2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) | 978-1-6654-3396-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/AVSS52988.2021.9663750

Bridging the Invisible and Visible World: Translation between RGB and IR Images through Contour Cycle GAN

Yawen Lu Guoyu Lu
Intelligent Vision and Sensing Lab
Rochester Institute of Technology, USA
{yl4280,luguoyu}@cis.rit.edu

Abstract

Infrared Radiation (IR) images that capture the emitted IR signals from surrounding environment have been widely applied to pedestrian detection and video surveillance. However, there are not many textures that appeared on thermal images as compared to RGB images, which brings enormous challenges and difficulties in various tasks. Visible images cannot capture scenes in the dark and night environment due to the lack of light. In this paper, we propose a Contour GAN-based framework to learn the cross-domain representation and also map IR images with visible images. In contrast to existing structures of image translation that focus on spectral consistency, our framework also introduces strong spatial constraints, with further spectral enhancement by illuminance contrast and consistency constraints. Designating our method for IR and RGB image translation, it can generate high-quality translated images. Extensive experiments on near IR (NIR) and far IR (thermal) datasets demonstrate superior performance for quantitative and visual results.

1. Introduction

Generative Adversarial Networks (GANs) [8] have been applied to tackle the problem of image-to-image translation in recent years [22, 16, 12, 13, 10, 26], such as summer-to-winter, photo-to-painting, day-to-night, and label-to-photo. Among all the image translation problems, visible and IR image translation has significant application scenarios. On one hand, once we can translate the thermal image to the visible image, nearly all the computer vision techniques that are not directly applicable to IR images can be applied, which is critical to invisible (e.g. nighttime and hidden objects) tasks of surveillance and tracking. On the other hand, once visible images are available, we can apply the visible images to estimate the temperature of the surrounding environment based on the translation.

However, due to the large discrepancy in spectral distribution and appearance variation, existing methods still cannot translate IR images and visible color images well. As it is known, infrared radiation images are sensitive to the temperature from surrounding objects and more robust to illumination changes. The textures and contexts in gray-scale IR images are much fewer compared with visible images from color sensors. In addition, visible images could be totally invalid in nighttime scenarios because the apertures of RGB cameras will stay magnified for a longer time to capture sufficient light. As a result, the captured images involve significant image blur. With the aforementioned issues, when applying the current GAN framework directly, the translated IR images are more likely gray-scale images and cannot reflect the spectral characteristic in real infrared images. Their illuminance and contrast also suffer huge gaps. Besides, the translated color images always suffer a considerable blur and lack fine details. As a result of the unsatisfactory results, we propose an image-to-image translation network designed to specifically transfer images between IR and visible modalities, as demoed in Fig. 1.

Figure 1. Left: Input thermal (first row) and color visible (second row) images. Right: Our cross-domain image translation result to color (first row) and thermal (second row) domains. Our CCGAN network is able to recover more details and keep scene boundaries clear with the proposed constraints.

In this paper, we propose a novel framework to address the cross-spectral image translation between infrared and color images with a conditional generative adversarial network, namely Contour Cycle GAN (CCGAN) as shown in

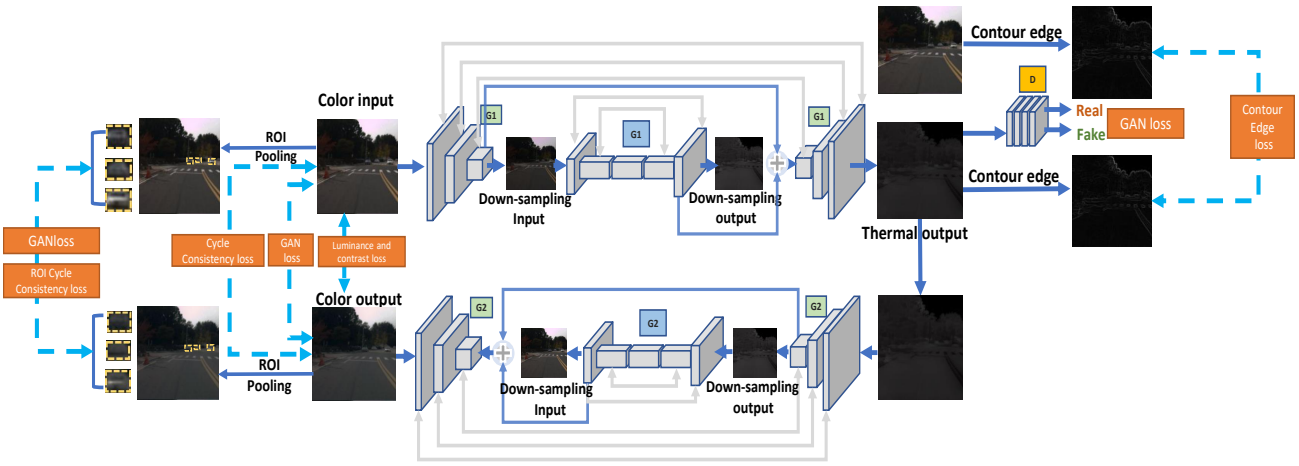


Figure 2. Overview of the proposed network between IR and visible modalities. The green G box and blue G box separately represent the entire image's generator and the down-sampled image generator. G1 and G2 are two separate generators for two directions.

