ReCoNet: Recurrent Correction Network for Fast and Efficient Multi-modality Image Fusion

Zhanbo Huang[1], Jinyuan Liu[2], Xin Fan[1], Risheng Liu[1,3], Wei Zhong[1], and Zhongxuan Luo[1]

[1] DUT-RU International School of Information Science & Engineering, Dalian University of Technology
[2] School of Software Technology, Dalian University of Technology
[3] Peng Cheng Laboratory
{zbhuang917,atlantis918}@hotmail.com,
{xin.fan,rsliu,zhongwei,zxluo}@dlut.edu.cn

Abstract. Recent advances in deep networks have gained great attention in infrared and visible image fusion (IVIF). Nevertheless, most existing methods are incapable of dealing with slight misalignment on source images and suffer from high computational and spatial expenses. This paper tackles these two critical issues rarely touched in the community by developing a recurrent correction network for robust and efficient fusion, namely ReCoNet. Concretely, we design a deformation module to explicitly compensate geometrical distortions and an attention mechanism to mitigate ghosting-like artifacts, respectively. Meanwhile, the network consists of a parallel dilated convolutional layer and runs in a recurrent fashion, significantly reducing both spatial and computational complexities. ReCoNet can effectively and efficiently alleviates both structural distortions and textural artifacts brought by slight misalignment. Extensive experiments on two public datasets demonstrate the superior accuracy and efficacy of our ReCoNet against the state-of-the-art IVIF methods. Consequently, we obtain a 16% relative improvement of CC on datasets with misalignment and boost the efficiency by 86%. The source code is available at https://github.com/dlut-dimt/reconet.

Keywords: Deep Learning, Multi-modality Image Fusion

1 Introduction

Infrared and Visible Image Fusion (IVIF) generates a fused image presenting complementary characteristics and having richer information than either modality. The generated image is visually appealing and more importantly favorable for practical applications such as video surveillance [31], remote sensing [29,30], and autonomous driving [6,40].

Conventional IVIF methods strive to find optimal representation of common features across modals and then to design appropriate weights for merging [17,28]. Recently, the community has witnessed the great success of deep learning in various artificial intelligent applications due to its strong ability in

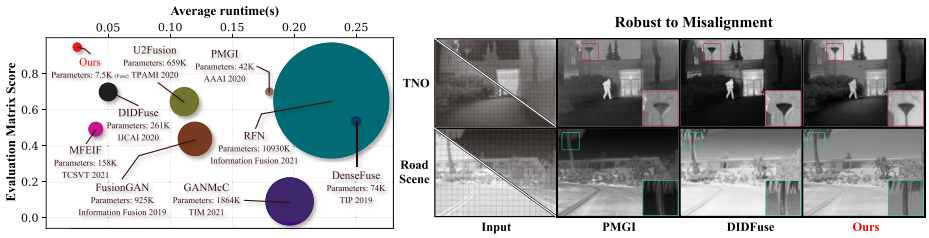


Fig. 1. Comparisons of computational complexity and robustness on the TNO and RoadScene datasets Our method outperforms all its counterparts with higher evaluation scores, lower average runtime, fewer training parameters, and is more robust to misalignment.

nonlinear fitting and feature extraction. Researchers employ deep networks to learn mutual features [23,21,27,41,26,44] or fusing strategies given training examples for IVIF [13,14,19,16]. These approaches can produce favorable fusion especially for human inspection in controlled scenarios, *e.g.*, fixed capturing devices and/or well-aligned input images [13,19]. Unfortunately, two vital issues still remain unresolved for existing IVIF methods in order to significantly foster subsequent Computer Vision (CV) tasks including object detection [24,47,46], tracking [12,43,2], and semantic segmentation [5,33,9].

First, existing IVIF approaches, either conventional [36,4] or deep-learning based ones [42,22,38], are typically sensitive to misalignment on input images. Slight shifts or deformations on one modality bring evident geometrical distortions on image structures and ghosting-like artifacts in the regions of textural details, as shown in Fig. 1, which substantially deteriorate downstream CV algorithms. Only a tiny fraction of works attempt to mitigate these unpleasant effects. Ma *et al.* [25] proposed total variation minimization that separately strengthens geometric structures in infrared images and preserves textures in visible inputs. However, these methods evidently smear details without exploiting complementary information between these two modalities. Additionally, its iterative optimizing process demands intensive gradient computations resulting in time-consuming fusion. Other deep-learning based methods [19,15,18,20,10] incorporate the attention/mask mechanism to bolster the misalignment's robustness, avoiding artifacts by reducing the weight of mismatched patches. Yet these attention/mask mechanisms have difficulty portraying correlations across different modalities, resulting in small, tiny artifacts in their fused results.

Second, the state-of-the-art methods demand a large space to store numerous network parameters and lag behind real time running, as illustrated in circles and time values in Fig. 1, though deep methods accelerate fusion with a large margin over conventional approaches. The major bottleneck lies in that these deep have to stack multiple layers of convolutional blocks to learn common features shared by infrared and visible images presenting significant differences on appearance. Meanwhile, training these huge networks require a large number of image pairs unavailable in practice.

