



MULTISENSOR IMAGE FUSION: DATASET, METHODS AND PERFORMANCE EVALUATION

MOHAMMED ZOUAOU LAIDOUNI

University of Defence Belgrade, Military Academy, mohammedz.laidouni@gmail.com

BOBAN BONDŽULIĆ

University of Defence Belgrade, Military Academy, bondzulici@yahoo.com

DIMITRIJE BUJAKOVIĆ

University of Defence Belgrade, Military Academy, dimitrijebujakovic@gmail.com

TOUATI ADLI

University of Defence Belgrade, Military Academy, adlitouati94@gmail.com

MILENKO ANDRIĆ

University of Defence Belgrade, Military Academy, andricsmilenko@gmail.com

Abstract: *Multisensor image fusion is a crucial research area aiming to enhance image clarity and comprehensibility by integrating information from multiple sensors. This paper presents a residual dense transformer (RDT) architecture for multisensor image fusion to address the challenges posed by the unique strengths and limitations of visual infrared (VIS), near-infrared (NIR), and long-wavelength infrared (LWIR) sensors. A comparative analysis is conducted with several state-of-the-art fusion methods using various objective evaluation indicators to assess the image fusion quality. We used a 313 triplet images collected from three datasets: TRICLOBS, MOFA, and MUDCAD, covering diverse environmental conditions such as foggy conditions and low illumination. Through the evaluation of the RDT and state-of-the-art fusion algorithms on this dataset, we observe that RDT achieve the best overall performance across multiple spectra image fusion. This work, thus, serves as a platform for developing and comparing new algorithms to deal with images from three sensors. which aids in the development of various applications such as object tracking, detection, and surveillance.*

Keywords: *Multisensor images; image fusion; multisensor dataset; residual dense transformer; deep leaning.*

1. INTRODUCTION

The development of image sensors and multisensor image fusion has been an important field of research for many years. By combining the complementary information from multiple image sensors about a same scene can produce a new single fused image with more clarity and understandability to be applied and enhance several applications such as object tracking, object detection, surveillance, military applications, facial analysis and recognition [1-4].

Sensor systems are categorized based on the wavelengths they capture and the three most used are visual infrared (VIS), near-infrared (NIR), and long-wavelength infrared (LWIR) [1]. To combine the strength of this sensors several algorithms have been proposed, including conventional algorithms and deep learning-based algorithms [5, 6]. Conventional algorithms according to their corresponding theories, can be classified into: multiscale transform-based algorithms, sparse representation-based algorithms, subspace-based methods, saliency-based algorithms, hybrid algorithms.

Considering the powerful feature representation capability of deep learning, researchers have proposed several algorithms such as CNN-, AE-, GAN-, and transformer-based algorithms. The basic components of CNN-, AE-, and GAN-based algorithms are convolutional layers which focus on local features and ignores some global information. To overcome this drawback, the transformer-based algorithms have been proposed to model the long-range dependency and capture the global context. However, these algorithms only stack transformer blocks without incorporating the global features of all previous blocks, which is essential for improving the fusion process by maximizing the use of all global information.

Furthermore, to evaluate the performance of image fusion algorithms across multiple spectra, many datasets have been used, including TNO [7], RoadScene [8], VIFB [9], KAIST [10], RGB-NIR [11]. However, current research on multisensor image fusion is suffering from several problems, such as the lack of a well-recognized multisensor image fusion dataset that can be used to compare image fusion performance across various spectral images of the same scene. As a result, it is common to observe that most of utilized dataset in

experiments in the literature usually contain only images from two different sensor (LWIR-VIS or NIR-VIS), which poses challenges in evaluating performance of fusion algorithms for multisensor image fusion.

To tackle the challenges mentioned above. In this work we use a residual dense transformer to maximize the use of the global features. Furthermore, we collect a multisensor image fusion dataset containing images from three sensors, including LWIR, NIR and VIS images. The images are collected from three available datasets TRICLOBS [12], MOFA [13] and MUDCAD [14] to facilitate the training and performance evaluation for multisensor image fusion. Thus, the main contributions of this work are summarized as follow:

- A residual dense transformer is used to maximize the use of the global features and produce a fused images with higher clarity and understandability.
- Training/testing dataset is created containing triplet images from LWIR, NIR and VIS spectra. These images are collected from three available datasets TRICLOBS, MOFA and MUDCAD, covering a wide range of environments, such as foggy condition and low illumination.
- Analysis and comparative evaluation of several state-of-art image fusion algorithms including traditional and advanced algorithms based on deep learning for the training/testing dataset.

