2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

Multispectral Domain Invariant Image for Retrieval-based Place Recognition

Daechan Han*[1], YuJin Hwang*[1], Namil Kim[2] and Yukyung Choi**[1]

***Abstract*— Multispectral recognition has attracted increasing attention from the research community due to its potential competence for many applications from day to night. However, due to the domain shift between RGB and thermal image, it has still many challenges to apply and to use RGB domain-based tasks. To reduce the domain gap, we propose multispectral domain invariant framework, which leverages the unpaired image translation method to generate a semantic and strongly discriminative invariant image by enforcing novel constraints in the objective function. We demonstrate the efficacy of the proposed method on mainly multispectral place recognition task and achieve significant improvement compared to previous works. Furthermore, we test on multispectral semantic segmentation and unsupervised domain adaptations to prove the scalability and generality of the proposed method. We will open our source code and dataset.**

I. Introduction

The RGB imaging sensor has been a key component of in the field of computer vision and robotics society over the past few decades, with diversified potential applications such as place recognition, object recognition, and autonomous driving. The majority of existing these methods mainly focus on color information, so that they usually fail to operate under illumination changing conditions. To overcome these limitations, Long-wave infrared (thermal) imaging sensors are used as solutions in many applications. Since thermal imaging sensors generate the image from the radiation of objects, the thermal image can capture clear objects and structural silhouettes under harsh conditions which can adversely affect the quality of the RGB image. Therefore, this advantage naturally triggers the combination of RGB and thermal camera (multispectral) to be complementary with each other.

Recently, multispectral tasks have received a lot of scholar attention [1], [2], [3], [4], [5], [6] and many multispectral datasets have introduced for various computer vision and robotics applications. Despite of these attentions, multispectral setting has bottlenecks to get through. Firstly, most popular benchmarks are based on RGB images without paired thermal images. This phenomenon indicates that the new paired dataset is needed to apply previous methods in the multispectral setting. Also, since training DNN models begin with curating data and its associated label, various large-scale datasets [7], [8], [9] have been introduced with object-level or pixel-wise annotations. Unfortunately, there is little labeled multispectral dataset. Intuitively, a basic solution is collecting a large-scale dataset and annotating all of them. However, the larger the numbers of collected data, the more expensive the expenditure to process them.

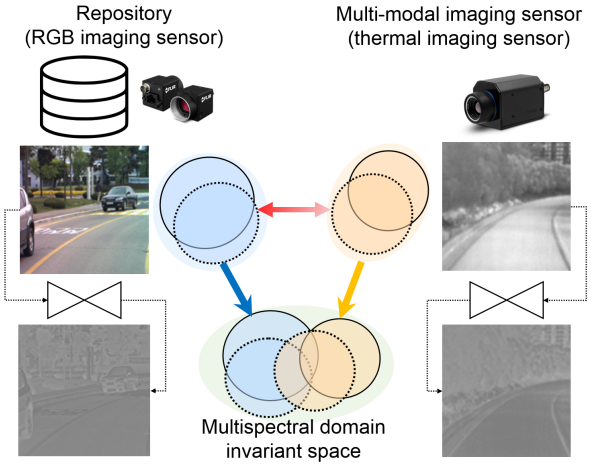

Fig. 1. We illustrate the multispectral domain invariant process. We depict a red arrow to domain shift between RGB source domains and thermal target domain. To apply the proposed multispectral invariant model, we can adapt both distinct domains into the shared latent space. Blue and yellow arrows denote the result of the proposed method in each modality. In the shared space, we can reuse a bunch of RGB image-based resources and adapt multispectral setting into various RGB domain-based applications.

To take full advantage of previous annotated dataset, multispectral image-to-image translation [3], [10], [11], [12] has become an active research area in recent times. Unlike general RGB-to-RGB image translation tasks, multispectral image-to-image translation has a fundamental issue. Despite of common features such as edges, silhouettes in RGB and thermal images, both images have a big differences due to each imaging principle [13], [14]. For example, the radiated heat from the car is only appeared in the thermal image, not in the RGB image. Therefore, we address this physical difference as the main challenge to reduce the domain discrepancy between RGB and thermal images.