Fig. 2. The CCGAN framework is applied to multi-scale generators and discriminators, which enforces the contour consistency in both holistic and local levels. Through this process, we can iteratively improve the translation effect by enforcing the cycle consistency. As IR images are lack of textures compared with RGB visible images, we explore the translation framework based on their consistency on the contours, which should appear on both modalities of images. Therefore, the layout distribution of the two modalities' images can be enforced to be consistent. Previous GAN-based structures mainly rely on spectral (intensity or color) consistency to guide the learning process. In contrast, our CCGAN explicitly introduces strict spatial consistency loss. Instead of constraining exactly corresponding contour pixels, we constrain based on a local window region, which can accommodate the variance of contour detection performance across different modalities.

To further extract finer details from both visible and IR images, we apply discriminators to differentiate the real image and fake image regions and apply ROI consistency for the translated back image to compare the ROI region pixel difference. An illuminance contrast loss is added to learn the correct mapping representation and further enhance the spectral similarity between different modalities. This is particularly helpful to enhance the spectral similarity due to the significant difference between color and thermal modalities.

To summarize, the contributions are as follows: 1) We introduce the specific problem of IR-Visible image translation and propose a newly designed learning framework targeted at this problem. 2) A Contour GAN framework with loss constraints on holistic and local regions to enforce the contour consistency aiming at IR-Visible translation. Unlike previous GAN structures relying on spectral contrast to lead the learning process, our GAN structure introduces spatial consistency constraints. 3) We explore the translation of IR and visible images in both directions, from visible image to IR image and from IR image to visible image, which can be applied in both sufficient and insufficient lighting (night) conditions. 4) The proposed method is evaluated both qualitatively and quantitatively on public datasets for far IR and near IR images with convincing results, which shows its potential ability in real-life applications. The translated color images from thermal images can be successfully applied to 3D reconstruction and other tasks based on thermal images. To the best of the author's knowledge, this is the first work to introduce spatial constraints in GAN-based translation tasks for cross-spectral images.

2. Related work

Image-to-image translation aims to learn a mapping representation to transform the input images to the target images in different domains. Recent success in this filed benefits from the development of GAN frameworks. Pix2pix [12] network used a "U-Net" architecture [23] for the generator and a convolutional "PatchGAN" classifier as the discriminator to make sure of the high-level similarity of the translated results. The advantage of using PatchGAN is that it has fewer parameters while still being applicable to large images without sacrificing the quality of the output. Pix2PixHD [27] further extended [12] to generate high-resolution photo-realistic images from semantic labels. Unlike methods mentioned above, dual learning was introduced to GAN to train the model alternatively on both sides, allowing translators to be trained from monolingual data only. It was applied by He et al. [9] to enable the translation network to learn from unpaired data by iteratively updating the two models at the same time until convergence. Following the similar structure, CycleGAN [29], DualGAN [28], DiscoGAN [13], SingleGAN [19], Drit++ [15] and RevGAN [26] were proposed to tackle the image translation problem by learning the mapping between different visual

domains jointly, each of them as a separate generative adversarial network by deploying cycle-consistency loss function as well as an adversarial loss function across diverse domains. However, these networks targeting for general translation purpose are not able to show satisfactory performance for the specific IR-visible domain translation.

IR images capture the reflected IR signals to generate images, which can further be used in abnormal behavior detection [25] and object tracking [7] [17]. Among different wavelength IR images, long-wave IR images, also called thermal image, indicate the temperature of the object's surface which have been applied to detect humans or animals [20, 1, 5], and search and rescue tasks [6], where light is usually not sufficiently supplied. Similar work was conducted by Fernandez et al. [5] to detect people in real-time on an autonomous mobile platform, and Cielniak et al. [3] to apply both visible and thermal cameras to track multiple people. Leveraging both visible images and thermal images through transfer learning [18], localization tasks have been conducted in the dark environment. Thermal sensors have been utilized to detect non-heat generating objects for robot navigation [4]. Spectral-spatial features [21] and shape features [24] are also extracted to classify thermal images. However, there has not been much work targeting the translation between the IR and visible images in the past. The proposed approach generates higher quality samples that are more stable than previous methods and can be applied to multiple real applications.

