DAE-Fuse: A Discriminative Autoencoder for Multi-Modality Image Fusion

Yuchen GUO[1,2*], Ruoxiang XU[1], Rongcheng LI[1], Zhenghao WU[4]
Supervisor: Prof. Weifeng SU[1,3]

[1] Computer Science and Technology Programme, Faculty of Science and Technology, BNU-HKBU United International College
[2] MMLab at Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[3] Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science
[4] School of Computer Science, University College Dublin
**Corresponding Student Author Tel: +86-17783210471, E-mail: r130026037@mail.uic.edu.cn*

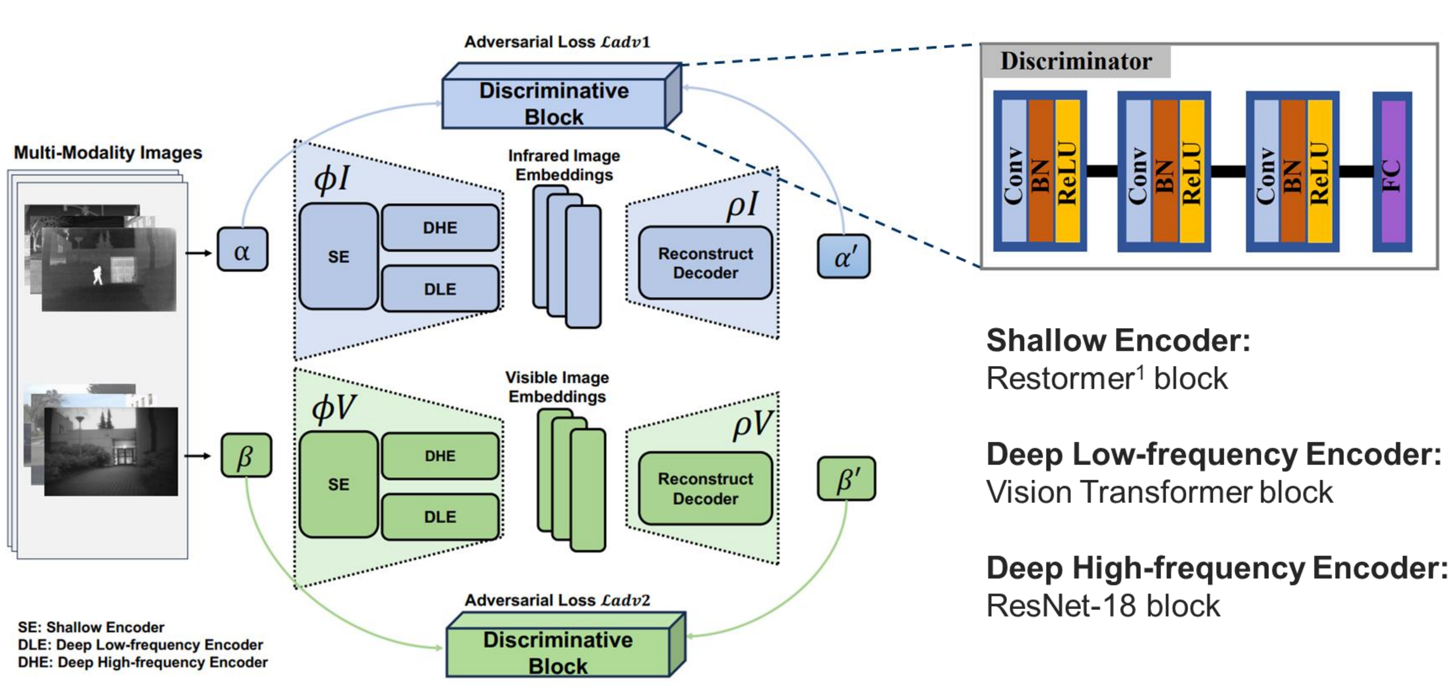
Abstract

Multi-modality image fusion aims to integrate complementary data information from different imaging modalities into a single image. Current methods generate either blurry fused images that lose fine-grained semantic information or unnatural fused images as perceptually cropped from inputs. Meanwhile, they are primarily optimized for a specific task. In this work, we propose a novel two phase discriminative autoencoder framework that generates sharp and natural fused images, termed DAE-Fuse. In the adversarial feature extraction phase, we introduce two discriminative blocks to the encoder-decoder architecture which provide an extra adversarial loss to better guide the feature extraction by reconstructing the source images. In the attention-guided cross-modality fusion phase, a cross-attention module is entailed to capture the complex correlation between modalities, then two discriminative blocks are further adapted to distinguish the structural differences between the fused output and the source inputs, injecting more naturalness to the output results. Extensive experiments across several public infrared-visible image fusion datasets and medical image fusion datasets demonstrate the superiority and generalizability of our method using both quantitative metrics and qualitative assessments. Additionally, our method also outperforms state-of-the-art methods on downstream object detection tasks.