This study addresses these two critical issues by developing a recurrent light network that effectively and efficiently corrects both structural distortions and textural artifacts brought by misalignment. Specifically, we train a micro registration module ($\mathcal{R}$) to predict deformation fields between input images. This module explicitly corrects distortions on geometrical structures caused by pixel shifts. We also learn attention maps from both modalities (σ_{ir} and σ_{vis}) that discover the salient regions in respective inputs. Hence, textures in the visible input weigh more in the fusion process while differentiate repeated patterns of high frequency caused by spatial offsets, thus implicitly attenuating ghosting artifacts. Catering for high efficiency, we design a parallel dilated convolutional layer (PDC) that learns contextual information with multiple scale receptive fields. We train one set of parameters of this simple PDC layer and recurrently run the network ($\mathcal{F}$) cascading the attention and lightweight PDC modules in the fusion workflow of Fig. 3. This recurrent process saves the space for network parameters and iteratively improves fusion quality. Fig. 1 demonstrates that our approach achieves higher numerical scores, lower computation costs, and fewer parameters on two public available datasets compared with the state-of-the-art. We summarize the our main contributions as follows:

- To our best knowledge, this is the first work to jointly learn deep networks for both registration and fusion on mid-wave infrared and visible images, which enables generating images robust to misalignment of sources.
- We design a deformation module to explicitly compensate geometrical distortions and an attention mechanism to mitigate remaining ghosting-like artifacts. This design properly tackles two different types of undesired effects occurring in structural and textural regions of a given scene, respectively.
- We develop a parallel dilated convolutional layer and a recurrent mechanism, significantly reducing both spatial and computational complexities.

2 The Proposed Method

In this section, we will introduce our motivation and the network architecture of our ReCoNet. In addition, the loss function is also illustrated in the following.

2.1 Motivation

In real-life scenarios, pixel-level registered infrared and visible images are unavailable caused by insuperable internal and external factors. As illustrated in Fig. 2, we show three typical factors that frequently occur in genuine acquisitions. (i) In most of the encapsulated devices, supposing the internal systems have been working for an extended period or in a high-temperature internal environment, the Complementary Metal-Oxide-Semiconductor (CMOS) produces noises into the image. (ii) For the server environments, *e.g.,* desert and tropical forest, the refraction of hot airflow may cause severe distortion on the source images. (iii) The bumpy roads, fast-moving objects, or non-synchronous multi-vision cameras may degenerate the source images [45], *e.g.,* motion blur and

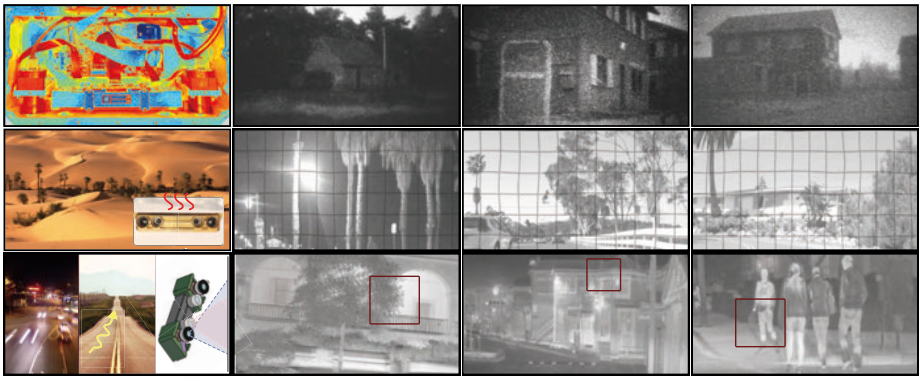


Fig. 2. Three representative misalignment situations that occur in actual scenarios.

transportation. Slight shifts or deformations on one modality bring evident geometrical distortions; few existing methods can overcome these issues because they only perform fusion on pixel-level registered pairs. Based on this observation, we raise a recurrent correction network for realizing IVIF, which has sufficient capacity to deal with sight misalignment source inputs.

Apart from that, most previous fusion approaches take every effort to strengthen the network with a crease of depth and width, achieving state-of-the-art performance. However, these catastrophic increases of network layer may lead to a significant requirement of computation and memory, thus making them difficult to apply them in the follow-up high-level computer vision tasks, *e.g.,* object detection, depth estimation, and object tracking. Consequently, a parallel dilated convolutional layer and a recurrent learning mechanism are sophisticatedly designed in our method to boost computational efficiency.

2.2 Micro Registration Module

The micro registration module $\mathcal{R}$ contributes to alleviate the slight misalignment errors cased by geometric distortions or scaling. It consist of two components: a deformation field prediction network $\mathcal{R}_\phi$ and a re-sampler layer $\mathcal{R}_S$. The deformation field ϕ is employ to represent the transformation, which allow our method to map images non-uniformly accurately.

Supposing given an infrared image x and a distorted visible image $\tilde{y}$, $\mathcal{R}_\phi$ aims to predict a deformation field $\phi_{\tilde{y}\to y} = \mathcal{R}_\phi(x, \tilde{y})$, describing how to align $\tilde{y}$ to y non-rigidly. The deformation field $\phi \in \mathbb{R}^{h\times w\times 2}$, in which each pair $\phi_{h,w} = (\Delta x_h, \Delta x_w) \in \mathbb{R}^2$ indicates the deformation offset for the (h, w) pixel $v_{h,w}$ in $\tilde{y}$. Our R mainly focuses on the fusion effect after registration, so that an U-Net like micro module is designed. The detailed architecture is given in the bottom-left conner of Fig. 3.

To apply geometric transformations to the image, we use a re-sampler layer $\mathcal{R}_S$ which takes the deformation field $\phi_{\tilde{y}\to y}$ generated by $\mathcal{R}_\phi$ and applies it to

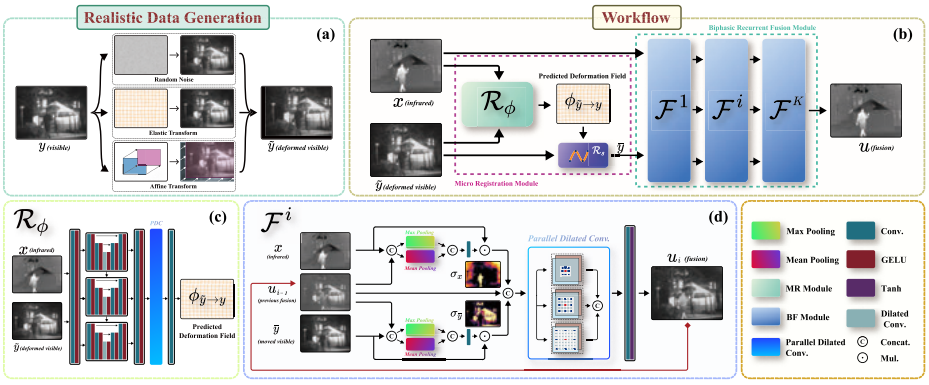


Fig. 3. Methodology framework: (a) pseudo-distortion data generation; (b) our ReCoNet workflow; (c) micro-registration (MR) module architecture; and (d) pipeline of biphasic-recurrent fusion (BF) module.

