comparing them with the homography associated with our ground-truth points and offsets of homologous points. It results in a loss closer to our desired result, especially in a multi-modality world.

1) Loss Expression:

$$L_{rmse} = RMSE(h'_{ir \to vis}, h_{ir \to vis}) \tag{3}$$

Where $h'_{ir \to vis}$ is the homography computed with the network's output of the points and offsets, and $h_{ir \to vis}$ is the one computed with our ground-truth homologous points.



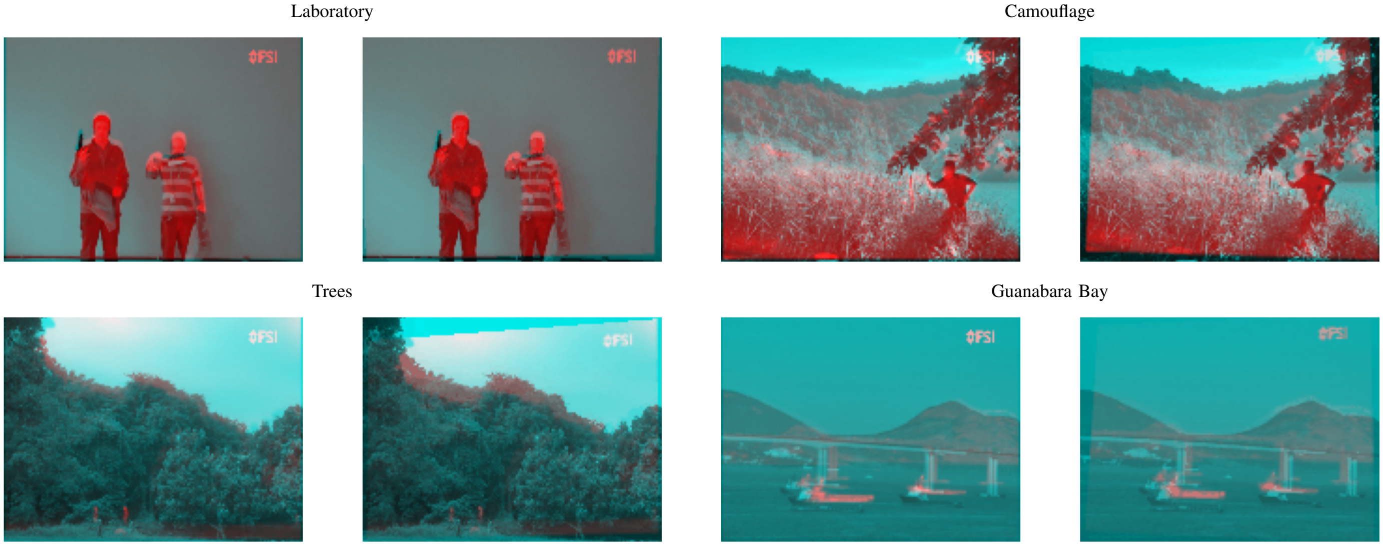
Laboratory

Camouflage

Trees

Bay

Fig. 2. For each scene results of the thermal to visible registration, thermal image (left), visible image (middle), ground truth (right).



