2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Efficient-PIP: Large-scale Pixel-level Aligned Image Pair Generation for Cross-time Infrared-RGB Translation

Jian Li[1], Kexin Fei[1], Yi Sun[1], Jie Wang[1], Bokai Liu[1], Zongtan Zhou[1], Yongbin Zheng[1], Zhenping Sun[1]*

***Abstract*— Generative models are gaining momentum in both academic and industrial applications driven by the availability of large-scale datasets, especially in tasks involving Image-to-Image Translation. Meanwhile, poor human perception of nighttime environment has led to a demand for translation from night-vision infrared to day-vision RGB images. However, collecting such cross-modal training data at the same time is impossible due to the thermal imaging properties of infrared cameras, the challenge lies in constructing image pairs during the day and at night respectively, where the requirement for data alignment poses significant difficulties. In this paper, we propose a Pixel-level aligned Image Pair generation framework *PIP* to explore efficient colorization of high-resolution infrared images. Specifically, we first construct a 3D high-precision point cloud map for the purpose of establishing the correlation between day and night scenes. Corresponding point clouds of modal images are collected simultaneously during data acquisition to obtain image sensor poses by Global Matching with the map, which allows us to calculate the transformation relationship from infrared to RGB image coordinate systems based on the sensor parameters and depth information of the map. Leveraging the relationship, the pixel values of RGB image is projected onto the infrared image followed by optimization as the colored image. Accordingly, we present a dataset *NUDT-PIP*, the first of its kind containing large-scale pixel-level aligned cross-time infrared-RGB image pairs of complicated real road scenes. Experimental results demonstrate the reliability and strong applicability of our dataset in Image-to-Image Translation. Our code will be released at *https://github.com/wjjjjyourFA/NUDT-PIP*.**

I. Introduction

In the field of assisted driving, robotics and surveillance, visible light cameras are regarded as the "eyes" of the machines, providing key input to achieve accurate perception, control, and decision-making. However, they may exhibit inferior performance under conditions of insufficient illumination, losing crucial texture and structural information. In this case, infrared cameras are employed as auxiliary imaging systems, utilizing the thermal imaging principles to stably provide visual signals with relatively clear structure, while infrared images defy our traditional cognition of the real world due to the difference in grayscale, appearance and contents caused by its unique imaging spectrum ranges. Generally, the infrared images are often used for night-vision enhancement by fusing with visible data [1], [2]. Since the object and its environment constantly emit and scatter thermal radiation, the resulting "ghost effect" leads to the failure of infrared images to provide detailed semantic information [3]. Some researchers turn to physical coloring for more direct perception. F Bao *et al.* [4] proposed HADAR to reconstruct texture and depth of infrared images as if it were day, though theoretically suggesting that exploiting heat signal could be a new frontier for scalable perception, its irrational results still differ from real world scenes significantly.

(a) (b) (c)

Fig. 1: Illustration of domain gap between infrared images taken during the day (b) and at night (c) due to the thermal imaging principles of infrared cameras.

A feasible method is using generative models to bridge infrared and RGB modalities. Due to the inherent difficulties of constructing registered (pixel-level aligned) datasets in Image-to-Image (I2I) translation tasks, efforts have been made to overcome the domain gaps between unpaired night-vision infrared and day-vision RGB images. Z Yu *et al.* [5] proposed a cross-modal region similarity matching model ROMA based on CycleGAN [6], which used structural knowledge of infrared image to enhance the correspondence between the two modalities. Although ROMA provides a solution for solving this problem to some extent, such elaborate design for restricted task seems too cumbersome. Moreover, evidence [6] shows that, style-wise supervised generative models of the same period are unable to achieve compelling results with any of the pixel-wise supervised ones on various tasks in extensive experiments. As a supplement, we have demonstrated that style-wise supervised models which perform well in structured scenes such as highways and cities, may underperform in unstructured complex ones in our experimental part. Theoretically, the pixel-wise supervised models benefit from clearer transition relationship in pixel-level aligned images always show an advantage. It is known that the performance of deep learning-based models heavily relies on large amounts of high-quality training data, constructing a large-scale dataset of registered night-vision infrared and day-vision RGB image pairs will be of great significance for generative models to achieve their performance ceilings.

Generally, most of the existing paired infrared-RGB image datasets are constructed in the following ways. The first is

All authors are with College of Intelligence Science and Technology, National University of Defense Technology, China. Z. Sun is the corresponding author. E-mail: sunzhenping@nudt.edu.cn