the distorted visible image $\tilde{y}$. The value of the transformed visible image $\bar{y}$ at pixel $v_{h,w}$ is calculated by the equation:

$$\bar{y}\left[v_{h,w}\right] = \tilde{y}\left[v_{h,w} + \phi_{h,w}^{\tilde{y}\to y}\right]. \tag{1}$$

2.3 Biphasic Recurrent Fusion Module

Contextual features (*e.g.,* edges, targets, and contours) play a vital role in the fusion process. However, with an increase of the network's depth, the contextual features degrade gradually, resulting in blurred targets and unclear details on the fusion results. To deal with this issue, previous works attempt to design various attention mechanisms or bring in enlarge the width of network (*e.g.,* adding dense or residual blocks). Actually, such aforementioned attention mechanisms have difficulty characterizing contextual features from the source images. The increasingly model architecture may lead to a significant requirement of computation and memory. Thus, we propose a biphasic recurrent fusion module to acquire high computational efficiency for sufficient contextual features representation at multiple scales.

Baphasic Attention Layer: To obtain the salient features and keep contextual consistency with the source images, a biphasic attention layer is proposed. It is composed of a max-pooling operation, an average-pooling operation, and a convolutional layer without bias. The maximum and average values of the pixels at each point of the two images are taken and combined as the input of the convolutional layer. Let $\mathcal{A}$ denote the biphasic attention layer, I_a and I_b as two input images, respectively, this process can be expressed as the following equation:

$$\mathcal{A}(I_a, I_b) = \theta_{\mathcal{A}} * \left[\texttt{max}(I_a, I_b), \texttt{avg}(I_a, I_b)\right],$$

where $*$ denotes the convolution operation, $\theta_{\mathcal{A}}$ denote the parameter of the convolutional layer in our attention layer, and we concat $\texttt{max}(I_a, I_b)$ and $\texttt{avg}(I_a, I_b)$ as the input of attention layer. As shown in Fig. 3, the network computes attention map σ_x and σ_y from input images group $\{x, u, \bar{y}\}$ according equation:

$$\sigma_{ir} = \mathcal{A}_x(x, u_i) \quad \sigma_{\bar{y}} = \mathcal{A}_{\bar{y}}(\bar{y}, u_i),$$

where $\mathcal{A}_x$ and $\mathcal{A}_y$ denote infrared and visible attention layer, respectively, while the u_i indicates the fused result of the last recurrence.

In addition, thanks to biphasic attention map, we can obtain more desired emphasis on contextual features and also make our method implicitly compatible with slight alignment errors by reducing the weight of the slightly distorted region such as non-smooth edge and ghosts.

Parallel Dilated Convolutional Layer: We develop a parallel dilated convolutional layer to extract features from the source images efficiently. A group of dilated convolutional layers with sawtooth wave-like dilated factors increases receptive field without losing neighboring information. Convolutions with the same kernel size 3×3 on three dilated paths have their receptive field with different dilated factors. As shown PDC in Fig. 3, the dilation rates are set as $1, 2, 3$ receptively. Thus, the three parallel convolutional paths have receptive fields of 3×3, 5×5, and 7×7.

To provide a formal description, let f_{in}^i denote the input for the dilated convolutional layers at the i-th recurrence. The output feature map f_{out}^i of the recurrent parallel dilated convolutional layers is gradually updated as follows:

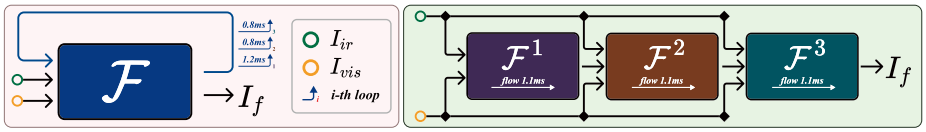
$$f_{out}^i = \{\mathcal{C}^k(f_{in}^i)\}_{k\in\{1,2,3\}}, \mathcal{C}(f_{in}^i) = \theta_{\mathcal{C}}^k * f_{in}^i + \mathbf{b}_{\mathcal{C}}^k,$$

where $\theta_{\mathcal{C}}^k$ and $\mathbf{b}_{\mathcal{C}}^k$ denote the parameter and bias of the convolutional layer with dilation rate equaling k.

Recurrent Learning: We raise a recurrent architecture to replace time-consuming multi-layer convolution to extract contextual features from a coarse-to-fine manner. We can reduce the computational complexity overhead of building the graph by partially reusing the computational graph for dynamically graph network frameworks such as PyTorch[32]. As shown in Fig. 4, compared to a series network structure, we will spend a little more time in the first loop used to build the graph than the no-loop structure, but for each subsequent loop, we will save about 27% of the time. Overall, our recurrent architecture reduces about 15% of the time, 33% parameters, and 42% GPU memory. Such recurrent learning allows ReCoNet to extract image features from contextual information and meet real-time standards($\geq 25fps$) [11]. Due to the parameters and memory reduction, our ReCoNet can be deployed on mobile devices.

