ROMA: Cross-Domain Region Similarity Matching for Unpaired Nighttime Infrared to Daytime Visible Video Translation

Zhenjie Yu  
Beijing Institute of Technology  
Beijing, China  
zjyu@bit.edu.cn

Kai Chen  
Yantai IRay Technologies Lt. Co.  
Shandong, China  
kai.chen@iraytek.com

Shuang Li∗  
Beijing Institute of Technology  
Beijing, China  
shuanglei@bit.edu.cn

Bingfeng Han  
Beijing Institute of Technology  
Beijing, China  
bfhan@bit.edu.cn

Chi Harold Liu  
Beijing Institute of Technology  
Beijing, China  
liuchi02@gmail.com

Shuigen Wang  
Yantai IRay Technologies Lt. Co.  
Shandong, China  
shuigen.wang@iraytek.com

Figure 1: Left: Display of the nighttime infrared, nighttime visible and translated daytime visible videos via ROMA for Highway and Monitor scenarios in InfraredCity, respectively. Right: Display of applications on object detection and video fusion for ROMA-translated results. Notably, the ROMA-translated daytime videos achieve superior detection performance compared with the corresponding nighttime infrared and visible videos, and the video fusion results between infrared and ROMA-translated videos are much sharper than the counterparts. The animated videos are best viewed via Adobe Acrobat, please zoom in for details.

ABSTRACT

Infrared cameras are often utilized to enhance the night vision since the visible light cameras exhibit inferior efficacy without sufficient illumination. However, infrared data possesses inadequate color contrast and representation ability attributed to its intrinsic heat-related imaging principle, which hinders its application. Although, the domain gaps between unpaired nighttime infrared and daytime visible videos are even huger than paired ones that captured at the same time, establishing an effective translation mapping will greatly contribute to various fields. In this case, the structural knowledge within nighttime infrared videos and semantic information contained in the translated daytime visible pairs could be utilized simultaneously. To this end, we propose a tailored framework ROMA that couples with our introduced cRoss-domain regiOn siMilarity mAtching technique for bridging the huge gaps. To be specific, ROMA could efficiently translate the unpaired nighttime infrared videos into fine-grained daytime visible ones, meanwhile maintain the spatiotemporal consistency via matching the cross-domain region similarity. Furthermore, we design a multiscale region-wise discriminator to distinguish the details from synthesized visible results and real references. Moreover, we provide a new and challenging dataset encouraging further research for unpaired nighttime infrared and daytime visible video translation, named InfraredCity, which is 20 times larger than the recently released infrared-related dataset IRVI. Codes and datasets are available here.

CCS CONCEPTS

- Information systems → Multimedia content creation.

KEYWORDS

Nighttime Infrared; Daytime Visible; Video-to-Video Translation

ACM Reference Format:
Zhenjie Yu, Kai Chen, Shuang Li, Bingfeng Han, Chi Harold Liu, and Shuigen Wang. 2022. ROMA: Cross-Domain Region Similarity Matching for Unpaired Nighttime Infrared to Daytime Visible Video Translation. In Proceedings of the 30th ACM International Conference on Multimedia (MM ’22), October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548221
1 INTRODUCTION

In real-world multimedia applications, visible light cameras are often leveraged to assist improving visual effects for various scenarios [27]. Unfortunately, their adaptability becomes even worse than the human biological vision system under extreme conditions, e.g., dark night or light exposure as shown in Fig. 1. In such cases, infrared sensors could take the place of visible cameras as an auxiliary imaging systems. Its heat-related imaging principle can stably provide visual signals with sufficient spatial or structural descriptions, however, the lack of detailed semantic information can not well satisfy the conventional cognition of this colorful world [46]. This makes the raw infrared data undesirable for direct utilization in practical tasks, such as autonomous driving or monitoring. Even so, infrared sensors are indispensable in real-world applications, especially for the dark night situation. Therefore, it is worthy to bridge the modality gaps between the infrared and visible data.

 Plenty of works [4, 23, 37, 44, 46] have thoroughly studied the data fusion for infrared and visible frames, while their results often visually retain the appearance of grayscales, which are still not distinct compared with visible ones. Besides, other methods [6, 15, 16, 34] try to transform infrared images to visible ones through different color mapping functions. However, they usually require complex manual interventions, which impose limitations in practical applications. With the development of deep learning based generative models, image-to-image translation methods [1, 21, 32, 45] have attracted appreciable attentions. They seek to achieve high-quality performance through powerful generative adversarial training techniques [19, 25]. However, the huge domain gaps between infrared and visible images make these methods incompetent in precisely preserving the proper infrared structure information and abundant visible semantic details. Moreover, nowadays most real-world applications provide feedback in the form of video signals. These image-level methods will lose their applicability due to lack of consideration for video temporal consistency.

To this end, video-to-video translation methods have recently taken a further step on the basis of their counterparts. For instance, [42, 43] synthesize target videos and predict future frames greatly rely on labeled pair data. However, like most video translation tasks, the nighttime infrared and daytime visible videos have no pixel-to-pixel aligned training data. Manual labeling is not only time-consuming and expensive, but also prone to introduce errors. Thus, some unpaired video-to-video translation methods [2, 3, 8, 24] loosen the requirements for paired data. In particular, [24] is a framework that adapts for daytime infrared and visible video translation. The huger domain gap between nighttime infrared and daytime visible data will inevitably diminish their translated performance, since the critical semantic structural correspondence across the two domains are largely overlooked. Moreover, their feature-level losses have integrated the domain-invariant content and domain-specific appearance together. Hence, the structural consistency and semantic details will be destroyed in the translated results, especially for tremendous domain discrepancy.