In this work, we mainly cover the place recognition problem [15], [16], [17]. Place recognition that is related with the localization is a canonical and an important problem in computer vision and robotics. In autonomous driving, it uses keeping track of ego vehicle position according to the previous surrounding environment, and it can aid the localization based on inertial sensor. This problem is highly suitable for domain shift scenarios, because map data (Google map) and pre-scanned image are mostly based on RGB imaging

*Equal contribution, +Corresponding author
[1]Daechan Han, Yoojin Hwang and Yukyung Choi are with School of Intelligent Mechatronic Engineering, Sejong University, South Korea {dchan,yjhwang,ykchoi}@rcv.sejong.ac.kr
[2]Namil Kim is with NAVER LABS, South Korea namil.kim@naverlabs.com

sensors. Therefore, an alternative sensor (thermal) is not fully exploiting these data, despite of the complementary advantage of this sensor.

To reduce the domain shift, we focus on extracting multispectral domain invariant image, including common information between RGB and thermal domain. To convert multispectral images to domain invariant images, we can reuse a lot of RGB images with annotations and we can apply newly captured image from alternative sensors to the RGB image-based algorithm in various applications. To do that, instead of using labeled pairs, we exploit advantages of a recent unpaired image-to-image translation. This method is conveniently unsupervised and practical in the real world scenario, rather than requiring a set of fully aligned multispectral images capturing the same place. Based on this method, we propose a novel and compatible framework to generate the domain invariant image corresponding to common semantic contents from multispectral domains. Motivated by the physical difference of both domains, we instead use the intermediate feature from the proposed domain invariant encoder as the domain invariant image rather than directly translating RGB images to thermal images. Moreover, we optimize the model to generate diverse and discriminative invariant image as the feature itself for place recognition task by the proposed constraints in the objective function.

We summarize our contributions as follows: 1) We propose a generalized and compatible framework with unpaired image-to-image translation to extract the domain invariant image between multispectral domains. 2) We test on multispectral place recognition benchmark and achieve better results compared to baseline methods. 3) We extend our method to multispectral semantic segmentation and RGB image-based unsupervised domain adaptation to demonstrate the generality and validity of the proposed framework.

II. Related work

A. Multispectral Visual Recognition

Multispectral system has been proposed for robust perception in all-day conditions. With the illumination invariant thermal sensor, multispectral system received a lot of attention in various applications such as object detection/segmentation/tracking [14], [4], [19], [20], [21], [22], [23], [24], depth estimation [1], [25], image enhancement [2] person re-identification [5], [26], and visual localization [27], [6]. Moreover, some large-scale datasets with fully or partial aligned multispectral pairs were introduced in all-day conditions [13], [14], [25]. Most of these works are focus on boosting the accuracy with complementary thermal images, which is intertwined with the multispectral fusion problem. In this work, we attempt to handle a new perspective of multispectral visual recognition on how to reuse existing RGB images with newly provided thermal images. To do that, we propose the multispectral domain invariant generation framework and apply to the multispectral place recognition as RGB images for reference and thermal images for query.

B. Image-to-Image translation

Image-to-Image translation problem is defined as transforming from domain **A** to **B** in another domain. Generative Adversarial Networks (GANs) [28] has shown exceptional performances on this problem using paired [29], [30] and unpaired [18], [31], [32] scenarios. This method have been widely used in the long-term place recognition task under severe appearance changes [33], [34], [35]. Most works leveraged well-conditioned images (*e.g., day*) on one domain to achieve strong performance on another domain with poor-conditioned images (*e.g., night, rain, snow*) with novelties of *n*-domain translation architectures [34], an objective function with image-aware discriminator [33], and encoder network [35]. These works were successfully adapting RGB source domains to various RGB target domains. However, due to different principles in RGB and thermal sensors, directly using translated image is not plausible for applying multispectral place recognition problem [3], [10], [11], [12]. Motivated by the study of [14], [13], we instead generate multispectral domain invariant image containing common information of both domains from the intermediate layer of the proposed model, and we use this output as the input of various applications rather than using the translated image.

