2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Multimodal Image Registration for GPS-denied UAV Navigation Based on Disentangled Representations

Huandong Li, Zhunga Liu, Yanyi lyu and Feiyan Wu

***Abstract*— Visual navigation plays an important role for Unmanned Aerial Vehicles(UAVs). In some applications, the landmark image and the real-time image may be heterogeneous, like near-infrared and visible images. In this work, we propose a multimodal image registration method to deal with near-infrared and visible images so that it can be applied to visual navigation system for the localization of UAVs in GPS-denied environments. At first, a new feature extraction strategy is developed to embed different modalities of images into the common feature space based on disentangled representations. Such common space is independent of the image modality, and this can eliminate the modality differences. Meanwhile, an intensity loss is introduced to measure the similarity of mono-modal images. In the proposed method, we can directly predict the transformation parameters and thus accelerates the localization of UAVs. Extensive experiments on synthetic datasets are conducted to demonstrate the validity of our method, and the experimental results show that the proposed method can effectively improve the localization accuracy.**

I. INTRODUCTION

Recently, Global Positioning System(GPS) navigation has been widely used for the localization of UAVs [1]. However, GPS is a passive navigation method, vulnerable to electromagnetic interference and noise signal, and can not be used in some scenes. Inertial navigation system(INS) is an active navigation method, which has the advantages of strong autonomy and robustness to external interference, but it has positioning errors that accumulate over time. With the development of computer vision technology, visual navigation has become an important navigation method for UAVs [2], [3], [4], [5]. In GPS-denied environments, in order to preserve the reliability of the navigation system, visual navigation system is often used as an aided navigation system to correct the error of INS.

Using landmark images preloaded on the UAVs in conjunction with real-time vertical view images is an available navigation method [6]. Utilizing the geographic information contained in the landmark images, we can get the position of UAVs. Visible images are effortless to access and usually serve as landmark images. However, visible images could be easily affected by weather, light, clouds, etc. The quality of the images obtained by visible sensors can not be guaranteed when UAVs perform missions at night or in complex climatic conditions. Some visual sensors such as the near-infrared sensors have strong night imaging ability and work well under bad weather conditions. Therefore, visual navigation system with visible landmark images and near-infrared real-time images is designed. The diagram of the visual navigation system is shown in Fig. 1. Multimodal image registration is a crucial step which refers to the process of aligning two images in space which are acquired by different sensors at different times.

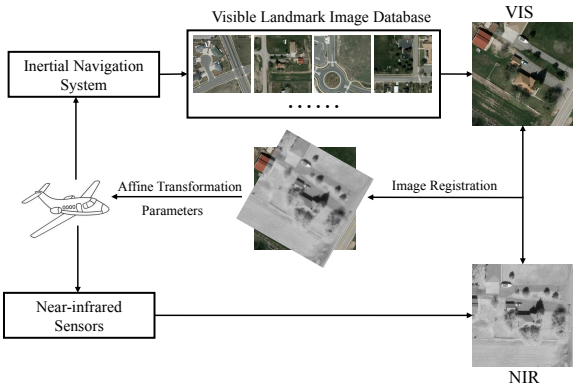

Fig. 1: Diagram of the visual navigation system. Visible landmark images containing geographic information are preloaded on the UAV. According to the INS, the vertical view image is captured by the near-infrared sensor when the UAV reaches near a landmark. Then image registration between the real-time image and the corresponding landmark image is performed. Once the transformation parameters are obtained, we could calculate the position of the UAV based on the affine transformation model.

For the navigation system of UAVs, the registration of visible images and near-infrared images faces many challenges. First, the imaging mechanism of visible sensors is different from that of near-infrared sensors. There is a large nonlinear gray distortion between them, which makes them look very different in appearance. Visible images are more colorful, more sensitive to light and contain more texture while near infrared images lack detail and texture information. Second, because of the drift error during flight, there is often a huge geometric difference between real-time images and landmark images, which makes the registration more difficult. Finally, the multimodal image registration method needs to be not only accurate but fast enough to ensure the localization process is real-time.

In order to solve the above problems, this paper proposes a fast multimodal image registration method for visible and near-infrared images based on disentangled representations. We make an assumption that images can be encoded into a shape latent code from shape space and an attribute latent code from attribute space. We reduce the modality

This work was supported in part by the National Natural Science Foundation of China under Grant U20B2067, Grant 61790552, and Grant 61790554; and in part by the Aeronautical Science Foundation of China under Grant 201920007001.