To tackle these challengings, we propose a novel one-sided end-to-end framework ROMA that couples with our introduced cross-domain region similarity matching technique (referred to as cross-similarity in the rest of this paper) to bridge the huge gap between nighttime infrared and daytime visible videos. Specifically, we focus on the domain-invariant structural information by optimizing three forms of cross-similarity maps to generate fine-grained translating results and keep temporal consistency. Besides, we propose a multiscale region-wise discriminator to enhance the detailed domain-specific style information.

As shown in Fig. 2, the cross-similarity map is calculated between input and synthesized output on the basis of a pre-trained ViT [11]. We first split the frames into several regions, and each region of the input could obtain similarity maps interacting with all regions of the output. Meanwhile, the similarity maps of the output regions could also obtain by interacting with the input regions. Notably, all the cross-similarity maps are calculated across domains, which could make the generative process focus on learning the content or structural correspondence between real and synthesized frames, while getting rid of the negative effects of different styles. Then we match the cross-domain region similarity maps calculated at the same location. This process is called global structural cross-similarity consistency, since it keeps the domain-invariant structural information to be the same under different appearances. Moreover, we take a step further to apply this pattern within local regions for fine-grained structure preservation, which is called local structural cross-similarity consistency. Since the cross-similarity map matching cuts off the influence of domain-specific style information, the similarity maps can also measure structural variants caused by scene movements. Thus, we optimize temporal cross-similarity consistency by minimize the distance of similarity maps that calculated from time \( t \) input, time \( t + 1 \) output and time \( t \) output, time \( t + 1 \) input.

As for domain-specific style enhancement, we propose a multiscale region-wise discriminator. We first extract the token embeddings of real references and synthesized results via ViT, then reshape and refine them into different scales. At last, the MLP layers are applied to distinguish real or fake depending on the concatenation of multiple representations.

Additionally and importantly, the nighttime infrared to daytime visible video translation is rarely studied due to the lack of a high-quality relevant dataset. In this paper, we offer a new and challenging dataset named InfraredCity, which consists of 9 long video clips including city, highway, and monitoring scenarios. All clips could be split into 579,984 frames in total. After manual selection, we additionally provide InfraredCity-Lite for research. In summary, our contributions are listed as below:
• A simple yet effective cross-domain region similarity matching technique is proposed, which could fully utilize the structure knowledge of nighttime infrared data, and enhance the structure correspondence between input and output, facilitating generating authentic and fluent daytime visible videos.
• We propose an efficient one-sided end-to-end framework ROMA, which performs cross-similarity matching and is coupled with our introduced multiscale region-wise discriminator. Besides, ROMA achieves superior performance compared to other state-of-the-art baselines on several datasets.
• The translated videos via ROMA could be further applied to real-world applications, such as object detection and video fusion. The promising results validate the effectiveness of ROMA for night vision scenarios.
• We provide new datasets for nighttime infrared to daytime visible video translation, i.e., InfraredCity and InfraredCity-Lite, encouraging further research on this area.

2 RELATED WORK

Infrared-to-Visible Translation. Since infrared-to-visible translation is an attractive strategy to enhance night vision perception, there are continuing studies on it. Generally, the infrared vision technique is often used for the context enhancement in nighttime vision by fusing it with the visible data [44, 44, 46]. Unfortunately, the nighttime visible image is often terrible on account of low-light conditions, which delivers limited knowledge to infrared data. Also, the gray fusion results are undesirable for human beings. Thus, [6, 34, 40, 41] regard infrared images as gray ones and attempt to generate visible images via colorization approaches. Similarly, [17, 26, 38, 39] utilize the GAN [19] module and attempt to affine the single-channel infrared image into the three-channel RGB result via colorization manners. Although these methods could generate colorful results, they prone to distort details without additional structural constraints. Moreover, these image-level methods impose limitations on the infrared-to-visible video translation task due to lack of consideration for temporal consistency.

Image-to-Image and Video-to-Video Translation. Image-to-image translation intends to learn a mapping from the source domain to the target domain. Pix2pix [33] explores the possibility of applying the deep network on image translation via the GAN framework [19] with paired datasets. Furthermore, to relax the requirements of paired datasets, CycleGAN [21] introduces the cycle consistency, which maintains the content during training. However, it requires auxiliary generators and discriminators for the reverse mapping, leading to more computation cost. To avoid this, [5, 14] adopt a one-sided framework and propose implicit structural consistency to replace cycle consistency. [7, 20, 31] propose structural consistency under the high-level semantic information. CUT [32] and its following algorithm [22] seek to maximize the mutual information between the two domains. Although these approaches guarantee spatial consistency in image-to-image translation tasks, they can not be directly applied to the video translation tasks due to the lack of consideration for temporal coherence.