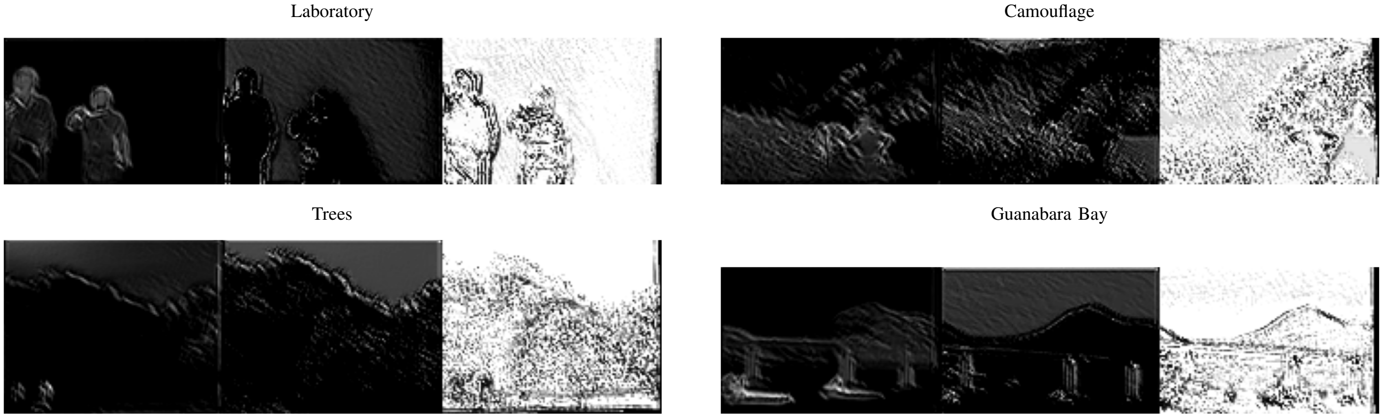
Laboratory

Camouflage

Trees

Guanabara Bay

Fig. 3. For each scene, thermal-visible fusion without registration (left), with the proposed registration (right).



Laboratory

Camouflage

Trees

Guanabara Bay

Fig. 4. For each scene, intermediate results of the thermal to visible alignment: thermal features (left), visible features (middle), mask image (right). See text for details.

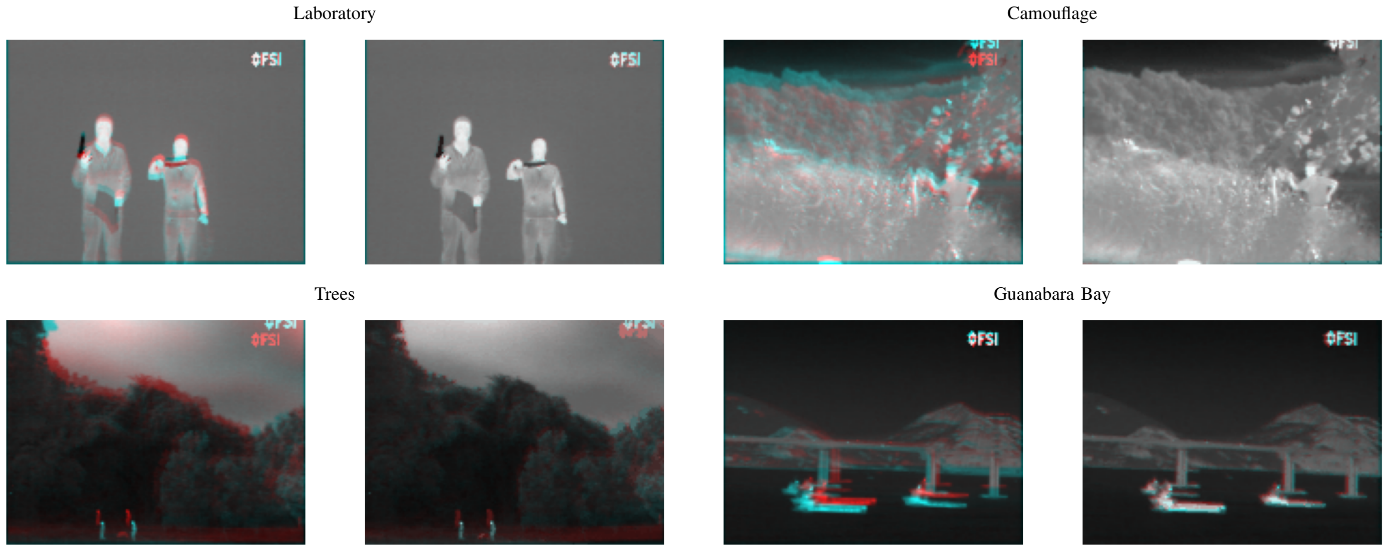


Fig. 5. For each scene, thermal to thermal fusion without registration (left), with the proposed registration (right).