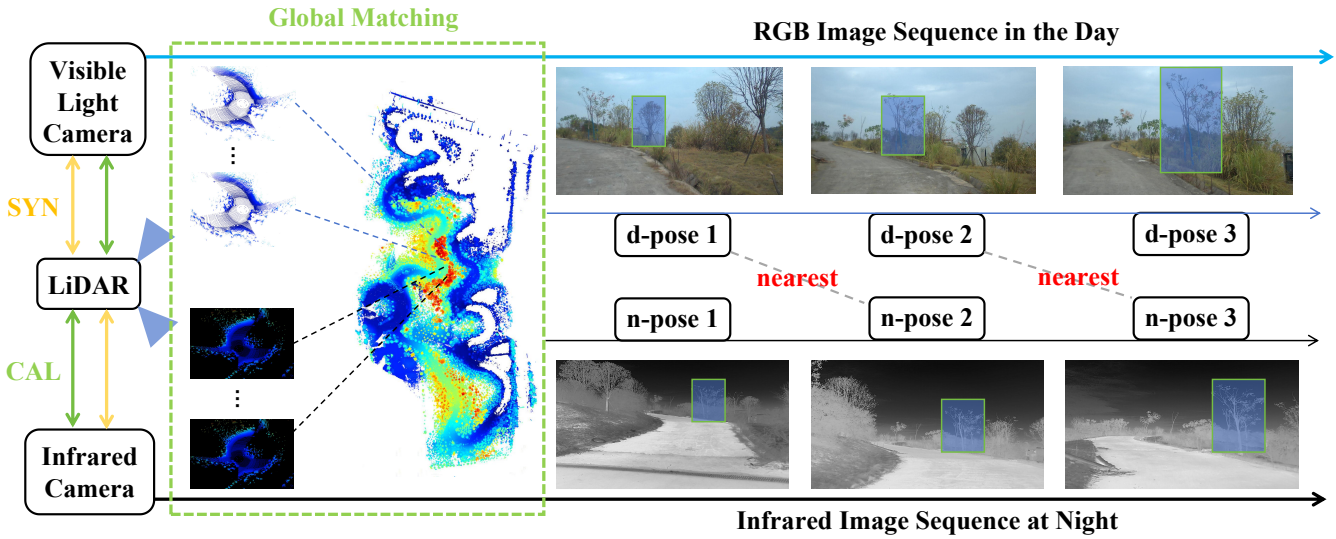

Fig. 2: After time calibration (CAL) and spatial synchronization (SYN) of the sensors, we collect synchronous modal image sequences and LiDAR points during the day and at night. Image sensor poses are obtained by Global Matching, where the nearest poses across time relate to the most similar infrared image and its corresponding RGB image.

collecting infrared and RGB images with binocular infrared-visible light camera at the same day time [7], which comes with some unavoidable defects. As is shown in Fig. 1, infrared images taken during the day and at night can be very different due to the heat-related imaging principle of infrared camera, models trained on day-vision infrared images can hardly produce satisfactory results on night-vision ones. Further, there are still perspective differences between cameras no matter how close they are put. In [8], a dataset of registered RGB and near-infrared (NIR) image pairs is introduced, the authors collect RGB and NIR images of the scene by fixing the tripod and replacing the camera. Similarly, a two-CCD camera system is employed in [9], where NIR and RGB spectra are split and directed to the dedicated CCD sensors respectively, the NIR and RGB images are recorded at the same time. Dataset obtained this way conquers the disadvantage of perspective differences, but still fails to collect registered data across time along with relatively small size, limited scenes and low efficiency. Other less used methods including physical coloring as is in [4] are often of poor quality. In general, it is empirically difficult to construct a cross-time pixel-level aligned image dataset directly, and there is no such datasets as far as we know.

Motivated by this observation, we introduce a cross-time image pair generation method *PIP*, which utilizes multi-sensors to construct the pixel-level aligned infrared-RGB dataset. We construct a 3D high-precision point cloud map in advance for the purpose of establishing the correlation between day and night scenes, then collect simultaneous modal image sequences and LiDAR points during the day and at night respectively. As we can see in Fig. 2, image sensor poses are obtained by matching LiDAR points with the pre-constructed 3D high-precision map. Naturally, the nearest poses relate to the most similar night-vision infrared image and day-vision RGB image. However, they are doomed to be unregistered (same object marked by the blue box in different location), as trajectories of the vehicle and the poses of the two cameras cannot be completely consistent across time. Notably, this misalignment cannot be eliminated by simple affine transformation due to the presence of depth in 3D scenes. To solve this problem, we propose to take LiDAR as medium to achieve cross-modal pixel alignment through point cloud projection and fill the pixel value of RGB image into the corresponding position of infrared image. Accordingly, we present a dataset *NUDT-PIP*, the first of its kind containing large-scale pixel-level aligned cross-time infrared-RGB image pairs of complicated real road scenes. Experimental results demonstrate the reliability and strong applicability of our dataset in I2I translation tasks.

In summary, the main contributions of this paper are as follows:

- An efficient cross-time image pair generation framework *PIP* is proposed, which can generate large-scale high-resolution data in speed and is extensible for any onshore scenes.
- A cross-time registered infrared-RGB dataset *NUDT-PIP* is proposed. To the best of our knowledge, it is the first pixel-level aligned night-vision infrared and day-vision RGB image dataset with real world scenes.
- The experimental results show that our dataset is reliable and practical with wide range of applications, which further promotes the development of I2I translation.

II. Related Work

A. Registered Cross-Domain Dataset Construction

Typically, the larger the training dataset is, the better performance of the deep learning model will be. For some

tasks, registered data can be collected manually or simply edited, the dataset for high-resolution reconstruction [20] can be constructed simply by down-sampling the images using bicubic interpolation with anti-aliasing enabled. Kim et al. [21] used the pre-trained face detector to collect anime faces, the selfie2anime dataset highly depends on the accuracy of the face detector. The labels of split dataset facades [22] are manually drawn with non-negligible limitations in size. Face datasets such as CelebA-HQ [23] usually have simple data distribution and small gap between the source and target domain, while models can perform well after reasonable training, there are very few application scenarios. For more challenging tasks as infrared-RGB image translation, the acquisition of registered data gets obviously harder, and the distribution gap between domains is larger. [7] proposed to put the infrared camera and the visible light camera close enough to obtain infrared and RGB image pairs at the same time. However, the effect of heat on infrared imaging is not taken into account, which is not conducive to the translation of nighttime infrared images to daytime RGB images in applications. Moreover, since the camera poses cannot be completely consistent, there still exists some noise in image pairs. Therefore, methods [24] are proposed to treat the infrared image as a gray image, attempting to generate the RGB image through coloring. Similarly, [25] uses the GAN [26] model to affine single-channel infrared images by shading. While these methods can produce colorful results, they tend to distort details in the absence of additional structural constraints, datasets obtained this way are often of poor quality and noisy. In addition, these image-level approaches lack consideration for time consistency, making it unsuitable for video translation tasks. Though attempts have been made to construct unpaired night-infrared and day-RGB images [5], the effect is still unsatisfactory.

