



# DVT-SLAM: Deep-Learning Based Visible and Thermal Fusion SLAM

Ruochen Wang<sup>1</sup>, Ling Pei<sup>2</sup>(✉), Lei Chu<sup>1</sup>, Qi Wu<sup>1</sup>, Tao Li<sup>1</sup>, Wenxian Yu<sup>1</sup>,  
and Xujun Guan<sup>2</sup>

<sup>1</sup> Shanghai Key Laboratory of Beidou Navigation and Location Services, Shanghai Jiao Tong University, Shanghai, China

<sup>2</sup> Shanghai Jiao Tong University, Shanghai, China  
ling.pei@sjtu.edu.cn

**Abstract.** The problem of visual odometry (VO) and localization in extreme illumination conditions is widely concerned. In this paper, we propose a novel SLAM algorithm namely DVT-SLAM (Deep-learning based Visible-Thermal SLAM). It focuses on the fusion of thermal infrared image and visible image which have complementary advantages in characteristics. With the contrastive learning and the measurement of mutual information between multi-modal images, the first part of DVT-SLAM is the DVT-GAN network to fuse visible-thermal images and generate pseudo visible images at night. Given the generated images, visual odometry is applied for pose estimation base. Extensive evaluations are performed on the Brno Urban Dataset, a multi-modal dataset containing different time and weather conditions in diverse scenarios. Series of experiments show that DVT-SLAM is a robustness and suitability solution for single visible camera failures, which can reduce positioning error by half and achieve superior SLAM performance.

**Keywords:** Visual SLAM · Thermal infrared camera · Deep learning

## 1 Introduction

In recent years, unmanned system and artificial intelligence have developed rapidly. Fast, reliable, and efficient positioning systems are the prerequisites for subsequent obstacle avoidance or decision-making tasks. Simultaneous Localization and Mapping (SLAM), as the key technology of unmanned system, would face challenges in extreme illumination conditions. Multi-modal SLAM system for complex environment has become a prevailing research topic.

Compared with visible camera, the thermal camera creates an image using infrared radiation. It is less affected by changing illuminance and can be a good supplement for traditional visible camera. To make full use of the infrared imaging characteristics, researchers have carried out a series of explorations. Tarek [1] and Beauvisage [2] did similar work. They adopted multi-spectral stereo matching to find the corresponding relationship between visible image and long-wave thermal image. Then they realized the Visual Odometry (VO) for pose estimation. Poujol et al. [3] used the discrete wavelet

transform method to fuse infrared and visible images for monocular VO. In 2020, Beauvisage et al. [4] proposed a new multi-modal monocular VO method, which tracked the features of two modalities at the same time. However, only the camera which performed better is used to estimate the motion. On the other hand, some researchers struggled with deep learning for processing images of two modalities. In 2016, Choi et al. [5] introduced a convolutional neural network to enhance low-resolution thermal image quality. Sun [6] et al. proposed FuseSeg network which could integrate visible and thermal images, and then get pixel-level semantic tags.

As illustrated above, some of these methods focus on the process of SLAM system, such as feature matching and keyframe extraction, which have no limitation to frame image quality. Other methods pay attention to the image fusion, without considering application in SLAM fields.

The main contributions of this paper are as follows:

Firstly, we propose a deep learning-based visible-thermal fusion SLAM algorithm DVT-SLAM. The traditional VO is replaced by a new VO with deep learning network to improve the positioning accuracy in low-illumination conditions. Secondly, we propose a multi-modal image fusion method based on Contrastive Learning (CL). The network uses infrared image and visible image at night to generate high-quality “pseudo visible” image, and the generated images are utilized for VO calculation. Thirdly, with mutual information, a contrastive loss function is introduced to reduce the size of network and improve the quality of generated fused images.

## 2 Visual SLAM

The classic visual SLAM system generally includes four main parts: front-end Visual Odometry (VO), back-end optimization, loop-closure detection, and mapping. Among these, VO is the process of estimating the 6 Degrees Of Freedom (DoF) ego-motion using only the input of cameras. The back-end module solves the drift problem and corrects the trajectory error through non-linear optimization.