The rest of the paper is organized as follows. Section 2 describes the databases used in this research. In Section 3, the framework of the residual dense transformer is presented, Next, the comparison performance of fusion methods is analyzed on the dataset. Finally, Section 4 concludes the paper.

2. DATABASES DESCRIPTION

In this research, three datasets are used, namely TRICLOBS, MOFA and MUDCAD. The general characteristics of these datasets are listed in Table 1.

Table 1. General characteristics of the datasets

Name	Image triplets	Image type	Resolution	Year
TRICLOBS	183	LWIR, NIR, VIS	480×640	2016
MOFA	84	LWIR, NIR, VIS	Various	2023
MUDCAD	46	LWIR, NIR, VIS	512×512	2023

TRICLOBS, MOFA and MUDCAD are multimodal image datasets contain images from the VIS, NIR and LWIR parts of the electromagnetic spectrum. TRICLOBS encompass outdoor scene representing various military and civilian surveillance scenarios at different locations [12]. MOFA dataset contain two main groups: indoor and outdoor scenery. The indoor scenery comprises of table top scenes and rooms with different objects category such as books, tools, electronics and toys. The outdoor scenery comprises of scenes from 8 different locations: university

campus buildings playground, crossroads, houses, garden, shore, grassland and pools [13]. MUDCAD dataset contain scene captured by an unmanned aerial vehicle in two different areas of the test site at the University of the Bundeswehr, Munich and under different environments, such as grassland, gravel and graveled soil, various bushes and trees, and both concrete and asphalt roads [14]. The representative images from TRICLOBS, MOFA and MUDCAD datasets are given in Figure 1.

Table 2. Train/test partition details of TRICLOBS, MOFA and MUDCAD datasets

Dataset	Train	Test
TRICLOBS	97	86
MOFA	65	19
MUDCAD	46	/

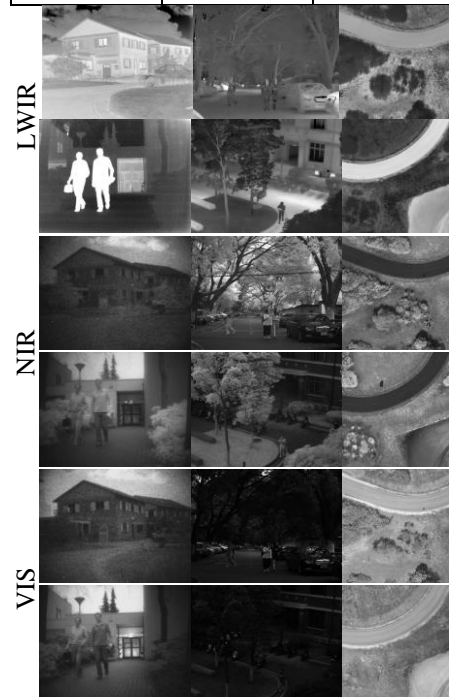


Figure 1. Representative images from TRICLOBS, MOFA and MUDCAD datasets

The dataset partition is provided in Table 2 including 208 triplets of LWIR, NIR and VIS images for training and 105 for testing. TRICLOBS, MOFA and MUDCAD are use in training set, while in testing set only TRICLOBS and MOFA are contributed. Each triplet of MOFA images has been registered to successful image fusion. As can be seen in Figure 1, the images cover a wide range of environments and conditions, such as indoor, outdoor, and remote sensing images which enrich the training and testing tasks.

3. METHOD

In this section, we first present the overall framework of the residual dense Transformer architecture and then we provide a detail about the loss function used for training.

3.1 Residual dense Transformer architecture

The overall framework of the RDT is shown in Figure 2 inspired by the work in RDN [15] and SDTFusion [16]. We design the RDT and it consists of three main modules: shallow feature extraction, global feature extraction, and reconstruction.

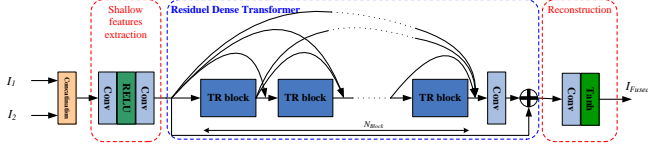


Figure 2. The general framework of the residual dense Transformer

At the beginning, the two-channel map is obtained by performing an element-wise concatenation on two inputs I_1 and I_2 , and then the two-channel map is fed to the first module to extract shallow features. After extracting the shallow features, they will be conducted through the RDT to extract the global features. The RDT consist of a multiple TR block densely connected and each TR block contain N_{TR} Swin Transformer. The Conv+RELU in the TR block is used to adapt the output channels of the previous TR block, while the Conv used in the RDT is used to fuse all the state of TR blocks. This architecture enables to retain more important information and improve the global feature extraction by incorporating the extracted features from all previous states. In the final module, convolutional and Tanh activation layers are used to integrate the channels information, and obtain the final fused image.