Corresponding author: Huandong Li(lihuandong@mail.nwpu.edu.cn).

All authors are with the School of Automation, Northwestern Polytechnical University, China.

differences by embedding both images into the shape space during the process of feature extraction. Then we propose an intensity loss which can constrain both the registration network and the disentangled representation model. In order to accelerate the localization process of UAVs, the method is designed to predict affine transformation parameters directly in an end-to-end way. Experiment results show that our method can effectively deal with the problem of large geometric deformation.

The contributions of this paper are summarized as follows:

1) Based on disentangled representations, a new effective feature extraction strategy is proposed to reduce the differences between modalities.
2) An intensity loss is proposed to facilitate the learning of both the registration network and the disentangled representation model.
3) Our proposed method has shown competitive performance on synthetic datasets and meet the requirements of accuracy and speed.

The rest of this paper is arranged as follows. Section II introduces the related literature. Section III describes the image registration method in detail. Quantitative and qualitative experiments are presented in Section IV. Finally, some conclusions are made in Section V.

II. RELATED WORKS

As an increasingly important way of navigation, visual navigation methods are first introduced in this section. Then we summarize several feature extraction methods based on image registration and matching. Finally, we introduce some translation-based image registration methods which are widely used in multimodal image registration recently.

Visual Navigation. There has been much literature on visual navigation. In [7], the authors proposed an assisted UAV localization method based on deep learning in GPS-denied cases. In [8], the authors localized UAVs using satellite images. They trained an autoencoder to encode images to feature vectors and used an inner-product kernel to compare them for matching. In [9], the authors proposed a framework for aerial image registration, and they applied it to UAV geolocalization. Compared with these schemes, our method mainly focuses on environments that are stricter. Since our real-time image and landmark image are images of two different modalities, the image registration process is more complicated and more crucial.

Feature Extraction. Compared with the traditional hand-crafted features, the learnable features extracted by neural networks are more robust and have stronger representation ability. Recently, the double branch networks, including Siamese networks and pseudo-Siamese networks, have been widely used to embed images into feature space. Generally, Siamese networks shared parameters between branches are suitable for tasks with similar inputs. In [10], [11], [12], [13], [14], [6], [15], the authors used Siamese networks to extract features from input images. Pseudo-Siamese networks with unshared parameters are often used to handle tasks with discrepant inputs. In [16], [17], [18], pseudo-Siamese networks were used for SAR and optical images registration or matching. Our proposed network considers the characteristics of Siamese networks and pseudo-Siamese networks. For visible images and near-infrared images, the low-level feature representations such as edges and texture are very different. A pseudo-Siamese network is first used to learn feature representations according to the characteristics of respective input modalities. At the same time, we embed both visible images and near-infrared images into the shape space so that the modality differences are reduced. Then a Siamese network is used to process features from shape space because the feature representations are very similar now.

Image Translation. In order to deal with the problem of multimodal image registration, many researchers have adopted the method of image translation, which means converting images of different modalities into the same modality. In [19], in order to match optical and SAR images, optical image blocks were used to generate SAR image blocks by conditional generative adversarial networks(cGANs) [20]. With the help of generated image blocks, the authors realized image blocks matching just by using conventional methods such as SIFT [21]. In [22], the authors proposed a similar method to perform image registration between visible and infrared images. These two-step methods mainly focus on generating images that are as similar as possible to real images in appearance. However, image translation itself is exactly a challenging task. When the quality of the generated image is poor, it is difficult to obtain ideal registration results. We reduce the effect of synthetic images quality on registration by mainly focusing on the feature space rather than the image space. We aim to get similar feature representations instead of just generating images. Besides, our proposed method can perform image translation and image registration simultaneously rather than separately.

III. METHODS

The multimodal image registration method must meet the requirements of accuracy and speed to be applied to the UAV navigation system. Based on disentangled representations, we reduce the differences between modalities and propose an intensity loss to improve the accuracy. As for speed, our proposed method can perform image registration by feature extraction, matching and regression in an end-to-end way, which avoids the complex post-processing process. In this section, we detail the proposed multimodal image registration method for visible and near-infrared images.