To make up for the deficiency, [42, 43] translate videos with hand-designed temporal consistency based on paired video datasets. However, it is almost impossible to collect the pixel-to-pixel paired videos for the nighttime infrared and daytime visible translation tasks. Thus, [3] explores unpaired video-to-video translation and proposes 3D Convolution as a pioneer. Recycle-GAN [2] utilizes cycle loss and recurrent loss to keep the consistency of spatial and temporal information. Similarly, mocyte-GAN [8] utilizes cycle loss for structural consistency and motion consistency loss based on optical flow to maintain the temporal coherence. Especially, I2V-GAN [24] is a tailored approach for daytime infrared-to-visible translation, which additionally proposes perceptual cyclic losses and similarity losses to enhance the spatiotemporal consistency compared with Recycle-GAN. Although I2V-GAN achieves acceptable results on infrared-to-visible translation, huger domain gaps between nighttime infrared and daytime visible data hinder its efficacy. Besides, these approaches perform their constraints mainly on image level and feature level, which will inevitably conflate video content and style information together in the optimization process due to the huge difference between two domains. Therefore, the structural correlation between each region in the infrared videos will be entangled with the domain style, leading to generating blurred and unpleasant visible videos.

3 THE PROPOSED METHOD

For an nighttime infrared video clip $\mathbb{X} = \{x_1, x_2, ..., x_N \mid x \in \mathbb{R}^{H \times W \times C}\}$ and a daytime visible video clip $\mathbb{Y} = \{y_1, y_2, ..., y_N \mid y \in \mathbb{R}^{H \times W \times C}\}$, we aim to guide the generator $\mathcal{G}$ to transfer $\mathbb{X}$ into the target-style clip $\hat{\mathbb{Y}} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_N \mid \hat{y} \in \mathbb{R}^{H \times W \times C}\}$, and $\hat{\mathbb{Y}}$ are used for denoting the number of total video frames from $\mathbb{X}$ and $\mathbb{Y}$, respectively. In particular, the translated results $\hat{\mathbb{Y}}$ should maintains the spatiotemporal consistency with $\mathbb{X}$, but converts the appearance appropriately as real visible video $\mathbb{Y}$. Notably, $\mathbb{Y}$ and $\hat{\mathbb{Y}}$ only share the same style information within their different scenes. This translation process is denoted as $\hat{\mathbb{Y}} = \mathcal{G}(\mathbb{X})$.

We begin this section by introducing our proposed cross-domain region similarity matching technique. Specifically, we propose three...
types of constraints from both spatial and temporal perspectives for video translation, which concentrate on the domain-invariant information and cut off the negative influence of domain-specific information. Then, we introduce a multiscale region-wise discriminator, which is applied to distinguish the details of domain-specific information between synthesized results and real references.

3.1 Cross-domain Region Similarity Matching

We take an input frame $x \in \mathcal{X}$ and the corresponding synthesized output $\hat{y} \in \bar{\mathcal{X}}$ as an example to illustrate. As shown in Fig. 2, we first divide $x$ and $\hat{y}$ into non-overlapping regions. The total number of regions from $x$ or $\hat{y}$ is $N_r$. Then the cross-similarity map of $i$-th source region $sr_i$ in $x$ can be calculated to associate to all regions of $\hat{y}$. Similarly, the cross-similarity map of $i$-th target region $tr_i$ is obtained by interacting with all the source regions. We formulate this process for cross-similarity maps calculation as following:

$$S_{sr_i} = u_i \cdot v_j^\top, \quad (1)$$
$$S_{tr_i} = v_i \cdot u_j^\top, \quad (2)$$

where $u_i, v_i \in \mathbb{R}^{1 \times d}$ are d-dimensional token embeddings for region $sr_i$ and region $tr_i$, $u_i, v_i \in \mathbb{R}^{d \times N_r}$ denote the transposed token embeddings of all $N_r$ non-overlapping regions in $x$ and $\hat{y}$, respectively.

Thus, $S_{sr_i}$ stands for the cross-similarity map of region $sr_i$ in $x$, and $S_{tr_i}$ stands for the cross-similarity map of region $tr_i$ in $\hat{y}$. After we obtain the cross-similarity collections $S_x = \{S_{sr_1}, S_{sr_2}, ..., S_{sr_{N_r}}\}$ and $S_{\hat{y}} = \{S_{tr_1}, S_{tr_2}, ..., S_{tr_{N_r}}\}$, we perform the similarity matching by minimizing their distance within multiple layers of the ViT. Specifically, by picking several representations from different layers, we get $S_x = \{S_x^1, S_x^2, ..., S_x^{N_l}\}$ and $S_{\hat{y}} = \{S_{\hat{y}}^1, S_{\hat{y}}^2, ..., S_{\hat{y}}^{N_l}\}$, which represent different views of $x$ and $\hat{y}$. Finally, the objective of cross-similarity matching between $x$ and $\hat{y}$ is defined as:

$$\mathcal{L} = \frac{1}{N_l} \sum_{i=1}^{N_l} d(\hat{S}_x^i, S_{\hat{y}}^i), \quad (3)$$

where $d(\cdot)$ is the cosine distance function which performs better than $L1$ and $L2$ according to numerous experimental results. As for $N_l$, it means the number of feature layers selected from the Feature Extractor. Based on the cross-similarity, we design three forms of constraints for spatiotemporal consistency shown in Fig. 3.