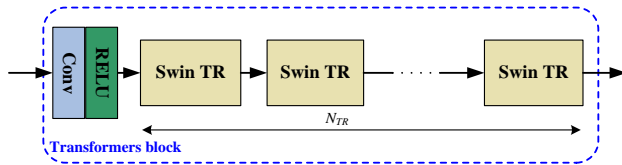


Figure 2. The structure of the TR block

3.2 Loss function

To unsupervised train the RDT for image fusion, a loss function composed of pixel sub-loss and gradient sub-loss are used, and is defined as:

$$L = \alpha \cdot L_P + \beta \cdot L_G \quad (1.1)$$

where α and β are weighted parameters that control the balance between the sub-losses. L_P and L_G are the pixel and gradient sub-loss, respectively. They can be formulated as:

$$L_G = \|\nabla I_F - \max\{\nabla I_1, \nabla I_2\}\|_1 \quad (1.2)$$

$$L_P = \|\|I_F - \text{mean}(I_1, I_2)\|_1 \quad (1.3)$$

4. MULTISENSOR IMAGE FUSION

In this section, we firstly present the methods used in the comparative evaluation on collected dataset in section 4.1, and then quantitative evaluation metrics are provided in section 4.2. In Sections 4.3-4.5 are presented a qualitative and quantitative performance comparison of LWIR/VIS, LWIR/NIR and NIR/VIS image fusion.

4.1 Comparison algorithms

In this research, we incorporated 6 published fusion algorithms, including two traditional algorithms FPDE [9] and LatLRR [9] and four advanced algorithms based on deep learning, namely, RFN-Nest [17], U2Fusion [18], SwinFuse [19] and DataFuse [20].

These algorithms cover many algorithm types. FPDE is multi-sensor image fusion which is based on fourth order partial differential equations to get approximation and detail information from source images, and then the principal component analysis was applied to obtain the optimal weights to fuse the approximation and detail information. LatLRR algorithm is based on latent low-rank representation, which decompose the images into low-rank parts (global structure) and salient parts (local structure), and then, the low rank parts are fused by weighted-average strategy while the salient parts are simply fused by sum strategy. RFN-Nest is autoencoder-based algorithm that utilizes a residual fusion network learned via training using visible and infrared image pairs to perform feature fusion. In addition, they utilized multiscale features in the encoder and nest connection in the decoder to improve the fusion performance.

U2Fusion is CNN-based algorithm which use a densely connected fusion network. In the training phase, U2Fusion automatically estimates an adaptive similarity degrees of source images by using feature extraction and information measurement to preserve this similarity between the fusion result and source images.

SwinFuse is a transformer-based algorithm that use a transformer-based encoder to model the long-range dependency and capture a global context.

DataFuse is transformer-based algorithm, which use a dual-attention residual module to examine the significant regions of source images, and a Transformer module to preserve global complementary information.

Table 3. Quantitative performance comparison of LWIR/VIS image fusion on TRICLOBS and MOFA