A. Overview

As shown in Fig. 2, our image registration framework consists of six independent components and three sub-networks. The six components are the two shape encoders E_x^s, E_y^s, the two attribute encoders E_x^a, E_y^a and the two generators G_x, G_y, respectively. The three sub-networks are the feature extraction network, the matching layer and the regression network, respectively. Disentanglement procedure is conducted in the process of feature extraction by image reconstruction and image translation to reduce modality

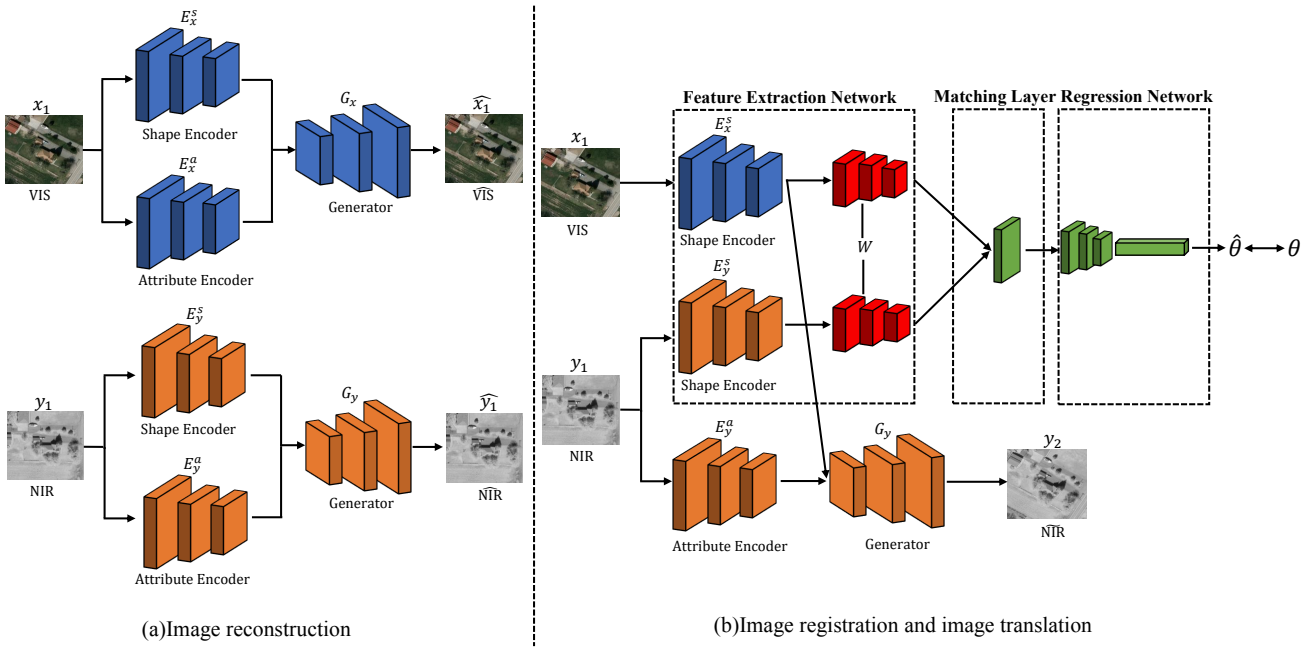

Fig. 2: Overview of the image registration framework. (a) The visible image and near-infrared image are encoded by shape encoder and attribute encoder. Then the reconstructed image is generated by generator. (b) We swap the shape encoder to perform image translation and generate an image. The generated image has the same "shape" as the visible image and the same "attribute" as the near-infrared image. Meanwhile, image registration is performed by three sub-networks. It is worth noting that the feature extraction network has partially shared parameters W.

differences. Then matching and regression are performed to predict the affine transformation parameters θ.

B. Disentangled Representations

As shown in Fig. 3, we make an assumption that images in different visual domains can be disentangled into a shared shape space and a specific attribute space. The shape space contains information that is invariant between modalities, such as geometric structure. On the contrary, the attribute space contains information that varies with modalities such as texture, brightness, and some detail. Since the shape space is domain-invariant, we can get modality-independent feature representations of images from different domains by mapping them into the common shape space. We can also generate an image utilizing a shape latent code from shape space and a attribute latent code from attribute space.