3. IR-Visible Image Translation

3.1. Cross-domain Cycle Consistency

Inspired by the cycle structure in [29], we further apply a multi-scale cycle structure as the basic learning framework. Under the basic cycle scheme, in order to capture both holistic and local information in both visible and thermal modalities, our network (illustrated in Fig. 2) consists of multi-scale generators (green G and blue G) and a discriminator (D) to enforce the model to learn more detailed information. Each local generator includes a convolutional front-end, three residual blocks and a deconvolution back-end, and each local discriminator is composed of four convolutional layers with increasing channels. To feed the shared content vector and domain-specific styles as input to learn a two-round representation mapping between Color-to-IR and IR-to-Color as a structural constraint for the entire network, we use the cross-domain consistency loss to force the G1(G2(IR)) to be close to the original IR image and G2(G1(Color)) to be close to the original color image. The consistency between the original images and the translated images can be expressed as:

$$\begin{aligned} L_{cycle} = E_{x\in ori(x)}[\|G2(G1(x)) - x\|_1] + \\ E_{y\in ori(y)}[\|G1(G2(y)) - y\|_1] \end{aligned} \tag{1}$$

where L_{cycle} represents the cross-domain cycle consistency loss. E is the loss expectation of all the training samples. x and y separately represent color and IR modality. $ori(x)$ and $ori(y)$ are the original color and IR image datasets. $G1(x)$ translates the input color to IR image and $G2(G1(x))$ further translates the IR to color modality, which is compared with the original color image in the L1 distance (same process for IR). L_{cycle} enforces the generated image in two directions to be as close to the original image as possible to guide the structure holistically.

3.2. Domain Adversarial Constraint

To learn transformation relationships between images in the source and target domains, we apply the adversarial loss to combine generator and discriminator pairs at the same time. The generators are to learn to transform IR image to synthetic color images $f_{IR-Color}$ and from color to synthetic IR image $f_{Color-IR}$. The synthetic images are then differentiated and evaluated with real images by discriminators. As a result, the global adversarial loss is able to simultaneously minimize the distribution difference of both generated data and real data. The total bi-direction adversarial loss and their separate loss are defined as follows:

$$\begin{aligned} &L_{GAN} = L_{GAN,IR\to C} + L_{GAN,C\to IR} \\ &L_{GAN,IR\to C} = E_{x\in C_{ori(x)}}[1 - logD_C(x)] \\ &+ E_{y\in IR_{ori(y)}}[logD_C(G(y))] \\ &L_{GAN,C\to IR} = E_{y\in IR_{ori(y)}}[1 - logD_{IR}(y)] \\ &+ E_{x\in C_{ori(x)}}[logD_{IR}(G(x))] \end{aligned} \tag{2}$$

L_{GAN} is the total adversarial loss, composed by color image generation adversarial loss $L_{GAN,IR\to C}$ and IR image generation adversarial loss $L_{GAN,C\to IR}$.

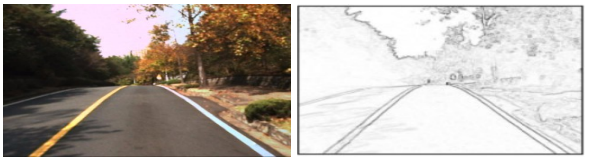
3.3. ROI Cycle Consistency

However, there are specific regions (ROI) that may affect more than others in the entire image translation effect. We apply existing salience detection to extract ROIs based on image intensity, and an ROI pooling layer to crop and resize bounding regions to the same size. ROI loss is then introduced to recover the local region more precisely. Global cycle consistency loss focus on the entire translated images without attention to finer details and textures [27]. However, specific regions are critical in the image modality translation process. To recover the local region more precisely, a loss term based on Region of Interest (ROI) is proposed here to further improve the generated image quality. The ROI cycle consistency loss between the ROI regions from the reconstructed images and the original input images can be formulated as follows:

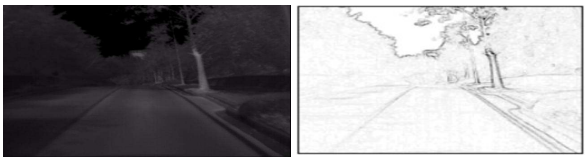
$$\begin{aligned} L^{ROI}_{cycle} = E_{x_{ROI}\in ori(x)}[\|G2(G1(x_{ROI})) - x_{ROI}\|_1] \\ + E_{y_{ROI}\in ori(y)}[\|G1(G2(y_{ROI})) - y_{ROI}\|_1] \end{aligned} \tag{3}$$

4. Contour Consistency

As color and IR images reflect signals in different wavelengths (RGB: 380-700 nanometers, thermal: 8-14 micrometers), the image intensity between these two modalities cannot correspond directly. To further build consistency between color and IR images, we apply the spatial contour information. The reason that we do not apply edge is mainly because color images usually contain much more textures, whose edges are not existing on the thermal images. With the contours to contain the network, shapes of objects inside the images can be retained, leading to a more precise translation effect. The detected contour result based on edge detection is shown as in Fig. 3.



(a) Color Image (b) Contour of color image.



(c) Thermal image (d) Contour of thermal image.

