Multimodal Deep Homography Estimation Using a Domain Adaptation Generative Adversarial Network

T. Pouplin[1], H. Perreault[1], B. Debaque[1], M-A. Drouin[2], 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*—Multimodal image registration is a challenging task. To begin with, the variation of parallax in the images makes the process intrinsically tricky. Additionally, due to phenomenology differences in modalities, the appearance of the same feature may vary significantly between the images making the registration laborious. To help mitigate these issues, we propose a two-step approach targeted at visible and infrared imagery. First, we train a generative adversarial network to learn the domain transfer function between the visible and the infrared domain, thereby mitigating the impact of the visual dissimilarity between the images. Second, we train a deep Siamese network to compute a homography in an unsupervised setting. Both elements are combined and trained sequentially. Our method is evaluated on a publicly available dataset. Our results show that the proposed method provides a reduction of more than 30% on average from the previous state-of-the-art, and outperforms several baselines and recent deep homography methods.**

I. Introduction

Image registration is an essential step of many computer vision applications that require images from different viewpoints to be transformed into the same coordinate system. This paper addresses the more challenging task of multimodal image registration. This task is arduous since the visual appearance of a feature can significantly vary between the modalities. In this paper, we focus on infrared and visible imagery.

Several registration methods have been proposed [1]–[6]. Some of those methods use a homography to establish the mapping between the images. A homography is a projective transformation that provides a valid mapping when the rigid transformation between the views is limited to a rotation or when the scene is composed of a single plane. For more general scenes and rigid transformations, a homography can still be used to approximate this mapping assuming that the translation between the views is small with respect to the distance of the scene.

We propose a domain transfer generative adversarial network (GAN) to facilitate the multimodal image registration prior to estimating a homography using a Siamese deep networks. We hypothesize that a domain shift from visible to infrared performed by a GAN will improve the performance of the deep homography estimator.

Most classical homography estimators are based on features such as corners, lines or more modern descriptors that are rotation invariant, illumination invariant and/or scale invariant [7]. The use of those detectors is challenging in the multimodality context since the appearance of matching features can vary significantly between modalities [8]. In recent work, Debaque *et al.* [9] made some progress using a deep homography estimator with a Siamese backbone. Their network could find some common features between the visible and the infrared domains, although it failed on some imagery. To deal with those pathological cases, we propose to use a domain transfer GAN to preprocess the visible image. The preprocessed image is then fed to a deep homography network that estimates the projective transformation between the visible and infrared images.

Fig. 1. Our model finds the homography matrix that best registers images from the visible and infrared spectrum. On the left is the visible image, in the middle the infrared, and on the right is a fusion of the visible and the warped infrared using an anaglyph.

In this paper, we exploit the recent success of GAN to solve various image-to-image translation problems that are somewhat related to ours. Notable examples include CycleGAN [10], Pix2Pix [11] and ThermalGAN [12]. These networks can perform style transfer, season change, day to night, edges to photo, etc. For certain tasks, GANs are known to hallucinate some image features. In our context, this is not an issue since the GAN-generated image is never meant to be displayed to users and are rather fed to a CNN. A change in illumination or texture of the image might affect the visual quality, but not necessarily the registration quality, once the image is transformed into the latent space.

We trained the proposed method using a visible and infrared public dataset which includes registered and unregistered imagery. We first use registered infrared and visible images to train the modality transfer GAN. In the second stage, we use the unregistered infrared and registered visible to estimate a homography between the two (as seen in Figure 1), and validate our process using handmade ground truth. Note that the ground truth is only used for testing, as the training is done in an unsupervised way.

The remainder of this paper is divided as follows. Section II presents the related work. An overview of the proposed approach is outlined in Section III. Finally, Section IV, V and VI presents the experiments, the discussion and the final remarks.

II. State of the art

A. Image registration

There are several families of methods that perform image registration. Some methods produce various geometric transforms directly [13]–[18], while others produce an elastic transformation using a displacement field [19]–[22]. Some authors change focus on other topics like the representation [23]–[26] or run-time [27]–[31]. Some other notable work in this field includes [32]–[36].