	Method	information error-based		information features-based					structure similarity-based				human perception-inspired		information theory-based
		RMSE	PSNR	AG	Q_{abf}	SCD	Q_P	Q_M	MS-SSIM	Q_C	Q_S	CC	CV	CB	NMI
TRICLOBS	FPDE	0.086	58,800	2,561	0.437	1,443	0.462	0.688	0.897	0.639	0.778	0.619	980,792	0.418	0.471
	LatLRR	0.088	58,690	2,630	0.450	1,526	0.368	0.627	0.900	0.657	0.807	0.620	1018,591	0.420	0.416
	RFN-Nest	0.094	58,471	2,487	0.430	1,590	0.390	0.596	0.929	0.594	0.782	0.614	905,660	0.413	0.492
	U2Fusion	0.086	58,802	2,153	0.331	1,423	0.405	0.584	0.892	0.615	0.790	0.621	881,889	0.456	0.523
	SwinFuse	0.112	57,643	3,355	0.381	1,633	0.376	0.560	0.891	0.579	0.537	0.536	1098,377	0.561	0.552
	DataFuse	0.140	56,681	4,388	0.226	1,444	0.240	0.428	0.829	0.349	0.491	0.583	999,955	0.319	0.369
RDT	0.107	57,845	4,111	0.605	1,586	0.484	0.702	0.928	0.749	0.839	0.586	738,462	0.456	0.540	
MOFA	FPDE	0.079	59,196	3,370	0.464	1,072	0.395	0.502	0.882	0.576	0.708	0.729	556,295	0.419	0.407
	LatLRR	0.081	59,094	3,245	0.483	1,266	0.381	0.488	0.896	0.520	0.719	0.735	364,366	0.454	0.376
	RFN-Nest	0.088	58,755	3,414	0.484	1,674	0.387	0.434	0.940	0.560	0.718	0.735	554,969	0.414	0.426
	U2Fusion	0.079	59,208	2,431	0.334	1,054	0.416	0.406	0.870	0.512	0.706	0.736	504,359	0.456	0.451
	SwinFuse	0.098	58,245	3,406	0.334	1,180	0.358	0.460	0.769	0.336	0.466	0.673	500,019	0.533	0.429
	DataFuse	0.126	57,146	3,564	0.358	1,160	0.297	0.315	0.881	0.504	0.594	0.676	676,522	0.323	0.475
RDT	0.084	58,934	5,114	0.590	1,522	0.446	0.648	0.947	0.567	0.750	0.709	266,446	0.471	0.398	

The train parameters of the RDT are adapted by the Adam optimizer with learning rate of 10^{-4} , and the epoch number is set to 30. The model parameter N_{TR} and N_{Block} are set to 2 and 3, respectively. The weighted parameter in the loss function α and β are set to 20 and 10, respectively. All experiments were conducted using a computer equipped with an NVIDIA RTX3060 GPU and i7-11800H CPU. Default parameters reported by the corresponding authors of each algorithm were employed. To obtain sufficient training samples for deep learning-based algorithms, the training data were randomly cropped into 64×64 patches, and then the models were retrained to perform the multisensor image fusion.

4.2. Evaluation metrics

To conduct a comprehensive quantitative evaluation, 14 frequently utilized objective evaluation indicators for image fusion are utilized, including, Root mean squared error (RMSE), Peak signal-to-noise ratio (PSNR) reveals the distortion during the fusion process at the pixel level, Average gradient (AG) measures the degree of clarity and sharpness, Objective gradient based image feature metric (Q_{abf}) calculates the edge information, which is transferred from source images to the fused image, the sum of correlations of differences (SCD) measures the maximum information of the fused images containing each source image. phase congruency-based image feature metric (Q_P) measures fusion performance by making comparisons within the local correlation between the feature maps of the fusion result and the source images. Multi-scale Scheme-Based image feature metric (Q_M) calculate the edge preservation value of the fused image across multiple scale, structure similarity (MS-SSIM) evaluates the structural loss and distortion of fused images from the human visual system's perspective, Cvejie's Metric (Q_C) use a local measurement of similarity between blocks of pixels in the input images and the fused images to estimate how well the important information in the source images is represented by the fused image. Piella's Metric (Q_S) utilizes local measures to estimate how well the salient information from the inputs is present in the fused images. correlation coefficient (CC) measures the degree of linear correlation between the fused images and the source images., Chen-Varshney metric (CV) and Chen-Varshney metric (CB)

are tow metrics to evaluate the image quality of the fused images based on the human visual system perception, normalized mutual information (NMI) calculates the mutual information transformed from the input images to the fusion result.

4.3. LWIR and VIS image fusion

Figure 4 illustrates two groups of LWIR and VIS image pairs from the TRICLOBS and MOFA datasets and the fused images generated through different algorithms. As can be seen in Figure 4, FPDE, LatLRR, U2Fusion and SwinFusion suffer from low contrast and fail to present some scene detail. RFN-Nest, DataFuse and RDT display a good contrast.