Figure 3. Color and thermal images and extracted contours.

Object contours are extracted using region-based active contour detection algorithm [14]. Once the contours are extracted, we measure the contour distance between the two images, as Fig. 4. Taking thermal to color image translation as an example, for a contour point on the thermal image, if there is also a contour point within the local window of the corresponding pixel, the distance between the two points is 0; otherwise, the contour distance for this pixel pair is 1. As the contours of thermal and color images may have small shifts, we use a window region to constrain the correspondences instead of using the exact corresponding pixels from both images. The overall contour loss is the sum of all the pixel loss. Each pixel's loss value is described in Eq. 4. The contour loss for the entire image is as Eq. 5.

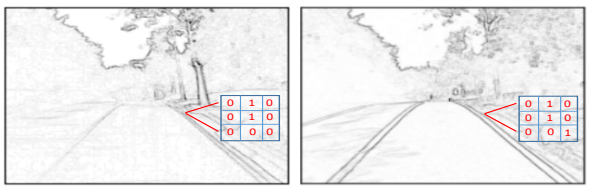


Figure 4. Distance calculation between corresponding pixels. Within a 3 $\times$ 3 window, each pixel value in the window is examined. If there is a corresponding contour pixel, we consider it is a correct contour translation for this pixel.

$$L_{pix}^{i,j} = \begin{cases} 1, & (\sum_{m=i-1}^{i+1}\sum_{n=j-1}^{j+1} I_{m,n}) == 0 \\ 0, & (\sum_{m=i-1}^{i+1}\sum_{n=j-1}^{j+1} I_{m,n}) > 0 \end{cases} \quad (4)$$

$$L_{contour} = \sum_{i=1}^{p}\sum_{j=1}^{q} I_{i,j} L_{pix}^{i,j} \quad (5)$$

where $L_{pix}^{i,j}$ is the contour loss for each pixel. For one contour pixel in the contour maps $I_{m,n}$, if in the corresponding window with coordinates $(i-1, j-1)$, $(i+1, j-1)$, $(i-1, j+1)$, $(i+1, j+1)$ in the color image, the sum is more than 0, that means there is a contour pixel in the window. Then the contour loss at this pixel is 0. Otherwise, the pixel's contour loss is 1. We sum up all the pixel contour loss in the image with the dimension of $p * q$ to form the image contour loss $L_{contour}$. If the pixel is not on the contour, its contour map value $I_{i,j}$ is 0, which generates 0 in the contour loss; otherwise, its contour map value is 1. The loss value for this pixel depends on the corresponding window. The same operation is for both thermal-to-color and color-to-thermal translation. The contour loss enforces the generated image to be spatially close to the original images.

5. Spectral Enhancement

As the thermal and color are two quite different modalities, the translated spectral values may also be quite different. We therefore enhance the spectral similarity between original and translated images. Within a cycle translation concept, illuminance consistency and its contrast in local regions can also enforce the local region to be similar to the original image. For a local patch, we expect the translated and the original images to be highly correlated and their average illumination strength to be close. At the same time, we expect the illuminance contrast within a patch between generated and original images to be similar as well, which represents the illuminance distribution and can be evaluated by the intensity variance. Our illuminance loss takes correlation relationship, average illuminance coefficient, and a contrast term into consideration, defined as follows for a local patch:

$$L_{local\_lx}(x, x') = \frac{\sigma_{x,x'}}{\sigma_x \sigma_{x'}} \cdot \frac{2\bar{x}\bar{x'}}{(\bar{x}^2 + \bar{x'}^2)} \cdot \frac{2\sigma_x \sigma_{x'}}{\sigma_x^2 + \sigma_{x'}^2} \quad (6)$$

where x and x' are the original image and translated image in the same modality. The first term is the correlation relationship between the original data and predicted data. The second term is to reduce the variation of average illuminance between x and x'. The third term is to measure the intensity contrast to guarantee that they are in similar distribution. Then we scan the entire image by sliding a 5-pixel dimensional square window through the entire image with

