Thermal-to-Color Image Translation for Enhancing Visual Odometry of Thermal Vision

Liyun Zhang[1], Photchara Ratsamee[2], Yuki Uranishi[3], Manabu Higashida[3], Haruo Takemura[3]

***Abstract*— A panoptic perception-based generative adversarial network for thermal-to-color image translation is proposed to demonstrate its potential as an image sequence enhancement for monocular visual odometry in blurry and low-resolution thermal vision. The pre-trained panoptic segmentation model is utilized to obtain the panoptic perception (i.e., bounding boxes, categories, and masks) of the image scene to guide the alignment between the object content codes of the original thermal domain and panoptic-level style codes sampled from the target color style space. A feature masking module further refines the style-aligned object representations for sharpening object boundaries to synthesize higher fidelity translated color image sequences. The extensive experimental evaluation shows that our method outperforms other thermal-to-color image translation methods in the image quality of translated color images. We demonstrate that the enhanced image sequences significantly improve the performance of monocular visual odometry compared with different competing methods including thermal image sequences.**

I. Introduction

Visual odometry (VO) and simultaneous localization and mapping (SLAM) are attractive areas because they are used flexibly in robotics, AR/VR and autonomous driving. Particularly, the monocular direct method VO can use the image sequence frames acquired by RGB cameras to achieve the motion trajectory estimation and semi-dense 3D reconstruction, it relieves the dependence on the extra sensors. However, tracking the RGB camera pose in poorly–illuminated conditions (e.g., night scenes) and in the presence of airborne obscurants (e.g., dust, fog and smoke) is still a challenge for current VO methods. In contrast, thermal cameras (e.g., Long Wave Infrared sensors) are considered to be one of the options that can see through the above situations.

However, because both feature-based methods [1], [2] and direct methods [3], [4] rely on image gradient-based key point extraction, which provides fewer high-quality points in thermal vision, as shown in Fig. 1 (a) middle row. Moreover, the low resolution and blurry attributions [5] of thermal images result in unsatisfactory performance in VO compared with color images, i.e., the estimated trajectory deviates significantly from Ground Truth as shown in Fig. 1 (a) bottom row. Inspired by the multi-frame GANs [6] that transformed night images to day images to improve resolution and alleviate the blurry problem for an impressive gain on VO performance. We aim to translate low-resolution and blurry thermal images to fine-grained color images to obtain a better VO performance compared with the traditional methods which directly use thermal image sequences in VO.

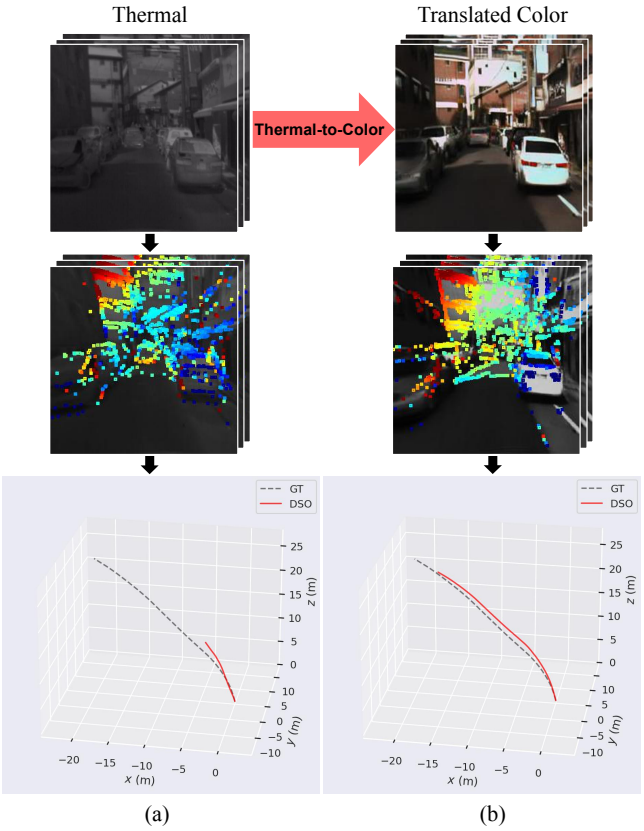

Fig. 1. A novel thermal-to-color image translation model enhances thermal vision for monocular VO. Our model translates consecutive low-resolution and blurry thermal images (a) to enhanced fine-grained color images (b) while preserving temporal and frame consistency. In the middle row, the extracted gradient-based key points by the state-of-the-art direct method VO Monocular Direct Sparse Odometry (DSO) are presented, (a) have significantly more than (b). In the bottom row, the estimated trajectory comparison is also shown. Due to the low image gradient, blurry and low resolution, typical VO such as DSO cannot achieve good tracking performance in thermal scenes. With the translated fine-grained color images from our method, the VO performance of DSO is notably improved.