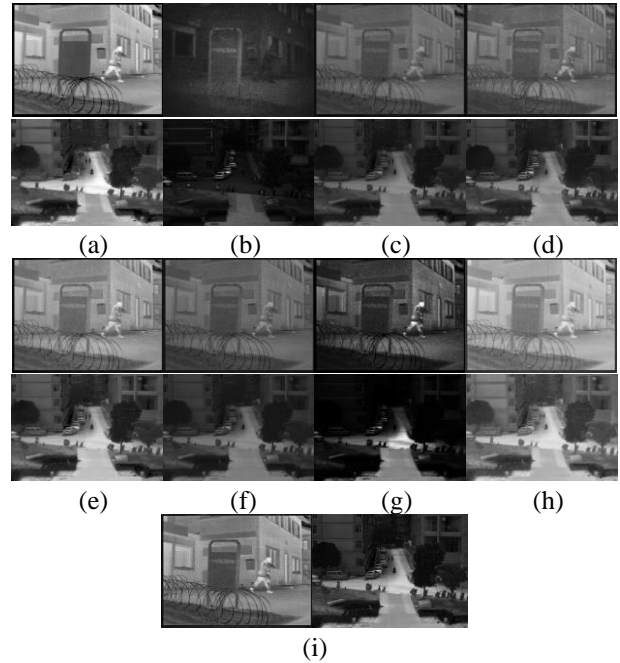


Figure 4. Qualitative performance comparison of LWIR/VIS image fusion on TRICLOBS and MOFA: (a) LWIR, (b) VIS, (c) FPDE, (d) LatLRR, (e) RFN-Nest, (f) U2Fusion, (g) SwinFusion, (h) DataFuse, (i) RDT

A quantitative comparison of RDT and several algorithms' performance on TRICLOBS and MOFA are shown in Table 3. The fused LWIR/VIS fused images are evaluated using 14 evaluation metrics, and for each metric, the best and the second-best methods are marked in red and green, respectively. RDT method shows the best overall performance in terms of information features, structure similarity and human perception. FPDE and U2Fusion show competitive performance in terms of information error, displaying the best and second-best performance in both RMSE and PSNR metrics.

4.4. LWIR and NIR image fusion

In Figure 5, two groups of LWIR and NIR image pairs

from the TRICLOBS and MOFA datasets are provided along with fused images. FPDE, LatLRR, RFN-Nest, U2Fusion, and SwinFusion suffers from low contrast. RDT method displays a good performance in combining information from the both image modalities.

Table 4 presents the quantitative performance comparison of LWIR/NIR images fusion. As can be seen, RDT method show the best overall performance in term of information features, structure similarity, human perception and information theory in both MOFA and TRICLOBS datasets. The second-best overall results are shown by RFN-Nest method. Furthermore, FPDE obtains the best performance in term of information error.

Table 4. Quantitative performance comparison of LWIR/NIR image fusion on TRICLOBS and MOFA

	Method	information error-based		information features-based					structure similarity-based				human perception-inspired		information theory-based
		RMSE	PSNR	AG	Q _{abr}	SCD	Q _P	Q _M	MS-SSIM	Q _C	Q _S	CC	CV	CB	NMI
TRICLOBS	FPDE	0.083	58,952	2,041	0.305	1,373	0.392	0.587	0.882	0.607	0.784	0.616	627,371	0.463	0.482
	LatLRR	0.105	57,949	2,923	0.418	1,418	0.392	0.633	0.900	0.664	0.759	0.548	388,914	0.369	0.546
	RFN-Nest	0.083	58,948	2,616	0.437	1,410	0.403	0.733	0.885	0.624	0.765	0.616	704,650	0.409	0.412
	U2Fusion	0.085	58,841	2,536	0.435	1,481	0.351	0.638	0.893	0.653	0.801	0.616	848,595	0.418	0.382
	SwinFuse	0.106	57,895	3,482	0.389	1,580	0.368	0.570	0.899	0.638	0.566	0.534	661,912	0.599	0.478
	DataFuse	0.093	58,519	2,824	0.492	1,605	0.412	0.645	0.938	0.653	0.790	0.596	499,178	0.406	0.473
	RDT	0.102	58,054	3,852	0.617	1,569	0.484	1.003	0.932	0.736	0.810	0.566	253,571	0.497	0.579
MOFA	FPDE	0.077	59,344	2,637	0.308	1,492	0.342	0.360	0.829	0.505	0.742	0.592	726,533	0.474	0.315
	LatLRR	0.103	58,047	3,472	0.397	1,499	0.336	0.337	0.838	0.569	0.678	0.507	928,844	0.391	0.425
	RFN-Nest	0.078	59,274	5,111	0.428	1,473	0.248	0.458	0.798	0.437	0.706	0.571	779,837	0.443	0.241
	U2Fusion	0.081	59,118	3,367	0.449	1,600	0.345	0.439	0.857	0.518	0.765	0.593	574,599	0.485	0.284
	SwinFuse	0.094	58,465	3,766	0.458	1,828	0.321	0.388	0.907	0.550	0.734	0.577	885,031	0.377	0.312
	DataFuse	0.101	58,123	3,874	0.347	1,726	0.303	0.402	0.821	0.429	0.583	0.573	801,136	0.541	0.296
	RDT	0.091	58,584	5,434	0.617	1,742	0.440	0.946	0.908	0.545	0.795	0.542	444,047	0.503	0.457