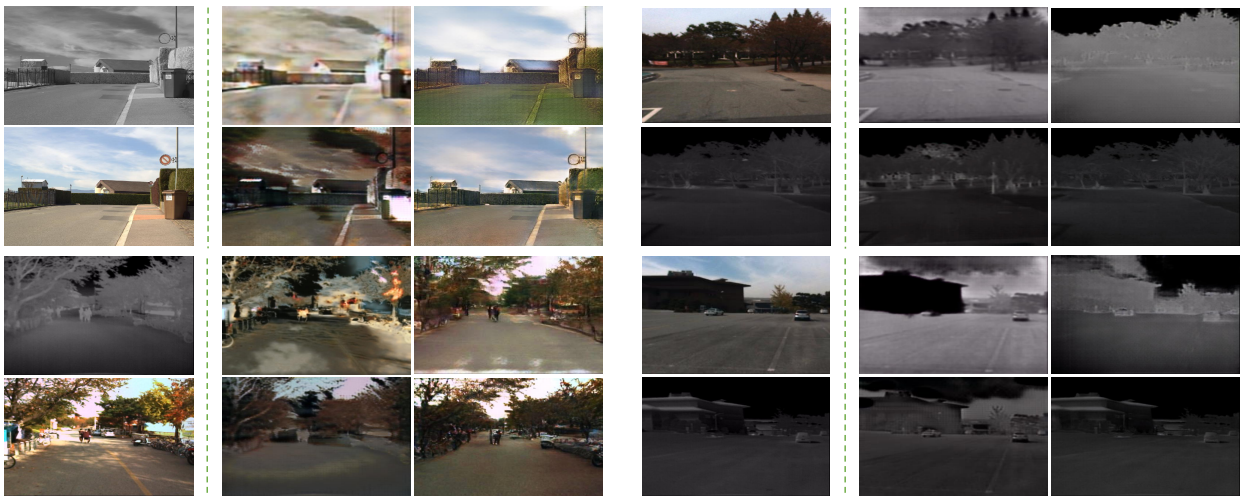


Figure 5. Visual comparison between our method and other state-of-the-art methods for IR-Visible and Visible-IR translation. For each sample, Left: Input IR/visible image (top) and the corresponding visible/IR ground truth image (bottom). Right: CycleGAN [29] output (top left); Pix2PixHD [27] output (top right); RevGAN [26] output (bottom left); Our output (bottom right). Compared with other methods, our method for cross-domain image translation generates more real and detailed images in challenging scenes.

a moving step size of 2. Assuming there are M steps in this process, the whole image illuminance loss is:

$$L_{lx} = \frac{1}{M}\sum_{i=1}^{M}(1 - L^{i}_{local\_lx}) \tag{7}$$

which averages all the local region illuminance consistency and contrast.

6. Experiments

6.1. Datasets

EPFL NIR-VIS dataset [2] contains totally 477 high-resolution images in 9 categories. These categories including country, field, forest, indoor, mountain, old building, street, urban and water captured by color camera and NIR camera at the same time. All the images are resized to be 256 $\times$256. We randomly choose 120 images for testing and the rest images of the dataset are used for training.

KAIST is a long-wave infrared (LWIR) benchmark for multi-spectral pedestrian detection. [11]. This dataset consists of around 95k color-thermal pairs (640x480, 20Hz) taken from a car during both day and night time. With a beam splitter-based hardware to physically align the two image domains, it does not need any post-processing. In this work, we randomly choose 20000 image pairs for training and another 2000 for testing.

6.2. Network Configuration

We train the network from scratch with Adam optimizer where $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The initial learning rate is set to be 0.0001, and we linearly decrease the rate to zero over the next 100 epochs. The LeakyReLu activation function is applied. Weights for input data are initialized from a Gaussian distribution with a mean of 0 and standard deviation of 0.02. We train all our models on an NVIDIA GTX1080Ti GPU with 11GB GPU memory. The weights of different losses in the combined objective function are set to be λ_{cycle}=1.0, λ_{GAN}=0.2, λ^{ROI}_{cycle}=1.0, λ_{lx}=0.5 and $\lambda_{contour}$ =1.0. Though the cycle consistency term plays a significant impact in the early stage, it becomes less stable in the late stage to generate images. Thus we progressively decrease the weight for the cycle consistency term after half of the entire training process.

Figure 6. Visual result of our method to translate color image to NIR domain, from input RGB images (top) to outputs (bottom).

6.3. Visual Evaluation

In this section, our method is compared with other state-of-the-art methods in multi-modal image translation: CycleGAN [29], Pix2PixHD [27] and RevGAN [26]. By comparing different methods trained on the same split, it is apparent in Fig. 5 that CycleGAN [29] has a serious wrong mapping problem in textures and colors. For Pix2PixHD [27] and RevGAN [26], though they perform better on mapping representation, they still lack sharp texture information and exist some blurriness in details. It can be observed that our method achieves the best visual performance in both

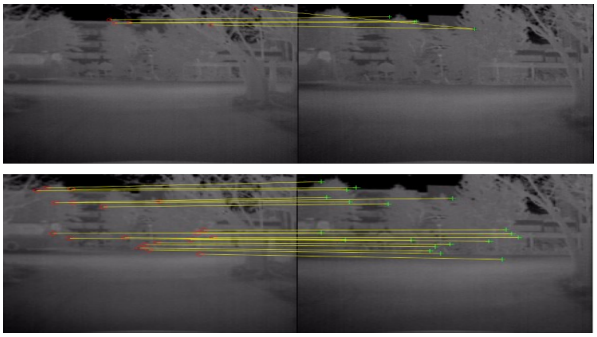


Figure 7. Feature matching without and with the support of image translation. (top row) SURF feature matching directly on thermal images. (bottom row) SURF feature matching on thermal images with the support of translated RGB images.

colorizing the thermal image and transferring the color image to thermal domain. Our method not only can learn a correct mapping representation between multi-spectral domain images but also preserve the objects' textures and boundaries, which attributes to the newly designed ROI loss and contour edge loss. Our additional image translation results from color to Near IR domain are shown in Fig. 6.