The conventional image translation approaches (e.g., MUNIT [7], BicycleGAN [8], TICCGAN [9]) extract image representation of input thermal image as content codes and combine sampled style codes from color space to generate the target color image. This is difficult to generate color images with fine-grained objects, whose sharp boundaries

[1]Liyun Zhang is with the Graduate School of Information Science and Technology, Osaka University, Osaka, Japan. liyun.zhang@lab.ime.cmc.osaka-u.ac.jp

[2]Photchara Ratsamee is with the Faculty of Robotics and Design, Department of System Design, Osaka Institute of Technology, Osaka, Japan. photchara@ime.cmc.osaka-u.ac.jp

[3]Yuki Uranishi, Manabu Higashida, and Haruo Takemura are with Cyber Media Center, Osaka University, Osaka, Japan.

are crucial to extract gradient-based key points for VO. To achieve this goal, the object's information of the input image needs to be sufficiently perceived as prior knowledge for tracking the style transformation of each object in the image translation process. In this paper, we propose a learning-based sequence enhancement method for monocular VO methods. We utilize the panoptic segmentation [10] model to obtain panoptic perception as prior knowledge (i.e., bounding boxes, categories, and masks). We use Region of Interest Align (RoIAlign) [11] to extract the object representations as object content codes. The object style codes are sampled from the color style space and aligned with the content codes for fine-grained image generation. A feature masking module refines object representations for sharpening object boundaries. The fine-grained generated color images can extract more gradient-based key points to estimate more accurate trajectories (as shown in Fig. 1 (b) middle and bottom rows) for improving VO performance.

II. Related Work

A. Thermal-to-Color Image Translation

Thermal-to-color image translation is to transform images of the thermal spectrum domain to the color domain, it only changes the image style but keeps the scene content unchanged. TIR2LAB [12] incorporated a Canny edge detector for the training of the network. TICCGAN [9] added perceptual loss [13] and total variation loss [14] to optimize networks. TIC-Pan [15] utilized a pan-sharpening method to merge low frequency with high-frequency representations. Compared to Pix2Pix [16] and CycleGAN [17], Pix2PixHD [18] can generate high resolution images by adopted a multi-scale discriminator and coarse2fine generator. PGGAN [19] utilized a progressive growth strategy to synthesize images from low to high resolution. AGGAN [20] and U-GAT-IT [21] extracted attention regions from the input image as image structure guidance to localize important content for a high-quality result. However, the above methods are difficult to translate low-resolution and blurry thermal images to fine-grained color images for satisfactory VO performance.

B. Object-Level Image Translation

The object-level image translation combines object perception (bounding boxes or masks) to generate sharper object boundaries. Instagan [22] incorporated a set of instance attributes for image translation. DA-GAN [23] learned a deep attention encoder to enable the instance-aware correspondences to be discovered consequently. SCGAN [24] and SalG-GAN [25] regarded saliency maps as an object perception. Shen et al.[26], Su et al.[27] and Chen et al.[28] combined an instance-aware feature with image-aware feature for higher quality image translation. However, they only use the instance-aware objects 'thing' without considering background semantic regions 'stuff'. Dundar et al.[29] proved panoptic perception makes generated images have higher fidelity and tiny objects in more detail. Therefore, we extract panoptic perception ('thing' + 'stuff') to make sampled panoptic-level style codes combine corresponding content codes for a higher fidelity panoptic-level thermal-to-color image translation.

C. Visual Odometry (VO)

Feature-based methods [1], [2] use feature matching to estimate the camera poses via minimizing the re-projection error. Direct methods [3], [4] directly optimize the photometric error without feature descriptors. In order to improve VO performance in challenging poorly conditions, NID-SLAM [30] replace intensities with a newly designed metric considering entropies in the frame. Alismail et al.[31] proposed a direct VO method with binary feature descriptor to enhance a poor light environment. Meanwhile, a few learning-based methods have been also proposed compared to the classical methods. Gomez et al.[32] use an LSTM-based neural network to enhance images and evaluated different methods on real-world static scenes. Multi-frame GANs [6] transformed low light images to refined images for a satisfactory SLAM performance. Compared to above methods, we use a panoptic-level thermal-to-color image translation neural network to generate fine-grained color images from thermal images. We validate results on DSO for VO performance comparison of extracted key points and estimated trajectories.

III. Methodology

A. Architecture

Our architecture is built upon a generator and discriminator. We deploy a supervised learning setting via the paired thermal and color images from the KAIST-MS dataset [33].