B. Image-to-Image Translation

I2I translation refers to the mapping between two different image domains, many related tasks in computer vision could be formulated as I2I problems. A natural approach to I2I translation is to learn the conditional distribution of the target images given the samples from the input domain. After generative adversarial networks (GANs) were proposed, the number of generative models for I2I translation tasks exploded. Represented by Pix2Pix [10], pixel-wise supervised I2I translation [11], [12] has been proved to be capable of generating more realistic images in some tasks. However, these methods must be trained on registered datasets to achieve good quantitative and qualitative performance, i.e. images in the source domain have their counterparts in the target domain. As registered datasets are difficult and in some cases impossible to obtain in practical applications, attempts were made to overcome the need for them. Represented by CycleGAN [6], researchers proposed style-wise supervised I2I translation [13], [14], where model structure were designed to learn the potential spatial mapping between the source and the target domains with a large number of unpaired datasets. Studies have shown that the mismatch between images is harmful to find such mapping relationship, resulting in increased iterations and instability in the training process. In the consensus, style-wise supervised models still struggle to achieve the performance of pixel-wise supervised models in general. Recently, diffusion models [15] have shown competitiveness in terms of high-quality image generation. The conditional diffusion models [16]–[18] treat I2I translation as conditional image generation and guide the diffusion to the target domain during the reverse denoising process. However, most conditional diffusion models have poor generalization which can only adapt to some specific applications where the source and the target domains have high similarity [19]. With the advent of large-model era, high-quality registered datasets are indispensable for the development of I2I translation tasks based on generative models. Our work will provide a solid foundation for night perception.

III. Methodology

A. Architecture

The data collection platform used is shown in Fig. 3. The vehicle (a modified Dongfeng AX7 car) is equipped with a Robosense 128-beam LiDAR (30Hz), a Xsens MTI300 IMU, a NovAtel SPAN-CPT GNSS/INS system, a OV10650 12mm trigger visible light camera (10Hz) and an infrared camera (30Hz).

Fig. 3: Our sensor platform.

Before data collection, time synchronization and LiDAR-camera calibration are carried out to obtain the parameters we need (described in details in the following sections). Then, we drive the vehicle on a fixed country ring road during the day (10-25 km/h) and at night (7-10 km/h) respectively, capturing corresponding modal images and point clouds of the environment synchronously, where we denote RGB image sequence as $\{I_d^i\}$, infrared image sequence as $\{I_n^j\}$, point cloud sequence as $\{S_d^i\}$ and $\{S_n^j\}$.

As is mentioned, the identification of pixel-wise correspondences between the acquired sequences of the two modal images is notably challenging due to the difficulty of maintaining identical car speeds and trajectories precisely across different periods. Further correction is required due to the inherent perspective difference and de-synchronization between the infrared and RGB cameras. We detail this in the following sections, where targeted modules are proposed to solve this problem.

Concretely, enough overlapping area between modal images is the premise of coloring, to establish the correspondence between cross-modal images, we construct a 3D high-precision point cloud map M and propose Global Pose Matching to identify the most similar infrared and RGB images measured by corresponding camera poses. Taking point cloud as medium, we calculate the transformation relationship from infrared to RGB image coordinate systems with sensor pose and calibration parameters.

As direct projection leads to sparse coloring results, we extract dense point cloud from the map M as the better medium. Due to projection dislocations and crossovers caused by LiDAR point accumulation, we carry out surface reconstruction and noise removal of object for optimization.

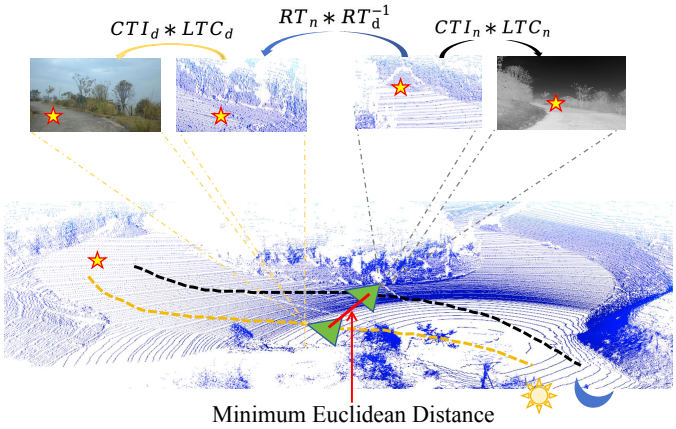

Fig. 4: Coordinate system relations between the world coordinate system, the day and night LiDAR coordinate systems, the infrared and RGB camera coordinate systems and the image coordinate systems. Taking point clouds as medium, cross-modal pixel correspondence is established from coordinate transformation.