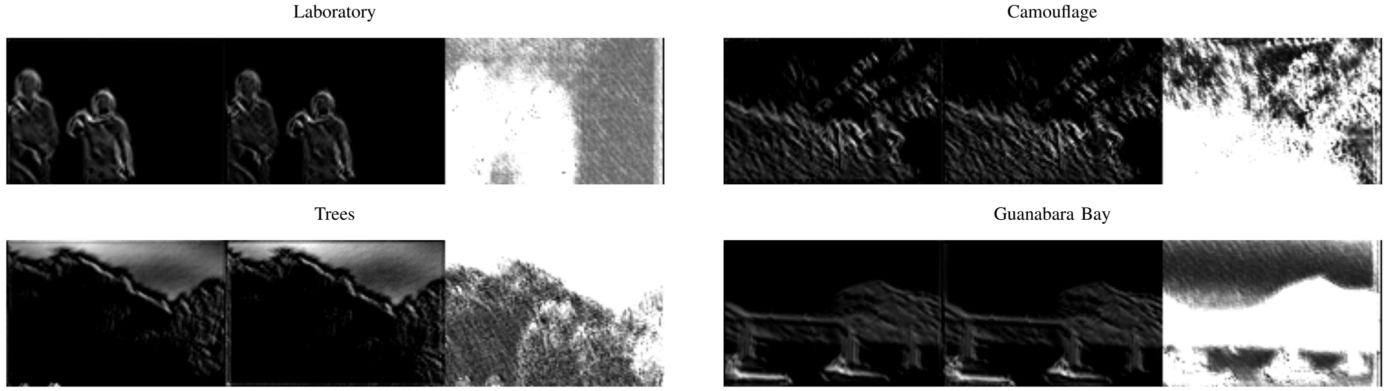


Fig. 6. For each scene, intermediate results of thermal to thermal registration: thermal features from an uncalibrated image (left), thermal features from a calibrated image (middle), mask image (right).See text for details.

V. Experiments

A. Datasets

The data used is from the Visible-Light and Infrared Video Database (VLIRVDIF) [15]. This dataset was initially conceived for image fusion but can also be used for image registration. It is composed of six scenes with multiple takes each. In the context of this paper, we use five of these six scenes (detailed in table IV), one of them not having enough thermal features to be usable for our task. Each scene is filmed in very different contexts (static/dynamic, indoor/outdoor, close/distant foreground), making this dataset suitable for thoroughly evaluating an image registration method. Both modalities are available in unsynchronized and unregistered, synchronized and unregistered, and finally synchronized and registered. The focus of this work is finding the homography to match an unregistered IR to a registered VI image.

B. Implementation details

We implement our model in Pytorch, using Pytorch Lightning for training and Hydra to set up our configuration files. An adam optimizer is used with a learning rate of 1e-4 and a batch size of 16.

TABLE IV
An overview of the sequences from the VLIRVDIF [15] dataset

Name	Distance	People	Light	Environment
Laboratory	Near	✓	Artificial	Indoor
Camouflage	Near & Far	✓	Sunlight	Outdoor
Trees	Far	✓	Sunlight	Outdoor
Guanabara Bay	Far	X	Night	Outdoor
Patio	Far	✓	Twilight	Outdoor

To train and evaluate our model, a small training and testing sets for each sequence are used, of length 200 and 50, respectively. Each sequence has a calibration phase at the beginning, which is skipped. The training and testing images are randomly selected in the remainder of the sequence. For more information about the exact list of training and testing images, please contact the authors. The default image resolution is 160x120, and patch size is 120x80. Several scales were tested, and these resolutions worked best.

Ground truthed homographies were computed from pairs of four points correspondences. Our evaluation metric is the average reprojection error between the ground truth and predicted homologous points. Using random homographies to train was not possible in our case, since rectified thermal and visible images are not available in this dataset.

The training speed is fast, averaging 0.03 seconds per image with a low-end GPU (4GB of memory). The inference speed on the other end averages 0.027 seconds per image.

VI. Results and discussion

A. Comparisons with existing solutions

Table V and VI show reprojection errors (reported as root mean square error or RMSE) for different thermal to visible sequences, respectively for a homography and an affine map. Table VII gives the reprojection errors for a homography for thermal to thermal sequences. Several sub experiments were conducted based on a different region of interest or patch size (120x80 and 120x95 pixels). Those regions are randomly selected during the training process to increase generalization performance.