The state-of-the-art approaches for VO can be divided into two classes, feature-based methods and direct methods. Feature-based methods minimise the geometric error between point of interest and its re-projected position in the image. Direct methods track camera poses by minimizing photometric errors.

In the field of traditional visual SLAM, the ORB-SLAM [7, 8] system is one of the most prevailing monocular visual SLAM systems, which well balances the accuracy and the computational demand. It extracts ORB features to achieve the ORB dictionary for VO and loop-closure detection. Then, with tracking thread, local BA (Bundle Adjustment) thread and global loop closure optimization threads to achieve fast and robust navigation.

## 3 Methodology

This section presents the image fusion algorithms evaluated in the DVT SLAM context. The overall framework of the DVT-SLAM system is shown in Fig. 1. It consists of two components. In the first part, the thermal and visible images generate pseudo

visible images through the DVT-GAN network. The second part is the overall architecture of DVT-SLAM, which uses the fusion images as the input image frames. The ORB features are extracted and key frames are determined to perform pose estimation. Finally, local map construction and loop-closure detection are completed through back-end optimization.

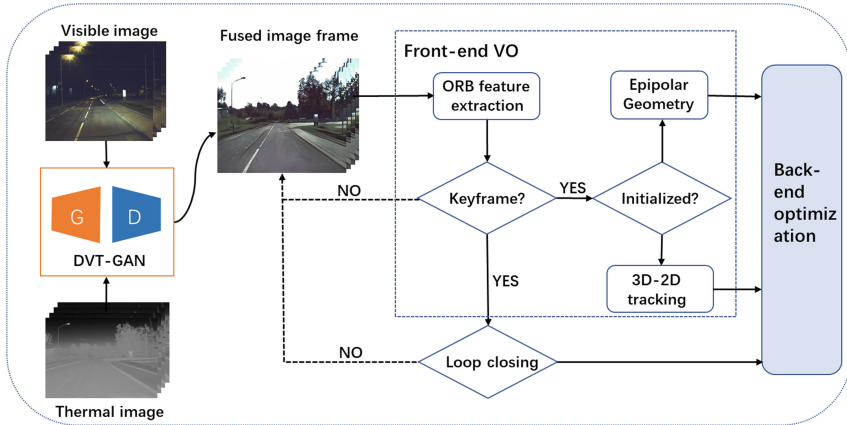


Fig. 1. Architecture overview of DVT-SLAM

### 3.1 Visible-Infrared Image Fusion Network DVT-GAN

#### 3.1.1 Mutual Information

In the fields of probability and information theory, Mutual Information (MI) is a measurement of the interdependence. It estimates the correlation between two random variables or sets of events. Similarly, MI could also be used to measure the correlation between two images. The concept of MI has gradually been adapted to multi-spectral stereo matching [9], multi-modal image registration, depth map calculation and other fields.

The mutual information of two random variables  $(X, Y)$  can be defined as:

$$MI(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right), \tag{1}$$

where  $p(x, y)$  represents the joint distribution function of  $X$  and  $Y$ ,  $p(x)$  and  $p(y)$  are the marginal probability distribution functions of  $X$  and  $Y$  respectively.

#### 3.1.2 The Architecture of DVT-GAN

To establish the relationship between infrared images and visible images, MI theory is adopted in the proposed DVT-GAN network. The scheme of the DVT-GAN network is presented in Fig. 2. It is designed on the basis of CycleGAN [10] and CUT [11] architecture. The input domain  $\mathcal{X} \subset \mathbb{R}^{H \times W \times C}$  represents the infrared image domain, and the

output domain  $\mathcal{Y} \subset \mathbb{R}^{H \times W \times 3}$  represents the visible image domain. The generator is composed of two parts, encoder and decoder. Then one discriminator is designed for leveraging adversarial training. Given infrared image  $x$  as input, the generated fused image can be expressed as  $\hat{y} = G(z) = G_{dec}(G_{enc}(x))$ . DVT-GAN learns image generation relationships only in one direction, that is, from thermal image to visible image.