B. Time Synchronization and Spatial Calibration

Since camera and LiDAR are different sensors, the image and point cloud sequences collected are unmatched without time synchronization and spatial calibration. We first design a synchronization system based on FPGA, which utilizes the infrared frame sync signal as trigger to synchronize the LiDAR, and trigger the exposure of the RGB camera sensor simultaneously, achieving joint sensor synchronization, which means, the cameras will be exposed when the LiDAR scanning line passes in front of the vehicle. It is known that LiDAR operates in rotation, the acquisition of a point cloud frame costs a fixed period, during which the vehicle is moving forward. Thus, the coordinate of each LiDAR point in the frame is relative to a moving car coordinate system, causing data distortion and dislocation. In order to achieve accurate correspondence, we carry out motion compensation for points within a frame based on the vehicle-mounted IMU, aligning all the points at a certain moment.

Spatial calibration solves the transformation relationship between LiDAR system and image coordinate system. The process is mainly divided into two parts. Transformation from LiDAR coordinate system to camera coordinate system is a rigid body transformation process, following ATOP [27], we use a Cross-Modal Object-matching Network (CMON) to estimate and optimize the LiDAR-Camera extrinsic parameters. While transformation from camera coordinate system to image coordinate system is a projection process, we use the Zhang's calibration method [28] to obtain the intrinsic parameters (including distortion parameters) of the cameras. The intrinsic parameters CTI and extrinsic parameters LTC of cameras obtained above describe the relations among the LiDAR, camera and image coordinate system.

C. High-Precision Map Construction and Global Pose Matching

We claim that the nearest sensor poses relate to the most similar night-vision infrared image and day-vision RGB image, where sensor poses of modal images must be solved. After time synchronization, modal image and point cloud frame at each timestamp describe the scene in the same sensor pose, which makes it possible for us to take point cloud as medium to identify the initial cross-time cross-modal image pair. To be specific, we scan the entire data collection area with LiDAR, and convert the points to the world coordinate system using IMU, then we employ offline SLAM [29] to complete the mapping and pose estimation. For each frame of point clouds S_d^i and S_n^j, their corresponding LiDAR pose P_d^i and P_n^j are obtained by joint optimization with NovAtel GNSS/INS. Note that the LiDAR data must be filtered as points hit on the vehicle body bring regular noise on the map road.

The coordinate of pose in the world coordinate system can be explicitly represented as $(x, y, z, \phi, \omega, \theta)$, where (x, y, z) denotes the LiDAR position and (ϕ, ω, θ) denotes the angles of deflection. The coordinate system relations and specific transformation parameters are shown in Fig. 4, where the coordinates of poses in the world coordinate system is the origin of the LiDAR coordinate system, we refer to this transformation as RT_n and RT_d. The poses (noted by trajectories in yellow and black) can hardly be completely consistent across time, which means,

$$\exists\ \epsilon > 0, \|P_n^j - P_d^i\|_2 > \epsilon, \forall i, j\ , \tag{1}$$

resulting in a difference in perspective between I_n^j and I_d^i. In order to keep as much overlap between cross-time point clouds as possible during transformation, for any given infrared image I_n^j with pose P_n^j, we use the Euclidean distance to calculate its nearest day pose P_d^i, and take its corresponding raw RGB image I_d^i as the initial matching image.

$$d(P_n^j, P_d^k) \geq d(P_n^j, P_d^i), \forall P_d^k, k \neq i\ , \tag{2}$$

where $d(P_n, P_d) = \|(x_n, y_n, z_n, \theta_n) - (x_d, y_d, z_d, \theta_d)\|_2$, θ denotes the yaw angle of the vehicle. As the vehicle may backtrack to its coming road during data collection, it is necessary to add θ in pose distance calculation to ensure that image pair related to the closest poses are captured in

the same direction. The image pairs obtained at this step guarantee the highest coverage of the scene, but they cannot achieve the desired pixel-level alignment. We propose to use Point Cloud Projection to find the corresponding pixel from I_d^i to colorize I_n^j.

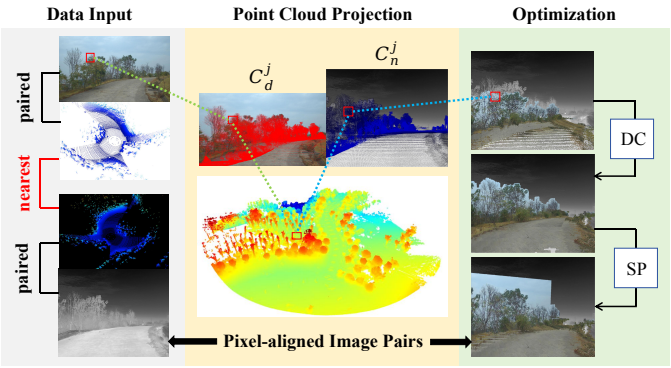

Fig. 5: After Point Cloud Projection, Optimization part consists of two continuous modules: DC module realizing the densification of point cloud through depth completion and SP module realizing the projection of sky region by value assignment.

D. Point Cloud Projection

Projecting the nighttime point cloud S_n^j onto the infrared image plane I_n^j, we get a sparse depth map Q_n^j with the same resolution as I_n^j,

$$S_n^j \xrightarrow{g} Q_n^j, where\ g = CTI_n * LTC_n\ . \tag{3}$$

Furthermore, we project S_n^j onto I_d^i as Q_d^j,

$$S_n^j \xrightarrow{f} Q_d^j, where\ f = CTI_d * LTC_d * RT_d^{-1} * RT_n\ , \tag{4}$$

where S_n^j is transformed to the day LiDAR coordinate system through pose transformation $RT_d^{-1} * RT_n$, and then projected onto the RGB image plane I_d^j. Notably, there is a pixel-to-pixel correspondence between Q_n^j and Q_d^j related to some point in S_n^j, where we should find the target RGB pixel value in I_d^i to fill into I_n^j. As S_n^j contains points behind the infrared camera, we convert S_n^j to the camera coordinate system with LTC_n to perform region filtering to prevent negative effects from the incorrect projected point in the above process. It is worth noting that we will find the projection points in Q_d^j and Q_n^j very sparse when the original point clouds S_n^j is directly projected, large number of pixels in I_n^j cannot be indexed. Therefore, we propose to take the dense point clouds D_n^j extracted from the high-precision map M according to pose P_n^j as the medium in practice.