1) Generator: As illustrated in Fig. 2, the input thermal image is extracted by a backbone module consisting of down-sampling residual blocks for obtaining image content codes C_{img} (size 32×32, dimension 256). Let $P = \left\{(category_i, bbox_i, mask_i)_{i=1}^{m}\right\}$ be a panoptic perception consisting of categories, bounding boxes, and masks, where m is the number of objects perceived from the pre-trained panoptic segmentation model and $category_i \in CAT$ (CAT defines 134 categories in the COCO-Panoptic dataset [10], here 'thing' has 80 categories and 'stuff' has 54 categories). C_{img} is cropped by RoIAlign [11] through object bounding boxes of $P(bbox_i)_{i=1}^{m}$ into object content codes $C_{obj} = \left\{C_{obj_i}\right\}_{i=1}^{m}$ (size 8×8, dimension 128). Define Z_{img} as image-level style codes (dimension 256) and $Z_{obj} = \left\{Z_{obj_i}\right\}_{i=1}^{m}$ as panoptic-level style codes (dimension 64), which are randomly sampled from normal distribution. The goal of the generator is to learn a generation function $G(\cdot)$, which is capable of translating thermal image t to a generated color image $c^{'}$ via a given (Z_{img}, Z_{obj}):

$$c^{'} = G(t|Z_{img}, Z_{obj}; \Theta_G) \tag{1}$$

where Θ_G are the parameters of the generation function.

We use a multilayer perceptron (MLP) network to process the panoptic-level style codes $Z_{obj} = \left\{Z_{obj_i}\right\}_{i=1}^{m}$ to dynamically generate the parameters $y = (y_\gamma, y_\beta)$ of Adaptive Instance Normalization (AdaIN) [34] layers, then C_{obj} are

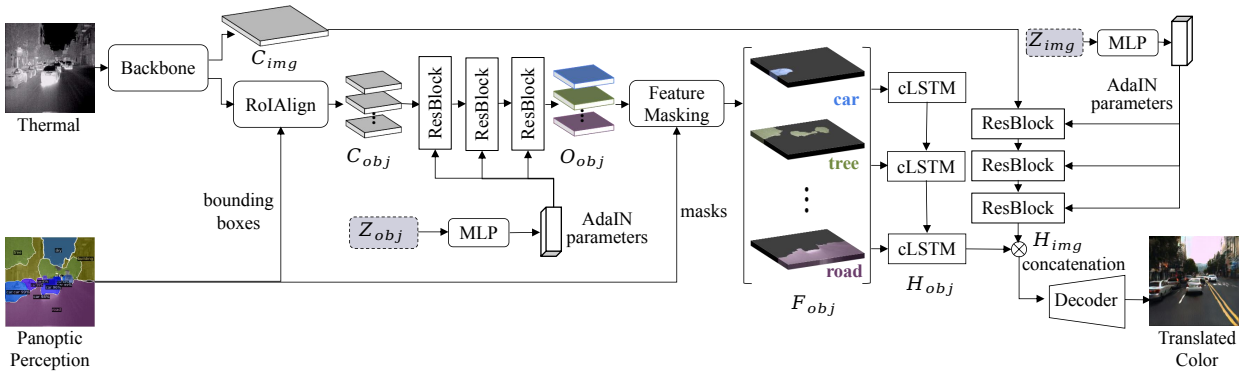

Fig. 2. Overview of our generator. The thermal image is extracted by the backbone consisting of down-sampling residual blocks to image-level representations, which are cropped by RoIAlign via bounding boxes from the panoptic perception of the pre-trained model to panoptic-level representations. Both representations are aligned with sampled color style codes from normal distribution. The redundant background information of panoptic-level style-aligned codes is removed by feature masking, the results are re-integrated by cLSTM to combine with image-level representations for color image generation.

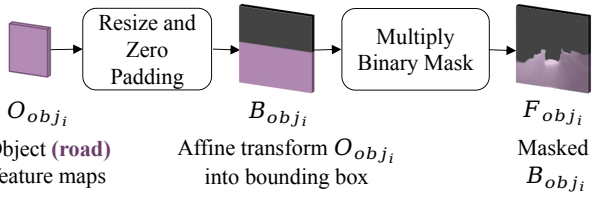

Fig. 3. The feature masking module removes the redundant background information outside the object contours from the object feature maps.