Thermal to visible matching is a much more challenging topic due to the differences in modalities. The difficulties arise especially when finding modality-invariant similarity metrics or common discriminative representations of both modalities. Despite that, most methods will rely on the same techniques for visible-to-visible matching [37]–[42]. Multiple works perform matching using a patch-based approach [40], [43]–[46]. On the other hand, other methods will focus on learning discriminative representations [37], [39], [47]–[49] Several authors have proposed deep-network methods to perform feature extraction and matching between pairs of images [30], [50]. Typically, these networks are not specialized for multimodal imagery.

B. GANs for modality transfer

In this work, we explored the use of GANs to perform modality transfer as part of our global architecture. GANs are widely used for cross modalities applications. These networks are able to find relations and consistency between domains. Those networks use a very general loss made by a discriminator that is trained at the same time as the generator. Hence, this generality made it a powerful tool for studying cross-modality in a wide range of domains with minor adaptation. Among all its possible applications, we mainly focus on image translation from one modality to another [10], [11], [51]–[57]. This application has recently gained lots of attention as it facilitates image visualization by combining multiple sources in medical imagery [51] or increases model accuracy by shifting the image to a modality better dealt with by a previous model [57].

In this paper, we focus on the task of domain shift for images taken by the same source from the infrared domain to the visible domain [12], [58]–[61], and conversely from the visible domain to the infrared domain [61]–[66]. A wide variety of GANs can perform such tasks as this kind of network has seen some evolution to better suit the distribution of the data at hand. Thus, CycleGAN is often used to deal with unpaired data [58], [60], [62], [66] while ConditionalGAN is used for datasets with paired images from both domains [11]. Finally, when further data are available, or more information on data distribution is known, more complex architectures can be used [63], [67].

III. Proposed Method

The proposed method performs the multimodal image registration by connecting a modality transfer GAN to a deep homography estimation network (see Figure 2).

A. Modality transfer

Given the good performance of the Deep Homography method on registration inside the infrared domain, we have chosen to translate images from the visible domain (Ivis) to the infrared (IR) domain. This newly generated must synthetically be as close as possible to an image taken from the same point, but with an IR camera instead. Once the translation is realized, we use the newly synthetically generated IR image (Isir) and the original IR image (Iir) to compute the homography matrix with the Deep Homography network. This homography is computed to register Iir with Isir, however, if the translation from the visible domain to the IR one has been performed correctly, the homography matrix is equal to the one describing the Iir to Ivis registration.

B. Generative adversarial network

To perform the image translation from the visible domain to the IR domain, we use generative adversarial networks. Because paired image datasets were available, we used pix2pix conditional GAN method.

The GAN is composed of a generator and a discriminator.

The generator (seen in Table II) must generate synthetic IR images with a visible image as input, while the discriminator must decide whether an image is a real IR image or a synthetically generated one. In the pix2pix model, the generator training loss is the combination of the GAN loss which is the mean square estimator of the difference between the discriminator prediction and the ground truth, and a pixel loss which is the L1 norm of the targeted paired IR image and the synthetically generated one.

The discriminator (seen in Table I) training loss is the mean of the real loss and fake loss. The real loss is the MSE of the difference between ground truth and discriminator predictions when given real IR images, and the fake loss is the MSE of the difference between ground truth and discriminator predictions when given synthetically generated IR images.

Both parts of the pix2pix model are trained simultaneously.