Our result has shown that the translated images help to significantly improve the feature matching performance for thermal images. The demo inputs and their corresponding results from our full pipeline are shown in Fig. 7. From Fig. 7, SURF features almost cannot match on original input thermal images. However, with the translated images, SURF features can be applied to thermal images bridging the RGB images translated from thermal images.

An ablation analysis is provided in Fig. 8 based on adding the ROI loss and contour edge consistency loss or vice versa, including CCGAN-with/without-ROI, CCGAN-with/without-contour consistency, and CCGAN full pipeline. It can be seen that our full method (CCGAN-full) captures and recovers finer details in specific regions (e.g., cars, bicycles, and traffic cones) and suffers less from blurriness compared with a partial implementation of our method CCGAN-w/o-ROI and CCGAN-w/o-Contour. Fig. 9 demonstrates the capability of our method in nighttime scenarios when visible images from RGB cameras are very dim and almost invalid because of insufficient light. Our translation method is able to recover the invisible light and texture of the RGB camera in the nighttime by translating thermal to color images, making the matching and detection tasks for thermal images possible.

In addition to scenes, we further verify our algorithm on living human face, as shown in Fig. 11. With a split of 80% images of the Tufts face thermal-RGB dataset for training and the rest for testing, we can observe that our method is able to be extended into humans. More visual results are provided in the supplementary video.

	NIR-Color				FIR-Color			
	PSNR	SSIM	COS	RMSE	PSNR	SSIM	COS	RMSE
CycleGAN [29]	9.2011	0.4722	0.9408	0.5842	9.5713	0.4758	0.8452	0.5537
MUNIT [10]	13.2140	0.5061	0.9456	0.5000	12.2981	0.5075	0.8160	0.4351
Pix2PixHD [27]	16.2137	**0.6271**	0.9620	0.4894	15.2596	0.5818	0.8674	0.4203
RevGAN [26]	14.9573	0.5894	0.9547	0.5041	14.1239	0.5482	0.8463	0.4736
Ours	**18.7115**	0.6166	**0.9861**	**0.4764**	**16.5169**	**0.6186**	**0.9388**	**0.3862**

Table 1. Average results on PSNR, SSIM, COS similarity, and RMSE on the testing dataset from IR to color domain. The best results are marked in bold.

	Color-NIR				Color-FIR			
	PSNR	SSIM	COS	RMSE	PSNR	SSIM	COS	RMSE
CycleGAN [29]	14.6093	0.6823	0.7824	0.2057	9.4033	0.3331	0.7507	0.3996
MUNIT [10]	15.3845	0.6852	0.8023	0.1784	11.6595	0.5243	0.7151	0.3657
Pix2PixHD [27]	18.0427	0.7919	0.8314	0.1609	16.9011	0.7203	0.8308	0.3477
RevGAN [26]	18.2430	0.8037	0.8528	0.1329	17.2903	0.7193	0.8433	0.2910
Ours	**21.9635**	**0.8205**	**0.8655**	**0.1197**	**18.0431**	**0.7841**	**0.8911**	**0.2396**

Table 2. Average results on PSNR, SSIM, COS similarity and RMSE on the testing dataset from color to IR domain. The best results are marked in bold.

	IR-Color		Color-IR	
	IS	FID	IS	FID
CycleGAN [29]	1.0	97.3	1.2	80.2
MUNIT [10]	1.4	75.9	1.6	58.3
Pix2PixHD [27]	1.6	56.2	1.9	36.4
RevGAN [26]	1.5	59.7	2.1	29.6
Ours	**1.7**	**39.2**	**2.2**	**21.5**

Table 3. Additional results on Inception score (IS, higher is better) and Frechet Inception Distance (FID, lower is better) on the testing split from IR to color and color to IR domain.

6.4. Quantitative Evaluation

To evaluate the effectiveness of our method quantitatively, we choose four commonly used measurement metrics for image quality evaluation, which are Root Mean Squared Error (RMSE), Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM), and COS Similarity (COS). RMSE evaluates a root difference between the two compared images, while PSNR indicates the level of losses. SSIM is a metric evaluating the similarity level to the human visual system that extracts useful information from images such as structure, illuminance, and contrast. COS similarity is defined as the average angular similarity between every generated RGB pixel and the corresponding ground truth image pixels. A comparison between our proposed method and other recent methods is shown in Table 1 and Table 2 for Far IR and Near IR images respectively. It can be observed from Table 1 and Table 2 that our method achieves the best performance in both Far IR and Near IR datasets. Though CycleGAN [29] achieves relatively good performance only in Color-Near IR conversion, it performs worse on the Far IR dataset. Compared with Pix2PixHD [27] and RevGAN [26], our method still enjoys an improvement benefiting from the designed Contour and ROI constraints on detail recovery. Results from our method demonstrate that our proposed method is capable of learning the correct mapping features and representation from the source to target domains, and enjoys a signif-