Global Cross-Similarity for Structural Consistency. Given input frames $\mathcal{X} = \{x_1, x_2, ..., x_N\}$ from source domain and their corresponding synthesized target frames $\bar{\mathcal{Y}} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_N\}$, the global cross-similarity matching performs as:

$$\mathcal{L}_g(G) = \frac{1}{N_l} \sum_{i=1}^{N} d(\Sigma_x(t), \Sigma_{\hat{y}}(t)), \quad (4)$$

where $\Sigma_x(t)$ and $\Sigma_{\hat{y}}(t)$ are the collections of cross-similarity maps of the $t$-th frames in $\mathcal{X}$ and $\bar{\mathcal{Y}}$ that obtained from different $l$ layers.

Local Cross-Similarity for Structural Consistency. Moreover, we propose a local optimization strategy to further improve the structural details for fine-grained translation results. For instance, as shown in Fig. 3, we first randomly pick $N_a$ areas from the input frame $x_t$. The size of an area is larger than a region. Then, the areas of $\hat{y}_t$ are automatically obtained with the same locations as their corresponding areas in $x_t$. After that, the local cross-similarity matching is performed within each areas pair as:

$$\mathcal{L}_l(G) = \frac{1}{N_l} \sum_{i=1}^{N_a} d(S_{sr_i}^l(t), S_{tr_i}^l(t)), \quad (5)$$

where $S_{sr_i}^l(t)$ and $S_{tr_i}^l(t)$ are the collections of cross-similarity maps of the $i$-th areas in $t$-th frames from $\mathcal{X}$ and $\bar{\mathcal{Y}}$ from different $l$ layers. Concretely, $N_a$ is equal to 64 in the experiment, i.e., we randomly pick 64 areas in each frame.

Cross-Similarity for Temporal Consistency. As shown in Fig. 2 and Fig. 4, cross-similarity can cut off the negative influences of domain-specific styles and focus on the domain-invariant structure. Thus, the similarity maps can apply to measure the structural variants caused by scene movements and further optimize the temporal consistency for video translation. To be specific, we utilize $u_i(t)$ and $v_i(t)$ to represent the $i$-th regions of $x_t$ and $\hat{y}_t$ at time $t$. $u_i(t+1)$ and $v_i(t+1)$ denote the token embeddings for all regions of $x_{t+1}$ and $\hat{y}_{t+1}$. The temporal cross-similarity is:

$$S_{sr_i}(t+1) = u_i(t) \cdot v_i(t+1)^\top, \quad (6)$$
$$S_{tr_i}(t+1) = v_i(t) \cdot u_i(t+1)^\top. \quad (7)$$

We formulate the multilayer temporal cross-similarity matching as:

$$\mathcal{L}_{tem}(G) = \frac{1}{N_l} \sum_{i=1}^{N_l} d(S_{sr_i}(t, t+1), S_{tr_i}(t, t+1)), \quad (8)$$

where $S_{sr_i}(t, t+1)$ is the collection $\{S_{sr_i}(t, t+1) | i = 1, ..., N_r\}$, and $S_{tr_i}(t, t+1)$ is the collection $\{S_{tr_i}(t, t+1) | i = 1, ..., N_r\}$. Moreover, this constraint can be extended to a generic version related to length of video fragment in each training step:

$$\mathcal{L}_{tem}(G) = \frac{1}{N_l} \sum_{i=1}^{N_l} \sum_{t'=t+\Delta t} d(S_{sr_i}(t, t'), S_{tr_i}(t, t')), \quad (9)$$

where $\Delta t$ is a hyper-parameter to adjust the length of training fragment, and we simply set it as 1 for fast implementation.

3.2 Multiscale Region-Wise Discriminator

In this paper, we design a multiscale region-wise discriminator to distinguish real or fake among visible references and synthesized target results, as shown in Fig. 5. Firstly, we get the token embeddings $T_y, T_{\hat{y}} \in \mathbb{R}^{N_r \times d}$ from the pre-trained ViT and reshape...
We train the network by minimizing the following losses:

\[
L_{adv}(G, D_r) = \mathbb{E}_{y \sim p_{data}(y)} \left[ \log D_r(y) \right] + \mathbb{E}_{x \sim p_{data}(x)} \left[ \log(1 - D_r(G(x))) \right],
\]

where \(D_r\) is our proposed multiscale region-wise discriminator, \(D_r(y) = MLP(T_y)\) and \(D_r(G(x)) = MLP(T_g)\).

### 3.3 Overall Optimization

We train the network by minimizing the following losses:

\[
\begin{align*}
L_D &= - \mathbb{E}_{y \sim p_{data}(y)} \left[ \log D_r(y) \right] \\
&\quad - \mathbb{E}_{x \sim p_{data}(x)} \left[ \log(1 - D_r(G(x))) \right], \\
L_G &= \mathbb{E}_{x \sim p_{data}(x)} \left[ \log(1 - D_r(G(x))) \right] + \lambda_1 \cdot L_g(G) + \lambda_2 \cdot L_f(G) + \lambda_3 \cdot L_{rem}(G),
\end{align*}
\]

where \(\lambda_1, \lambda_2\) and \(\lambda_3\) are tradeoff parameters.