TABLE I
Details discriminator architecture

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

C. GAN's training

The training of the GAN is performed using a dataset of registered visible and IR images. The image registration step for the training dataset is crucial as we want to limit the deformation between the original visible images and synthetically generated ones. We used 256x256 pixels images with a

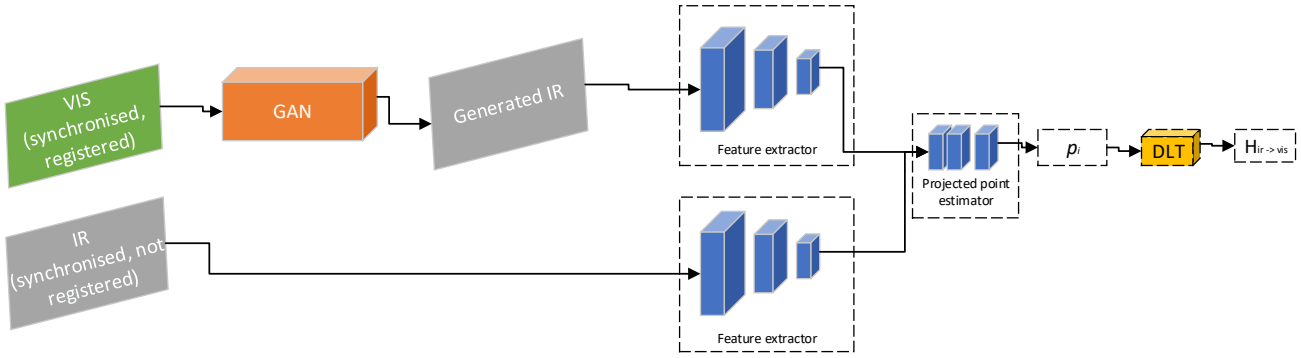

Fig. 2. Our network architecture. The terms VIS and IR represent the visible and infrared images respectively.

TABLE II
DETAILS GENERATOR ARCHITECTURE

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

TABLE III
AN OVERVIEW OF THE SEQUENCES FROM THE VLIRVDIF [69] DATASET

Sequence	**Distance**	**People**	**Light**	**Env.**	**Background**
Laboratory	Near	✓	Artificial	Indoor	Wall
Camouflage	Near & Far	✓	Sunlight	Outdoor	Forest
Trees	Far	✓	Sunlight	Outdoor	Forest
Guanabara Bay	Far	X	Night	Outdoor	Water
Patio	Far	✓	Twilight	Outdoor	Building

training dataset size of 1200 and a test dataset size of 300. The batch size is 10. We usually let the training reach 500 epochs, which takes around 4 hours with a Quadro K2200 (4 GB of memory).

D. GAN's results

Figure 3 shows examples of image translation from the visible to IR domain accomplished with GAN. Synthetically generated images are almost identical to IR images; most of the differences come from a blurring effect that sometimes appears or a loss of texture. We found that the edges are particularly well preserved which, according to intermediate results from the feature extractor, is what matters the most.

E. Deep Homography Estimation

Following the GAN, the image pair is fed to a CNN. The entire architecture can be seen in Figure 2. The feature extractor is a ResNet34 [68], which is not pretrained, and the predicted point estimator is composed of fully connected layers. Using the predicted points, we compute a homography matrix. To solve that task differently, we use the deep linear transformer (DLT), a homography solver from [50].

F. Unsupervised Learning

Our deep homography network is trained end-to-end using an unsupervised triplet loss. The main criteria are the similarity between reprojected features and, thus, alignment of both images. Even though they are not needed for training, ground truth is still required for validating our approach. They were computed by manually matching points and shown in Figure 4.

1) Loss Expression: The training loss is taken from [50].

$$\min_{f,m,h} L_n(I'_{ir}, I_{vi}) + L_n(I'_{vi}, I_{ir}) - \lambda L + \mu||\mathcal{H} - \mathcal{I}|| \quad (1)$$

where I'_{ir} and I'_{vi} are the reprojected feature maps from the visible and the infrared image, I_{ir} and I_{vi} the patches with no reprojection. $\mathcal{H}$ is the multiplication of homography matrices to encourage their symmetry, with $\mathcal{I}$ being the identity matrix. The idea is that we will switch the features of I_{ir} and I_{vi} and compute the opposite homography, $\mathcal{H}_{vis->ir}$. As we want $\mathcal{H}_{ir->vis}$ and $\mathcal{H}_{vis->ir}$ to be inverse, their multiplication should be as close as possible to the identity matrix $\mathcal{I}$. The weights are equal to $\lambda = 2.0$ and $\mu = 0.01$. L is a term that optimizes for discriminative features and L_n is the loss between the reprojected I' and I.

