Letter

S2-Net: Self-Supervision Guided Feature Representation Learning for Cross-Modality Images

Shasha Mei, Yong Ma, Xiaoguang Mei, Jun Huang, and Fan Fan

Dear Editor,

This letter focuses on combining the respective advantages of cross-modality images which can compensate for the lack of information in the single modality. Meanwhile, due to the great appearance differences between cross-modality image pairs, it often fails to make the feature representations of correspondences as close as possible. In this letter, we design a cross-modality feature representation learning network, S2-Net, which is based on the recently successful detect-and-describe pipeline, originally proposed for visible images but adapted to work with cross-modality image pairs. Extensive experiments show that our elegant formulation of combined optimization of supervised and self-supervised learning outperforms state-of-the-arts on three cross-modal datasets.

Establishing the local correspondences between two images, as a primary task, is the premise of various visual applications, including target recognition, visual navigation, image stitching, 3D reconstruction and visual localization [1]. The conventional matching methods are based on the handcrafted local feature descriptors [2]–[4] to make the representation of two matched features as similar as possible and as discriminant as possible from that of unmatched ones. Over the recent years, the deep learning-based methods have achieved significant progress in general visual tasks, and have also been introduced into the field of image matching. The current approaches are mostly based on a two-stage pipeline that first completes the extraction of keypoints and then encodes the patches centered on the keypoints into descriptors, thus referred to as the detect-then-describe methods. In the field of cross-modality image matching, the detect-then-describe methods have been widely used with a manual detector to detect and an adapted deep learning network to perform description [5]. For example, a cross-spectral local descriptor, Q-Net [6], uses a quadruplet network to map input image patches from two different spectral bands to a common Euclidean space. SFc-Net [7] adopts the Harris corner detector for candidate feature point detection and then gets correspondences by a Siamese CNN.

Despite this apparent success, it is an inevitable disadvantage of this paradigm that the global spatial information is discarded during the description process, which happens to be essential for cross-modality images. In contrast to it, the detect-and-describe framework for visible images uses a network to simultaneously perform feature point extraction and descriptor construction [8], [9]. This approach postpones the detection process without missing high-level information of images. Additionally, the detection stage is tightly coupled with the description so as to detect pixels with locally unique descriptors that are better for matching. Undoubtedly, it is promising to introduce the framework into cross-modality image matching, however, challenges come up due to the huge heterogeneity. To be specific, it is difficult to optimize the model for cross-modality images with extreme geometric and radiometric variances.

Self-supervised learning (SSL), which helps the model obtain easy invariance with augmented data, is one of the most popular techniques in natural language processing and computer vision. As for local feature representation learning, the well-known Superpoint [10] proposed a novel Homographic Adaptation procedure, which is a form of self-supervision, to tackle the ill-posed problem of keypoint extraction. Nevertheless, the SSL technics have not been introduced into cross-modality scenario, while current methods are devoted to obtaining supervised signals from labeled data instead. Since the learning becomes harder for cross-modality images due to the serious radiometric variances, it is desirable to introduce SSL into this task. In fact, among the challenges faced by cross-modality descriptors, excluding inter-modal invariance, other necessities including geometric invariance as well as robustness to noise and grayscale variations can be well-addressed by SSL.

In this work, we explore the possibility of using SSL, based on the recent success of the detect-and-describe methods, but adapted to work with cross-modality image pairs. Although the cross-modality images are heterogeneous and quite different in appearance, they still have some similar semantic information, such as the shape, structure, and topological relationship. The detect-and-describe methods retains the global spatial information which is rather crucial for our task. As for the optimization problem, we provide an effective solution for the application of the detect-and-describe methods to the cross-modality domain. More precisely, we propose our novel architecture of joint training with supervised and self-supervised learning, termed S2-Net, which takes full advantage of SSL to improve matching performance without extra labeled data, as illustrated in Fig. 1. Self-supervision simulates the feature representation learning of images in the same modalities. Since the task of training image representations of the same modality is relatively easier compared to different modalities, self-supervision plays a guiding role in the training process. Also, we design a loss function that combines both supervised and self-supervised learning and optimally balances the guidance of the two optimization methods. To the best of our knowledge, S2-Net is the first algorithm that introduces the SSL technique into cross-modality feature representation, and sufficient experiments have demonstrated the great effectiveness of our work.