Let $X \subset R^{H\times W\times 3}$ be the visible image domain and $Y \subset R^{H\times W\times 3}$ be the near-infrared image domain. $x_1 \in X$ and $y_1 \in Y$ are the visible image and near-infrared image to be registered, respectively. According to the assumption, x_1 can be mapped into the shape space S and the visible attribute space A_x. We design a visible image shape encoder E_x^s and a visible image attribute encoder E_x^a to implement this nonlinear mapping.

$$s_x = E_x^s(x_1), a_x = E_x^a(x_1) \tag{1}$$

Similarly, we can decompose image y_1 into the shape space S and the near-infrared attribute space A_y by a near-infrared image shape encoder E_y^s and a near-infrared image attribute encoder E_y^a.

$$s_y = E_y^s(y_1), a_y = E_y^a(y_1) \tag{2}$$

Here, $s_x, s_y \in S$ are the shape latent codes, and $a_x \in A_x, a_y \in A_y$ are the attribute latent codes. Then, we design a visible image generator G_x and a near-infrared image generator G_y to implement image generation for the two domains, respectively. As shown in Fig. 2(a), image reconstruction is performed to make sure the latent codes contain enough information to recover input images.

$$\hat{x_1} = G_x(s_x, a_x), \hat{y_1} = G_y(s_y, a_y) \tag{3}$$

The image reconstruction loss L_{recon} is used to measure the error between reconstructed images and input images. L_{recon} is formulated as follows:

$$\begin{aligned} L_{recon} &= L_{recon}^x + L_{recon}^y \\ &= ||x_1 - \hat{x_1}||_1 + ||y_1 - \hat{y_1}||_1 \\ &= ||x_1 - G_x(s_x, a_x)||_1 + ||y_1 - G_y(s_y, a_y)||_1 \end{aligned} \tag{4}$$

Here, $||\cdot||_1$ represents the $L1$ norm. To realize the disentanglement of the shape latent code and the attribute latent code, as shown in Fig. 2(b), we swap the shape latent code to perform image translation.

$$y_2 = G_y(s_x, a_y) \tag{5}$$

Here, y_2 has the same attribute code with y_1 and the same shape code with x_1. Therefore, $y_2 \in Y$, is a near-infrared

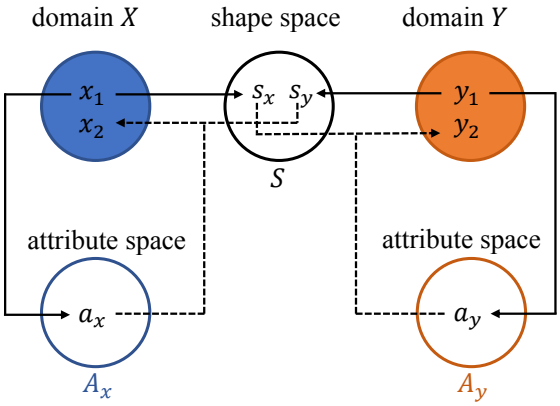

Fig. 3: Illustration of the assumption. There is a sample x_1 in domain X and a sample y_1 in domain Y. x_1 can be encoded into a shape latent code s_x and an attribute latent code a_x. Similarly, y_1 can be encoded into s_y and a_y. Then s_y and a_x can be used to generate a sample x_2 that belongs to domain X. Sample y_2 that belongs to domain Y can be generated by s_x and a_y.

image and spatially aligned with the visible image x_1. The assumption in Fig. 3 is bidirectional, but we just perform unidirectional translation based on Eq. (5). In fact, our experiment results show that using the synthetic near-infrared images only can get better registration results. Translating images from near-infrared image domain to visible image domain is more difficult than from visible image domain to near-infrared image domain in our cases, so we just perform unidirectional translation.

Since translated image y_2 and image x_1 are spatially aligned, we can align image y_1 with image y_2 using an affine transformation T_θ based on ground truth θ. The two aligned images should be the same in overlap area. The translation loss L_{trans} is defined as follows:

$$L_{trans} = ||T_\theta(y_1)' - y_2'||_1 \tag{6}$$

where $(\cdot)'$ represents the overlap area of two images, $||\cdot||_1$ represents the $L1$ norm.

C. Image Registration

As shown in Fig. 2(b), our image registration network consists of three sub-networks: feature extraction network, matching layer, and regression network. The whole image registration network can be trained in an end-to-end way.