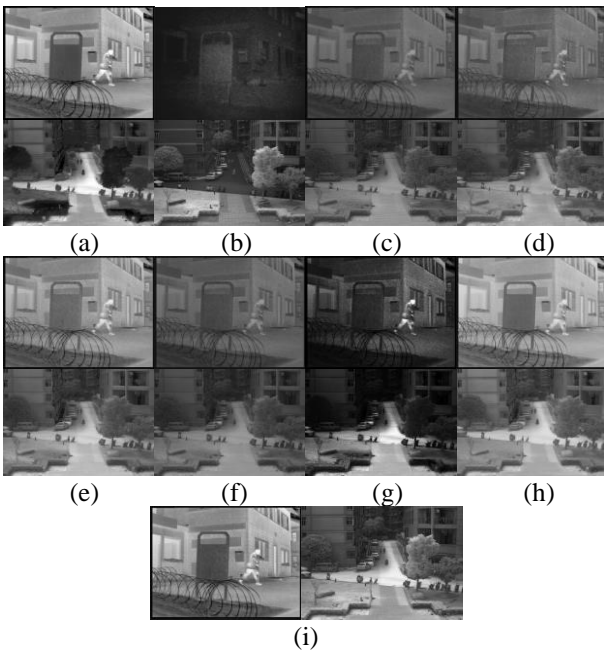


Figure 5. Qualitative performance comparison of LWIR/NIR image fusion on TRICLOBS and MOFA: (a) LWIR, (b) NIR, (c) FPDE, (d) LatLRR, (e) RFN-Nest, (f) U2Fusion, (g) SwinFusion, (h) DataFuse, (i) RDT

4.5. NIR and VIS image fusion

Figure 6 provides qualitative comparison of several fusion algorithms on two groups NIR and VIS image pairs from the TRICLOBS and MOFA datasets. In the case of TRICLOBS image, all methods provide an unclear scene to a certain degree. In other hand, in MOFA image, DataFuse and RDT methods can describe the general details about the scene.

Table 5 shows the quantitative comparisons between RDT and the state-of-the-art algorithms. As one can see, RDT method shows the best overall performance in terms of information features, structure similarity, human perception and information theory in both MOFA and TRICLOBS datasets. RFN-Nest method shows the second overall performance, and displays the best results in terms of information error.

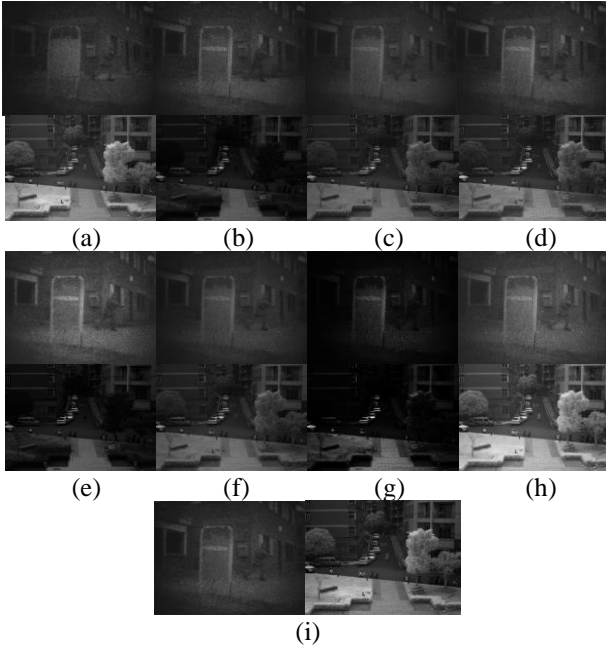


Figure 6. Qualitative performance comparison of NIR/VIS image fusion on TRICLOBS and MOFA: (a) NIR, (b) VIS, (c) FPDE, (d) LatLRR, (e) RFN-Nest, (f) U2Fusion, (g) SwinFusion, (h) DataFuse, (i) RDT

4.6 Ablation study

In this section, ablation studies are performed to verify the effectiveness of RDT architecture, and the loss function design. The results of the ablation experiments on the LWIR+VIS image fusion is shown in Table 6. It is observed that removing the residual or the dense connection from the architecture degrade the overall performance which prove the importance of incorporating the features of the previous states. Furthermore, it is noticed that removing any term of the loss function change the global performance. Eliminating the gradient term allow the model to have better performance in information error-based metrics, in the other hand, without the intensity term the model can achieve better performance in some information features, structure similarity based-metrics. However, only the both terms allow the model to have a balance and achieve better overall performance.