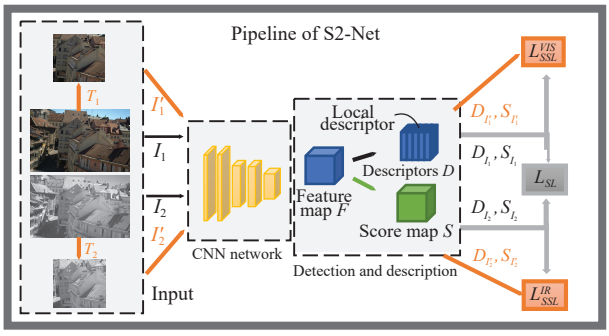


Fig. 1. Our proposed S2-Net for cross-modality images.

Method: In this section, our proposed technique of self-supervision-guided optimization will be explained in detail.

1) Framework of self-supervision guided optimization: We propose S2-Net, a general framework aims to make the detect-and-describe methods suitable for cross-modality image matching. To train the basic framework, relevant constraints for single modality images are proposed. However, the lack of strong supervision in these constraints, e.g., which point should be the key point, always troubles the training. Moreover, it is common knowledge that the difference between pairs in the same modality is much smaller than cross modalities. To this end, it is promising to introduce the mono-

Corresponding author: Fan Fan.

Citation: S. S. Mei, Y. Ma, X. G. Mei, J. Huang, and F. Fan, "S2-Net: Self-supervision guided feature representation learning for cross-modality images," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 10, pp. 1883–1885, Oct. 2022.

The authors are with the Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: mss_1998@whu.edu.cn; mayong@whu.edu.cn; meixiaoguang@gmail.com; junhwong@whu.edu.cn; fanfan@whu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2022.105884

modality self-supervised learning to guide the cross-modality training. As illustrated in Fig. 1, based on the basic framework, the other two branches with augmented cross-modality images for self-supervised learning are introduced for joint training. It should be noted that our approach only changes the training process that the original pair of images inside a batch becomes three pairs, which is equivalent to tripling the batch size, so the training time also becomes three times the original, but the testing time remains the same.

Suppose we take the image pair I as an example, and it is randomly tranformed to image I' with the tranform $T(\cdot)$. Speficically, we focus on two major transforms, geometric transform and gray transform, which are the most natural discrepancies across modalities. As for geometric transforms, we mainly consider the random rotating $[-10°, 10°]$, random scaling $[1, 1.2]$ and quadrangular random projection of a ratio $[0, 0.2]$, which are denoted as $T_s(\cdot)$, $T_r(\cdot)$ and $T_q(\cdot)$, respectively. Moreover, to encourage the model to obtain better generalization on the gray variation, we also add a random noise $T_n(\cdot)$, where the noise follows the normal distribution with the mean of 0 and the variance of 0.01, and a Gaussian blur $T_b(\cdot)$ with the kernel size of $[3, 3]$ and the standard deviation of $[0.01, 1]$ as well as a random gray inverting $T_i(\cdot)$ that inverts the gray scale larger than a random threshold 0.5. Therefore, the total transform can be described in the cascading sub-transforms as

$$T(\cdot) = T_s \times T_r \times T_q \times T_n \times T_b \times T_i(\cdot). \tag{1}$$

Independent random transform T_{vis} and T_{oth} would be conducted on raw visible and the other modal images for two parallel self-supervised learning.

2) Overall loss function: The input is processed by the network to generate the feature map, and then the descriptors and score maps are calculated. The overall loss function is expressed as follows:

$$L = L_{SL} + \lambda(L_{SSL}^{VIS} + L_{SSL}^{OTH}) \tag{2}$$

where L_{SL} is the supervised learning (SL) loss, L_{SSL}^{VIS} is the SSL loss for visible images, and L_{SSL}^{OTH} is the SSL loss for the other modality images. λ is the weighting coefficient of the loss functions, whose influence on the matching results is experimentally shown in the analysis of the weighting coefficients.