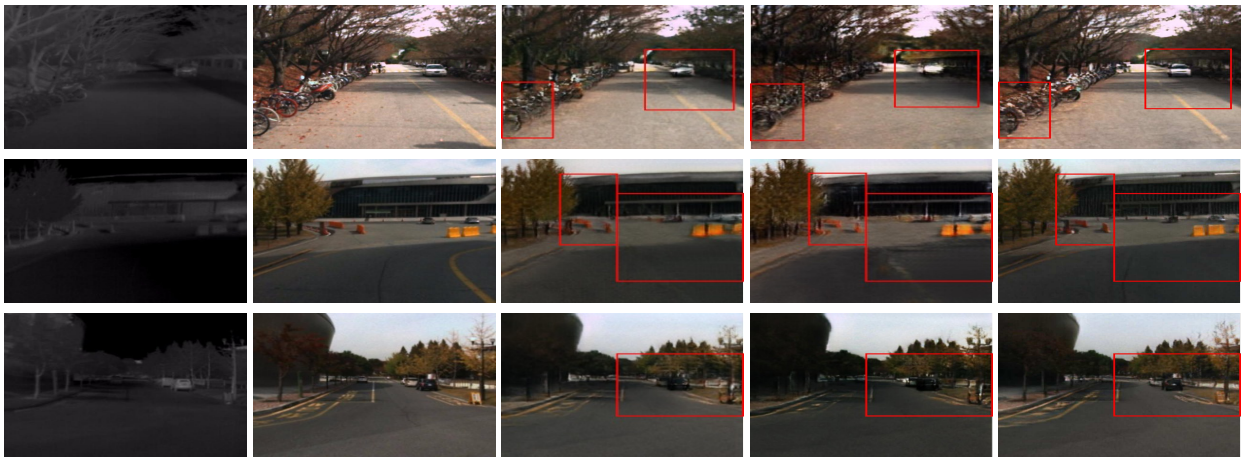


Figure 8. Ablation analysis of each core contribution in our network. From left to right: Source input thermal image; Ground truth color images; Result without the proposed contour loss; Result without the proposed ROI loss; Result from our full pipeline.

Figure 9. Example images of our translated result on night scenarios. Left to right: Input thermal image at nighttime; Real color image at nighttime; Our translated image. It can be observed that our method has the ability to estimate images under low light.

icant improvement compared with the MUNIT [10], CycleGAN [29], Pix2PixHD [27] and RevGAN [26], especially

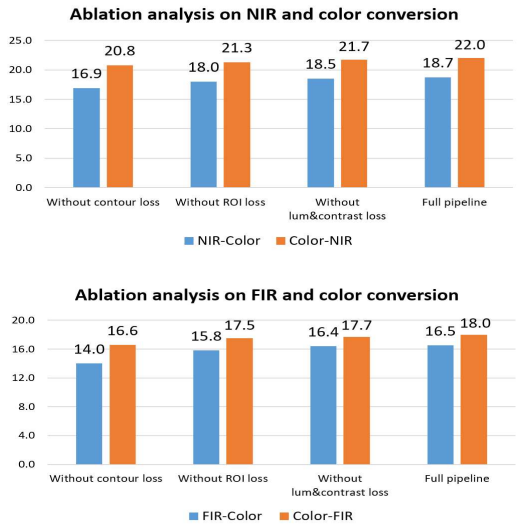


Figure 10. Quantitative PSNR comparisons on ablation analysis of NIR-Color and FIR-Color conversion.

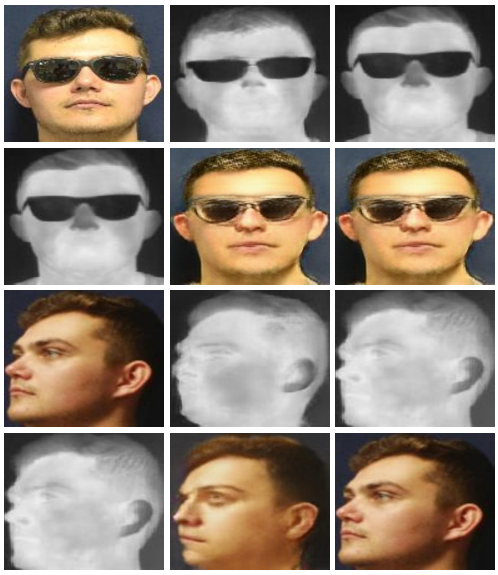


Figure 11. Additional results of RGB-thermal face image translation on Tufts face thermal-RGB dataset. Left to right in each row: input raw RGB / thermal images; our estimated thermal / RGB images; corresponding ground truth thermal / RGB images.

in terms of PSNR and RMSE. In addition to local geometric measurements (PSNR/SSIM/COS/RMSE) above, we also provide average results on Inception score (IS) and Frechet Inception Distance (FID) to measure the quality of generated images by calculating the corresponding feature vectors, as shown in Table 3. A higher IS score and lower FID indicates better-quality images.