As a baseline and comparison to our approach, we implemented and tested several existing solutions and compared them in the same settings. Four methods were used to perform point matching: SIFT [32], ORB [47], SOSNet [54] and CNN matching [14]. Then, three different approaches were used to select the best points to keep: RANSAC [16], USAC [44] and MAGSAC [4]. Results are shown in table V. Due to the multimodality nature of our experiments, these baselines often fail to find any matches in the images. When this happens, ** appears in the cell V. ORB did not find any matches inside our twenty pixels threshold to be considered inliers, and so was not reported.

B. Comparison Supervised Vs Unsupervised

Generally speaking, our method converge but within a few pixels and sometimes less. Offsets can reach four pixels for more complex, textured, and panoramas scenes. See Figure 2, Figure 3 and Figure 5 for a visual appreciation on anaglyphs showing residuals offsets. The supervised homography solution gives the best results for small patches, but loses its prevalence for locations with short distances. The possible causes may come from finer details that may be better expressed with the unsupervised loss under more substantial geometric constraints. Thermal-to-thermal matching is better solved in an unsupervised fashion, where the expressiveness of thermal feature produce more straightforward gradient directions. It also proves that such registration approach is possible with thermal scenes.

C. General discussion on experiments

In most cases, the algorithm can learn similar features and put a degree of importance on each, as seen in Figure 4. However, explaining the discrepancy between thermal and visible features is complex. One possible explanation to homography estimates with high errors is combining the same feature extraction network for different modalities is harder to solve. However, trials not reported here demonstrate that it is not the case. We believe that a deeper analysis is needed to explain this behaviour fully.

VII. Conclusion

This paper addressed the challenge of registering thermal and visible images with small baseline parallax correction. A supervised and an unsupervised deep homography models are presented and evaluated on an open thermal and visible dataset. Comparisons to thermal to thermal registration and affine estimation is reported as a comparative study. The primary advantage of deep homography approaches is an improved spatial stability, ensuring almost imperceptible jitters (less than a pixel) when images are fused.

Future work will include a study of network convergence in the context of a multi-scale variant of the proposed network. Also, an analysis of the network parts that learn invariant feature will be conducted, to understand better what image content is most influential in the homography estimation.

References

[1] Aguilera, C. A., Aguilera, F. J., Sappa, A. D., Aguilera, C., and Toledo, R. Learning cross-spectral similarity measures with deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2016), pp. 1–9.

[2] Aguilera, C. A., Sappa, A. D., Aguilera, C., and Toledo, R. Cross-spectral local descriptors via quadruplet network. *Sensors 17*, 4 (2017), 873.

[3] Arar, M., Ginger, Y., Danon, D., Bermano, A. H., and Cohen-Or, D. Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 13410–13419.

[4] Barath, D., Matas, J., and Noskova, J. Magsac: marginalizing sample consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 10197–10205.

[5] Beaupre, D.-A., and Bilodeau, G.-A. Siamese cnns for rgb-lwir disparity estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2019), pp. 0–0.

[6] Beaupre, D.-A., and Bilodeau, G.-A. Domain siamese cnns for sparse multispectral disparity estimation. In *2020 25th International Conference on Pattern Recognition (ICPR)* (2021), IEEE, pp. 3667–3674.

[7] Brown, L. G. A survey of image registration techniques. *ACM Comput. Surv. 24*, 4 (dec 1992), 325–376.

[8] Calin, O. *Deep learning architectures*. Springer, 2020.

[9] Cui, S., Ma, A., Wan, Y., Zhong, Y., Luo, B., and Xu, M. Cross-modality image matching network with modality-invariant feature representation for airborne-ground thermal infrared and visible datasets. *IEEE Transactions on Geoscience and Remote Sensing* (2021).

[10] Deshpande, B., Hanamsheth, S., Lu, Y., and Lu, G. 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)* (2021), IEEE, pp. 1905–1909.

[11] DeTone, D., Malisiewicz, T., and Rabinovich, A. Deep image homography estimation. *arXiv preprint arXiv:1606.03798* (2016).

[12] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[13] Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., and Brox, T. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (2015), pp. 2758–2766.

TABLE V

Homography reprojection errors (in pixels) given as mean and standard deviation (std) of the RMSE for different methods. When applicable the percentage of inlier is given (N/A stands for not applicable). ** indicates a large error.