Supervised learning is used in the original detect-and-describe methods, and also retained in our solution. Given a corresponding image pair I_{vis} and I_{oth}, and U denotes their mapping function. That is, for a pixel $p = (x_i, y_i)$ in I_{vis}, $U_{(x_i,y_i)}$ is the corresponding pixel of p in I_{oth}. Then SL loss can be expressed as follows:

$$L_{SL} = L_f(I_{vis}, I_{oth}, U) \tag{3}$$

where L_f denotes the original loss function of a network $f(\cdot)$ in detect-and-describe methods for visible images. It is different from each method, but the inputs are fixed as a pair of corresponding images and the ground-truth correspondences between them. As a part of the overall loss, SL loss utilizes the labeled dataset to optimize the network.

Obtaining a large number of images with known correspondences requires a significant cost. In this case, SSL is a reliable and efficient solution, and we proposed the SSL loss as

$$\begin{aligned} L_{SSL}^{VIS} &= L_f(I_{vis}, I'_{vis}, T_{vis}) \\ L_{SSL}^{OTH} &= L_f(I_{oth}, I'_{oth}, T_{oth}). \end{aligned} \tag{4}$$

For visible images, I'_{vis} is obtained from I_{vis} by transformation T_{vis}, so as to form a corresponding pair with I_{vis} as the input to the network $f(\cdot)$. Also, the same operation is performed on the other modal images.

Experiments: In order to demonstrate the effectiveness of our proposed approach, we selected D2-Net [8] and R2D2 [9], two classic detect-and-describe methods for visible images, and two handcrafted descriptors of scale-invariant feature transform (SIFT) [3] and radiation-variation insensitive feature transform (RIFT) [4] to compare the performance of our self-supervision on these methods. To better evaluate the performance on cross-modality images, we also compared CMM-Net [11], which designed a novel network for feature representations of thermal infrared and visible images.

1) Implementation details: All the experiments were implemented on a computer with NVIDIA RTX 3090 GPU. As for D2-Net, we fine-tuned the overall network for 100 epochs instead of fine-tuning the last layer of the dense feature extractor (conv4_3). The network was optimized using Adam with a fixed learning rate of 10^{-4} and weight decay of 10^{-5}. For each pair, we selected a random 256× 256 crop centered around one correspondence with a batch size of 1. In R2D2, the learning rate is 0.0001, the weight decay is 0.0005 and the baze size is 2 with the input pairs cropped to 192×192. The data augmentation is performed through the random flipping, random rotating $(90°, 180°, 270°)$ and random noise blurring. Moreover, the image pairs of the two datasets are co-registered with pixels aligned without offsets, so we carried out random rotating $[-10°, 10°]$, random scaling $[1, 1.2]$ and random projection of a ratio $[0, 0.2]$.

2) Experimental datasets: The matching of thermal infrared (TIR) and visible images is a typical cross-modality problem, so we perform our experiments on RoadScene dataset [12], which is comprised of 221 aligned thermal infrared and visible images. This dataset was split in a testing dataset with 43 image pairs from different scenes and a training dataset from the remaining 178 pairs. We also perform our experiments on a public registered RGB-NIR scene dataset [13], which consists of 477 pairs in 9 scenes captured in RGB and near-infrared (NIR). We randomly select 171 pairs for testing (19 per scene) and train on the rest. In addition, we conduct experiments on OS dataset, which is a high-resolution dataset of co-registered optical and SAR patch-pairs [14]. And we select training set from 512×512 pairs for training (2011 pairs) and testing set from 512×512 pairs for testing (424 pairs).

3) Evaluation metrics and comparison results: Three evaluation metrics, number of correspondences in the extracted points (NC), number of correct matches (NCM) and the correctly matched ratio (CMR) are used to evaluate the different methods quantitatively. NC indicates the repeatability of extracted interest points and NCM is crucial for the image registration. CMR is computed as

$$CMR = \frac{NCM}{NC} \times 100\%. \tag{5}$$

In the testing process, we vary the number of points extracted from both images, which is denoted as K, and record the evaluation results on each method. Specifically, we set $K = 1024$, $K = 2048$ and $K = 4096$. The results obtained are listed in Tables 1−4.