Fig. 10 shows a quantitative ablation analysis for each key component in our designed framework. We observe the highest PSNR in our full pipeline on all of the four domain transfer scenarios, compared with partial constraints without either of them (contour loss, ROI loss, illuminance and contrast loss).

7. Conclusion

We propose CCGAN, a GAN framework targeting IR and visible image translation. We design the network based on a multi-scale structure with constraints dedicated to IR and visible image translation, which preserves the shared properties between these two image modalities. The proposed method is able to learn the mapping representations between different image modalities. In addition to the spectral constraint, the framework introduces spatial constraint in image translation tasks through contour consistency. The transformation from the visible image to infrared thermal image makes it possible to predict the temperature of the object surface for inspection and surveillance tasks. The transformation from infrared images to visible images makes it possible to apply existing computer vision algorithms on thermal images such as image matching and 3D reconstruction. Our method improves the image translation performance on both Far IR and Near IR datasets.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Award No. 2105257.

References

[1] M. Bertozzi, A. Broggi, P. Grisleri, T. Graf, and M. Meinecke. Pedestrian detection in infrared images. In *IVS*, 2003. 3

[2] M. Brown and S. Süsstrunk. Multi-spectral sift for scene category recognition. In *CVPR*, 2011. 5

[3] G. Cielniak, T. Duckett, and A. J. Lilienthal. Data association and occlusion handling for vision-based people tracking by mobile robots. *RAS*, 58(5):435–443, 2010. 3

[4] W. L. Fehlman and M. K. Hinders. *Mobile robot navigation with intelligent infrared image interpretation*. Springer Science & Business Media, 2009. 3

[5] A. Fernández-Caballero, J. C. Castillo, J. Martínez-Cantos, and R. Martínez-Tomás. Optical flow or image subtraction in human detection from infrared camera on mobile robot. *RAS*, 58(12):1273–1281, 2010. 3

[6] A. Garcia-Cerezo, A. Mandow, J. Martinez, J. Gómez-de Gabriel, J. Morales, A. Cruz, A. Reina, and J. Seron. Development of alacrane: A mobile robotic assistance for exploration and rescue missions. In *SSRR*, 2007. 3

[7] E. Gebhardt and M. Wolf. Camel dataset for visual and thermal infrared multiple object detection and tracking. In *AVSS*. IEEE, 2018. 3

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 1

[9] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma. Dual learning for machine translation. In *NIPS*, 2016. 2

[10] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, pages 172–189, 2018. 1, 6, 7

[11] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *CVPR*, 2015. 5

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1, 2

[13] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, pages 1857–1865, 2017. 1, 2

[14] S. Lankton and A. Tannenbaum. Localizing region-based active contours. *TIP*, 17(11), 2008. 4

[15] 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. *IJCV*, 128(10), 2020. 2

[16] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017. 1

[17] Q. Liu, X. Li, Z. He, N. Fan, D. Yuan, and H. Wang. Learning deep multi-level similarity for thermal infrared object tracking. *Transactions on Multimedia*, 2020. 3

[18] G. Lu, Y. Yan, L. Ren, P. Saponaro, N. Sebe, and C. Kambhamettu. Where am i in the dark: Exploring active transfer learning on the use of indoor localization based on thermal imaging. *Neurocomputing*, 173:83–92, 2016. 3

[19] Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, and L. Lin. Single view stereo matching. In *CVPR*, 2018. 2

[20] H. Nanda and L. Davis. Probabilistic template based pedestrian detection in infrared videos. In *IVS*, 2002. 3

[21] B. Pan, Z. Shi, and X. Xu. Longwave infrared hyperspectral image classification via an ensemble method. *International Journal of Remote Sensing*, 38(22):6164–6178, 2017. 3

[22] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1

[23] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2

[24] P. Saponaro, S. Sorensen, A. Kolagunda, and C. Kambhamettu. Material classification with thermal imagery. In *CVPR*, pages 4649–4656, 2015. 3

[25] K. Van Beeck, K. Van Engeland, J. Vennekens, and T. Goedemé. Abnormal behavior detection in lwir surveillance of railway platforms. In *AVSS*. IEEE, 2017. 3

[26] T. F. van der Ouderaa and D. E. Worrall. Reversible gans for memory-efficient image-to-image translation. In *CVPR*, 2019. 1, 2, 5, 6, 7

[27] 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 *CVPR*, 2018. 2, 3, 5, 6, 7

[28] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017. 2

[29] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2, 3, 5, 6, 7