2.4 Loss Functions

The total loss function $\mathcal{L}_{\text{total}}$ of our network is accomplished to two loss term, the fusion loss $\mathcal{L}_{\text{fuse}}$ and the registration loss $\mathcal{L}_{\text{reg}}$. The fusion loss ensures the network to generate the fused result with better effects and rich information, while



(a) Recurrent (b) Series

Fig. 4. Efficiency comparison of our recurrent and series architectures with the same floating point operations (FLOPs).

the registration loss contributes to constrain and refine image distortion caused by misalignment. We train our network minimize the following loss function:

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{fuse}} + (1-\lambda)\,\mathcal{L}_{\text{reg}}, \tag{2}$$

where λ is a trade-off parameter.

The fusion loss consists of two loss terms. Structure similarity $\mathcal{L}_{\text{SSIM}}$ is employed to maintain the over structure from light, contrast and structure information aspect, while $\mathcal{L}_{\text{pixel}}$ is used to balance the pixel intensity of the two source images. Therefore, $\mathcal{L}_{\text{fuse}}$ is expressed as:

$$\mathcal{L}_{\text{fuse}} = \gamma \mathcal{L}_{\text{SSIM}} + (1-\gamma)\,\mathcal{L}_{\text{pixel}}, \tag{3}$$

where γ is the weight of two loss items. Specifically, we constrain our fused result to have the same fundamental architecture as the source images, and hence the $\mathcal{L}_{\text{SSIM}}$ loss is defined as:

$$\mathcal{L}_{\text{SSIM}} = (1 - SSIM\,(u,x)) + (1 - SSIM\,(u,y))\,. \tag{4}$$

Similarly, the fused results should balance the pixel intensity distribution from both infrared and visible images, and the pixel loss can be formulated as:

$$\mathcal{L}_{\text{pixel}} = \|u-x\|_1 + \|u-y\|_1\,, \tag{5}$$

where $\|\cdot\|_1$ denotes the $l1$-norm.

Apart from that, the registration loss $\mathcal{L}_{\text{reg}}$ also plays a key role in correcting the distortion, which can be expressed as:

$$\mathcal{L}_{\text{reg}} = \eta \mathcal{L}_{sim} + (1-\eta)\,\mathcal{L}_{\text{smooth}}, \tag{6}$$

where $\mathcal{L}_{sim}$ denotes similarity loss, and $\mathcal{L}_{\text{smooth}}$ is a smoothing loss that targets ensure to generate a smooth deformation. η is trade-off parameter in balancing the two terms.

More precisely, $\mathcal{L}_{sim}$ is calculated as:

$$\mathcal{L}_{\text{sim}} = \|\phi_{\tilde{y}\to y} - (-\phi_{y\to\tilde{y}})\|_2^2\,, \tag{7}$$

where $\phi_{\tilde{y}\to y}$ denotes deformation field and $\phi_{y\to\tilde{y}}$ expresses the generated random deformation field, respectively. As our framework mainly focuses on the fusion

effect after registration, the $-\phi_{y\to\tilde{y}}$ is roughly used as the ground truth of the deformation field for our register to converge. These slight errors introduced in this process will be eliminated in our recurrent fusion mechanism.

For each voxel p in 2D spatial domain Ω, the $\mathcal{L}_{smooth}$ can be specifically defined as:

$$\mathcal{L}_{\text{smooth}} = \sum_{p\in\Omega} \|\nabla\phi\,(p)\|_1^1\,, \tag{8}$$

where ∇ denotes the approximate spatial gradients using differences between neighboring voxels.

3 Experiments and Results

We first introduce the datasets, evaluation metrics and training details. Then we compare the proposed method against a broad range of the eight state-of-the-arts methods (*i.e.,* DenseFuse [13], FusionGAN [26], RFN [14], GANMcC [27], MFEIF [19], PMGI [44], DIDFuse [16] and U2Fusion [41]) on aligned/misaligned dataset, respectively. Besides, we also provide the complexity evaluation, the mean opinion score analysis and the extensive ablation experiments. All experiments are conducted with Pytorch on a computer with Nvidia V100 GPU.

3.1 Dataset and Preprocessing

Dataset: Both our aligned and misaligned fusion experiments are conducted on the TNO [37] and RoadScene [41] datasets. We generate infrared images with different degrees of distortion by randomly using deformation field. In each aligned/misaligned IVIF experiment, we randomly selected 20/180 pairs of images their corresponding TNO/RoadScene datasets as training samples.
Evaluation Metrics: We employ three existing statistical metrics including standard deviation (SD), entropy (EN) and correlation coefficient (CC), to comprehensively evaluates the quality of the fused images from different aspects.
Training Details: The λ, γ and η are set as 0.6, 0.28, and 0.78, respectively. The Adam optimizer updates the parameters with the learning rate of 0.001 and a total epoch of 300. The micro registration $\mathcal{R}_\phi$ and the biphasic recurrent fusion module $\mathcal{F}$ are jointly trained.

3.2 Results on Aligned Dataset

Qualitative Comparisons: Fig. 5 shows eight representative fused images generated by different models. Visual inspection shows that, our method has obvious advantages over the comparative models. Although other methods achieve meaningful fused results, they still remain problems, such as unclear thermal targets (see the green box in the Fig. 5 of DenseFuse, RFN and U2Fusion), blurred details (see the red box in the Fig. 5 of GANMcC and DIDFuse). On the contrary, our method can generate visual-friendly fused results with clear target, distinct contrast and abundant details.

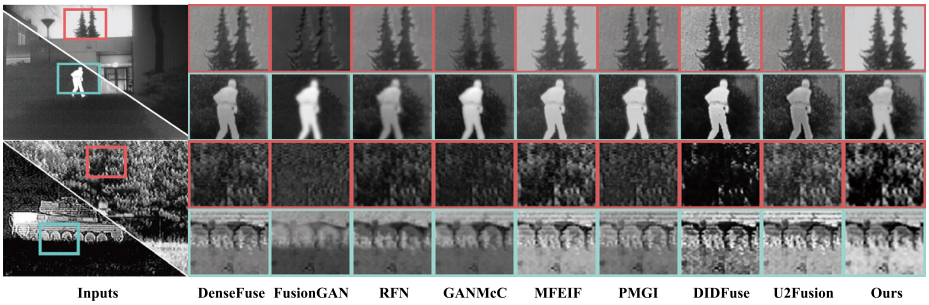


Fig. 5. Visual comparisons of our ReCoNet with state-of-the-art methods on the aligned TNO dataset.