### 3.4 Discussion

In nighttime infrared to daytime visible translation, the key challenge is to well preserve the structural details of infrared data. Perceptual loss [12, 29] and PatchNCE loss [32] have been studied to constrain the structural consistency. Unfortunately, these losses do not explicitly decouple structural and style information. The quality of results is limited by different domain styles. Furthermore, F/LSeSim [45] proposes a self-similarity strategy, which is enhanced by a learnable network, to represent the domain-invariant structural information. However, the data augmentation manner utilized to train the small network is not suitable for reducing the stylistic effects of infrared data. We experimentally analyze the effect of style on similarity and visualize the cosine similarity between irrelevant regions in the Fig. 4. Experiments indicate that the influence of infrared style is negative for self-similarity matching. On the contrary, our cross-similarity shows its ability of cutting off negative influences of style, which is tailored for representing the domain-invariant structural details from infrared data.

### 4 INFRADEDCITY DATASET

| InfraredCity | Total Frame |
|--------------|-------------|
| Nighttime Infrared | 201,856 |
| Nighttime Visible | 178,698 |
| Daytime Visible | 199,430 |

| InfraredCity-Lite | Infrared Train | Infrared Test | Visible Train | Total |
|-------------------|---------------|--------------|--------------|-------|
| City | clear | 5,538 | 1,000 | 5,360 | 15,180 |
| City | overcast | 2,282 | 1,000 | 2,282 | 4,204 |
| Highway | clear | 4,412 | 1,000 | 4,412 | 6,463 | 15,853 |
| Highway | overcast | 2,978 | 1,000 | 2,978 | 4,956 | 9,912 |

We utilize the binocular infrared color camera (DTC equipment) to capture nighttime infrared and visible videos. Additionally, a visible camera is adopted during the day to capture daytime visible videos in the same scenes, while having no sync with nighttime ones. It is well-known that deep models require massive amounts of training data. Our InfraredCity dataset consists of 201,856 nighttime infrared frames, 178,698 nighttime visible frames, and 199,430 daytime visible frames, detailed in Tab. 1. The InfraredCity is about 20 times larger than the most relative dataset IRVI [24] which is provided for daytime infrared-to-visible video translation. Besides, we capture these infrared videos at night from three scenes (city, highway, and monitoring scenarios). Specifically, City and Highway are captured under clear and overcast weather conditions. The dataset is more challenging than the IRVI since the domain gaps between nighttime infrared and daytime visible videos are much larger, raising requirements for preserving structural information of infrared videos covered by the gray appearance.
To facilitate comparison with other methods, we select parts of the InfraredCity dataset to build the InfraredCity-Lite dataset, which contains 41,839 frames in total. We design the InfraredCity-Lite for three forms: Single, Double and Triplet to be in line with the input requirements of most image/video translation methods. The selection strategy is detailed in the supplementary material. Several examples of each scene are shown in Fig. 6.

We comprehensively select four popular infrared related datasets (IRVI [24], VOT2019-RGBTIR [30], FLIR [13] and KAIST [18]) for comparison and details are shown in Tab. 2. As shown in Tab. 2, we performed a detailed comparison in terms of dataset size, number of video clips, and their respective primary purposes.

### 5 EXPERIMENTS

#### 5.1 Datasets

**InfraredCity-Lite** is collected for nighttime infrared to daytime visible translation. This dataset contains 37,339 training and 4500 testing frames. The resolution of both infrared and visible videos is 256 × 256. Our experiments are mainly based on it.

**IRVI** is a widely-popular dataset for infrared-to-visible translation, consisting of 22,080 training and 2272 testing frames. The videos are captured during the day, requiring models to align the infrared video to visible results. The resolution is 256 × 256.

**5.2 Experiment Setup**

**Evaluation Metrics.** We first use the standard Fréchet Inception Distance (FID) [29] to compare the distribution of synthesized daytime visible frames with the distribution of real daytime visible frames from the features space. In the standard setting, these features are estimated by an InceptionV3 [9] pre-trained on the ImageNet [10] dataset. The FID evaluation from the perspective of feature distributions can effectively reflect whether the objects (e.g., buildings, cars, etc.) in each frame from the generated results are similar to the ones in the real videos. The lower value of FID is better.

**Implementation Details.** We design our ROMA as a one-sided framework, consisting of a generator and a discriminator. Following [21], we adopt the encoder-decoder architecture as our backbone network and apply this setting to all methods in our experiments for a fair comparison. The \( l_{size} \) of our multiscale region-wise discriminator is set as \([3, 5, 7]\) for diverse receptive fields. For local structural consistency, the number of areas \( N_a \) is 64 and the size of areas is \( 75 \times 75 \). For the balance of quality improvement and computation cost, \( \Delta t \) is set as 2 for \( L_{term} \). For the constraints of ROMA, the hyperparameter \( \lambda_1, \lambda_2 \) and \( \lambda_3 \) are set as 5.0, 5.0 and 1.0.

#### 5.3 Comparisons with Other Methods

We compare our method with several state-of-the-art methods of unpaired image-to-image and video-to-video translation, i.e., CUT [32], CycleGAN [21], F/SeSim [45], Recycle-GAN [2], Mocycle-GAN [8] and I2V-GAN [24]. Notably, we only use the nighttime infrared and the daytime visible videos as inputs for training. The daytime visible videos are only utilized for comparisons.