As is shown in Fig. 5, RGB points are correctly assigned to the corresponding positions in the infrared image, but still presents a sparse state, especially on the road surface with a relatively close distance, where pixels can hardly get covered even through interpolation. Moreover, dense point clouds accumulating multiple frames brings wrong projection, as the points should be obscured behind the target may be incorrectly colored due to the holes of the LiDAR points on the front surface of the object. Therefore, it is necessary to further densify D_n^j so that it can index as many pixels as possible on I_d^i.

E. Optimization

It turns out that there remains several problems, namely region missing and wrong projection caused by accumulation as well as the inherent sparsity and distance limitations of point clouds. We propose to solve these problems with two successive modules. Firstly, we aim to estimate the surface of the objects in the current viewing angle by filtering to avoid point cloud view confusion caused by perspective, which is quite similar to depth completion task. Thus, we borrowed the method in [30] to densify the sparse depth map Q_n^j with basic image processing operations. In Depth Completion (DC) module, we use the Minimum Filter with a custom mask to bias the value of the blank area towards the surrounding points with the shortest distance, the holes are continued to be filled by morphology closing and dilatation. Further, Median Filter and Gaussian Filter are used for noise removal and smoothness, final completion result can be given by inverting the depth value.

As LiDAR returns no point cloud in the sky, there is no corresponding pixel in Q_d^j for sky region in I_n^j, in Sky Projection (SP) module, we manually assign a large value L for sky region, where

$$D_n^j(p) \ll L, \forall D_n^j(p) \in D_n^j\ , \tag{5}$$

the L arises from the fact that the distance to the sky is considered infinite, its surface can be approximated as a plane, large value can effectively distinguish the sky from other terrestrial regions. As we can see in Fig. 5, our method can achieve a good effect with the above modules.

IV. Dataset Analysis

A. Dataset Description

In the generation of the cross-time registered images, we utilize a large amount of multi-modal data, which is also provided in NUDT-PIP. As is shown in Fig. 6, the first two columns are night-vision infrared images and their pixel-level aligned colored results. The scene is quite complicated varying in sandy and cement road with turns and uphill slopes, detailed textures such as branches, leaves and sandy pavements are well synthesized with a high proportion of valid and coherent region. The corresponding sparse and dense LiDAR point clouds with relevant important parameters are included, containing points 70m area ahead. For ease of presentation, we project the point cloud onto its corresponding modal image. As part of infrared image cannot be colored due to the difference between night-vision infrared image and its adjacent day-vision RGB image, we also provide mask of the valid region. There are 5595 cross-time image pairs in total with a resolution of 1024x1280, and the number of LiDAR points is collectively referred to as resolution, other statistics are detailed in Tab. 1. Actually,

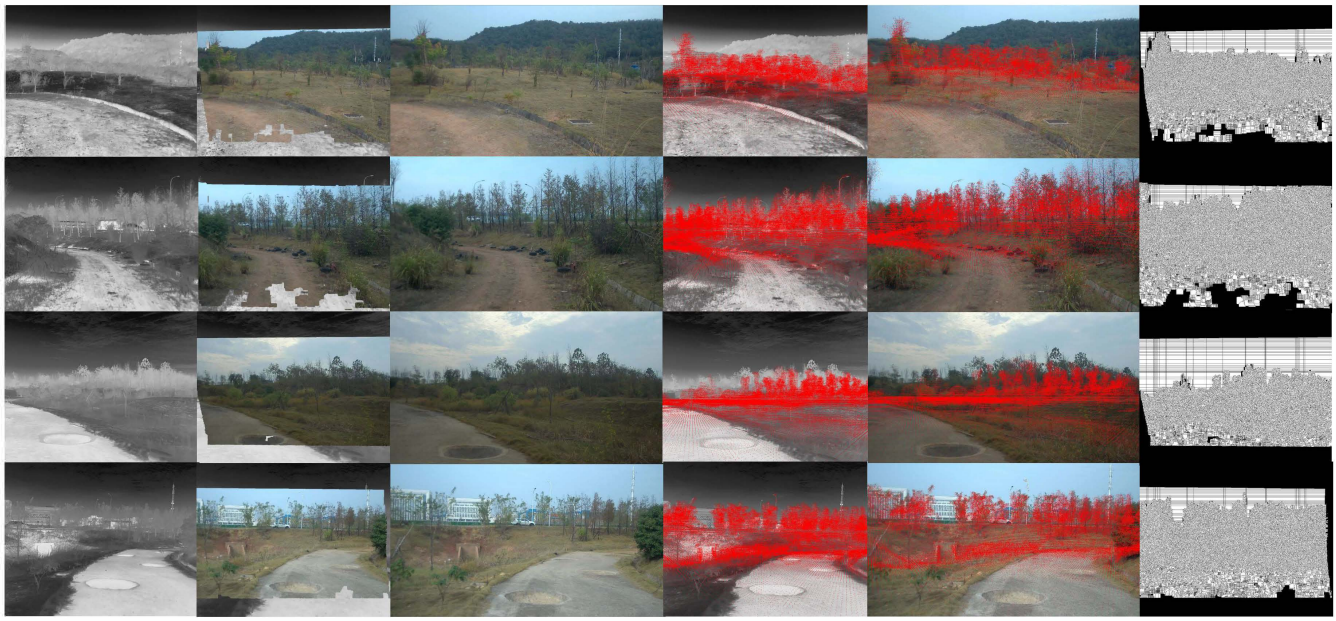

Fig. 6: Examples of NUDT-PIP. Each column displays the night-vision infrared image, its coloring results, adjacent day-vision RGB image, the corresponding cross-time dense point clouds, and the region mask from the left to the right.