Table 5. Quantitative performance comparison of NIR/VIS image fusion on TRICLOBS and MOFA

	Method	information error-based		information features-based					structure similarity-based				human perception-inspired		information theory-based
		RMSE	PSNR	AG	Q _{abr}	SCD	Q _P	Q _M	MS-SSIM	Q _C	Q _S	CC	CV	CB	NMI
TRICLOBS	FPDE	0.027	63.957	1.746	0.425	0.728	0.435	0.735	0.968	0.691	0.919	0.932	116.863	0.585	0.558
	LatLRR	0.071	59.676	2.572	0.420	1.419	0.423	0.680	0.964	0.621	0.832	0.917	114.542	0.525	0.566
	RFN-Nest	0.027	63.960	1.904	0.452	0.685	0.433	0.756	0.966	0.721	0.918	0.932	122.073	0.564	0.556
	U2Fusion	0.032	63.140	2.130	0.422	1.021	0.387	0.734	0.959	0.691	0.909	0.926	127.170	0.540	0.508
	SwinFuse	0.137	56.859	3.482	0.170	0.540	0.109	0.517	0.616	0.443	0.452	0.143	2075.968	0.339	0.310
	DataFuse	0.043	61.904	2.527	0.505	1.300	0.500	0.764	0.920	0.541	0.844	0.871	94.497	0.635	0.782
	RDT	0.032	63.160	2.827	0.564	0.973	0.564	1.074	0.977	0.778	0.906	0.915	68.518	0.594	0.676
MOFA	FPDE	0.053	61.208	3.164	0.476	0.770	0.579	0.445	0.920	0.662	0.803	0.842	166.741	0.583	0.526
	LatLRR	0.087	58.903	5.664	0.555	1.198	0.547	0.264	0.940	0.473	0.556	0.759	152.950	0.479	0.363
	RFN-Nest	0.053	61.219	3.953	0.634	0.806	0.619	0.551	0.940	0.780	0.847	0.841	186.156	0.557	0.538
	U2Fusion	0.055	60.941	4.404	0.590	1.179	0.580	0.540	0.945	0.737	0.848	0.843	184.616	0.572	0.479
	SwinFuse	0.062	60.419	4.992	0.375	0.470	0.503	0.336	0.751	0.493	0.576	0.746	575.546	0.675	0.536
	DataFuse	0.089	58.712	3.874	0.304	0.278	0.218	0.391	0.627	0.556	0.579	0.476	1451.403	0.448	0.255
	RDT	0.055	61.005	5.543	0.719	1.276	0.675	1.189	0.983	0.745	0.890	0.784	86.608	0.599	0.703

Table 6. Ablation study on MOFA and TRICLOBS

Dataset	Method	information error-based		information features-based					structure similarity-based				human perception-inspired		information theory-based
		RMSE	PSNR	AG	Q _{abr}	SCD	Q _P	Q _M	MS-SSIM	Q _C	Q _S	CC	CV	CB	NMI
Architecture	RDT	0.1030	58.0442	4.2940	0.6021	1.5743	0.4769	0.6921	0.9318	0.7158	0.8227	0.6086	652.2285	0.4587	0.5139
	w/o D	0.1043	57.9885	4.3190	0.5908	1.5867	0.4751	0.6205	0.9298	0.6998	0.8103	0.6062	668.8243	0.4556	0.5126
	w/o R	0.1032	58.0359	4.2702	0.5883	1.5723	0.4451	0.6990	0.9278	0.6999	0.8202	0.6007	624.0815	0.4601	0.5216
Loss	RDT	0.1030	58.0442	4.2940	0.6021	1.5743	0.4769	0.6921	0.9318	0.7158	0.8227	0.6086	652.2285	0.4587	0.5139
	w/o G	0.0949	58.3888	3.1802	0.3786	1.2577	0.2705	0.5660	0.7419	0.5417	0.7093	0.5250	1323.2103	0.4477	0.5239
	w/o P	0.1054	57.9486	4.5344	0.5799	1.6083	0.4356	0.4946	0.9396	0.6034	0.7835	0.6143	597.8346	0.4513	0.4579