The shape encoders are not only used for disentanglement but applied to feature extraction. As mentioned in Section II, we first use pseudo-Siamese networks to deal with the input images. The visible image shape encoder E_x^s has the same structure as the near-infrared image shape encoder E_y^s but shares no parameters with E_y^s. The two branches of such network can extract features according to the characteristics of the input, respectively. Based on our disentangled representation model, we can get high-level semantic features s_x and s_y after the shape encoders. s_x and s_y are similar because they are in the same feature space and contain no information related to modalities. Then we utilize the Siamese network to perform further feature extraction. The advantage of Siamese networks is that they can guarantee the features are always in the same feature space.

Given two feature maps output by the feature extraction network, the purpose of the matching layer is to calculate the similarity scores between the two feature maps. Given two L2-normalized feature maps $F_x, F_y \in R^{H\times W\times C}$, we can get a score map S after matching layer. The matching process is performed by global vector dot products. Compared with concatenation [23] and subtraction [24], this method is better at solving large geometric difference problems, so it is suitable for our cases. Given the score map S output by the matching layer, regression network can predict the affine transformation parameters $\hat{\theta}$. More details about matching layer and regression network can be found in [14]. After getting the affine transformation parameters $\hat{\theta}$, we follow [14] to define the grid loss as follows:

$$L_{grid} = \frac{1}{N}\sum_{i=1}^{N} d(T_\theta(g_i), T_{\hat{\theta}}(g_i)) \tag{7}$$

where N is the number of grid points, $d(\cdot)$ represents the Euclidean distance and g_i is the grid point.

Grid loss is directly calculated by the position of grid points, it does not take image itself into account. Based on the disentangled representation model, we translate image x_1 into image y_2. Since both y_1 and y_2 are near-infrared images, the similarity between y_1 and y_2 can be measured by a simple metric for mono-modal image. We define the intensity loss L_{inten} as follows:

$$L_{inten} = ||T_{\hat{\theta}}(y_1)' - y_2'||_1 \tag{8}$$

where $(\cdot)'$ represents the overlap area, $||\cdot||_1$ represents the $L1$ norm, $T_{\hat{\theta}}$ is affine transformation using predicted parameter $\hat{\theta}$. Compared with grid loss, intensity loss is not only related to $\hat{\theta}$ but related to the generated image intensity. This means that the intensity loss can simultaneously facilitate the learning of both registration network and disentangled representation model.

The total loss of our image registration framework is defined as follows:

$$L = \lambda_1 L_{recon} + \lambda_2 L_{trans} + \lambda_3 L_{grid} + \lambda_4 L_{inten} \tag{9}$$

IV. EXPERIMENTS

In this section, the proposed method is evaluated on synthetic datasets. Our proposed method is compared with several methods including the baseline method GEOCNN [14], two translation-based methods GEOCNN+MUNIT [25] and GEOCNN+DRIT++ [26]. Quantitative and qualitative experiment results demonstrate the effectiveness of our method.

A. Datasets

Since little research on multimodal visual localization of UAVs has been done, there are no available datasets for our method. VEDAI [27] is an aerial image dataset that contains 1271 pairs of registered visible and near-infrared images. All the images are captured from a vertical view at the same height. We select some pairs of images with

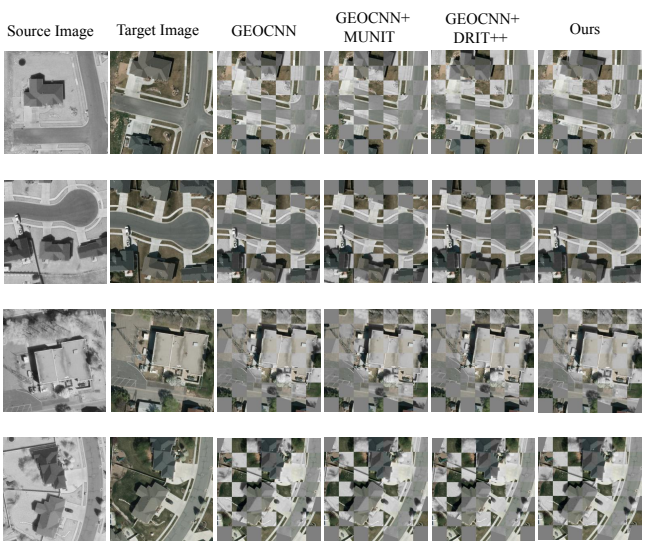

Fig. 4: Qualitative results. The checkerboard mosaic image of each sample based on four methods is shown.

distinguishable road features from VEDAI. Then we transform the visible images using a random affine transformation that includes translation and rotation. The corresponding affine transformation parameters are saved as labels. We crop the central area of visible images and near-infrared images and use them as landmark images and real-time images, respectively. Finally, we get three datasets with different complexity, which are VEDAI-1, VEDAI-2 and VEDAI-3. The overlapping ratio of the paired images is 60%$\sim$100%, 70%$\sim$100%, and 80%$\sim$100%, respectively. Each dataset contains 9600 pairs of images with a size of 256×256 pixels.