C. Place recognition

Place recognition is a canonical problem in computer vision and robotics field, and aims to identify an ego-vehicle position from pre-capturing images. One way is obtaining positions through image retrieval technique that finds the most similar image of pre-captured repository to the current query image. A traditional method widely used is representing every image to global descriptors by using BOW [36] or VLAD [15] with hand-crafted local feature descriptors. The process is firstly extracting local features [37],and describe a global descriptor via the residual from local feature to the estimated clustering *k* centers. Recently, NetVLAD [16] was proposed to reformulate the VLAD process into DNN framework. The main difference is replacing hand-crafted features with DNN features and learning cluster centers in the end-to-end manner. Since the current and past images has different viewing points, weather and illumination conditions, the invariance is critical to estimating the accurate position. In this paper, we address another invariant problem for place recognition that is related to reduce the domain gap between the past captured image (RGB) and the current captured image (thermal).

III. Proposed Method

We first define the multispectral domain invariant image problem and relevant notations to our work. We use data from two domains: the source domain **A** (RGB) and the target domain **B** ($thermal$). Source and target domain contained images in the same trajectory, but both domains were captured at different times and different camera poses. As shown in Fig. 2, we set the base model as CycleGAN [18]. One reason is that we assume that RGB images on the source domain is hardly captured in exactly same pose of thermal

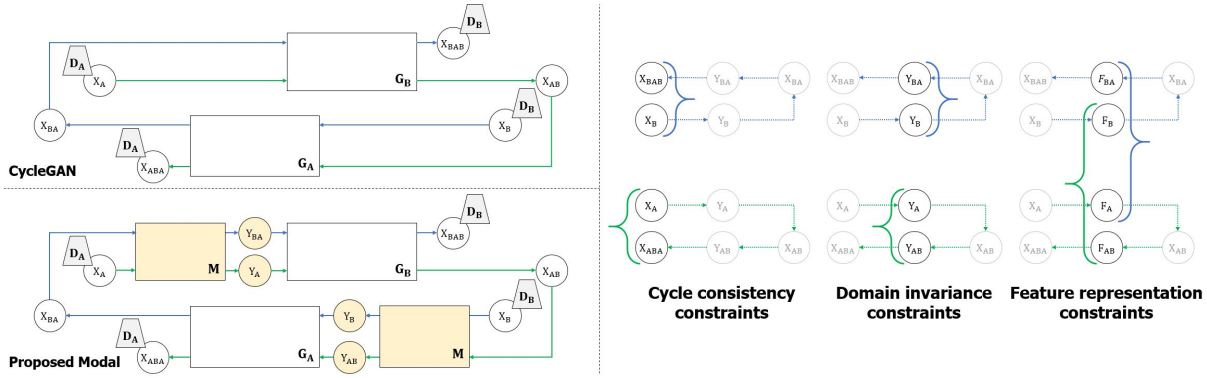

Fig. 2. Architecture of our model and topology of constrains. X is image and Y is invariant image. (Left) The proposed framework is based on CycleGAN [18] as unpaired image translation model. We use the same architecture of Generator ($\mathbb{G}$) and discriminator ($\mathbb{D}$) in CycleGAN. Our multispectral domain invariant model ($\mathbb{M}$) can be easily applied into CycleGAN-style framework and is automatically trained via the almost same training process of base model. (Right) During training, we can encourage the model to generate semantic and discriminative invariant image through proposed constraints (except for cyclic consistency).

images on the target domain in the real world. Another reason is that many unpaired image translation model are mainly motivated by key concepts to utilize unpaired training data so that successful adaptation to CycleGAN can be compatible for other unpaired image translation methods. We denote X as the image, Y as the domain invariant image, F as the intermediate feature map of the proposed *multispectral domain invariant (MDI) model*, and notation $_{AB}$ means that A domain converts to B domain and vice versa. Note that we only formulate domain **A** into following paragraphs for convenience of description.

A. Multispectral Domain invariant representation

1) System architecture: Our framework is based on CycleGAN model, consisting of generators ($\mathbb{G}_\mathbb{A}$,$\mathbb{G}_\mathbb{B}$) and discriminators ($\mathbb{D}_\mathbb{A}$,$\mathbb{D}_\mathbb{B}$) with *Multispectral Domain Invariant model* ($\mathbb{M}$). The proposed invariant model is the core part of the total framework. All generators and discriminators work for each domains, while the proposed invariant model work for both domains to extract common features from both domains. Given unpaired RGB and thermal images, the invariant model outputs corresponding invariant images as shown in Fig. 3.