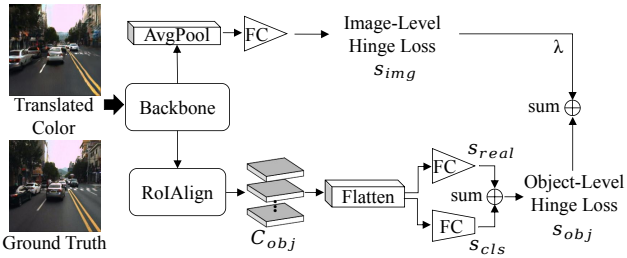

Fig. 4. Overview of our discriminator. The backbone and RoIAlign extract translated or ground truth image to image-level and panoptic-level representations, which are pooled and flattened respectively, and then processed by different fully connected (FC) layers to a fusion hinge loss consisting of image-level realness score s_{img} and object-level realness scores s_{real} with category projection scores s_{cls}.

processed by the residual blocks with AdaIN layers. The parameters of AdaIN layers fuse panoptic-level style with content to translate the different objects in the target image.

$$AdaIN(x_i, y) = y_{\gamma,i}\left(\frac{x_i - \mu(x_i)}{\sigma(x_i)}\right) + y_{\beta,i} \tag{2}$$

where x_i is each feature map of C_{obj}, which is normalized separately and then scaled and biased using the corresponding scalar components from style y. The μ and σ are channel-wise mean and standard deviation, γ and β are AdaIN parameters generated from Z_{obj}. We obtained the style-aligned object representation $O_{obj} = \{O_{obj_i}\}_{i=1}^{m}$.

$$O_{obj} = AdaIN(C_{obj}, Z_{obj}) \tag{3}$$

Image-level style codes Z_{img} are also processed by MLP to generate AdaIN parameters, which fuse the image-level style with image content codes C_{img} by residual blocks with AdaIN layers to obtain a hidden representation H_{img}.

As illustrated in Fig. 3, in our feature masking module, O_{obj} contains m object feature maps $\{O_{obj_i}\}_{i=1}^{m}$. Since the object bounding boxes $P(bbox_i)_{i=1}^{m}$ define the size and location of each object in the original image, we firstly affine transform each object feature maps O_{obj_i} into its corresponding original bounding box, secondly we do zero padding outside each bounding box in the image to obtain new object feature maps $B_{obj} = \{B_{obj_i}\}_{i=1}^{m}$. To remove the redundant background information outside the object contour, we further refine B_{obj} via object masks $M = P(mask_i)_{i=1}^{m}$ for more precise object boundaries. Compared with the Convolutional Feature Masking (CFM) layer [35] using the pixel projection method, after affine transformation the size of each feature map in B_{obj} is the same as masks M, therefore we only need to align B_{obj} and M along the category sequence of $1 \sim m$ and multiply to mask the values outside of object contour. Finally, we can obtain finer object feature maps $F_{obj} = \{F_{obj_i}\}_{i=1}^{m}$.

$$F_{obj} = B_{obj} \cdot M \tag{4}$$

We feed F_{obj} into three layers convolutional Long-Short-Term Memory (cLSTM) module, the number of channels in each layer is 256, 128, 128, respectively. The cLSTM is a multi-layer convolutional LSTM network, it can better preserve spatial information. It integrates each object feature maps $\{F_{obj_i}\}_{i=1}^{m}$ along object sequence of $1 \sim m$ to obtain fused hidden representation H_{obj}. We concatenate H_{obj} with H_{img} as H, which is up-sampled by decoder consisting of up-sampling residual blocks to synthesize color image c'.

2) Discriminator: As illustrated in Fig. 4, our discriminator consists of image-level and object-level classifiers. Similar to generator, the translated image is encoded by the backbone as image content codes C_{img}, which is refined by RoIAlign [11] as object content codes $C_{obj} = \{C_{obj_i}\}_{i=1}^{m}$ via bounding boxes $P(bbox_i)_{i=1}^{m}$. The image-level classifier consists of a global average pooling and one-output fully

connected (FC) layer to process C_{img} to obtain a scalar realness score s_{img}. The object-level classifier consists of a flatten layer and two FC layers. One FC layer processes C_{obj} to compute a realness score for each object, denoted by $s_{real} = \{s_{real_i}\}_{i=1}^{m}$. Another FC layer computes a category projection score [36], [37], [38] for each object, denoted by $s_{cls} = \{s_{cls_i}\}_{i=1}^{m}$, which is the inner product between category embedding (transforming each category of $P(category_i)_{i=1}^{m}$ to a corresponding latent vector sampled from normal distribution) and linear projection (using a FC layer) of down-sampled C_{obj}. Therefore, the overall object-level loss of an object is $s_{obj_i} = s_{real_i} + s_{cls_i}$. The discriminator will be denoted by $D(\cdot, \Theta_D)$ with parameters Θ_D.

$$(s_{img}, s_{obj_i}, \cdots, s_{obj_m}) = D(I; \Theta_D) \tag{5}$$

Given an image I (ground truth c or generated $c^{'}$), the discriminator computes the prediction score for the image and the average scores for objects.