B. Implementation Details

The structure of our shape encoders, attribute encoders and generators are consistent with MUNIT [25]. The rest of the feature extraction network contains a pooling layer, a convolutional layer, an activation function ReLU [28] and a pooling layer in turn. Our regression network contains two convolutional layers and one fully connected layer. There is a batch normalization layer [29] and an activation function ReLU after each convolutional layer. The four hyperparameters in Eq. (9) are set to $\lambda_1 = 10, \lambda_2 = 1, \lambda_3 = 10, \lambda_4 = 1$, respectively. To save training costs, instead of an exact overlap area, we used a square area of 50×50 pixels in the centre of the image to calculate Eq. (6) and Eq. (8).

All the experiments are performed on an NVIDIA GeForce RTX 3090 GPU with 24-GB memory. We use the Pytorch [30] framework to implement the proposed network. The network is trained using Adam [31] optimizer with an initial learning rate of 10^{-4} for 400 epochs. The learning rate is reduced by 0.5 every 150 epochs. A total of 8000 pairs of images are used to train the model and the remaining 1600 pairs are used for testing.

C. Evaluation Criterion

In order to measure the error of localization, we use the center point of images to approximate the position of the UAVs. Then we define the RMSE, which means the Euclidean distance between the center point after ground truth transformation and the center point after predicted transformation. In addition to RMSE, we use three evaluation criteria which are normalized mutual information(NMI) [32], the average probability of correct keypoint(PCK) [33] and FPS to evaluate our method. As for PCK, we uniformly select 100 points on each image as key points, and the hyperparameter α is set to 0.05. RMSE, NMI and PCK measure the accuracy. FPS measures the running time.

D. Comparison with Baseline Approaches

In this part, our proposed method is compared with three methods using the four metrics on the three synthetic datasets. For the two translation-based methods, image translation is performed first, and then GEOCNN is applied to register images.

Qualitative results are shown in Fig. 4. We randomly select four pairs of images from test sets and predict the affine parameters by four methods. Source image is transformed with predicted parameters to align with target image. We can intuitively estimate the quality of registration results from the checkerboard mosaic images. Our method obtains more consistent mosaic images most of the time while the other three methods get unsatisfactory results. This means our proposed method gets higher registration accuracy.

Quantitative results are shown in Table I. Unlike [19] and [22], the two translation-based methods show almost no improvement in all metrics. The reason for the unsatisfactory registration results mainly focuses on the poor quality of the generated images. Complex natural scenes and high resolution of images both lead to a decrease in translation performance. As described in Section II, these two-step registration methods are limited by the quality of the generated images. Compared with these two separate translation-based methods, our proposed method focuses on the alignment on the feature level rather than the image level. Therefore, poor generated images do not have much impact on our registration process. Our method achieves 1.154, 0.393, and 0.467 in RMSE, NMI, and PCK on VEDAI-1, which is a huge improvement compared with the baseline method GEOCNN. This means that our proposed method can improve accuracy effectively. Since VEDAI-2 and VEDAI-3 are relatively simple, the performance of each method on each metric is almost improved compared with the results on VEDAI-1. Similarly, our method is still the one that has the best performance, and the two translation-based methods are still not ideal. Comparing the results on all datasets, our method has the most significant improvement on the VEDAI-1 dataset, which also shows that our method is suitable for cases of large geometric deformations.