The proposed invariant model consists of residual blocks [38] similar to the generator architecture of CycleGAN. There are two distinct differences. One is that we do not resize the feature map in the model to prevent the artifacts such as loss of subtle information via down-sampling layer and the checkerboard pattern artifacts of up-sampling layer. Another difference is that we design *self-abstraction* module before the output layer. To avoid redundant details such as patterns and textures from RGB images on the invariant image, we normalize the feature map by own mean and variance like whitening and then truncate the negative variable by ReLU.

$$F_{norm}(A) = \max\left\{\frac{Y_A - \mu_A}{\sqrt{\sigma_A^2 + \epsilon}}, 0\right\} \tag{1}$$

This process (Eq. (1)) can effectively regularize the feature map to contain only common information without unnecessary details.

2) Consistency constraints: To satisfy the semantic domain consistency, we train the model by two types of consistency losses. Firstly, to guarantee that the model can recover each domain image from the domain invariant image, we employ the cyclic consistency constraints [18]. The cyclic loss (Eq. (2)) encourage that the invariant image contains semantic information that can be useful to recover RGB and thermal image respectively. We exploit the same cyclic loss as the original paper for both domain A and B.

$$L_{cycle}(A) = \|\mathbb{G}_A(\mathbb{M}(\mathbb{G}_B(\mathbb{M}(X_A)))) - X_A\|_1 \tag{2}$$

Another constraint is that outputs from RGB and thermal image should have same invariant information. Therefore, we exploit the pixel-wise loss (Eq. (3)) that minimizes the distinction of invariant images from both domains. Through this constraint, the invariant image can keep only semantic features that simultaneously occur in RGB and thermal images among all features that need to reconstruct the original image.

$$L_{domain}(A) = \|\mathbb{M}(\mathbb{G}_B(\mathbb{M}(X_A))) - \mathbb{M}(X_A)\|_1 \tag{3}$$

3) Feature representation constraints: Consistency alone encourages invariant image to bias generating multispectral images. In order to obtain good coverage for place recognition task, we require task-specific constraints. We define two types of constraints as discriminability and diversity. For discriminability constraint, we use structural similarity index (SSIM) [39] that has been widely employed as a metric to evaluate image-processing algorithms. Compared to the pixel-wise metric, SSIM evaluates images accounting for changes in local structure. In place recognition, the global descriptor is usually made by a set of local features, so that the similarity of locality is important for the better feature representation. We set a single scale SSIM loss (Eq. (4)) with a 3×3 block filter instead a Gaussian.

$$L_{disc}(A) = \frac{1}{N}\sum_{(i,j)} \frac{(1 - SSIM(Y_A, Y_{AB}))}{2}. \tag{4}$$

Fig. 3. Examples on domain invaraint images from RGB and thermal respectively. Left to right: RGB image, thermal image, RGB's invariant image, thermal's invariant image. All examples are randomly selected.

where $Y_A = \mathbb{M}(X_A)$, $Y_{AB} = \mathbb{M}(\mathbb{G}_B(\mathbb{M}(X_A)))$ are denoted.

Mode collapse is one challenging issue of GAN-based methods, which the generator collapses to produce limited varieties of outputs. If the manifold of domain invariant models is collapsed, the resulted image is not distinguishable for place recognition to find the most similar pre-scanned image. Especially, many on-road scenarios of autonomous driving have similar appearances such as lane, road, central vanishing points, or surrounding vehicle, so that the diversity is critical to estimating the accurate position. To handle this issue, we utilize triplet loss (Eq. (5)) where anchor input as F_A is compared to a positive F_{AB} and a negative F_B and vice versa. Note that we instead use the intermediate feature of the invariant model, because of preventing to conflict the consistency and diversity constraints.

$$L_{divs}(A) = \max\{d(F_A, F_{AB}) - d(F_A, F_B) + \alpha, 0\} \quad (5)$$

where $\alpha = 1$, $F_A = \mathbb{M}_L(X_A), Y_A = \mathbb{M}_U(\mathbb{M}_L(X_A))$ are denoted, and $\mathbb{M}_L$ and $\mathbb{M}_U$ are separated in exactly halfway between the input and output layer.

By making invariant images discriminative and diverse, our method indirectly encourages them to become highly specialized. This happens automatically, without enforcing such geometric properties explicitly. This intuition is strongly supported by our experiments.