unlimited number of data can be generated with our method in theory. Notably, our image resolution is about 4 times larger than ROMA [5], and our scenes are much more complicated, which makes it quite challenging.

TABLE I: The structure of NUDT-PIP.

Period	Data	Frames	Resolution
Day	RGB Images	36521	1024x1980
	Raw LiDAR Data	36509	∼2e5
	Dense LiDAR Data	5094	∼2e6
Night	Infrared Images	108352	1024x1280
	Raw LiDAR Data	36110	∼2e5
	Dense LiDAR Data	5595	∼2e6
Generated	Infrared Images	5595	1024x1280
	Colored Images	5595	1024x1280
	Masks	5595	1024x1280

B. Error Analysis

To illustrate the reliability of NUDT-PIP, we randomly select 100 pairs of infrared-colored-raw RGB images and carry out confidence analysis with three metrics named *Point Error*, *Distance Error* and *Angle Error*. Specifically, we label two explicit target points $p_1 = (x_1, y_1)$ and $p_2 = (x_2, y_2)$ on the image, where p_i^{inf} and p_i^{col} represent the same position of the infrared and colored images respectively (p_i^{col} will be replaced by p_i^{raw} when calculating metrics of the raw RGB images). As is shown in Fig. 8(b), the target point is usually picked on a fixed reference object, such as a building or a tree trunk to avoid errors caused by non-algorithmic reasons. The calculation of metrics are as follows:

$$\textit{Point Error} = \frac{e(p_1^{inf}, p_1^{col})}{r} \ , \tag{6}$$

$$\textit{Distance Error} = \frac{\left| e(p_1^{inf}, p_2^{inf}) - e(p_1^{col}, p_2^{col}) \right|}{max\{e(p_1^{inf}, p_2^{inf}), e(p_1^{col}, p_2^{col})\}} \ , \tag{7}$$

$$\textit{Angle Error} = \left| \theta(p_1^{inf}, p_2^{inf}) - \theta(p_1^{col}, p_2^{col}) \right| \ , \tag{8}$$

where $e(p_1, p_2) = \|p_1 - p_2\|_2$, $\theta(p_1, p_2) = arctan(\frac{x_2 - x_1}{y_2 - y_1})$, and r denotes the resolution of images. The metrics evaluate the pixel alignment accuracy and affine transformation degree of the generated image, in which way the overall "matching degree" of the colored image can be observed. The results of error analysis are shown in Fig. 8(a)(c)(d), the *Point Error* of colored images (noted in blue) is basically within 3% of the image resolution, which is significantly lower than 40% of the raw RGB image (noted in green). Similarly, the *Distance Error* is within 10%, and the *Angle Error* is under $\frac{0.14}{2\pi}$, both of which are greatly improved. During data pre-processing such as motion compensation and LiDAR-camera calibration, the defects of hardware platform would inevitably bring certain errors to the generated data. Considering that there also exists some bias in manual annotation, we claim that our dataset is reliable.

V. Experiments

In this section, we conduct infrared to RGB image translation experiments on NUDT-PIP dataset to illustrate its usability. Further, aiming at the effect of registered data on model performance, we mainly tested on two generative models Pix2Pix and CycleGAN to prove the necessity of constructing pixel-level aligned dataset.