**Results of InfraredCity-Lite.** As shown in Tab. 3, we obtain state-of-the-art performance, which indicates that ROMA can substantially improve the quality of generated daytime visible frames. Especially, F/SeSim which only focuses on maintaining structure achieves similar results with the tailored infrared-to-visible translation method, i.e., I2V-GAN. It confirms the key challenge of maintaining the structural information of infrared videos. Moreover, compared with them, our ROMA outperforms them by 22.3% on average relatively on the FID metric. Especially, 35.5% improved relatively on the clear Highway scene. These improvements indicate the advantages of ROMA for generating excellent daytime visible results, especially for preserving domain-invariant structure.

| Name                | Frame | Clip | Main Task               |
|---------------------|-------|------|-------------------------|
| InfraredCity        | 579,984 | 9    | video translation       |
| InfraredCity-Lite   | 41,839 | 9    |                         |
| IRVI [24]           |       |      |                         |
| VOT2019 (RGBTIR) [30] | 20,083 | 60   | object tracking         |
| FLIR [13]           | 4,224 | 1    | object detection        |
| KAIST (DAY ROAD) [18] | 16,176 | 9    |                         |

| Method               | FID  |
|----------------------|------|
|                      | Traffic | Monitoring |
|                      | sub-1 | sub-2 | sub-3 | sub-4 | all |
| CycleGAN             | 0.5739 | 1.4448 | 0.6731 | 2.2294 | 2.0535 | 1.9793 | 1.0893 |
| F/SeSim              | 0.6714 | 1.4027 | 0.8056 | 2.1497 | 1.0359 | 1.6266 | 0.8792 |
| Recycle-GAN          | 0.4321 | 1.5135 | 0.5452 | 1.6665 | 0.9467 | 1.9594 | 0.9232 |
| Mocycle-GAN          | 0.7911 | 1.5556 | 0.9847 | 2.5013 | 1.1048 | 2.1711 | 1.0515 |
| I2V-GAN              | 0.5255 | 1.6680 | 0.7521 | 2.0387 | 1.2959 | 1.8518 | 1.0609 |
| Ours w/o \( D_p \)    | 0.4061 | 1.3427 | 0.3383 | 1.6905 | 0.6155 | 1.5241 | 0.9036 |
| Ours w/o \( L_p \)    | 0.3873 | 1.3367 | 0.3756 | 1.6163 | 0.6084 | 1.5830 | 0.7935 |
| Ours w/o \( L_t \)    | 0.3824 | 1.2759 | 0.2723 | 1.5064 | 0.5453 | 1.5402 | 0.7437 |
| Ours w/o \( L_{term} \) | 0.3633 | 1.2527 | 0.2849 | 1.5034 | 0.5518 | 1.4733 | 0.7465 |
| Ours                 | 0.3467 | 1.2301 | 0.2485 | 1.4765 | 0.5188 | 1.4438 | 0.7334 |

| Method               | FID  |
|----------------------|------|
|                      | City | Highway | Monitor |
|                      | clear | overcast | all | clear | overcast | all |
| CUT                  | 0.5809 | 0.5607 | 0.6086 | 0.4544 | 0.5133 | 0.4799 | 0.4809 | 0.9785 |
| CycleGAN             | 0.6299 | 0.5879 | 0.7125 | 0.4787 | 0.5849 | 0.4920 | 0.4204 | 0.8129 |
| F/SeSim              | 0.4984 | 0.5369 | 0.4834 | 0.5108 | 0.5288 | 0.4809 | 0.2724 | 0.8984 |
| Recycle-GAN          | 0.5942 | 0.5974 | 0.5969 | 0.5173 | 0.5998 | 0.5101 | 0.3431 | 0.9433 |
| Mocycle-GAN          | 0.5117 | 0.5346 | 0.5011 | 0.5029 | 0.5976 | 0.4791 | 0.3163 | 0.7298 |
| I2V-GAN              | 0.5052 | 0.5574 | 0.6469 | 0.5064 | 0.5305 | 0.4515 | 0.2872 | 0.7019 |
| Ours w/o \( D_p \)    | 0.4765 | 0.5358 | 0.4994 | 0.4623 | 0.4926 | 0.4903 | 0.2635 | 0.7322 |
| Ours w/o \( L_p \)    | 0.4601 | 0.5296 | 0.4028 | 0.4133 | 0.3970 | 0.3924 | 0.2493 | 0.6015 |
| Ours w/o \( L_t \)    | 0.4222 | 0.5244 | 0.4117 | 0.3935 | 0.4601 | 0.3992 | 0.2174 | 0.6913 |
| Ours w/o \( L_{term} \) | 0.4295 | 0.5013 | 0.4699 | 0.3872 | 0.4583 | 0.4115 | 0.2251 | 0.5751 |
| Ours                 | 0.4018 | 0.5149 | 0.3929 | 0.3325 | 0.3823 | 0.3444 | 0.2002 | 0.5488 |
ROMA: Cross-Domain Region Similarity Matching for Unpaired
Nighttime Infrared to Daytime Visible Video Translation

Figure 7: Qualitative comparisons of different methods in InfraredCity-Lite. Our ROMA has obvious advantages in detail.

Figure 8: Translation results of Highway. ROMA takes advantage of preserving structural and temporal consistency. We highlight key error-prone regions via red bounding boxes.