B. Optimization and Inference

A full objective is:

$$\begin{aligned} L(A) = &\{L_{adv}(A) + L_{cycle}(A)\} \\ &+ \{\lambda_0 L_{domain}(A) + \lambda_1 L_{disc}(A) + \lambda_2 L_{divs}(A)\} \end{aligned} \quad (6)$$

where the first bracket is the objective function of CycleGAN and L_{adv} denotes adversarial loss. The second bracket is the proposed objective function to generate multispectral domain invariant images and λ_0, λ_1 and λ_2 are set to 1, 0.1 and 0.1 respectively. Inheriting from CycleGAN, optimizing the objective function results in solving a mini-max problem. All lambdas are decided experimentally.

$$\mathbb{M}^* = \min_{\mathbb{M},\mathbb{G}} \max_{\mathbb{D}} L(A).$$

After training the model, we solely use $\mathbb{M}$ to generate domain invariant images as the input of place recognition, semantic segmentation and other applications.

IV. Experimental Results

In this section, we evaluate the proposed method in various experimental settings. To demonstrate the effectiveness of our model, we first conduct the experiment in multispectral place recognition task. The scenario is that RGB images were captured along the trajectory, and a new thermal sensor currently captures an image in the same route. However, there is no RGB and thermal pair having the same viewing point and camera poses. Moreover, to validate the generality of the proposed method, we conduct two other experiments: multispectral semantic segmentation and RGB image-based unsupervised domain adaptation benchmark. In each experiment, we select one domain as the source domain and another domain as the target domain. We denote "*source only*" as the performance of target domain of a model trained by the source domain, and "*target only*" as that of a model trained by the target domain. These two metrics serve as baselines for the lower and upper bound performance in this experiment. We do not tune a set of hyper-parameters, so we use the same values of the place recognition experiment.

A. Place Recognition

1) Dataset: We use the KAIST dataset [14] for training and evaluation of place recognition. It provides well-aligned RGB and thermal pairs with a high accuracy PS/IMU data in 6 sequences recorded at different times of a day, including day, night, sunset, and sunrise. RGB images captured at night time are not properly used to the translation model due to the bad visibility, so that we selected two sequences (5 AM, 9 AM) captured on the west route of the campus. We expect that since the thermal image is robust for illumination changes, this setting could cover not only day(RGB)-to-day(thermal) conditions and also day(RGB)-to-night(thermal) to some extent. With synchronizing to GPS timestamps, we sampled 1332 RGB images of 9 AM for reference images, and 3061 thermal images of 5 AM for querying images.

2) Experimental setup: All images are resized to 128×128 for training the model. Similar to [18], we set a single image as mini-batch and data augmentations including random resizing and cropping during training. Training is run for 200 epochs with constant learning rate (2e-4) in the first half of total epoch and decreasing linearly to zero during

TABLE I

RESULTS ON MULTISPECTRAL PLACE RECOGNITION. WE COMPARE OUR DOMAIN INVARIANT IMAGE WITH ORIGINAL AND TRANSFORMED PAIR.

Methods	1.00m	2.00m	3.00m	5.00m	10.0m	1.00m	2.00m	3.00m	5.00m	10.0m
Source: RGB / Query: Thermal										
	Top-1 accuracy					Top-5 accuracy				
DSIFT+BOW [17]	3.7	5.4	7.9	10.8	16.3	11.6	14.4	17.7	21.8	29.6
DenseVLAD [15]	4.3	6.1	8.1	10.1	12.6	12.9	15.5	18.1	21.2	26.2
NetVLAD [16]	1.2	1.9	2.8	4.2	7.1	6.6	10.9	17.0	25.8	44.9
Pix2pixHD [29]	3.8	5.5	8.0	10.9	14.6	13.5	16.4	19.9	23.4	28.6
Excavate [35]	5.4	7.7	10.1	12.3	15.5	15.4	18.7	22.1	24.8	27.9
CycleGAN [18]	1.2	2.1	2.8	2.8	3.5	4.1	5.0	6.2	7.7	9.8
Ours	12.2	18.3	24.7	30.6	37.4	36.3	42.5	48.4	53.0	57.6
Source:Thermal / Query:RGB										
Pix2pixHD [29]	1.0	1.7	2.0	2.5	3.2	4.9	6.4	7.4	8.6	10.4
Excavate [35]	10.8	15.4	20.6	26.3	32.3	28.6	34.1	39.3	43.1	47.1
CycleGAN [18]	1.4	2.0	2.7	3.1	3.7	5.4	6.6	7.8	8.6	10.0
Ours	25.5	36.9	47.9	56.1	65.8	60.9	66.9	72.9	76.7	81.1

TABLE II

RESULTS ON MULTISPECTRAL SEMANTIC SEGMENTATION, USING MFNET AND A MODIFIED SEGNET AS THE BASE NETWORK.