IV. EXPERIMENTS

A. Datasets

The experiments were performed on the open-source Visible-Light and Infrared Video Database (VLIRVDIF) [69]. For both modalities, the images are available as either synchronized, unsynchronized, registered, or unregistered, with all possible permutations of those. Since the task explored in this work is multimodality homography estimation, we use the synchronized unregistered IR and synchronized registered visible images. Although this dataset was designed for image fusion, it is also suitable for image registration. The scenes are filmed in various settings in Brazil, including indoor and outdoor, with close foreground, distant foreground, natural light, indoor light, water as background, a forest as background, etc. We formally detail the scenes in Table III.

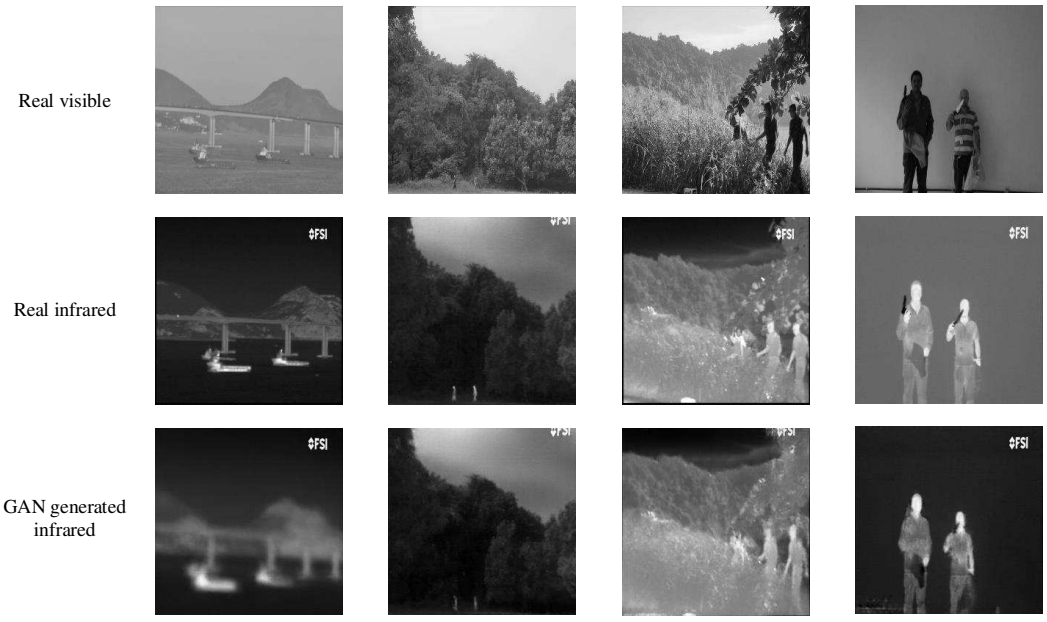

Fig. 3. For each scene, from top to bottom, real visible image, real IR image, GAN generated IR image.

B. Implementation details

The model is implemented in PyTorch [70] using PyTorch Lightning [71] and Hydra [72]. We follow the training and evaluation protocols from [9]. To evaluate our results, ground-truth homographies are manually computed.

V. Results and discussion

A. Comparisons with existing solutions

To validate our approach, we compare with four state-of-the-art methods from the literature, namely SOSNet [73], CNN matching [74], SIFT [75] and Debaque *et al*. [9]. All of those methods except the latter were used in conjuncture with RANSAC [76], USAC [77] and MAGSAC [78].

The quantitative results of our method are shown in Table IV. The error is expressed in the percentage of mean pixel error on average image length. For example, if the image is 256×256 and the mean pixel error is 5, the error percentage would be $(5/256) * 100 = 1.95\%$. We show a clear improvement over Debaque *et al*. [9] thanks to using the preprocessed images of our domain adaptation GAN. Only the method SOSnet on the "Trees" sequence outperform the proposed approach. Some visual examples of our results can be seen in Figure 5.