Quantitative Comparisons: Subsequently, the quantitative results on 15/40 image pairs of the TNO/RoadScene dataset are shown in Fig. 6. Obviously, our method reach the values for two metrics (SD and EN), followed by DIDFuse and U2Fusion. For the CC metric, our method only follows behind FGAN by a narrow margin on the TNO dataset.

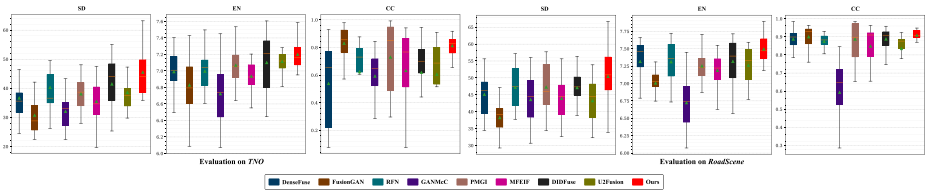


Fig. 6. Quantitative comparisons with eight IVIF methods on TNO and RoadScene datasets, respectively. In the boxes, the orange lines and the green tangles denote medium and mean values.

3.3 Results on Slightly Misaligned Dataset

Qualitative Comparisons: As our method has the ability to fuse image pairs with slightly misalignment, we further test its fusion performance against other state-of-the-art methods on sight misaligned TNO and RoadScene datasets, respectively, which is shown in Fig. 7. Obviously, other methods suffer structural distortion or undesirable halos on their fused results. By comparison, our method overcomes the limitation of undesirable artifacts caused by misalignment in image pairs to a certain degree. This mainly benefits from the structure refinement and recurrent attention module in the training process.

Quantitative Comparisons: As shown in Fig. 8, we evaluated the CC metric of these methods on selected 20 images from TNO/RoadScene dataset with four different tranform: random noise, elastic transform, affine transform and mixed

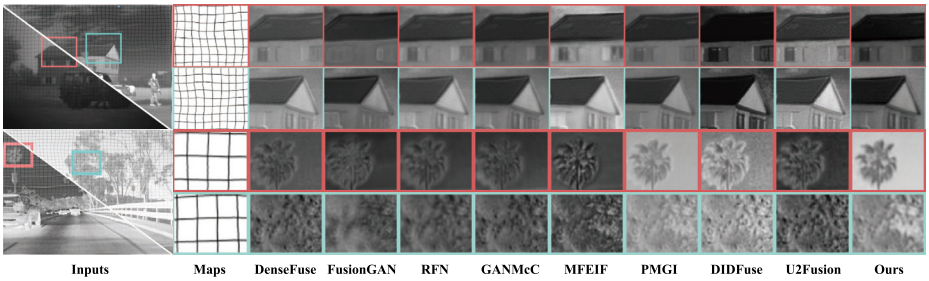


Fig. 7. Visual comparison of our method with eight state-of-the-art methods on the slightly misaligned TNO and Roadscene datasets.

transform. It is easy to notice that as the transform applied to the input images, the scores of DenseFuse, PMGI, DIDFuse and U2Fusion drop down dramatically. As the MFEIF that uses the Attention mechanism, it exhibits some resistance to random noise. Since FusionGAN is a gradient transfer-based method, the elastic transform does not disturb it much. On the contrary, our method has a strong ability to deal with all four transforms.

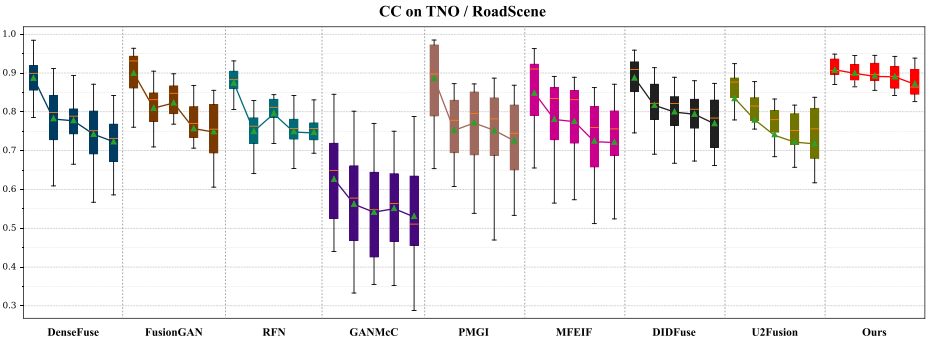


Fig. 8. CC matrix comparison with eight IVIF methods on the TNO and RoadScene dataset. The five scores in each method group represent, from left to right: original, datasets with random noise, with elastic transform, with affine transform dataset and with mixed transform dataset, respectively.

3.4 Computational Complexity Analysis

As shown in Table. 1, a complexity evaluation is introduced to evaluate the efficiency of our method from three aspects, *i.e.,* training parameters, FLOPs and runtime. It worth to pointing that our method have the fastest average running speed and minimum size. This indicates the efficiency of our ReCoNet, which can serve practical vision tasks well.

Table 1. Computational efficiency comparison with a series of completive CNN-based methods, the value is tested on GPU.