Method		Laboratory			Camouflage			Trees			Guanabara Bay		
		mean	std	% inlier	mean	std	% inlier	mean	std	% inlier	mean	std	% inlier
SOSnet	RANSAC	**	**	**	11.43	3.8	2.3	6.77	3.26	68.61	**	**	**
	MAGSAC	**	**	**	8.71	3.6	76	4.10	1.29	97	11.03	3.79	12.52
	USAC	**	**		10.22	4.3	10	**3.87**	1.73	82.17	12.55	4.85	0.17
SIFT	RANSAC	**	**	**	**	**	**	16.94	2.83	1.86	**	**	**
	MAGSAC	**	**	**	18.60	0	0.05	15.5	2.84	1	**	**	**
	USAC	**	**	**	**	**	**	16.60	2.52	0.96	**	**	**
CNN-matching	RANSAC	**	**	**	**	**	**	**	**	**	**	**	**
	MAGSAC	**	**	**	**	**	**	14.64	0	0.05	**	**	**
	USAC	**	**	**	**	**	**	16.79	2.1	0.12	**	**	**
Unsupervised 120x80 patch		3.15	0.03	N/A	4.3	0.0002	N/A	7.83	0.04	N/A	6.36	0.004	N/A
Supervised 120x80 patch		**2.78**	0.04	N/A	**2.34**	0.03	N/A	**5.38**	0.01	N/A	**4.03**	0.036	N/A
Unsupervised 120x95 patch		3.03	0.02	N/A	5.49	0.05	N/A	9.48	0.04	N/A	7.55	0.009	N/A
Supervised 120x95 patch		2.99	0.04	N/A	3.11	0.03	N/A	7.2	0.01	N/A	5.30	0.03	N/A

TABLE VI

Affine reprojection errors (in pixels) given as mean and standard deviation (std) of the RMSE for our methods.

Method	Laboratory		Camouflage		Trees		Guanabara Bay	
	mean	std	mean	std	mean	std	mean	std
Unsupervised 120x80 patch	**2.15**	0.1	**2.51**	0.10	7.74	0.05	6.63	0.004
Supervised 120x80 patch	3.01	0.13	3.01	0.13	8.51	0.05	**3.96**	0.02
Unsupervised 120x95 patch	2.73	0.02	5.49	0.02	9.25	0.07	7.34	0.01
Supervised 120x95 patch	2.99	0.04	3.23	0.03	**6.2**	0.06	3.98	0.01

TABLE VII

Thermal unregistered to thermal registered homography reprojection errors (in pixels) given as mean and standard deviation (std) of the RMSE for our methods.

Method	Laboratory		Camouflage		Trees		Guanabara Bay	
	mean	std	mean	std	mean	std	mean	std
Unsupervised 120x80 patch	**0.62**	0.004	**0.96**	0.003	**4.81**	0.008	**1.07**	0.003
Supervised 120x80 patch	2.71	0.02	2.44	0.13	5.73	0.02	3.99	0.006
Unsupervised 120x95 patch	1.11	0.005	1.24	0.003	6.44	0.00	2.90	0.005
Supervised 120x95 patch	2.81	0.03	3.46	0.03	7.69	0.02	4.89	0.02

[14] Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., and Sattler, T. D2-net: A trainable cnn for joint detection and description of local features. *arXiv preprint arXiv:1905.03561* (2019).

[15] Ellmauthaler, A., Pagliari, C. L., Silva, E. A., Gois, J. N., and Neves, S. R. A visible-light and infrared video database for performance evaluation of video/image fusion methods. *Multidimensional Syst. Signal Process. 30*, 1 (jan 2019), 119–143.

[16] Fischler, M. A., and Bolles, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM 24*, 6 (1981), 381–395.

[17] Georgiou, T., Liu, Y., Chen, W., and Lew, M. A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision. *International Journal of Multimedia Information Retrieval 9*, 3 (2020), 135–170.

[18] Hartley, R., and Zisserman, A. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[19] Haskins, G., Kruger, U., and Yan, P. Deep learning in medical image registration: a survey. *Machine Vision and Applications 31*, 1 (2020), 1–18.