Methods	unlabeld	car	pedestrian	bike	curve	car stop	guardrail	color cone	bump	class avg.	mIoU
MFNet [23]											
Source Only	93.2	7.2	8.5	0.	0.	0.	0.	0.	0.	12.1	5.4
CycleGAN [18]	95.6	11.0	17.1	33.4	0.	0.	0.	0.	0.	17.4	6.1
Ours	93.5	24.8	13.7	3.7	26.9	0.	0.	0.	52.0	23.9	21.9
Target Only	95.7	67.6	78.0	58.7	64.4	10.0	0.	0.	36.4	45.7	39.4
SegNet [40]											
Source Only	92.9	7.8	24.6	0.	0.	3.2	0.	0.	0.	14.3	10.2
CycleGAN [18]	93.2	28.6	44.6	0.	0.	6.6	0.	0.	0.	19.2	10.7
Ours	93.4	43.6	29.8	0.	11.0	0.	0.	0.	42.2	24.4	26.3
Target Only	96.4	72.5	81.3	56.9	58.6	0.	0.	0.	3.6	41.0	36.2

the second half. After training the model, we extract the domain invariant image of every RGB and thermal image at 9 AM and 5 AM and then apply these images to the DenseVLAD [15] framework to describe a global descriptor. We use the number of cluster centers (K) as 1024 and the default SIFT setting in VLFeat library [41].

3) Evaluations: We measure both top-1 and top-5 retrieval accuracy under different distance thresholds (1.0 ,2.0, 3.0, 5.0, 10.0 meters). If any of these top-N ranked images are within the certain distance threshold of the query image, we counted it as a successful localization. For fair comparison, we select various baseline methods such as Dense SIFT+BoW [36], DenseVLAD [15], NetVLAD [16], CycleGAN [18], Pix2pixHD [29], and Excavate [35]. We use the same parameters of DenseSIFT framework across all baseline models. DenseSIFT + BoW, DenseVLAD, NetVLAD are trained by source domain images at 9 AM and apply them to query domain images at 5 AM for place recognition. Other translation models are trained by using the same dataset pairs of the proposed method. After training them, CycleGAN and Pix2pixHD test on original RGB images at 9 AM and the transformed RGB images given thermal images at 5 AM and vice versa. Since Excavate is based on extracting intrinsic features given input images, we apply it to the same protocol of the proposed method for place recognition.

As shown in Table I, our method outperforms all of the baselines by a large margin in every distance threshold. Compared to image translation baseline models, we achieve a significant improvement. This is because the transformed image fails to recover visual details of RGB images due to different properties of RGB and thermal sensor. This argument is strongly supported by Pix2pixHD since this method is based on supervised learning that usually generates a better image than unpaired image translation methods. Compared to Excavate, we prove that multispectral domain is more challenging and the proposed feature representation constraints are effective to extract discriminative and diverse domain invariant image. In the reverse setting (Source:Thermal, Query:RGB), our method shows the best performance among other methods. One interesting point is that the averaged accuracy of the reverse setting is higher than the normal setting. From this perspective, we found that the translation from the higher contextual domain to the lower contextual domain is more feasible to generate the image, and the invariant images from RGB images seem to have more discriminative power to distinguish most similar images in the same set.