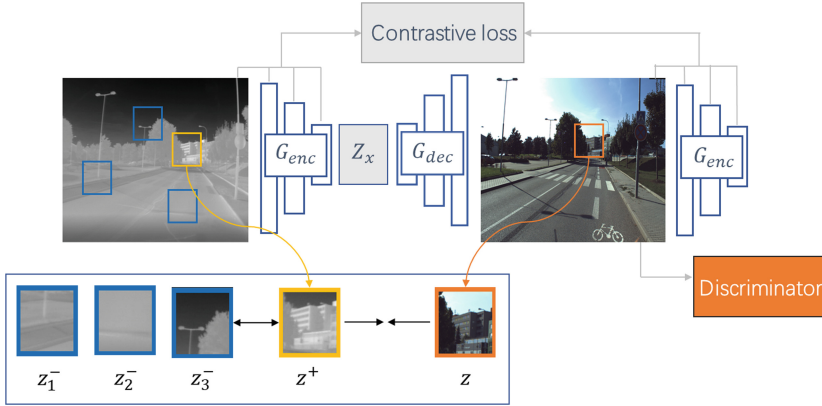


Fig. 2. The Architecture of DVT-GAN network

Different from CycleGAN, which contains two generators and discriminators for cyclic domain migration, the DVT-GAN network uses diverse loss functions to maximize the similarity of images between input and output domains.

**Adversarial Loss.** The adversarial loss is used to make the generated pseudo visible image similar to the real visible image visually. The adversarial loss can be expressed as follows:

$$\mathcal{L}_{GAN}(G, D, X, Y) = \mathbb{E}_{y \sim Y} \log D(y) + \mathbb{E}_{x \sim X} \log(1 - D(G(x))) \quad (2)$$

**Contrastive Loss.** Traditionally, cycle consistency loss is used to prevent contradictions between generators, it can be defined as:

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{y \sim Y} \|G(F(y)) - y\|_1 + \mathbb{E}_{x \sim X} \|F(G(x)) - x\|_1 \quad (3)$$

However, cycle consistency loss requires a reverse generator and an additional discriminator to migrate the generated images back to the input domain  $\mathcal{X}$ , which enlarges the network and training time cost.

Li et al. [12] proved that cycle consistency loss is the upper bound of conditional entropy  $H(X|Y)$ . According to the formula between MI and conditional entropy,  $MI(X, Y) = H(X) - H(X|Y)$ , the maximization of MI is equivalent to the minimization of conditional entropy. Generally, the problem of minimizing cycle consistency loss can be simplified as minimizing MI.

Based on the method of contrast learning, noise contrastive estimation is used to maximize the MI between the input domain and the output domain. As is shown in

Fig. 2, the infrared image and visible image are encoded into feature vectors. Multi-modal image patches at the same position are taken as positive samples while others are taken as negative ones. Patch-level image loss is computed to generate a fused image as realistic as possible.

### 3.2 DVT-SLAM Based on the Visible-Infrared Fusion Image

The front-end of SLAM estimates camera pose by iteratively optimizing the rotation and translation relationship between adjacent frames. Among these VO methods, feature-based VO, as is shown in Fig. 3, optimizes the geometric error:

$$T_{k-1,k} = \arg \min_T \sum_i \left\| u'_i - \pi(p_i) \right\|^2 u_i I_k, \tag{4}$$

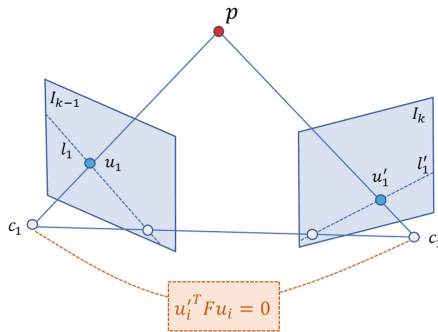


Fig. 3. Indirect method of visual SLAM

where  $u_1$  is any pixel in the image  $I_{k-1}$ , its projection in the space is  $p$ ,  $u'_1$  represents the coordinate pixel in  $I_k$  which is the projection of  $p$ .