Methods	DenseFuse	FusionGAN	RFN	GANMcC	PMGI	MFEIF	DIDFuse	U2Fusion	Ours$_F$	Ours$_{R-F}$
SIZE(M)	0.074	0.925	10.93	1.864	0.042	0.158	0.261	0.659	**0.007**	0.209
FLOPs(G)	48.96	497.76	-	1002.56	745.21	25.32	18.71	366.34	**1.162**	12.54
TIME(s)	0.251	0.124	0.238	0.246	0.182	0.045	0.055	0.123	**0.024**	0.052

3.5 Mean Opinion Score Analysis

We selected 20 typical image pairs from each dataset (*i.e.,* aligned/misaligned TNO/RoadScene) for the subjective experiment. Ten computer vision researchers rated the fused images' overall visual perception, target clarity, and detail richness. Fig. 9 shows the sorted mean opinion score of all methods after normalization. Note that our method gets the highest rate for both groups, indicating the outstanding visual perception effects.

We conduct the additional subjective experiment on the aligned/misaligned TNO/Roadscene dataset of these eight IVIF methods, in which we select 20 typical image pairs from each dataset. The misaligned datasets are generated by transforming the infrared image with three kinds of transformation methods (*i.e.,* affine, elastic and both of them). We have found ten computer vision researchers, to provide a score from three aspects (*i.e.,* overall visual perception, target clarity and richness of details) for the fused image. Fig. 9 shows the sorted mean opinion score (MOS) of all methods after normalization, in which the shade of the color indicates the level of the score (yellow: the best, purple: the worst). Note that our method acquires the highest score towards all the testing image pairs, which indicates our method is more in line with the human visual system.

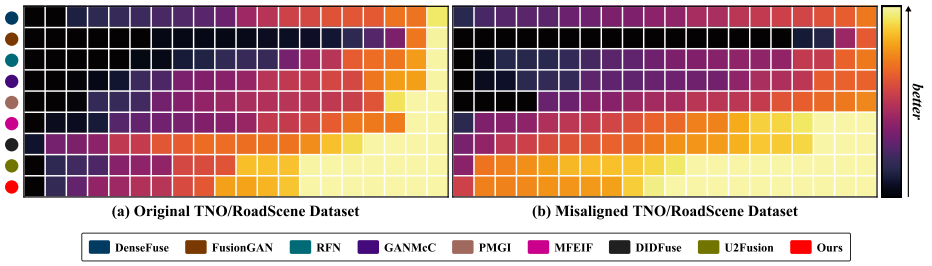


Fig. 9. Heat maps of MOS towards all methods on 20 typical image pairs from aligned and slightly misaligned datasets, respectively. Note that our method achieves more significant advantages when fusing the misaligned pairs.

3.6 Ablation Studies

Discuss The Iteration in Attention Module: Fig. 10 exhibits the effect of iterations in recurrent attention learning the on the fusion result. According to the fused results, we discover that with the increase of the number of iteration in our attention module, the fused results tend to achieve a better visual effects.Both texture details and targets are become more clearly. This mainly benefits from the progressive recurrent attention module, which allows each iteration to have a positive effect on the fused result.

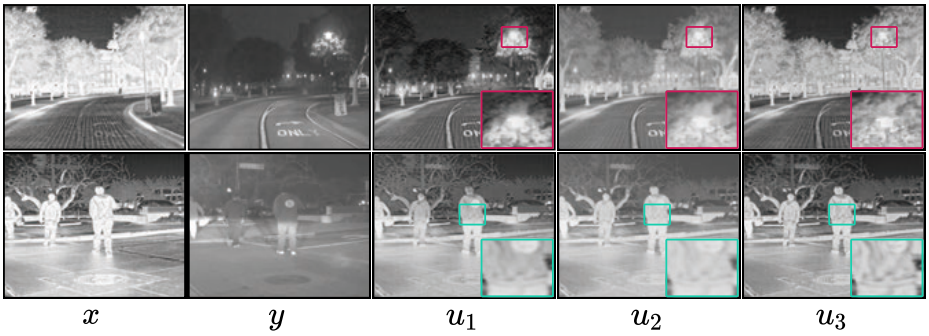


Fig. 10. A step-by-step visual result of our recurrent learning mechanism.

Ablation of Our Attention Mechanism To validate the benefit of our attention module, we pick out the of attention and corresponding ablation study in Fig. 11. We can discover that our attention module perceive the most discriminative regions (*i.e.,* targets in the infrared image and details in the visible images) form the source images, and hence the fused results keep more meaningful information.

Ablation of Our Deformable Alignment Module: To investigate the effect of deformable alignment module, we present the visual results of with/without deformable alignment module in Fig. 12. Obvious that the unfavorable artifacts appears on the fusion result without the attention module (see road sign in the second row and flagpole in the bottom row). In contrast, our method can overcome ghosting halos and structure distortion to a certain degree.

3.7 Applications in Related Tasks

This section experiments with our ReCoNet in conjunction with a series of related follow-up applications on the RoadScene dataset covering day and night scenarios.

Salient Object Detection: Extracting critical information about a target scene under a harsh environment is a challenging task. We carry out our experiment based on U^2-Net[34]. Taking an example from Fig. 13, the bright light

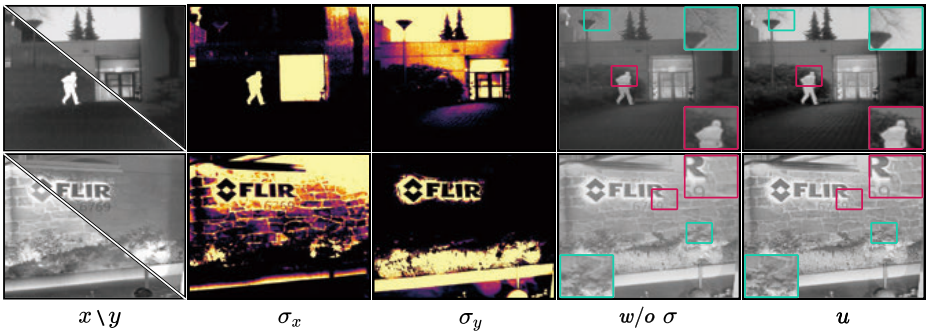


Fig. 11. Qualitative results on discussing biphasic attention layer.