B. Semantic Segmentation

To demonstrate our methods applicability to real-world adaptation settings, we evaluate a domain invariant image in multispectral semantic segmentation task. We use the multispectral semantic segmentation dataset [23] which consists of

Fig. 4. Examples of invariant images in MNIST↔USPS. Left to right: MNIST, USPS, MNIST's invariant image, USPS's invariant image

Source Target	USPS MNIST	MNIST USPS
Source only	96.3	76.9
SE [42]	97.3	98.1
DCRN [17]	73.7	91.8
G2A [43]	90.8	92.5
ADDA [44]	90.1	89.4
SBADA-GAN [45]	**97.6**	95.4
CyCADA [46]	95.7	94.8
Ours	94.9	**98.7**
Target only	97.8	99.6

TABLE III

RESULTS FOR UNSUPERVISED DOMAIN ADAPTATION ON DIGIT CLASSIFICATION.

1569 pixel-wise labeled RGB and Thermal images with eight classes obstacles commonly encountered during driving. For the source domain, we use the RGB image and use the paired thermal image as the target domain.

Similar to place recognition experiments, we first train the domain invariant encoder by using a training sample of the dataset. Then we extract the invariant feature image, and then use them to train the semantic segmentation model. All configurations are evaluated with the class-wise pixel accuracy and the mean Intersection-over-Union (mIoU) metric. For the validity, we follow the procedure of [23] and test on two types of segmentation models as MFNet [23] and a modified version of SegNet [40] respectively. We report our results in Table II, alongside results of the source only and target only models.

Our method clearly improves upon both metrics mIoU of the source only model, and also some improvements competing target only model. Even with the same training procedure and settings as in the place recognition experiments, our method is extremely effective at adapting the most common classes in the dataset, and show the better mean accuracy than image translation baseline method.

C. Domain Adaptation

To further evaluate the generality of the proposed model, we perform experiments on a small-scale unsupervised domain adaptation. We use MNIST↔USPS on digits recognition. Note that we choose these datasets because our domain invariant image is a single-channel image as same as these datasets. For fair comparison against recent state-of-the-art methods ([42], [17], [43], [44], [45], [46]), we conduct experiments on the same network architectures as in SE [42] for this task. Same as the semantic segmentation task, we train to extract the domain invariant feature image by using both datasets, and then we use the extracted image to train the digit classification model.

Constraints				Accuracy
Cyclic consistency	Domain consistency	Diversity	Discriminability	Top-1 in $3.0m$
✓				2.8
✓	✓			16.7
✓	✓	✓		23.6
✓	✓	✓	✓	24.7

TABLE IV

ABLATION STUDY FOR CONSTRAINTS.

In MNIST to USPS direction, we achieve accuracy close to the performance of fully supervised setting on the target domain[1] (in Table III). Moreover, we achieve a substantial margin of improvement over the most recent works. However, in the opposite direction, we achieve an accuracy lower than the performance on the source domain as similar to most recent works, we think that the number of USPS dataset is relatively small, allowing us to achieve improved performance by adapting from MNIST. For qualitative analysis, Fig. 4 visualizes both input images and the generated images of MNIST and USPS respectively. We can see that our model successfully highlights the digit and attenuates the background to extract the common semantic region from both images.

V. DISCUSSIONS & CONCLUSIONS

In this paper, we have introduced multispectral domain invariant framework with the proposed architecture and some novel constraints to make the output image contain more semantic and discriminative feature. Our results show that the proposed method significantly outperforms previous works on various real-world applications. In Table IV, the combination of proposed constraints is constantly boosting the accuracy in the place recognition task. Since our framework is based on CycleGAN, it is highly compatible wit other unpaired image translation methods. Since the quality of the invariant image depends on that of a translated image to some extent, we expect that the better-qualified model will have room for improvement.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2018R1C1B5086415).

REFERENCES

[1] N. Kim, Y. Choi, S. Hwang, and K. In So, "Multispectral transfer network: Unsupervised depth estimation for all-day vision," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[2] Y. Choi, N. Kim, S. Hwang, and I. S. Kweon, "Thermal image enhancement using convolutional neural network," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[1]The results for the previous works have been partially copied from the original papers, and we experimented the released code in other works. We used the same architecture for source/target model and the same experimental setup as SE [42].

[3] V. V. Kniaz, V. A. Knyaz, J. Hladuvka, W. G. Kropatsch, and V. Mizginov, "Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset," in *European Conference on Computer Vision (ECCVW)*, 2018.

[4] C. Li, D. Song, R. Ton, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," in *British Machine Vision Conference (BMVC)*, 2018.