In this part, we use visible-infrared fusion images as input frames to optimize VO and realize the DVT-SLAM system based on the ORB-SLAM framework. In the tracking thread, we extract keypoints in the fused images instead of the original low-illumination visible or infrared image. Then through the back-end local mapping and loop closing thread, we achieve global BA optimization.

## 4 Experiments and Results

### 4.1 Dataset

Brno Urban Dataset [14] is intentionally proposed for autonomous driving and mapping tasks. It was recorded in Brno D-Czech Republic for more than 350 km distance. The dataset provides multimodal data from sensors including WUXGA visible camera, infrared camera, LiDAR, and Inertial Measurement Unit (IMU). Ground truth was recorded by an centimetre-accuracy differential RTK GNSS receiver which produce

an accurate pose estimate. Camera calibration parameters including their intrinsic and extrinsic parameters can also be obtained. The characteristics of visible camera and thermal infrared camera used in the experiment are shown in Table 1.

**Table 1.** Camera characteristics of Brno datasets

Model	FLIR Tau 2	WUXGA RGB camera
Resolution	640 × 512px	1920 × 1200px
Max. frame rate	30 Hz	10 Hz
Spectral band	7.5–13.5 μm	Visual spectrum

## 4.2 Visible - Thermal Image Fusion

### A. Experiment Implements

We conducted experiments on the Brno Urban dataset. 3675 daytime infrared images and 3491 daytime visible images in different scenarios are collected to train the DVT-GAN network. Training images from visible domain and thermal domain are unpaired. After establishing the conversion relationship between infrared and high-resolution visible images, infrared images are put into the DVT-GAN network to generate pseudo visible images at night. The resolution of the generated fused image is 512 \* 512.

We trained the DVT-GAN network on a server with an Intel E3–1230 V2 (3.3 GHz) CPU and TITAN V graphics card, including 12-GB graphics memories.

### B. Fusion Results

The image comparison is shown in Fig. 4. The left column depicts the origin night visible image; the thermal images of the same scenarios are shown in the middle column and the generated fused images are presented in the right column. Visually, the generated fused image realizes “pseudo visible” effect, that is, it combines the structural features of infrared image and RGB texture features.

### C. Evaluation Metrics

We first evaluate the proposed method in terms of the fused image quality.

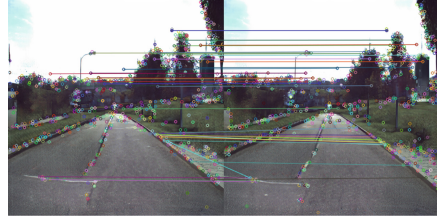
FID (Frechet Inception Distance score) [10] is a measurement of image diversity and authenticity statistically. It calculates the distance between the generated image distribution and real image distribution to measure image similarity. The formula can be expressed as:

$$FID(x, g) = \|\mu_x - \mu_g\|_2^2 + Tr(\Sigma_x + \Sigma_g - 2\sqrt{\Sigma_x \Sigma_g}), \quad (5)$$

where  $x$  is the real image and  $g$  is the generated image.  $\mu$  and  $\Sigma$  respectively represent the mean value and covariance matrix of the feature vector set, generated by pre-trained



**Fig. 4.** Images of DVT-GAN



**Fig. 5.** SIFT matching results of fused images

Inception Net-V3 network.  $Tr$  represents the trace of the matrix. Lower FID means closer distributions, thus indicates higher image generation quality and better diversity.