According to the FPS, GEOCNN is the fastest method. Our proposed method is slightly slower and achieves 153.8 FPS. Since the translation-based method is two-step, these two methods are the slowest. In fact, during testing, the attribute encoders and generators are removed, and our network structure is consistent with GEOCNN. The difference

TABLE I: QUANTITATIVE RESULTS

Methods	VEDAI-1			VEDAI-2			VEDAI-3			FPS
	RMSE	NMI	PCK	RMSE	NMI	PCK	RMSE	NMI	PCK	
GEOCNN	1.543	0.360	0.321	1.055	0.373	0.510	0.593	0.384	0.784	**191.0**
GEOCNN+MUNIT	1.652	0.361	0.292	1.457	0.360	0.330	0.786	0.375	0.593	56.3
GEOCNN+DRIT++	1.764	0.355	0.246	1.302	0.366	0.340	0.806	0.374	0.632	62.5
Ours	**1.154**	**0.393**	**0.467**	**0.881**	**0.398**	**0.602**	**0.487**	**0.401**	**0.832**	153.8

TABLE II: ABLATION STUDY

disentanglement	intensity loss	VEDAI-1			VEDAI-2			VEDAI-3		
		RMSE	NMI	PCK	RMSE	NMI	PCK	RMSE	NMI	PCK
✓	✓	1.154	0.393	0.467	0.881	0.398	0.602	0.487	0.401	0.832
✓	✗	1.278	0.369	0.412	0.989	0.372	0.546	0.625	0.379	0.776
✗	✗	1.609	0.359	0.293	1.219	0.362	0.417	0.683	0.376	0.723

in FPS between GEOCNN and our method is caused by the use of different backbones for feature extraction.

E. Ablation Study

In this part, we perform experiments on the three generated datasets based on three different experimental configurations to verify the effectiveness of two components: the disentanglement component and intensity loss.

- Configuration 1 uses the whole network we proposed.
- Configuration 2 uses the network without intensity loss.
- Configuration 3 uses a network just contains feature extraction network, matching layer and regression network. It only performs image registration.

Table II shows the results of ablation study. By comparing configuration 2 with configuration 3, we find that all results of configuration 2 are better than configuration 3. This means that the disentanglement component can improve accuracy effectively. Based on the disentangled representation model, we transform images of different modalities into the same feature space to reduce differences and thus improve the image registration accuracy. By comparing configuration 1 with configuration 2, we can verify the effectiveness of intensity loss. We can find that all the metrics are improved after using the intensity loss. The RMSE is significantly improved with a 22.1% improvement on VEDAI-3. Intensity loss can facilitate the learning of both the disentangled representation model and the registration network. Compared with grid loss, intensity loss takes image intensity into account and thus improves the accuracy.

F. Visualization of Feature Maps

Though we have shown improvement of using disentangled representations, it is uncertain that the improvement is due to the reduction of differences between modalities. In this section, we design an experiment to prove that the disentangled representations does reduce the differences between modalities. We select several pairs of registered images from VEDAI dataset and crop the central area as input images. Since the input images are registered, the two feature maps output by two branches of the feature extraction network should be the same if the differences are reduced. As shown in Fig. 5, we compare the results of configurations 2 and 3. The only difference between them is using disentangled representations or not. The size of feature maps is resized to 16×16 pixels with single channel. Configuration 2 which uses disentangled representations gets more similar feature maps. This means that our proposed disentangled representations does reduce the modality differences.

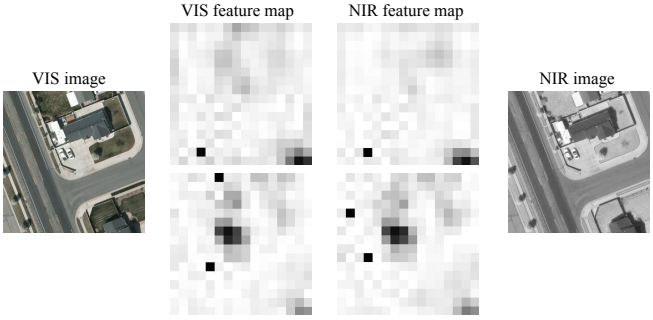

Fig. 5: Feature map visualization of a sample. The top row is the results of configuration 2. The bottom row is the results of configuration 3.

V. CONCLUSION

This paper presents a fast and accurate multimodal image registration method for UAV visual navigation system based on visible images and near-infrared images. Based on disentangled representations, we propose a new feature extraction strategy. We convert the visible images and near-infrared images into the same feature space to reduce the modality differences at feature level. Then we propose a simple intensity loss, which can make better use of image information. Compared with other translation-based methods, our proposed method is less affected by the quality of the generated images. Extensive experiments have shown that our proposed method can effectively improve localization accuracy. Then we prove the effectiveness of each component by ablation study. Compared with existing methods, our method offers higher localization accuracy with less reduction in speed.