B. General discussion on experiments

As shown in Table IV, the proposed GAN allows to significantly reduce the registration error. Note that the GAN may cause artifacts like blurring or hallucinations that could induce inaccuracy in the localization of features.

VI. Conclusion

This paper addressed the problem of multimodal image registration as well as visible to infrared image generation using a GAN network. The GAN is trained and evaluated on a public visible and infrared video database. Our method was compared with a similar deep homography network which does not use a GAN, as well as several other baseline methods. It was demonstrated, in our experimental setup, that using a GAN-based domain transfer function before feeding the imagery to the deep homography network significantly reduces the registration error.

Future works might include a deeper study of comparative metrics for evaluating multimodal registration methods. An ideal metric would be invariant to domain change, but discriminant to viewpoint change. Additionally, focus on better generalization would also be useful for most applications, in order to apply the same model to different environmental settings. An analysis of feature localization errors caused by GAN artifacts could also be done, and how to mitigate those errors.

References

[1] T. Georgiou, Y. Liu, W. Chen, and M. Lew, "A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision," *International Journal of Multimedia Information Retrieval*, vol. 9, no. 3, pp. 135–170, 2020.

[2] T. Mouats, N. Aouf, D. Nam, and S. Vidas, "Performance evaluation of feature detectors and descriptors beyond the visible," *Journal of Intelligent & Robotic Systems*, vol. 92, no. 1, pp. 33–63, 2018.

[3] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 23–79, 2021.

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

[5] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image and vision computing*, vol. 21, no. 11, pp. 977–1000, 2003.

[6] L. G. Brown, "A survey of image registration techniques," *ACM Comput. Surv.*, vol. 24, p. 325–376, dec 1992.

[7] S. A. K. Tareen and Z. Saleem, "A comparative analysis of sift, surf, kaze, akaze, orb, and brisk," in *2018 International conference on computing, mathematics and engineering technologies (iCoMET)*, pp. 1–10, IEEE, 2018.

TABLE IV

REPROJECTION ERRORS OF DIFFERENT METHODS, SHOWN IN PERCENTAGE OF PIXELS. WHEN METHODS FAIL TO FIND ENOUGH MATCHES TO COMPUTE A HOMOGRAPHY, WE USE ** TO REPLACE THE ERROR PERCENTAGES.

Method	Laboratory	Camouflage	Trees	Guanabara Bay
SOSNet + RANSAC	**	8.16%	49.01%	**
SOSNet + MAGSAC	**	6.22%	2.93%	7.88%
SOSNet + USAC	**	7.30%	**2.76%**	8.96%
SIFT + RANSAC	**	**	12.10%	**
SIFT + MAGSAC	**	13.29%	11.07%	**
SIFT + USAC	**	**	11.86%	**
CNN-matching + RANSAC	**	**	**	**
CNN-matching + MAGSAC	**	**	10.46%	**
CNN-matching + USAC	**	**	11.99%	**
Debaque *et al*. [9]	2.25%	3.07%	5.59%	4.54%
Ours	**1.29%**	**2.84%**	3.50%	**2.58%**

[8] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Information Fusion*, vol. 45, pp. 153–178, 2019.

[9] B. Debaque, H. Perreault, J.-P. Mercier, M.-A. Drouin, R. David, B. Chatelais, N. Duclos-Hindie, and S. Roy, "Thermal and visible image registration using deep homography," in *2022 25th International Conference on Information Fusion (FUSION)*, pp. 1–8, IEEE, 2022.

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[12] V. V. Kniaz, V. A. Knyaz, J. Hladuvka, W. G. Kropatsch, and V. Mizginov, "Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

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

[14] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "Deepmatching: Hierarchical deformable dense matching," *International Journal of Computer Vision*, vol. 120, no. 3, pp. 300–323, 2016.