Meanwhile, metrics of feature points characteristics is evaluated. SIFT (Scale-invariant feature transform) is used to detect local features and find the feature matching relationship between two images. Oriented FAST and Rotated BRIEF (ORB) performs as an efficient and viable alternative to SIFT while being almost two orders of magnitude faster. The larger the number of feature points, the higher the ability for matching. The comparison results of the image quality evaluation are shown in Table 2.

**Table 2.** Image quality evaluation

Image	Visible image (night)	Infrared image (night)	Fused image (CycleGAN)	Fused image (DVT-GAN)
FID	199.29	223.12	191.29	<b>148.04</b>
ORB	<b>500</b>	391	456	<b>500</b>
Matched (ORB)	84	41	31	<b>89</b>
SIFT	<b>1433</b>	154	352	<b>848</b>
Matched (SIFT)	15	27	21	<b>38</b>

As to the high noise in visible images at night, only a few keypoints can achieve good matching despite a large amount of SIFT feature points. Compared with the original image, the number of SIFT feature points and matched feature points in generated fusion images has been improved to a certain extent, as shown in Fig. 5.

### 4.3 Dvt-Slam

In this section, we further evaluate the position and navigation performance of DVT-SLAM.

We select four night-time sequences in the Brno dataset, as Table 3 shows. Each sequence contains visible images and their corresponding infrared images. Sequences 1 and 2 are collected in the urban area, while sequences 3 and 4 are collected in suburbs.

**Table 3.** Numbers of images and distance for each sequence

	Seq_1	Seq_2	Seq_3	Seq_4
Number of images	280	500	100	100
Distance traveled (m)	110	133	98.8	53.8

Absolute trajectory error (ATE) calculates the Euclidean distance between each estimated camera position and the closest GPS ground truth in time. In Table 4, we calculate the Root Mean Squared Error (RMSE) of ATE. The first column in Table 4 shows that visible failure occurs in low-illumination conditions, which may be caused by insufficient feature points, mismatching, and wrong extraction of keyframes. In the last column, the pseudo images generated by DVT-GAN help DVT-SLAM perform better. Compared with thermal SLAM, RMSE reduces by half on average. This result verifies our previous conjecture.

**Table 4.** Errors obtained from all sequences

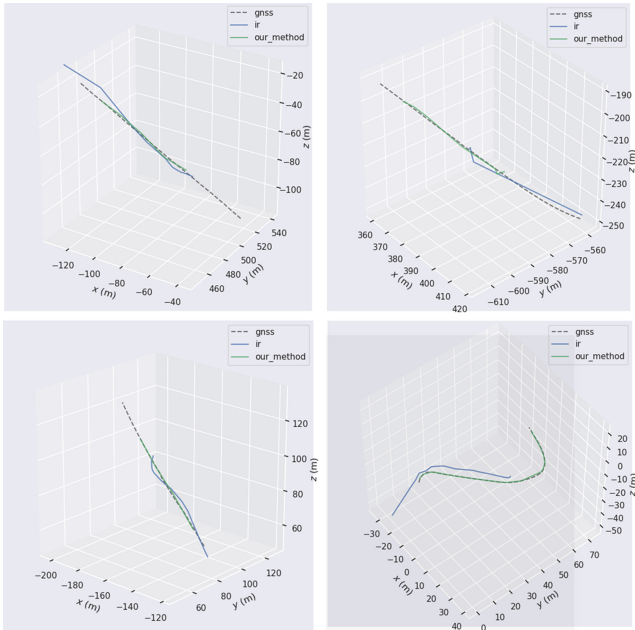
	Visible SLAM	Thermal SLAM	DVT-SLAM
	Sequence 1		
RMSE(m)	\	\	<b>16.04</b>
	Sequence 2		
RMSE(m)	\	24.68	<b>12.638</b>
	Sequence 3		
RMSE(m)	\	18.43	<b>5.13</b>
	Sequence 4		
RMSE(m)	\	41.13	<b>9.48</b>

“\” represents for the failure of SLAM.