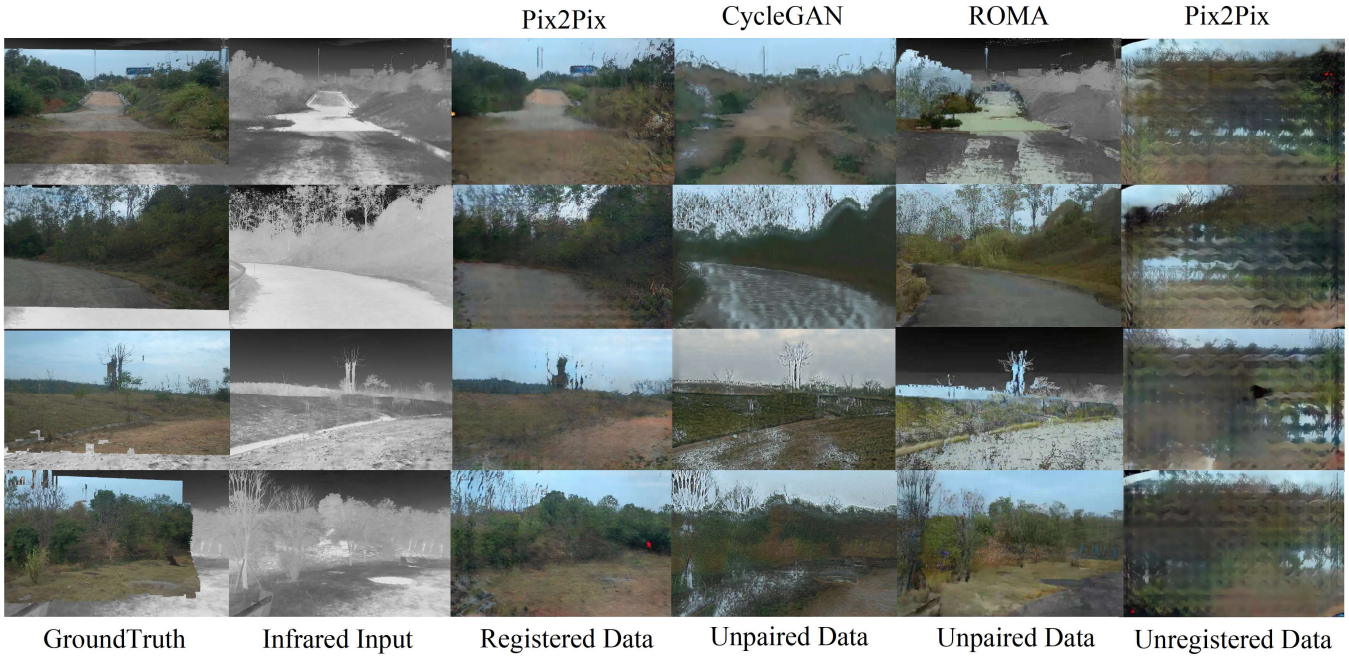

Fig. 7: Translation results of models on different inputs. Each column displays the colored ground-truth images, infrared images, translation results of Pix2Pix on registered data, results of CycleGAN and ROMA on unpaired data and of Pix2Pix on unregistered data from left to the right.

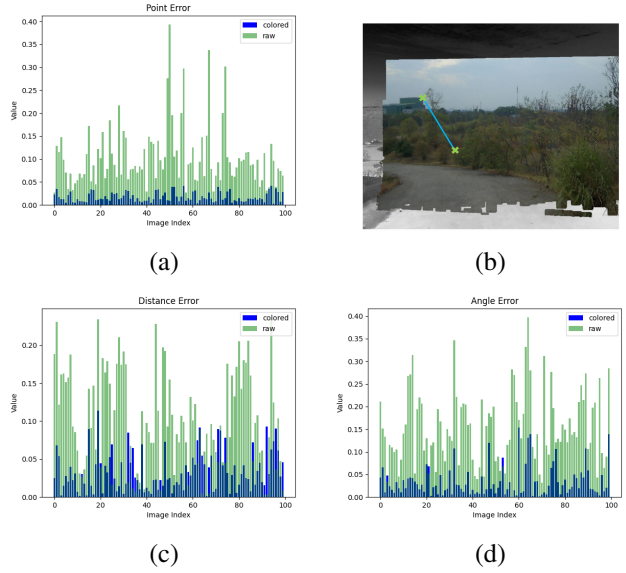

Fig. 8: Error analysis on 100 image pairs randomly selected from NUDT-PIP, including Point Error(a), Distance Error(c) and Angle Error(d).

A. Experimental Setup

Datasets 5595 registered infrared-colored image pairs are used as training data of pixel-wise supervised Pix2Pix, and we deliberately disrupt the order of image pairs as unpaired data used for training style-wise supervised CycleGAN and ROMA. In addition, infrared images and raw RGB images with nearest poses are also taken as unregistered data for Pix2Pix training to explore the the effect of "matching degree" on model performance, where "matching degree" is equivalent to a kind of data noise. We point out that the noise contained in image pairs gets greater from registered to unregistered, and then to unpaired data. Before the training starts, we scale the image to the same size to avoid shift of aligned pixels.

Implementation Details All experiments are carried out on NVIDIA RTX4090. We keep the original network structure configuration and loss functions of Pix2Pix and CycleGAN, following the authors, we use mini-batch SGD and apply the Adam solver with initial learning rate of 0.0002, momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$, where learning rate is fixed in the first 100 epochs, and then decline linearly to 0 afterwards. Notably, if trained with the unprocessed colored data, models may generate random results in part of the sky region which could not be covered by our method. Thus, we apply online data augmentation to improve it, including cropping, rotation, normalization, etc. We also introduce the mask into loss function to avoid the negative effect of uncolored region, only the valid part of training image is involved in back-propagation.

B. Performance on NUDT-PIP

In this section, we aim to show experimentally that registered data is of significance for I2I translation based on generative models. We trained Pix2Pix on registered data and CycleGAN on unpaired data where images were disrupted. For comparison, we trained unregistered data on Pix2Pix which consisted of the infrared image and their corresponding raw RGB images of nearest poses. The results

are shown in Fig. 7, as we can see, Pix2Pix trained on registered data can well learn the texture distribution of the target domain in general, generating realistic RGB images across all scenes. Detailed structure such as the tiny sign next to the uphill slope in the translated results can be clearly identified. Various pavement materials are accurately colored, and the transition from sandy to cement road is quite smooth. In general, our dataset provides a satisfactory translation guidance, which means that error within relatively small range has a controllable effect on the generative model.

For Pix2Pix trained on data with unregistered images, the tone of the scene is natural, while the structure gets unacceptable chaotic, it is inferred that the noise caused by structure dislocation is highly disadvantageous for the model to learn the corresponding distribution. Designed for training with unpaired data, CycleGAN shows a stronger resistance to the above structural noise, its translation results grasp the structure of the input infrared images. However, their textures are incorrectly colored, with only a vague differentiation between the trees and the road. For comparison, we also trained our dataset on ROMA, the results tells that style-wise supervised models with good performance on structured environments such as highways and cities are not sufficient for complex scenarios. All of the above indicate that training with unpaired data will have a certain negative impact on generative models in I2I translation. For challenging I2I translation tasks, building a large number of registered datasets is still crucial, as it significantly affects model performance on corresponding task.