Results of IRVI. Our main evaluation is thus against the state-of-art I2V-GAN on IRVI. The results of the comparison are presented on Tab. 4. Notably, our ROMA achieves the top-ranked FID performance on all scenes and outperforms I2V-GAN 26.2% on average relatively, surpassing by 57.9% relatively on the sub-2 scene. This again validates our effectiveness in maintaining infrared structure.

Ablation Study. We conduct ablation experiments on both two datasets to study how each design of ROMA influences the whole framework and the results are shown in Tab. 3 and Tab. 4. Our region-wise cross-similarity matching constraints ($L_g$, $L_l$ and $L_{tem}$) have great contributions to improving the quality of generated visible results. Notably, the improvement of multiscale region-wise discriminator is not inferior to cross-similarity.

Qualitative Comparisons. Furthermore, we make a qualitative comparison with other approaches on the InfraredCity-Lite dataset. From Fig. 7 and Fig. 8, it is observed that the synthesized daytime visible frames by ROMA have better visual performance compared with other methods. The constraints of other approaches help the generator learn the structural information comfortably at the beginning but heavily as the training process proceeds due to the prominent difference between source and target styles. In contrast, our cross-similarity matching technique gets rid of the negative influence of style all the time, which helps the generator to learn structural information consistently and comfortably.

Furthermore, as depicted in Fig. 8, the baselines all generate visible frames whose overall color is somewhat inaccurate and especially disappoint us on the generated truck. In contrast, ROMA demonstrates pleasant results which are more analogous to the style of real daytime visible videos. Besides, favorable temporal coherence in our results can be observed in these results.

More comparisons are displayed in the supplementary material.
Figure 10: Comparisons of vehicle detection results on nighttime infrared, nighttime visible and translated daytime visible. Detection tasks are performed by the pre-trained YOLOv3 [36] model. Our goal is to generate the translated results as similar as possible to actual visible ones, especially in detail.

Training Time. Besides, we make a comparison on efficiency since it is critical for applications in reality. Fig. 9 (a) showcases the efficiency of the different methods under a unified standard. Our ROMA achieves the best performance compared with other approaches. Especially, to obtain the best FID scores of other approaches, only approximately 11.7 hours are cost by ROMA, which is about 11 times faster than F/LeSeSim. Besides, I2V-GAN consists of many hand-designed constraints from feature level and pixel level for spatial and temporal consistency. These bring improvements yet arduous convergence. In ROMA, the designed domain-invariant representation, cross-similarity not only maintains structure well but also helps the generator learn the structural information candidly. The results of Fig. 9 (b) make a further demonstration of the efficiency of our ROMA. It indicates that ROMA gains excellent details even in the early stage of training.

Table 5: Comparison of YOLO scores (%) for vehicle detection.

| Scenes               | Nighttime Infrared | Nighttime Visible | I2V-GAN      | ROMA (Ours) |
|----------------------|--------------------|-------------------|--------------|-------------|
| AP                   | 25.0               | 26.1              | 32.2         | 50.1        |

5.4 Object Detection

Object detection is a fundamental and core problem in computer vision. Studies [28, 35, 36] achieve remarkable improvements with the advantages of large-scale annotated datasets. However, the object detection model is often vulnerable to data variance, especially when operating at night. In such cases, the translation from stable nighttime infrared to daytime visible is an ideal solution.

For further evaluating the quality of generated results, we utilize the YOLO score to evaluate the generated results. From Fig. 10, we can observe that the translated daytime videos indeed favor the detection. Besides, the improvements of our ROMA are detailed on Tab. 5 via YOLO scores. The more vivid the generated vehicles and videos are, the more accurate the object detection will be. Our

5.5 Video Fusion

Generally, the infrared (IR) vision technique is adopted for the context enhancement in nighttime vision by fusing it with the visible (VI) image [23, 27, 46]. However, IR/VI image fusion methods are only ideals at dawn when the visible camera can still capture the relatively clear visible scene. Additionally, it is challenging for fusing infrared videos with visible ones since they are not paired at the pixel level and videos contain the nature of temporal coherence.

On the contrary, our translated daytime visible results are clear and match the infrared input videos at the pixel level. Observed from Fig. 11, the fusion results of nighttime infrared frames and translated daytime visible frames are more clear and semantic compared with the ones of nighttime infrared and nighttime visible frames.

6 CONCLUSION

In this paper, we introduce a tailored framework ROMA to translate the unpaired nighttime infrared videos into fine-grained daytime visible ones via the proposed cross-domain region similarity matching technique, which effectively transferring the structural knowledge of the infrared data and preserving the spatiotemporal consistency. To further enhance the reality of translated videos, a multiscale region-wise discriminator is designed, and extensive experiments validate that ROMA obtains state-of-the-art performance when producing unambiguous daytime visible videos. Moreover, tests on nighttime object detection and video fusion tasks demonstrate that ROMA can generate reliable results for night vision applications. Besides, we provide a challenging dataset for nighttime infrared to daytime visible video translation, i.e., InfraredCity, and we hope this will encourage more researches in this area.

ACKNOWLEDGEMENT

This paper was supported by National Key R&D Program of China (No. 2021YFB3301503), and also supported by the National Natural Science Foundation of China under Grant No. U21A20519.
REFERENCES

[1] 2020. PCGAN. Perceptual cyclic-synthesized generative adversarial networks for thermal and NIR to visible image transformation. Neurocomputing.