B. Loss Function

The full objective comprises three loss functions:

Firstly, we utilize image-level and object-level fusion hinge version [39] of standard adversarial loss [40] to train (Θ_G, Θ_D),

$$l_k(I) = \begin{cases} \min(0, -1 + s_k); & \text{if } I \text{ is ground truth } c \\ \min(0, -1 - s_k); & \text{if } I \text{ is generated } c^{'} \end{cases} \tag{6}$$

where $k \in \{img, obj_i, \cdots, obj_m\}$. The overall loss is $l(I) = \lambda \cdot l_{img}(I) + \frac{1}{m}\sum_{i=1}^{m} l_{obj_i}(I)$ with trade-off parameter λ (1.0 used in experiment) in fusion hinge losses between image-level and object-level. We define the losses for the discriminator and generator respectively [38],

$$\begin{aligned} L_{\text{adv}}(\Theta_D, \Theta_G) &= -\mathop{\mathbb{E}}_{(I)\ p_{all}(I)}[l(I)] \\ L_{\text{adv}}(\Theta_G, \Theta_D) &= -\mathop{\mathbb{E}}_{(I)\ p_{fake}(I)}[D(I; \Theta_D)] \end{aligned} \tag{7}$$

where minimizing $L_{\text{adv}}(\Theta_D, \Theta_G)$ tries to tell the discriminator to distinguish ground truth and translated images, but minimizing $L_{\text{adv}}(\Theta_G, \Theta_D)$ tries to fool discriminator by translating fine-grained images. $p_{all}(I)$ represents all of ground truth and translated images, and $p_{fake}(I)$ represents translated images.

Secondly, we use $L_1^{img} = \left\| c' - c \right\|_1$ to penalize the L_1 difference between the translated image $c^{'}$ and ground truth c, $\|_1$ calculates the L1 norm.

Thirdly, we use L_{p} to alleviate the problem that translated images are prone to producing distorted textures [18]. It is beneficial to keep textures in high-level space through the ReLU activation of the VGG-19 network [13],

$$L_{\text{p}} = \Sigma_k \frac{1}{C_k H_k W_k} \sum_{i=1}^{H_k} \sum_{j=1}^{W_k} \left\| \phi_k(c^{'})_{i,j} - \phi_k(c)_{i,j} \right\|_1 \tag{8}$$

where $\phi_k(\cdot)$ represents feature representations of the kth max-pooling layer in VGG-19 network, and $C_k H_k W_k$ represents the size of feature representations.

Therefore, the final loss function is defined as:

$$L_{\text{total}} = \lambda_1 L_{\text{adv}} + \lambda_2 L_1^{img} + \lambda_3 L_{\text{p}} \tag{9}$$

where λ_i are the parameters balancing different losses.

C. Implementation Details

As for L_{total}, we empirically set $\lambda_1 \sim \lambda_3$ to 0.1, 1, 10, respectively. Model parameters were initialized using the Orthogonal Initialization method [41]. The spectral normalization [42] is used in the layers of both generator and discriminator to stabilize the GANs training. We used leaky-ReLU with a slope of 0.2 in the activation function. We trained our model using the Adam optimizer [43] with $\beta_1 = 0$ and $\beta_2 = 0.9$. The learning rates were set to 10^{-4} for the generator and 0.005 for the discriminator. We set 400,000 iterations for training on four NVIDIA V100 GPUs.

IV. Experiments and Results

A. Dataset and Training

We trained and evaluated the thermal-to-color image translation model using the KAIST-MS dataset [33], which contains 11,610 thermal-color image pairs for training and 2,541 thermal images for evaluation. For panoptic perception in the training, we perceive it from the paired color images via pre-trained Panoptic FPN model [44] on COCO-Panoptic dataset [10]; in the inference, it is perceived from input images via pre-trained Panoptic FPN model on a compact our contributed dataset of thermal panoptic segmentation, the source data[1] are the pairs of thermal and color images from partial KAIST-MS [33] dataset.

B. Evaluation Setting

We compared results of different methods in aspects of image quality and VO performance. For baselines, MUNIT [7], BicycleGAN [8] and TICCGAN [9] belong to image-level image translation; SCGAN [24] and INIT [26] belong to object-level image translation. For an adequately fair comparison, we add panoptic perception to image-level baselines, where panoptic perception (as an additional features channel) is concatenated with image features for training, hence we call them MUNIT+Seg, BicycleGAN+Seg and TICCGAN+Seg. For metrics, we chose Inception Score (IS) [45], Fréchet Inception Distance (FID) [46] and Diversity Score (DS) instead of Peak signal-to-noise ratio (PSNR) and structure similarity index (SSIM) [47] to evaluate image quality because traditional PSNR and SSIM deviate from human visual perception [48] for images generated by GANs learning models. We adopted Absolute Pose Error (APE) to evaluate VO performance, it consists of the Median Absolute Translational Error t_{rel} and Absolute Rotational Error r_{rel} as proposed in the KITTI Odometry benchmark [49]. We run the DSO [4] method and calculate APE scores to evaluate the trajectory similarity between the thermal image sequences and translated color image sequences from different approaches. We summarized the results in qualitative and quantitative aspects respectively.