REFERENCES

[1] J. Kwak and Y. Sung, "Autonomous uav flight control for gps-based navigation," *IEEE Access*, vol. 6, pp. 37947–37955, 2018.

[2] T. Wang, K. Celik, and A. K. Somani, "Characterization of mountain drainage patterns for gps-denied uas navigation augmentation," *Machine Vision and Applications*, vol. 27, no. 1, pp. 87–101, 2016.

[3] D. Costea and M. Leordeanu, "Aerial image geolocalization from recognition and matching of roads and intersections," *arXiv preprint arXiv:1605.08323*, 2016.

[4] W. Teng, K. Celik, and A. K. Somani, "Characterization of mountain drainage patterns for gps-denied uas navigation augmentation," in *International Conference on Pattern Recognition*, 2014.

[5] Y. Tang, Y. Hu, J. Cui, F. Liao, M. Lao, F. Lin, and R. S. Teo, "Vision-aided multi-uav autonomous flocking in gps-denied environment," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 1, pp. 616–626, 2018.

[6] T. Wang, Y. Zhao, J. Wang, A. K. Somani, and C. Sun, "Attention-based road registration for gps-denied uas navigation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 4, pp. 1788–1800, 2020.

[7] M. H. Mughal, M. J. Khokhar, and M. Shahzad, "Assisting uav localization via deep contextual image matching," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 2445–2457, 2021.

[8] M. Bianchi and T. D. Barfoot, "Uav localization using autoencoded satellite images," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1761–1768, 2021.

[9] A. Nassar, K. Amer, R. ElHakim, and M. ElHelw, "A deep cnn-based framework for enhanced aerial imagery registration with applications to uav geolocalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1513–1523, 2018.

[10] Y. Xu, J. Li, C. Du, and H. Chen, "Nbr-net: A nonrigid bidirectional registration network for multitemporal remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[11] W. Ma, J. Zhang, Y. Wu, L. Jiao, H. Zhu, and W. Zhao, "A novel two-step registration method for remote sensing images based on deep and local features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4834–4843, 2019.

[12] 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 8, pp. 3028–3042, 2019.

[13] Y. Ye, T. Tang, B. Zhu, C. Yang, B. Li, and S. Hao, "A multiscale framework with unsupervised learning for remote sensing image registration," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[14] 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.

[15] R. Fan, B. Hou, J. Liu, J. Yang, and Z. Hong, "Registration of multiresolution remote sensing images based on l2-siamese model," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 237–248, 2020.

[16] Y. Ge, Z. Xiong, and Z. Lai, "Image registration of sar and optical based on salient image sub-patches," in *Journal of Physics: Conference Series*, vol. 1961, p. 012017, IOP Publishing, 2021.

[17] L. Zhou, Y. Ye, T. Tang, K. Nan, and Y. Qin, "Robust matching for sar and optical images using multiscale convolutional gradient features," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[18] H. Zhang, L. Lei, W. Ni, T. Tang, J. Wu, D. Xiang, and G. Kuang, "Explore better network framework for high-resolution optical and sar image matching," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2021.

[19] N. Merkle, S. Auer, R. Müller, and P. Reinartz, "Exploring the potential of conditional adversarial networks for optical and sar image matching," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 6, pp. 1811–1820, 2018.

[20] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[22] L. Wang, C. Gao, Y. Zhao, T. Song, and Q. Feng, "Infrared and visible image registration using transformer adversarial network," in *2018 25th IEEE International Conference on Image Processing*, pp. 1248–1252, 2018.

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

[24] A. Kanazawa, D. W. Jacobs, and M. Chandraker, "Warpnet: Weakly supervised matching for single-view reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3253–3261, 2016.

[25] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision*, pp. 172–189, 2018.

[26] H.-Y. Lee, H.-Y. Tseng, Q. Mao, J.-B. Huang, Y.-D. Lu, M. Singh, and M.-H. Yang, "Drit++: Diverse image-to-image translation via disentangled representations," *International Journal of Computer Vision*, vol. 128, no. 10, pp. 2402–2417, 2020.

[27] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *Journal of Visual Communication and Image Representation*, vol. 34, pp. 187–203, 2016.

[28] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 315–323, 2011.

[29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, pp. 448–456, 2015.

[30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] C. Studholme, D. L. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3d medical image alignment," *Pattern Recognition*, vol. 32, no. 1, pp. 71–86, 1999.

[33] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2012.