[15] M. Arar, Y. Ginger, D. Danon, A. H. Bermano, and D. Cohen-Or, "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*, pp. 13410–13419, 2020.

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

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

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

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

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

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

[22] G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: a survey," *Machine Vision and Applications*, vol. 31, no. 1, pp. 1–18, 2020.

[23] G. Wu, M. Kim, Q. Wang, B. C. Munsell, and D. Shen, "Scalable high-performance image registration framework by unsupervised deep feature representations learning," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1505–1516, 2015.

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

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

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

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

[28] B. D. d. Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum, "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*, pp. 204–212, Springer, 2017.

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

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

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

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

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

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

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

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

[37] B. Deshpande, S. Hanamsheth, Y. Lu, and G. Lu, "Matching as color images: Thermal image local feature detection and description," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1905–1909, IEEE, 2021.

[38] N. Yang, Y. Yang, P. Li, and F. Gao, "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*, vol. 11913, p. 1191303, International Society for Optics and Photonics, 2021.

[39] Y. Lu and G. Lu, "Superthermal: Matching thermal as visible through thermal feature exploration," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2690–2697, 2021.

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

[41] S. Cui, A. Ma, Y. Wan, Y. Zhong, B. Luo, and M. Xu, "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.

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

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

[44] C. A. Aguilera, A. D. Sappa, C. Aguilera, and R. Toledo, "Cross-spectral local descriptors via quadruplet network," *Sensors*, vol. 17, no. 4, p. 873, 2017.

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

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

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

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

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

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

[51] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical image analysis*, vol. 58, p. 101552, 2019.

[52] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE international conference on computer vision*, pp. 2849–2857, 2017.

[53] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *European conference on computer vision*, pp. 319–345, Springer, 2020.

[54] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *International conference on machine learning*, pp. 1857–1865, PMLR, 2017.

[55] H. Tang, D. Xu, N. Sebe, and Y. Yan, "Attention-guided generative adversarial networks for unsupervised image-to-image translation," in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2019.

[56] P. Perera, M. Abavisani, and V. M. Patel, "In2i: Unsupervised multi-image-to-image translation using generative adversarial networks," in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 140–146, IEEE, 2018.

[57] R. Abbott, N. M. Robertson, J. M. del Rincon, and B. Connor, "Unsupervised object detection via lwir/rgb translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 90–91, 2020.

[58] T. Sun, C. Jung, Q. Fu, and Q. Han, "Nir to rgb domain translation using asymmetric cycle generative adversarial networks," *IEEE Access*, vol. 7, pp. 112459–112469, 2019.

[59] M. S. Uddin, J. Li, *et al.*, "Generative adversarial networks for visible to infrared video conversion," *Recent Advances in Image Restoration with Applications to Real World Problems*, 2020.

[60] K. Yun, K. Yu, J. Osborne, S. Eldin, L. Nguyen, A. Huyen, and T. Lu, "Improved visible to ir image transformation using synthetic data augmentation with cycle-consistent adversarial networks," in *Pattern Recognition and Tracking XXX*, vol. 10995, p. 1099502, SPIE, 2019.

[61] S. Li, B. Han, Z. Yu, C. H. Liu, K. Chen, and S. Wang, "I2v-gan: Unpaired infrared-to-visible video translation," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3061–3069, 2021.

[62] A. Mehri and A. D. Sappa, "Colorizing near infrared images through a cyclic adversarial approach of unpaired samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[63] P. L. Suárez, A. D. Sappa, and B. X. Vintimilla, "Infrared image colorization based on a triplet dcgan architecture," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 18–23, 2017.

[64] A. Berg, J. Ahlberg, and M. Felsberg, "Generating visible spectrum images from thermal infrared," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1143–1152, 2018.

[65] T. Zhang, A. Wiliem, S. Yang, and B. Lovell, "Tv-gan: Generative adversarial network based thermal to visible face recognition," in *2018 international conference on biometrics (ICB)*, pp. 174–181, IEEE, 2018.