[1] https://segments.ai/panoptic/visible/

Fig. 5. Comparison of image quality for translated images from different approaches including corresponding original thermal and ground truth images.

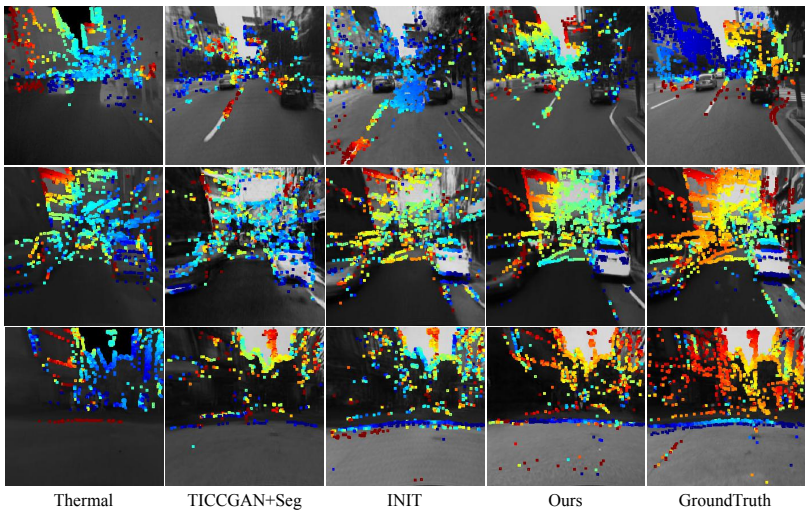

Fig. 6. The extracted key points with depth maps from translated images in VO for different approaches (This figure is best visualized in color format).

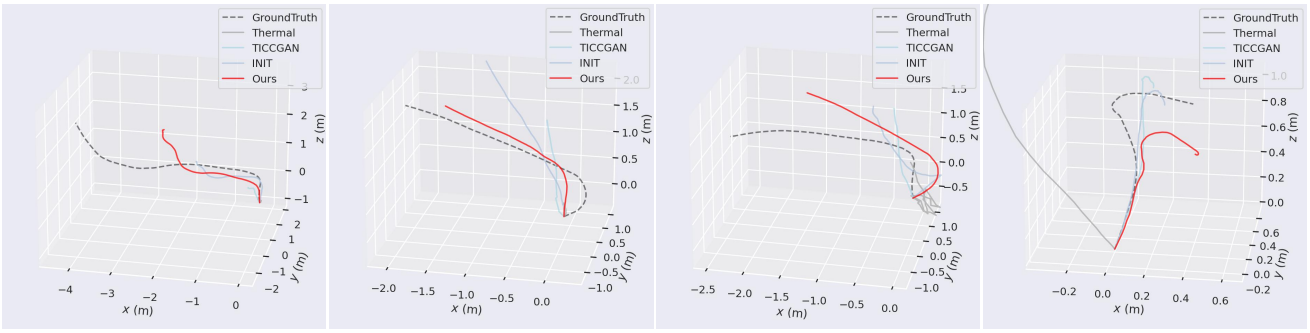

Fig. 7. The VO performance of translated sub-sequences for different approaches, estimated trajectories of different approaches are compared.

C. *Qualitative results*

For image quality, Fig. 5 presents the image translation results from different approaches. Referring to the ground truth, our model can translate more realistic, fine-grained and colorful images; the results of other methods are more blurry, distorted, and the colors are not bright. Also, our translated objects have better sharpness, a more natural color style, and display diversity (e.g., the appearance of cars); translated objects of other methods have not satisfactory sharpness, the style is far from the ground truth, and insufficient diversity. For comprehensive scene similarity with ground truth, our

results are also relatively more robust compared to other methods. For VO performance, we use two methods (i.e., TICCGAN+Seg [9] and INIT [26]) with the highest performance to compare with our results. We run the DSO method on both original thermal image sequences and translated color image sequences through different approaches. As shown in Fig. 6, the extracted key point comparison for translated color images from different methods including original thermal and ground truth shows that our results are more than baselines, they can be robustly obtained and more consistent with ground truth. Fig. 7 shows that our method can estimate more accurate trajectories compared to the results of other methods. Note that we measure the absolute trajectory error because the raw KAIST-MS dataset [33] was published by some discontinuous video clips and test sequences are relatively short, i.e., less than 50m. As shown in different sub-graphs, our results consistently show the estimated trajectories that match the expected orientation and pose on each sub-sequence compared with other methods.