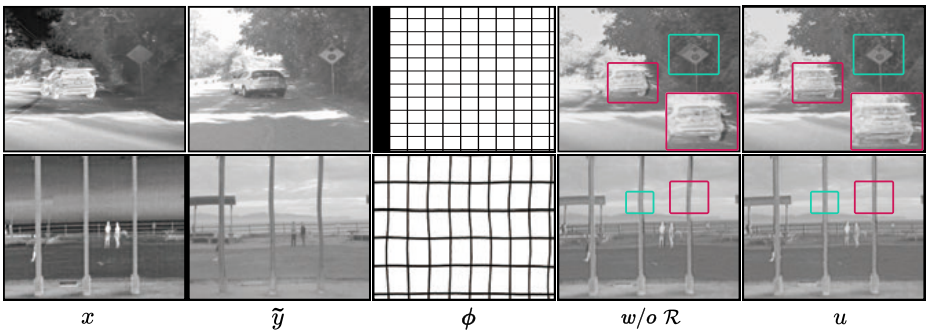


Fig. 12. Ablation study about the effect of our micro registration module.

from the opposite headlights causes the car to be invisible. Under poor lighting conditions, we confirm that the infrared information can detect more desirable areas, but some portion with low thermal radiation is easily ignored. Moreover, current methods focus more on infrared information(*e.g.,* PMGI), which cannot estimate the main natural object and visible details (*e.g.,* DIDFuse) that introduce unwanted artifacts. In contrast, our method estimates the whole region without artifacts.

Depth Estimation: Indeed, recent algorithms for depth estimation [39,8] are trained on daytime road datasets (*e.g.,* KITTI [7] and CityScapes [3]), resulting in a disconnect between daytime and nighttime sceneries. The second row of Fig. 13 illustrates the depth maps calculated by MiDaS [35] on the recent efficiency fusion methods. Note that the depth maps from visible images and other approaches render apparent deficiencies, in which the wayside trees are misestimated. By comparison, our method can accurately estimate the depth map for diverse tree shapes, thereby providing a new auxiliary option for real-world depth estimation.

Object Detection: As a non-trivial byproduct, we also provide an improvement in our approach for detecting targets using the well-known YoloV4 [1].

In the last row of Fig. 13, our method accurately detects all three targets, and only our fusion results correctly detect the pedestrian on the left. These results demonstrate that our method significantly impacts the object detection task.

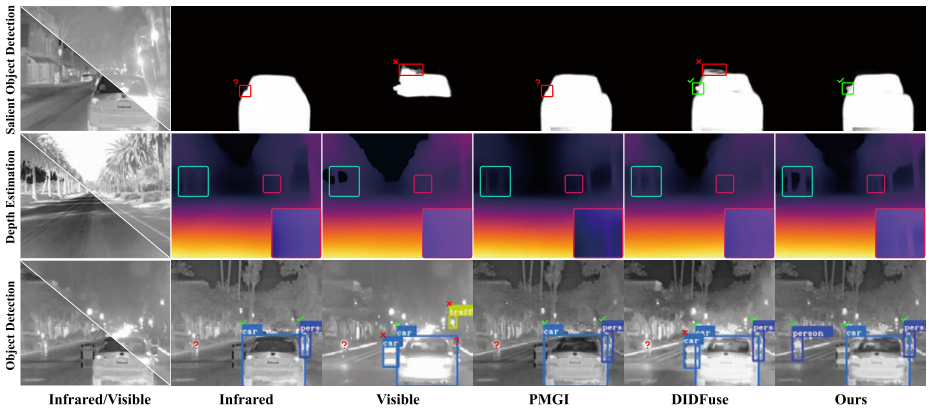


Fig. 13. Visual comparisons of multiple practical vision tasks.

4 Conclusion

In this paper, we propose an innovative network based on biphasic recurrent attention learning, which robustly and efficiently realizes IVIF take in an end-to-end manner. We first design a micro registration module to coarse estimate the distortion caused by misalignment. Then, a biphasic recurrent learning network successfully merges the source images and removes other remaining ghosting halos or artifacts. Furthermore, we also employ the parallel dilated convolutional and share calculation graph in our recurrent network to achieve high computational efficiency. Both subjective and objective experimental results reveal that our ReCoNet has significant superiority against the state-of-the-art methods with high efficiency. In addition, our ReCoNet also can deal with misalignment image pairs to a certain degree.

Acknowledgments: This work is partially supported by the National Key R&D Program of China (2020YF-B1313503), the National Natural Science Foundation of China (Nos. 61922019, 61906029 and 62027826), and the Fundamental Research Funds for the Central Universities.