Figure 7 shows the comparison of the trajectories by DVT-SLAM on different images. The trajectories are aligned with singular values since the monocular SLAM has no scale. Even though some drift can be noticed for fused images in Fig. 7, the overall trajectories of DVT-SLAM are estimated more properly compared with Thermal SLAM. In general,



the usage of pseudo images results in quite stable positioning effects. These results demonstrate the suitability of the proposed visible-thermal fusion approach for SLAM.



**Fig. 7.** The comparison of the trajectories between Thermal SLAM and DVT-SLAM

## 5 Conclusion

This paper proposed a visible-thermal fusion SLAM algorithm called DVT-SLAM. It combined self-supervised image generation network and visual SLAM system. The DVT-GAN network is designed for generating pseudo-visible fused images in low-illumination conditions. Moreover, high-quality fused images were utilized for monocular visual SLAM. The results showed that compared with the original visible and infrared images, the quality of the generated fusion image was improved. At the same time, better positioning and navigation effects could be obtained on the DVT-SLAM algorithm, and the trajectory error of the experiment on the dataset reduced by half.

In future, it is necessary to optimize the architecture of image fusion network DVT-GAN. The current method may have difficulty solving the problems of pseudo visible images generation in more complicated conditions. Even if the number of datasets involved is enlarged, additional work could be carried out to highlight the benefits.

## References

1. Mouats, T., Aouf, N., Sappa, A.D., et al.: Multispectral stereo odometry. *IEEE Trans. Intell. Transp. Syst.* **16**(3), 1210–1224 (2014)

2. Beauvisage, A., Aouf, N., Courtois, H.: Multi-spectral visual odometry for unmanned air vehicles. In: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 001994–001999 IEEE (2016)
3. Poujol, J., Aguilera, C.A., Danos, E., et al.: A visible-thermal fusion based monocular visual odometry. In: Reis, L., Moreira, A., Lima, P., Montano, L., Muñoz-Martinez, V. (eds.) Robot 2015: Second Iberian Robotics Conference. Advances in Intelligent Systems and Computing, vol. 417. Springer, Cham pp. 517–528 (2016)
4. Beauvisage, A., Ahiska, K., Aouf, N.: Multimodal tracking framework for visual odometry in challenging illumination conditions. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 11133–11139. IEEE (2020)
5. Choi, Y., et al.: Thermal image enhancement using convolutional neural network. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE (2016)
6. Sun, Y., Zuo, W., Yun, P., et al.: FuseSeg: semantic segmentation of urban scenes based on RGB and thermal data fusion. IEEE Trans. Autom. Sci. Eng. (2020)
7. Mur-Artal, R., Tardós, J.D.: Orb-slam2: an open-source slam system for monocular, stereo, and RGB-D cameras. IEEE Trans. Rob. **33**(5), 1255–1262 (2017)
8. Campos, C., Elvira, R., Rodríguez, J.J.G., et al.: ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM. arXiv preprint [arXiv:2007.11898](https://arxiv.org/abs/2007.11898), (2020)
9. Krotosky, S.J., Trivedi, M.M.: Mutual information based registration of multimodal stereo videos for person tracking. Comput. Vis. Image Underst. **106**(2–3), 270–287 (2007)
10. Zhu, J.Y., Park, T., Isola, P., et al.: 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. Park, T., Efros, A.A., Zhang, R., Zhu, J.: Contrastive learning for unpaired image-to-image translation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX, pp. 319–345. Springer, Cham (2020)
12. Li, C., Liu, H., Chen, C., Pu, Y., Chen, L., Heno, R., Carin, L.: Alice: Towards understanding adversarial learning for joint distribution matching. In: Advances in Neural Information Processing Systems (2017)
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: NIPSs, pp. 6626–6637 (2017)
14. Ligocki, A., Jelinek, A., Zalud, L.: Brno urban dataset-the new data for self-driving agents and mapping tasks. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 3284–3290. IEEE (2020)