[20] Hirschmuller, H. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (2005), vol. 2, pp. 807–814 vol. 2.

[21] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

[22] Jaderberg, M., Simonyan, K., Zisserman, A., et al. Spatial transformer networks. *Advances in neural information processing systems 28* (2015).

[23] Jiang, X., Ma, J., Xiao, G., Shao, Z., and Guo, X. A review of multimodal image matching: Methods and applications. *Information Fusion 73* (2021), 22–71.

[24] Jin, X., Jiang, Q., Yao, S., Zhou, D., Nie, R., Hai, J., and He, K. A survey of infrared and visual image fusion methods. *Infrared Physics & Technology 85* (2017), 478–501.

[25] Kaur, H., Koundal, D., and Kadyan, V. Image fusion techniques: a survey. *Archives of Computational Methods in Engineering 28*, 7 (2021), 4425–4447.

[26] Koopman, T., Martens, R., Gurney-Champion, O. J., Yaqub, M., Lavini, C., de Graaf, P., Castelijns, J., Boellaard, R., and Marcus, J. T. Repeatability of ivim biomarkers from diffusion-weighted mri in head and neck: Bayesian probability versus neural network. *Magnetic resonance in medicine 85*, 6 (2021), 3394–3402.

[27] Krebs, J., Mansi, T., Delingette, H., Zhang, L., Ghesu, F. C., Miao, S., Maier, A. K., Ayache, N., Liao, R., and Kamen, A. Robust non-rigid registration through agent-based action learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2017), Springer, pp. 344–352.

[28] Krishnan, P. T., Balasubramanian, P., and Jeyakumar, V. Histogram matched visible and infrared image registration for face detection. In *IEEE EUROCON 2021-19th International Conference on Smart Technologies* (2021), IEEE, pp. 222–226.

[29] Li, J., Hu, Q., and Ai, M. Rift: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Transactions on Image Processing 29* (2020), 3296–3310.

[30] Liang, T., Jin, Y., Gao, Y., Liu, W., Feng, S., Wang, T., and Li, Y. Cmtr: Cross-modality transformer for visible-infrared person re-identification. *arXiv preprint arXiv:2110.08994* (2021).

[31] Liao, R., Miao, S., de Tournemire, P., Grbic, S., Kamen, A., Mansi, T., and Comaniciu, D. An artificial agent for robust image registration. In *Proceedings of the AAAI conference on artificial intelligence* (2017), vol. 31.

[32] Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision 60*, 2 (2004), 91–110.

[33] Lu, Y., Fan, Y., Lv, J., and Stafford Noble, W. Deeppink: reproducible feature selection in deep neural networks. *Advances in neural information processing systems 31* (2018).

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

[35] Ma, J., Jiang, X., Fan, A., Jiang, J., and Yan, J. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision 129*, 1 (2021), 23–79.

[36] Ma, J., Ma, Y., and Li, C. Infrared and visible image fusion methods and applications: A survey. *Information Fusion 45* (2019), 153–178.

[37] Mallat, S. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374*, 2065 (2016), 20150203.

[38] Miao, S., Piat, S., Fischer, P., Tuysuzoglu, A., Mewes, P., Mansi, T., and Liao, R. Dilated fcn for multi-agent 2d/3d medical image registration. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2018), vol. 32.

[39] Mok, T. C., and Chung, A. Fast symmetric diffeomorphic image registration with convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 4644–4653.

[40] Mouats, T., Aouf, N., Nam, D., and Vidas, S. Performance evaluation of feature detectors and descriptors beyond the visible. *Journal of Intelligent & Robotic Systems 92*, 1 (2018), 33–63.

[41] Ogino, Y., Shibata, T., Tanaka, M., and Okutomi, M. Coaxial visible and FIR camera system with accurate geometric calibration. In *Thermosense: Thermal Infrared Applications XXXIX* (2017), P. Bison and D. Burleigh, Eds., vol. 10214, International Society for Optics and Photonics, SPIE, pp. 319 – 324.

[42] Poursaeed, O., Yang, G., Prakash, A., Fang, Q., Jiang, H., Hariharan, B., and Belongie, S. Deep fundamental matrix estimation without correspondences. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (2018), pp. 0–0.