Table 1. Number of Correspondences in the Extracted Points on the Two Datasets

Method	RoadScene dataset			RGB-NIR dataset		
	K = 1024	K = 2048	K = 4096	K = 1024	K = 2048	K = 4096
SIFT	294	788	1537	322	743	1681
RIFT	443	**1046**	1295	351	845	1575
CMM-Net	176	195	195	155	438	1146
D2-Net	183	265	265	170	475	1283
R2D2	184	511	1669	448	953	2015
D2-Net+ SSL (ours)	**449** (↑145.4%)	1033 (↑289.8%)	1538 (↑480.4%)	223 (↑31.2%)	557 (↑17.3%)	1424 (↑11.0%)
R2D2+SSL (ours)	243 (↑32.1%)	668 (↑30.7%)	**2015** (↑20.7%)	**474** (↑5.8%)	**1013** (↑6.3%)	**2165** (↑7.4%)

Combining the three evaluation metrics, it can be seen that on RoadScene dataset, D2-Net with SSL achieves the best performance, and R2D2 with SSL ranks second. The SIFT algorithm, performs the worst of all since it requires texture details that differ across modalities. It should be specially noted that the original R2D2 achieves fairly good results among the compared methods, nevertheless, we improve it quite a bit. This is due to the fact that the original R2D2 algorithm takes repeatability and reliability into account in its loss, which is not available in D2-Net. So, with SSL, the performance of D2-Net has been extremely boosted. The relevant visualization results are shown in the first column of Fig. 2. As for optical and SAR images, the R2D2 and D2-Net algorithms are not able to obtain

Table 2. Number of Points Correctly Matched on the Two Datasets

Method	RoadScene dataset			RGB-NIR dataset		
	$K = 1024$	$K = 2048$	$K = 4096$	$K = 1024$	$K = 2048$	$K = 4096$
SIFT	3	6	6	141	286	541
RIFT	36	57	66	111	200	298
CMM-Net	29	31	31	36	95	233
D2-Net	12	14	14	90	214	508
R2D2	30	60	157	234	458	825
D2-Net+ SSL (ours)	**92** (↑666.7%)	**177** (↑1157.1%)	**233** (↑1564.3%)	127 (↑41.1%)	272 (↑27.1%)	590 (↑16.1%)
R2D2+ SSL (ours)	50 (↑66.7%)	93 (↑55.0%)	228 (↑45.2%)	**248** (↑6.0%)	**486** (↑6.1%)	**903** (↑9.5%)

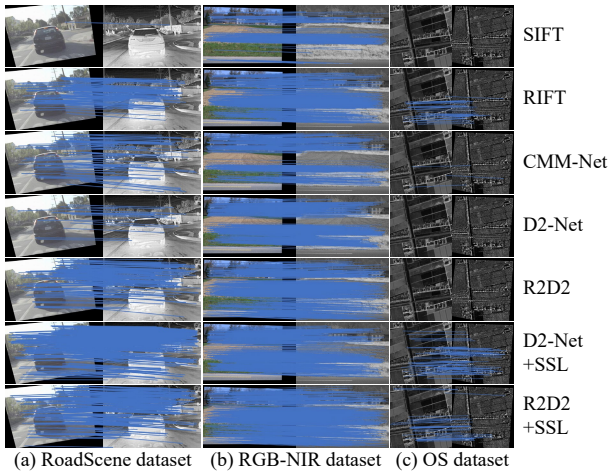
Table 3. Ratio of Correct Matches on the Two Datasets