[5] M. Ye, X. Lan, J. Li, and P. C. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[6] T. Mouats, N. Aouf, D. A. Sappa, C. Aguilera, and R. Toledo, "Multispectral stereo odometry," *IEEE Transactions on Intelligent Transportation Systems (TITS)*, vol. 16, pp. 1210–1224, 2015.

[7] D. Jia, D. Wei, S. Richard, L. Li-Jia, L. Kai, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[8] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," in *The International Journal of Robotics Research (IJRR)*, 2013.

[9] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Doll'ar, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV)*, 2014.

[10] C. Devaguptapu, N. Akolekar, M. M. Sharma, and V. N. Balasubramanian, "Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2019.

[11] A. Nyberg, A. Eldesokey, D. Bergstrom, and D. Gustafsson, "Unpaired thermal to visible spectrum transfer using adversarial training," in *European Conference on Computer Vision (ECCVW)*, 2018.

[12] A. Berg, J. Ahlberg, and M. Felsberg, "Generating visible spectrum images from thermal infrared," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2018.

[13] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon, "Kaist multi-spectral day/night data set for autonomous and assisted driving," *IEEE Transactions on Intelligent Transportation Systems (TITS)*, vol. 19, pp. 934–948, 2018.

[14] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baselines," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[15] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[16] R. Arandjelovi, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *European Conference on Computer Vision (ECCV)*, 2016.

[18] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *International Conference on Computer Vision (ICCV)*, 2017.

[19] H. Choi, S. Kim, K. Park, and K. Sohn, "Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks," in *IEEE International Conference on Pattern Recognition (ICPR)*, 2016.

[20] J. Liu, S. Zhang, S. Wang, and D. Metaxas, "Multispectral deep neural networks for pedestrian detection," in *British Machine Vision Conference (BMVC)*, 2016.

[21] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster r-cnn for robust multispectral pedestrian detection," in *CoRR*, 2018.

[22] D. Guan, Y. Cao, J. Liang, Y. Cao, and M. Y. Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," in *CoRR*, 2018.

[23] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[24] L. Chenglong, Z. Chengli, H. Yan, T. Jin, and W. Liang, "Cross-modal ranking with soft-consistency and noisy labels for robust rgb-t tracking," in *European Conference on Computer Vision (ECCV)*, 2018.

[25] T. Wayne, S. Philip, S. Scott, K. Abhishek, O. Michael, P. Brian, S. Kelly, and K. Chandra, "Cats: A color and thermal stereo benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[26] D. T. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," vol. 17, pp. 605–634, 2017.

[27] Y. Choi, N. Kim, K. Park, S. Hwang, J. S. Yoon, and I. S. Kweon, "All-day visual place recognition: Benchmark dataset and baseline," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2015.

[28] G. Ian J., P.-A. Jean, M. Mehdi, X. Bing, W.-F. David, O. Sherjil, C. Aaron, and B. Yoshua, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[29] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[30] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[31] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain imageto-image translation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[32] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[33] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. V. Gool, "Night-to-day image translation for retrieval-based localization," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[34] A. Asha, A. Eirikur, T. Radu, and G. Luc Van, "Combogan: Unrestrained scalability for image domain translation," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2018.

[35] X. Jian, W. Chunheng, S. Cunzhao, and X. Baihua, "Excavate condition-invariant space by intrinsic encoder," in *CoRR*, 2019.

[36] S. Lazebnik, C. Schmid, and J. Ponce, "beyond bags of features spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[37] A. Relja and Z. Andres, "Three things everyone should know to improve object retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[38] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[39] W. Zhou, B. Alan Conrad, S. Hamid Rahim, and S. Eero P., "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing (TIP)*, vol. 88, pp. 303–338, 2004.

[40] S. Lazebnik, C. Schmid, and J. Ponce, "beyond bags of features spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[41] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[42] G. French, M. Mackiewicz, and M. Fisher, "Self-ensembling for visual domain adaptation," in *International Conference on Learning Representations (ICLR)*, 2018.

[43] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, "Generate to adapt: Aligning domains using generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[44] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[45] P. Russo, F. M. Carlucci, T. Tommasi, and B. Caputo, "From source to target and back: symmetric bi-directional adaptive gan," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[46] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *The International Conference on Machine Learning (ICML)*, 2018.