[43] Quan, D., Liang, X., Wang, S., Wei, S., Li, Y., Huyan, N., and Jiao, L. Afd-net: Aggregated feature difference learning for cross-spectral image patch matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 3017–3026.

[44] Raguram, R., Chum, O., Pollefeys, M., Matas, J., and Frahm, J.-M. Usac: A universal framework for random sample consensus. *IEEE transactions on pattern analysis and machine intelligence 35*, 8 (2012), 2022–2038.

[45] Revaud, J., Weinzaepfel, P., Harchaoui, Z., and Schmid, C. Deepmatching: Hierarchical deformable dense matching. *International Journal of Computer Vision 120*, 3 (2016), 300–323.

[46] Rocco, I., Arandjelovic, R., and Sivic, J. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 6148–6157.

[47] Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision* (2011), Ieee, pp. 2564–2571.

[48] Shamsafar, F., Woerz, S., Rahim, R., and Zell, A. Mobilestereonet: Towards lightweight deep networks for stereo matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2022), pp. 2417–2426.

[49] Shen, Z., Dai, Y., and Rao, Z. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 13906–13915.

[50] Simonovsky, M., Gutiérrez-Becker, B., Mateus, D., Navab, N., and Komodakis, N. A deep metric for multimodal registration. In *International conference on medical image computing and computer-assisted intervention* (2016), Springer, pp. 10–18.

[51] Sokooti, H., Vos, B. d., Berendsen, F., Lelieveldt, B. P., Išgum, I., and Staring, M. Nonrigid image registration using multi-scale 3d convolutional neural networks. In *International conference on medical image computing and computer-assisted intervention* (2017), Springer, pp. 232–239.

[52] Song, H., Xu, W., Liu, D., Liu, B., Liu, Q., and Metaxas, D. N. Multi-stage feature fusion network for video super-resolution. *IEEE Transactions on Image Processing 30* (2021), 2923–2934.

[53] Takahata, T. Coaxiality evaluation of coaxial imaging system with concentric silicon–glass hybrid lens for thermal and color imaging. *Sensors 20*, 20 (2020).

[54] Tian, Y., Yu, X., Fan, B., Wu, F., Heijnen, H., and Balntas, V. Sosnet: Second order similarity regularization for local descriptor learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 11016–11025.

[55] Toth, C. Photogrammetric computer vision: statistics, geometry, orientation and reconstruction. *Photogrammetric engineering & remote SenSing 83*, 10 (2017), 661–662.

[56] Truong, P., Danelljan, M., and Timofte, R. Glu-net: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 6258–6268.

[57] Vos, B. D. d., Berendsen, F. F., Viergever, M. A., Staring, M., and Išgum, I. End-to-end unsupervised deformable image registration with a convolutional neural network. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2017, pp. 204–212.

[58] Wang, J., and Zhang, M. Deepflash: An efficient network for learning-based medical image registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 4444–4452.

[59] Wu, G., Kim, M., Wang, Q., Munsell, B. C., and Shen, D. Scalable high-performance image registration framework by unsupervised deep feature representations learning. *IEEE Transactions on Biomedical Engineering 63*, 7 (2015), 1505–1516.

[60] Xu, H., and Zhang, J. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 1959–1968.

[61] Yang, G., Manela, J., Happold, M., and Ramanan, D. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 5515–5524.

[62] Yang, N., Yang, Y., Li, P., and Gao, F. Research on infrared and visible image registration of substation equipment based on multi-scale retinex and asift features. In *Sixth International Workshop on Pattern Recognition* (2021), vol. 11913, International Society for Optics and Photonics, p. 1191303.

[63] Yang, X., Kwitt, R., Styner, M., and Niethammer, M. Quicksilver: Fast predictive image registration–a deep learning approach. *NeuroImage 158* (2017), 378–396.

[64] Zhang, J., Wang, C., Liu, S., Jia, L., Ye, N., Wang, J., Zhou, J., and Sun, J. Content-aware unsupervised deep homography estimation. In *European Conference on Computer Vision* (2020), Springer, pp. 653–669.

[65] Zitova, B., and Flusser, J. Image registration methods: a survey. *Image and vision computing 21*, 11 (2003), 977–1000.