[2] Bansal Aayush, Ma Shugao, Ramanan Deva, and Sheikh Yaser. 2018. Recycle-GAN: Unsupervised Video Retargeting. (2018).

[3] Dina Bashkirova, Ben Usman, and Kate Saenko. 2018. Unsupervised Video-to-Video Translation. CoRR (2018).

[4] Durga Prasad Bavarisetty, Gang Xiao, Junhao Zhao, Ravindra Dhilli, and Gang Liu. 2019. Multi-scale Guided Image and Video Fusion: A Fast and Efficient Approach. Circuits Syst. Signal Process. (2019), 5576–5605.

[5] Sagie Benaim and Lior Wolf. 2017. One-Sided Unsupervised Domain Mapping. In NeurIPS. Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 752–762.

[6] Sheng Bin, Sun Hanqui, Magnor Marcus, and Li Ping. 2014. Video Colorization Using Parallel Optimization in Feature Space. IEEE Trans. Cir. and Sys. for Video Technol. (2014).

[7] QiFeng Chen and Vladlen Koltun. 2017. Photographic Image Synthesis with Cascaded Reﬁnement Networks. In ICCV. 1520–1529.

[8] Y Chen, Y Pan, T Yao, X Tian, and T Mei. 2019. Mocycle-GAN: Unpaired Video-to-Video Translation. In ACCM/MM.

[9] Szegedy Christian, Vanhoucke Vincent, Ioffe Sergey, Shlens Jon, and Wojna Zbigniew. 2016. Rethinking the inception architecture for computer vision. In CVPR. 2818–2826.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In CVPR. 248–255.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aozhu Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In ICLR.

[12] Alexey Dosovitskiy and Thomas Brox. 2016. Generating Images with Perceptual Similarity Metrics based on Deep Networks. In NeurIPS. Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 658–666.

[13] FLIR. 2018. FREE FLIR Thermal Dataset for Algorithm Training. https://www.flir.com/oem/adas/adas-dataset-form/

[14] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. 2019. Geometry-Consistent Generative Adversarial Networks for One-Sided Unsupervised Domain Mapping. In CVPR. 2427–2436.

[15] Raj Kumar Gupta, Alex Yong Sang Chia, Deepu Rajan, Ee Sin Ng, and Zhiyong Chanyong Jung, Gihyun Kwon, and Jong Chul Ye. 2022. Exploring Patch-wise Semantic Relation for Contrastive Learning in Image-to-Image Translation Tasks. CoRR (2022).

[16] Joseph Redmon. 2018. YOLOv3: An Incremental Improvement. CoRR (2018).

[17] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In ICLR.

[18] Shuang Li, Bingfeng Han, Zhongyi Yu, Chi Harold Liu, Kai Chen, and Shuigen Wang. 2021. i2V-GAN: Unpaired Infrared-to-Visible Video Translation. In ACM MM, Heng Tao Shan, Yueting Zhuang, John R. Smith, Yang Yang, Pablo Cesar, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 3061–3069.

[19] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. Video-to-Video Synthesis. In NeurIPS.

[20] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2019. Few-shot Video-to-Video Synthesis. In NeurIPS. 1153–1160.

[21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In ECCV. 21–37.

[22] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. 2018. The Contextual Loss for Image Transformation with Non-aligned Data. In ECCV. 800–815.

[23] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. 2018. The Contextual Loss for Image Transformation with Non-aligned Data. In ECCV. 800–815.

[24] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2019. Few-shot Video-to-Video Synthesis. In NeurIPS. 1153–1160.

[25] Wei Liu, Vijay John, Erik Blasch, Zheng Liu, and Ying Huang. 2018. IR2VL: Enhanced Night Environmental Perception by Unsupervised Thermal Image Translation. In CVPR.

[26] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. CoRR (2018).

[27] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. CoRR (2018).

[28] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In ECCV. 21–37.

[29] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. 2018. The Contextual Loss for Image Transformation with Non-aligned Data. In ECCV. 800–815.

[30] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. 2018. The Contextual Loss for Image Transformation with Non-aligned Data. In ECCV. 800–815.

[31] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. 2018. The Contextual Loss for Image Transformation with Non-aligned Data. In ECCV. 800–815.

[32] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. 2018. The Contextual Loss for Image Transformation with Non-aligned Data. In ECCV. 800–815.

[33] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. 2018. The Contextual Loss for Image Transformation with Non-aligned Data. In ECCV. 800–815.

[34] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. 2018. The Contextual Loss for Image Transformation with Non-aligned Data. In ECCV. 800–815.

[35] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. 2018. The Contextual Loss for Image Transformation with Non-aligned Data. In ECCV. 800–815.

[36] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. 2018. The Contextual Loss for Image Transformation with Non-aligned Data. In ECCV. 800–815.

[37] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. 2018. The Contextual Loss for Image Transformation with Non-aligned Data. In ECCV. 800–815.

[38] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. 2018. The Contextual Loss for Image Transformation with Non-aligned Data. In ECCV. 800–815.

[39] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. 2018. The Contextual Loss for Image Transformation with Non-aligned Data. In ECCV. 800–815.

[40] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. 2018. The Contextual Loss for Image Transformation with Non-aligned Data. In ECCV. 800–815.