References

1. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
2. Brasó, G., Leal-Taixé, L.: Learning a neural solver for multiple object tracking. In: IEEE CVPR. pp. 6247–6257 (2020)
3. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
4. Du, Q., Xu, H., Ma, Y., Huang, J., Fan, F.: Fusing infrared and visible images of different resolutions via total variation model. Sensors **18**(11), 3827 (2018)
5. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: CVPR. pp. 3146–3154 (2019)
6. Gao, H., Cheng, B., Wang, J., Li, K., Zhao, J., Li, D.: Object classification using cnn-based fusion of vision and lidar in autonomous vehicle environment. IEEE TII **14**(9), 4224–4231 (2018)
7. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR) (2013)
8. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR (2017)
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
10. Jiang, Z., Li, Z., Yang, S., Fan, X., Liu, R.: Target oriented perceptual adversarial fusion network for underwater image enhancement. IEEE Transactions on Circuits and Systems for Video Technology (2022)
11. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., ˇCehovin Zajc, L., Vojir, T., Hager, G., Lukezic, A., Eldesokey, A., et al.: The visual object tracking vot2017 challenge results. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 1949–1972 (2017)
12. Lan, X., Ye, M., Shao, R., Zhong, B., Yuen, P.C., Zhou, H.: Learning modality-consistency feature templates: A robust rgb-infrared tracking system. IEEE TIE **66**(12), 9887–9897 (2019)
13. Li, H., Wu, X.J.: Densefuse: A fusion approach to infrared and visible images. IEEE TIP **28**(5), 2614–2623 (2018)
14. Li, H., Wu, X.J., Kittler, J.: Rfn-nest: An end-to-end residual fusion network for infrared and visible images. Information Fusion **73**, 72–86 (2021)
15. Li, J., Huo, H., Li, C., Wang, R., Feng, Q.: Attentionfgan: Infrared and visible image fusion using attention-based generative adversarial networks. IEEE Transactions on Multimedia **23**, 1383–1396 (2020)
16. Li, P.: Didfuse: deep image decomposition for infrared and visible image fusion. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. pp. 976–976 (2021)
17. Li, S., Kang, X., Hu, J.: Image fusion with guided filtering. IEEE TIP **22**(7), 2864–2875 (2013)
18. Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., Luo, Z.: Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5802–5811 (2022)

19. Liu, J., Fan, X., Jiang, J., Liu, R., Luo, Z.: Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. IEEE TCSVT (2021)
20. Liu, J., Shang, J., Liu, R., Fan, X.: Attention-guided global-local adversarial learning for detail-preserving multi-exposure image fusion. IEEE Transactions on Circuits and Systems for Video Technology pp. 1–1 (2022). https://doi.org/10.1109/TCSVT.2022.3144455
21. Liu, J., Wu, Y., Huang, Z., Liu, R., Fan, X.: Smoa: Searching a modality-oriented architecture for infrared and visible image fusion. IEEE Signal Processing Letters **28**, 1818–1822 (2021)
22. Liu, R., Liu, J., Jiang, Z., Fan, X., Luo, Z.: A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion. IEEE TIP **30**, 1261–1274 (2021). https://doi.org/10.1109/TIP.2020.3043125
23. Liu, R., Liu, Z., Liu, J., Fan, X.: Searching a hierarchically aggregated fusion architecture for fast multi-modality image fusion. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1600–1608 (2021)
24. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
25. Ma, J., Chen, C., Li, C., Huang, J.: Infrared and visible image fusion via gradient transfer and total variation minimization. Information Fusion **31**, 100–109 (2016)
26. Ma, J., Yu, W., Liang, P., Li, C., Jiang, J.: Fusiongan: A generative adversarial network for infrared and visible image fusion. Information Fusion **48**, 11–26 (2019)
27. Ma, J., Zhang, H., Shao, Z., Liang, P., Xu, H.: Ganmcc: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. IEEE TIM **70**, 1–14 (2020)
28. Ma, J., Zhou, Z., Wang, B., Zong, H.: Infrared and visible image fusion based on visual saliency map and weighted least square optimization. Infrared Physics & Technology **82**, 8–17 (2017)
29. Nencini, F., Garzelli, A., Baronti, S., Alparone, L.: Remote sensing image fusion using the curvelet transform. Information Fusion **8**(2), 143–156 (2007)
30. Palsson, F., Sveinsson, J.R., Ulfarsson, M.O.: Multispectral and hyperspectral image fusion using a 3-d-convolutional neural network. IEEE Geoscience and Remote Sensing Letters **14**(5), 639–643 (2017)
31. Paramanandham, N., Rajendiran, K.: Infrared and visible image fusion using discrete cosine transform and swarm intelligence for surveillance applications. Infrared Physics & Technology **88**, 13–22 (2018)
32. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019), `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`
33. Pu, M., Huang, Y., Guan, Q., Zou, Q.: Graphnet: learning image pseudo annotations for weakly-supervised semantic segmentation. In: ACM MM. pp. 483–491. ACM (2018)
34. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. vol. 106, p. 107404 (2020)

35. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2020)
36. Shreyamsha Kumar, B.: Image fusion based on pixel significance using cross bilateral filter. Signal, image and video processing **9**(5), 1193–1204 (2015)
37. Toet, A.: The tno multiband image data collection. Data in brief **15**, 249 (2017)
38. Wang, D., Liu, J., Fan, X., Liu, R.: Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. arXiv preprint arXiv:2205.11876 (2022)
39. Wang, L., Zhang, J., Wang, Y., Lu, H., Ruan, X.: Cliffnet for monocular depth estimation with hierarchical embedding loss. In: ECCV. Springer (2020)
40. Xiao, Y., Codevilla, F., Gurram, A., Urfalioglu, O., López, A.M.: Multimodal end-to-end autonomous driving. IEEE Transactions on Intelligent Transportation Systems (2020)
41. Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H.: U2fusion: A unified unsupervised image fusion network. IEEE TPAMI (2020)
42. Xu, H., Ma, J., Yuan, J., Le, Z., Liu, W.: Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19679–19688 (2022)
43. Xu, T., Feng, Z.H., Wu, X.J., Kittler, J.: Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking. IEEE TIP **28**(11), 5596–5609 (2019)
44. Zhang, H., Xu, H., Xiao, Y., Guo, X., Ma, J.: Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In: AAAI. vol. 34, pp. 12797–12804 (2020)
45. Zhang, L., Zhu, X., Chen, X., Yang, X., Lei, Z., Liu, Z.: Weakly aligned cross-modal learning for multispectral pedestrian detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5127–5137 (2019)
46. Zhang, X., Ye, P., Leung, H., Gong, K., Xiao, G.: Object fusion tracking based on visible and infrared images: A comprehensive review. Information Fusion **63**, 166–187 (2020)
47. Zhao, J.X., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: Egnet: Edge guidance network for salient object detection. In: CVPR. pp. 8779–8788 (2019)