[66] K. K. Babu and S. R. Dubey, "Pcsgan: Perceptual cyclic-synthesized generative adversarial networks for thermal and nir to visible image transformation," *Neurocomputing*, vol. 413, pp. 41–50, 2020.

[67] Y. Luo, D. Pi, Y. Pan, L. Xie, W. Yu, and Y. Liu, "Clawgan: Claw connection-based generative adversarial networks for facial image translation in thermal to rgb visible light," *Expert Systems with Applications*, vol. 191, p. 116269, 2022.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

[70] A. e. a. Paszke, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., 2019.

[71] W. Falcon et al., "Pytorch lightning," *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, vol. 3, 2019.

[72] O. Yadan, "Hydra - a framework for elegantly configuring complex applications." Github, 2019.

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

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

[75] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[76] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[77] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J.-M. Frahm, "Usac: A universal framework for random sample consensus," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 2022–2038, 2012.

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

APPENDIX
ADDITIONAL FIGURES

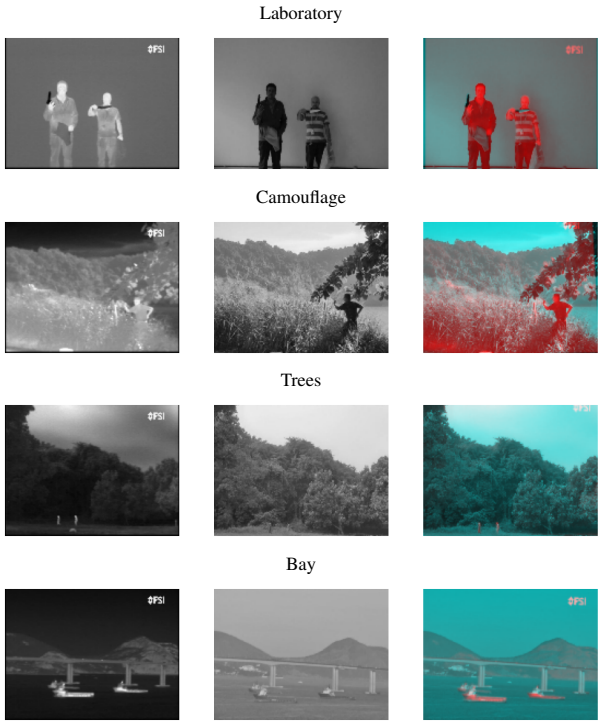

Fig. 4. For every sequence, here is an example of our thermal to visible ground-truth registration presented in the form of an anaglyph.

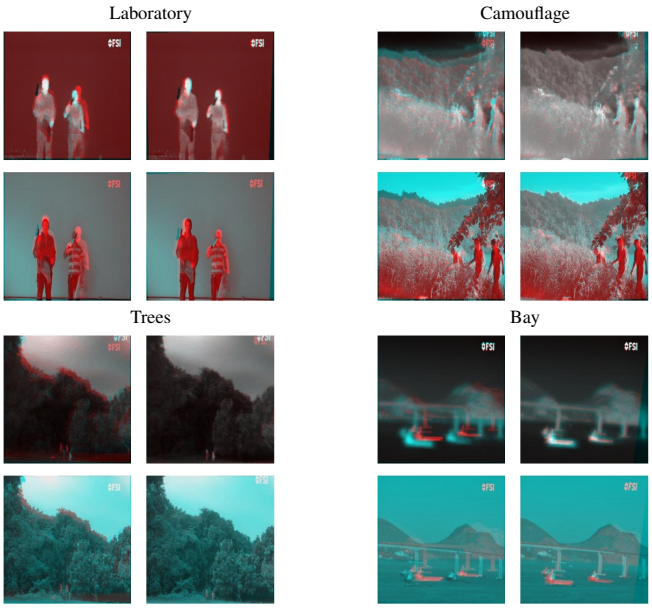

Fig. 5. For each scene, anaglyph between thermal and GAN generated image before registration (top left) and after registration (top right), anaglyph between the thermal and visible generated image before registration (bottom left) and after registration (bottom right).