TABLE I

DIFFERENT BASELINES AND OUR APPROACH ARE COMPARED ON IMAGE QUALITY (EVALUATED BY IS, FID AND DS METRICS). HIGHER IS, HIGHER DS, AND LOWER FID IS BETTER.

method	IS ↑	FID ↓	DS ↑
MUNIT+Seg [7]	2.29 ± 0.18	98.56 ± 0.41	0.46
BicycleGAN+Seg [8]	2.61 ± 0.30	98.83 ± 0.58	0.47
SCGAN [24]	2.59 ± 0.14	96.41 ± 0.14	0.39
TICCGAN [9]	2.67 ± 0.50	90.15 ± 0.67	0.48
INIT [26]	2.70 ± 0.62	83.27 ± 0.34	0.37
Ours	$\mathbf{2.85 \pm 0.38}$	$\mathbf{72.74 \pm 0.21}$	**0.54**

TABLE II

ABSOLUTE TRANSLATIONAL ERROR (T_{REL}) AND ROTATIONAL ERROR (R_{REL}) METRICS (LOWER IS BETTER) ARE TO EVALUATE THE VO PERFORMANCE OF TRANSLATED SEQUENCES FOR DIFFERENT METHODS.

	Thermal		TICCGAN		INIT		Ours	
seq.	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
01	2.30	2.00	1.24	1.57	1.05	1.56	**0.89**	**1.16**
02	2.45	1.81	1.75	2.79	1.59	2.86	**1.28**	**1.49**
03	2.37	3.14	1.31	1.29	1.20	1.38	**1.03**	**1.10**
04	3.30	2.56	2.44	2.23	2.18	2.46	**2.05**	**1.71**

D. Quantitative results

For image quality, the IS, FID and DS in Table I show that our approach achieved superiority in image quality of translated images compared to other approaches. Our method overall outperforms baselines since we avoid losing too much information in the translation. The higher IS and lower FID of our method verified that our translated color images have higher fidelity and sharpened object information. The higher DS demonstrated that our method can show more flexibility and high robustness when the scene is invariant, e.g., generated objects. For VO performance, Table II (TICCGAN [9] with +Seg, just not shown) validated that our method can estimate more accurate trajectories compared to other methods. We measure APE (absolute translational error t_{rel} + rotational error r_{rel}) between translated color (including original thermal) image sequences and ground truth image sequences, and our results are significantly lower than others, this is consistent with the estimated trajectory comparison results in Fig. 7. Especially for image sequences in discontinuous environments, our sub-set error remains consistently lowest, which demonstrates the robustness of our method to environmental scene adaptation. In contrast, the results of other methods are not very stable.

TABLE III

THE PERFORMANCE OF MODELS REMOVING DIFFERENT LOSSES AND MODULES ARE COMPARED IN IMAGE QUALITY (EVALUATED BY IS, FID AND DS) AND VO (EVALUATED BY T_{REL} AND R_{REL}).

method	IS ↑	FID ↓	DS ↑	t_{rel} ↓	r_{rel} ↓
w/o L_{obj}	2.24 ± 0.17	110.4 ± 0.46	0.47	2.03	2.10
w/o L_1^{img}	2.64 ± 0.42	104.3 ± 0.13	0.45	1.43	1.63
w/o L_p	2.66 ± 0.16	97.1 ± 0.27	0.42	1.50	1.60
w/o M_{msk}	2.53 ± 0.73	101.6 ± 0.65	0.47	1.41	1.67
w/o M_{clstm}	2.63 ± 0.59	103.2 ± 0.70	0.42	1.81	1.78
full model	$\mathbf{2.85 \pm 0.38}$	$\mathbf{72.74 \pm 0.21}$	**0.54**	**1.31**	**1.37**

E. Ablation Study

We demonstrate the necessity of losses and modules (L_{obj}: object-level hinge loss; L_1^{img}: image reconstruction loss; L_p: perceptual loss; M_{msk}: feature masking; M_{clstm}: cLSTM) by comparing IS [45], FID [46] and DS [50] for image quality and APE for VO performance. As shown in Table III, Removing L_{obj} and L_1^{img} have lower IS, DS, and higher FID due to generating low fidelity image and objects with fewer variations and L_{obj} compute the category projection scores for objects. The t_{rel} and r_{rel} values rose because the DSO method is affected by gradient points through the boundaries of objects. Removing L_p, the model produces distorted textures to inevitably decrease image quality and VO performance. Removing any M_{msk} or M_{clstm} module decreased the overall performance, which demonstrates their necessities. Because M_{msk} sharpens object boundaries and M_{clstm} sequentially integrates different objects back into image.