Method	RoadScene dataset			RGB-NIR dataset		
	$K = 1024$	$K = 2048$	$K = 4096$	$K = 1024$	$K = 2048$	$K = 4096$
SIFT	3	6	6	141	286	541
RIFT	36	57	66	111	200	298
CMM-Net	29	31	31	36	95	233
D2-Net	12	14	14	90	214	508
R2D2	30	60	157	234	458	825
D2-Net+ SSL (ours)	**92** (↑666.7%)	**177** (↑1157.1%)	**233** (↑1564.3%)	127 (↑41.1%)	272 (↑27.1%)	590 (↑16.1%)
R2D2+ SSL (ours)	50 (↑66.7%)	93 (↑55.0%)	228 (↑45.2%)	**248** (↑6.0%)	**486** (↑6.1%)	**903** (↑9.5%)

Table 4. Three Evaluation Metrics on OS Dataset

Method	NC			NCM			CMR (%)		
	$K = 1024$	$K = 2048$	$K = 4096$	$K = 1024$	$K = 2048$	$K = 4096$	$K = 1024$	$K = 2048$	$K = 4096$
SIFT	150	493	1475	0	0	0	\	\	\
RIFT	**220**	**707**	1232	8	13	17	3.52	1.82	1.33
CMM-Net	143	329	374	3	3	3	2.02	1.10	1.05
D2-Net+SSL (ours)	208	634	**1728**	**10**	**22**	**37**	**4.54**	**3.37**	**2.12**
R2D2+SSL (ours)	81	308	1070	2	6	17	2.86	1.88	1.34

correctly matched pairs, and therefore the results of them are not presented in Table 4. The multi-modal descriptor, RIFT, performs well among the comparison algorithms. Nevertheless, our method achieves the best results as shown in the third column of Fig. 2 and in Table 4.



(a) RoadScene dataset (b) RGB-NIR dataset (c) OS dataset

Fig. 2. Experimental results of S2-Net and the state-of-the-art image matching methods for the three datasets.

And on RGB-NIR dataset, since the difference between visible and thermal infrared images is much more significant than that with near-infrared images, SSL does not improve the performance of the original method as much as on the RoadScene dataset. And it is reasonable that SIFT achieves a good accuracy. However, the performance of R2D2 with SSL ranks best above all methods, as depicted in the middle column of Fig. 2. The guiding effect of self-supervision in the training process is rather beneficial when learning modality-invariant feature representations.

Conclusion: In this article, we propose S2-Net, which introduces the self-supervised learning in the training learn the modality-invariant feature representation. After performing experiments on three datasets, it can be demonstrated that our strategy significantly improves the networks' capability of feature representation for cross-modality images, including the detection and description.

Acknowledgments: This work was supported by the National Natural Science Foundation of China (NSFC) (62003247, 62075169, 62061160370).

References

[1] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 23–79, 2021.

[2] J. Ma, Z. Li, K. Zhang, Z. Shao, and G. Xiao, "Robust feature matching via neighborhood manifold representation consensus," *ISPRS J. Photogramm. Remote Sens.*, vol. 183, pp. 196–209, 2022.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[4] J. Li, Q. Hu, and M. Ai, "RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform," *IEEE Trans. Image Process.*, vol. 29, pp. 3296–3310, 2019.

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

[6] 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. DOI: 10.3390/s17040873.

[7] H. Zhang, W. Ni, W. Yan, D. Xiang, J. Wu, X. Yang, and H. Bian, "Registration of multimodal remote sensing image based on deep fully convolutional neural network," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 12, no. 8, pp. 3028–3042, 2019.

[8] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-Net: A trainable CNN for joint description and detection of local features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8092−8101.

[9] J. Revaud, C. De Souza, M. Humenberger, and P. Weinzaepfel, "R2D2: Reliable and repeatable detector and descriptor," *Adv. Neural Inf. Process Syst.*, vol. 32, pp. 12405–12415, 2019.

[10] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2018, pp. 224−236.

[11] 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 Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.

[12] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, 2022.

[13] M. Brown and S. Süsstrunk, "Multi-spectral sift for scene category recognition," in *Proc. IEEE Int. Conf. Comput. Vis.* 2011, pp. 177−184.

[14] Y. Xiang, R. Tao, F. Wang, H. You, and B. Han, "Automatic registration of optical and SAR images via improved phase congruency model," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 5847–5861, 2020.