4. CONCLUSION

In this paper, we present a RDT architecture for image fusion task. Furthermore, we collected set of 313 triplet images (LWIR, NIR and VIS) from 3 different datasets, TRICLOBS, MOFA and MUDCAD to build a novel training/testing dataset. This benchmark facilitates better understanding of the state-of-the-art image fusion approaches across 3 spectral bands, and can provide a platform for developing new algorithms which deal with multisensor images technology. An experiment is carried out to evaluate the performance of the RDT method compared to the state-of-the-art fusion algorithms on the collected dataset. The experiment is conducted through different image type fusion and illustrates that the RDT achieves the best overall performance.

Further research of this work, may include extending this dataset and implementing image fusion algorithms which deal with the three-sensor image fusion. Thus, it will help for selecting which algorithm is adequate for fusing specific sensors.

References

- [1] Li, B., Xian, Y., Zhang, D., Su, J., Hu, X., Guo, W.: Multi-Sensor Image Fusion: A Survey of the State of the Art, *Journal of Computer and Communications* 100 (2021) 86–97.
- [2] Shopovska, I., Jovanov, L., Philips, W.: Deep visible and thermal image fusion for enhanced pedestrian visibility, *Sensors*, 19 (17) (2019) 3727–3746.
- [3] Zhang, J., Ding, Y., Yang, Y., Sun, J.: Real-time defog model based on visible and near-infrared information, 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Seattle, WA, 2016.
- [4] Coffey, V.C.: Seeing in the dark: Defense applications of IR imaging, *Optics and Photonics News*, 22 (4) 2011 26–31.
- [5] Zhang, X., Demiris, Y.: Visible and Infrared Image Fusion Using Deep Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (8) (2023) 10535 - 10554.
- [6] Ma, J., Ma, Y., Li, C.: Infrared and Visible Image Fusion Methods and Applications: A Survey. *Information Fusion*, 45 (2018) 153-178.
- [7] TNO Image Fusion Dataset. Accessed: Oct. 10, 2022. [Online]. Available: https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029
- [8] Xu, H., Ma, J., Li, Z., Jiang, J.: FusionDN: A unified densely connected network for image fusion, *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020
- [9] Zhang, X., Ye, P., Xiao, G.: VIFB: A visible and infrared image fusion benchmark, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2020
- [10] Hwang, S., Park, J., Kim, N., Choi, Y., Kweon, I.S.: Multispectral pedestrian detection: Benchmark dataset and baseline. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA, 2015
- [11] Brown, M., Susstrunk, S.: Multispectral SIFT for scene category, *CVPR 2011*, Colorado Springs, CO, USA, 2011
- [12] Toet, A., Hogervorst, M.A., Pinkus, A.R.: The TRICLOBS dynamic multi-band image data set for the development and evaluation of image fusion methods. *PLoS One*. 11 (12) (2016) e0165016.
- [13] Xiao, K., Kang, X., Liu, H., Duan, P.: MOFA: A novel dataset for Multi-modal Image Fusion Applications. *Information Fusion* 96 2023 144-155.
- [14] Hupel, T., Stütz, P.: Measuring and Predicting Sensor Performance for Camouflage Detection in Multispectral Imagery, *Sensors*, 23(19) 2023 8025.
- [15] Zhang, Y., Kong, Y., Zhong, B., Tian, Y., Fu, Y.: Residual Dense Network for Image Super-Resolution. *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018
- [16] Pang, S., Huo, H., Liu, X., Zheng, B., Li, J.: SDTFusion: A split-head dense transformer based network for infrared and visible image. *Infrared Physics & Technology*, 128 (2024).
- [17] Li, H., Wu, X., Kittler, J.: RFN-Nest: An end-to-end residual fusion network for infrared and visible images, *Information Fusion*, 73 2021 72-86
- [18] Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H.: U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (1) 2020 502-518.
- [19] Wang, Z., Chen, Y., Shao, W., Li, H., Zhang, L.: SwinFuse: A Residual Swin Transformer Fusion Network for Infrared and Visible Images, *IEEE Transactions on Instrumentation and Measurement*, 71 2022 1-12
- [20] Tang, W., He, F., Liu, Y., Duan, Y., Si, T.: DATAFuse: Infrared and visible image fusion via dual attention transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33 (7) 2023 3159–3172.