V. CONCLUSION

In this paper, we presented a novel thermal-to-color image translation method that outperforms baselines, verified that the image quality of its translated color image is significantly better than baselines and the translated image sequences can intuitively enhance VO performance compared to baselines or thermal image sequences. Furthermore, It can also enhance object recognition performance and is expected to be used in night vision for autonomous driving and robotic systems.

ACKNOWLEDGMENT

This work has been partly supported by the KAKENHI Fund for the Promotion of Joint International Research (fostering joint international research (B) No. 20KK0086) and the Mohamed Bin Zayed International Robotics Challenge (MBZIRC) Grant.

REFERENCES

[1] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[2] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *2007 6th IEEE and ACM international symposium on mixed and augmented reality*. IEEE, 2007, pp. 225–234.

[3] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[4] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[5] M. A. Marnissi, H. Fradi, A. Sahbani, and N. E. B. Amara, "Thermal image enhancement using generative adversarial network for pedestrian detection," in *25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 6509–6516.

[6] E. Jung, N. Yang, and D. Cremers, "Multi-frame gan: image enhancement for stereo visual odometry in low light," in *Conference on Robot Learning*. PMLR, 2020, pp. 651–660.

[7] 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 (ECCV)*, 2018, pp. 172–189.

[8] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Multimodal image-to-image translation by enforcing bi-cycle consistency," in *Advances in neural information processing systems*, 2017, pp. 465–476.

[9] X. Kuang, J. Zhu, X. Sui, Y. Liu, C. Liu, Q. Chen, and G. Gu, "Thermal infrared colorization via conditional generative adversarial network," *Infrared Physics & Technology*, vol. 107, p. 103338, 2020.

[10] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9404–9413.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

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

[13] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[14] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[15] F. Almasri and O. Debeir, "Robust perceptual night vision in thermal colorization," *arXiv preprint arXiv:2003.02204*, 2020.

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

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

[18] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

[19] 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.

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

[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] S. Mo, M. Cho, and J. Shin, "Instagan: Instance-aware image-to-image translation," *arXiv preprint arXiv:1812.10889*, 2018.

[23] S. Ma, J. Fu, C. W. Chen, and T. Mei, "Da-gan: Instance-level image translation by deep attention generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5657–5666.

[24] Y. Zhao, L.-M. Po, K.-W. Cheung, W.-Y. Yu, and Y. A. U. Rehman, "Scgan: Saliency map-guided colorization with generative adversarial network," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[25] L. Jiang, M. Xu, X. Wang, and L. Sigal, "Saliency-guided image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 509–16 518.

[26] Z. Shen, M. Huang, J. Shi, X. Xue, and T. S. Huang, "Towards instance-level image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3683–3692.

[27] J.-W. Su, H.-K. Chu, and J.-B. Huang, "Instance-aware image colorization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7968–7977.

[28] T. Chen, W. Xiong, H. Zheng, and J. Luo, "Image sentiment transfer," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4407–4415.

[29] A. Dundar, K. Sapra, G. Liu, A. Tao, and B. Catanzaro, "Panoptic-based image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8070–8079.

[30] G. Pascoe, W. Maddern, M. Tanner, P. Piniés, and P. Newman, "Nid-slam: Robust monocular slam using normalised information distance," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1435–1444.

[31] H. Alismail, M. Kaess, B. Browning, and S. Lucey, "Direct visual odometry in low light using binary descriptors," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 444–451, 2016.

[32] R. Gomez-Ojeda, Z. Zhang, J. Gonzalez-Jimenez, and D. Scaramuzza, "Learning-based image enhancement for visual odometry in challenging hdr environments," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 805–811.

[33] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1037–1045.

[34] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[35] J. Dai, K. He, and J. Sun, "Convolutional feature masking for joint object and stuff segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3992–4000.

[36] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[37] T. Miyato and M. Koyama, "cgans with projection discriminator," *arXiv preprint arXiv:1802.05637*, 2018.

[38] W. Sun and T. Wu, "Image synthesis from reconfigurable layout and style," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 531–10 540.

[39] J. H. Lim and J. C. Ye, "Geometric gan," *arXiv preprint arXiv:1705.02894*, 2017.

[40] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[41] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2013.

[42] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.

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

[44] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6399–6408.

[45] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, pp. 2234–2242, 2016.

[46] S. Ravuri and O. Vinyals, "Classification accuracy score for condi-

tional generative models," *Advances in neural information processing systems*, vol. 32, 2019.

[47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[48] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[49] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[50] B. Zhao, L. Meng, W. Yin, and L. Sigal, "Image generation from layout," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8584–8593.