VI. Conclusion

In this paper, we proposed a brand-new method PIP to colorize the night-vision infrared image utilizing data from multi-sensors, achieving efficient high-resolution pixel-level aligned cross-time data generation. NUDT-PIP is the first large-scale registered night-vision infrared and day-vision RGB image dataset with complicated scenes to the best of our knowledge. The experimental results show that our dataset is reliable and practical in wide range of applications.

References

[1] X. Zhang, P. Ye, and G. Xiao, VIFB: A visible and infrared image fusion benchmark, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. pp. 468–478.

[2] Z. Zhou, M. Dong, X. Xie, and Z. Gao, Fusion of infrared and visible images for night-vision context enhancement, *Applied Optics*, 2016. pp. 6480–6490.

[3] M. He, Q. Wu, K. Ngan, F. Jiang, F. Meng and L. Xu, Misaligned RGB-infrared object detection via adaptive dual-discrepancy calibration, *Remote Sensing*, 2023, pp. 4887.

[4] F. Bao, X. Wang, S. Sureshbabu, G. Sreekumar, L. Yang, V. Aggarwal, et al, Heat-assisted detection and ranging, *Nature*, 2023, pp. 743-748.

[5] Z. Yu, K. Chen, S. Li, B. Han, C. Liu and S. Wang, ROMA: cross-domain region similarity matching for unpaired nighttime infrared to daytime visible video translation, *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5294-5302.

[6] J. Zhu, T. Park, P. Isola, and A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2223-2232.

[7] S. Li, B. Han, Z. Yu, C. Liu, K. Chen, and S. Wang, I2V-GAN: Unpaired Infrared-to-visible video translation, *Proceedings of the ACM International Conference on Multimedia*, 2021, pp. 3061-3069.

[8] M. Brown, S. Süsstrunk, Multi-spectral sift for scene category recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 177–184.

[9] M. Limmer, H. Lensch, Infrared colorization using deep convolutional neural networks, *IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016, pp. 61–68.

[10] P. Isola, J. Zhu, T. Zhou, and A. Efros, Image-to-image translation with conditional adversarial networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125-1134.

[11] S. Srivastava, F. Jurie and G. Sharma, Learning 2d to 3d lifting for object detection in 3d for autonomous vehicles, *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 4504-4511.

[12] X. Li, S. Zhang, J. Hu, L. Cao, X. Hong, X. Mao, et al, Image-to-image translation via hierarchical style disentanglement, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8639-8648.

[13] T. Park, A. Efros, R. Zhang and J. Zhu, Contrastive learning for unpaired image-to-image translation, *European Conference on Computer Vision (ECCV)*, 2020, pp. 319-345.

[14] M. Zhao, F. Bao, C. Li and J Zhu, Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations, *Advances in Neural Information Processing Systems (NeurIPS)*, 2022, pp. 3609-3623.

[15] J. Ho, A. Jain, and P. Abbeel, Denoising diffusion probabilistic models, *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 6840–6851.

[16] G. Batzolis, J. Stanczuk, C. Schonlieb, and C. Etmann, Conditional image generation with score-based diffusion models, arXiv preprint arXiv:2111.13606, 2021.

[17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, High-resolution image synthesis with latent diffusion models, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10684–10695.

[18] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, et al, Palette: Image-to-image diffusion models, arXiv preprint arXiv: 2111.05826, 2021.

[19] M. Yi, J. Sun, Z. Li, On the generalization of diffusion model, arXiv preprint arXiv:2305.14712, 2023.

[20] C. Saharia, J. Ho, W. Chan, T. Salimans, D. Fleet, and M. Norouzi, Image super-resolution via iterative refinement, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022, pp. 4713-4726.

[21] J. Kim, M. Kim, H. Kang, and K. Lee, U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation, arXiv preprint arXiv:1907.10830, 2019.

[22] R. Tyleček and R. Šára, Spatial pattern templates for recognition of objects with regular structure, *Pattern Recognition*, 2013, pp. 364-374.

[23] T. Karras, T. Aila, S. Laine and J. Lehtinen, Progressive growing of gans for improved quality, stability, and variation, arXiv preprint arXiv:1710.10196, 2017.

[24] B. Sheng, H. Sun, M. Magnor, and P. Li, Video colorization using parallel optimization in feature space, *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2013, pp. 407-417.

[25] P. Suárez, A. Sappa and B. Vintimilla, Infrared image colorization based on a triplet DCGAN architecture, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 18-23.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al, Generative adversarial networks, *Communications of the ACM*, 2020, pp. 139-144.

[27] Y. Sun, J. Li, Y. Wang, X. Xu, X. Yang and Z. Sun, ATOP: An attention-to-optimization approach for automatic LiDAR-camera calibration via cross-modal object matching, *IEEE Transactions on Intelligent Vehicles*, 2022, pp. 696-708.

[28] Z. Zhang, A flexible new technique for camera calibration, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2000, pp. 1330–1334.

[29] T. Shan, B. Englot, D. Meyers, W. Wang, C. Ratti, and D. Rus, Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping, *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 5135-5142.

[30] J. Ku, A. Harakeh and S. Waslander, In defense of classical image processing: Fast depth completion on the cpu, *Conference on Computer and Robot Vision*, 2018, pp. 16-22.