KEYWORDS: Multi modality image fusion; discriminative autoencoder; adversarial learning; multi-level feature extraction

Methodology

Training Phase One: adversarial feature extraction phase

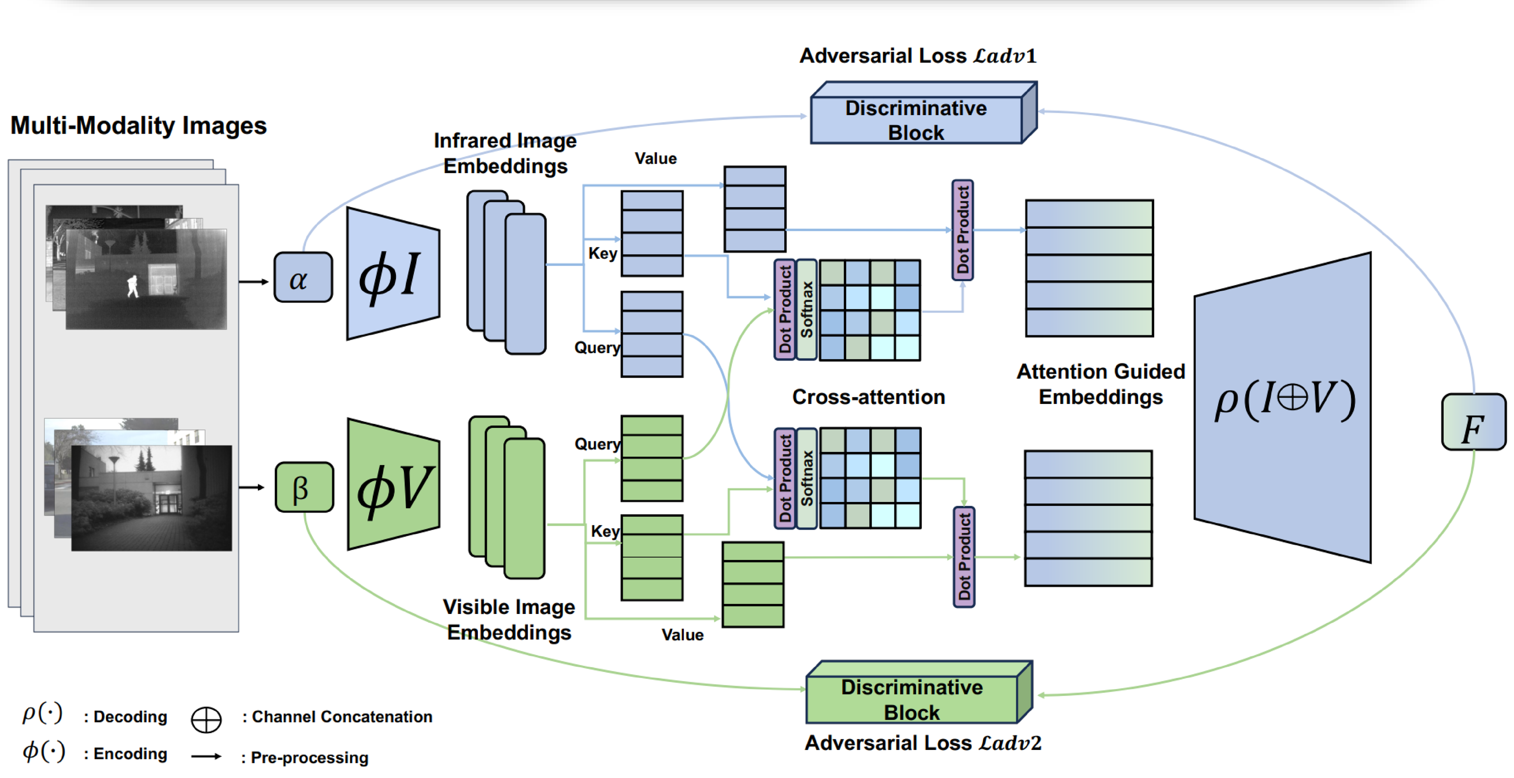


Shallow Encoder: Restormer[1] block

Deep Low-frequency Encoder: Vision Transformer block

Deep High-frequency Encoder: ResNet-18 block

Training Phase Two: attention-guided cross-modality fusion phase



Dataset & Settings

- **Framework**: PyTorch
- **GPU**: one NVIDIA A100
- **Optimizers**:
  - **Autoencoder**: Adam
  - **Discriminators**: RMSprop
- **Hyperparameters**:
  - **Phase 1**: 80 epochs; **Phase 2**: 140 epochs
  - **Learning rate**: 1e-4, decreasing by 0.5 every 20 epochs

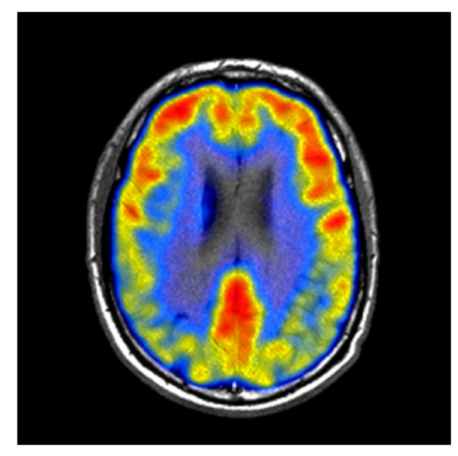
Table 1: Datasets. Unit: number of image pairs. Symbol × indicates not used.

Task	Dataset	Train	Validate	Test
IVIF	MSRS	1083	×	361
	RoadScene	×	50	50
	TNO	×	×	40
MIF	MRI-CT	×	×	21
	MRI-PET	×	×	42
	MRI-SPECT	×	×	73
MMOD	M3FD	×	×	4200

Evaluation Methods

Qualitative:

via output images

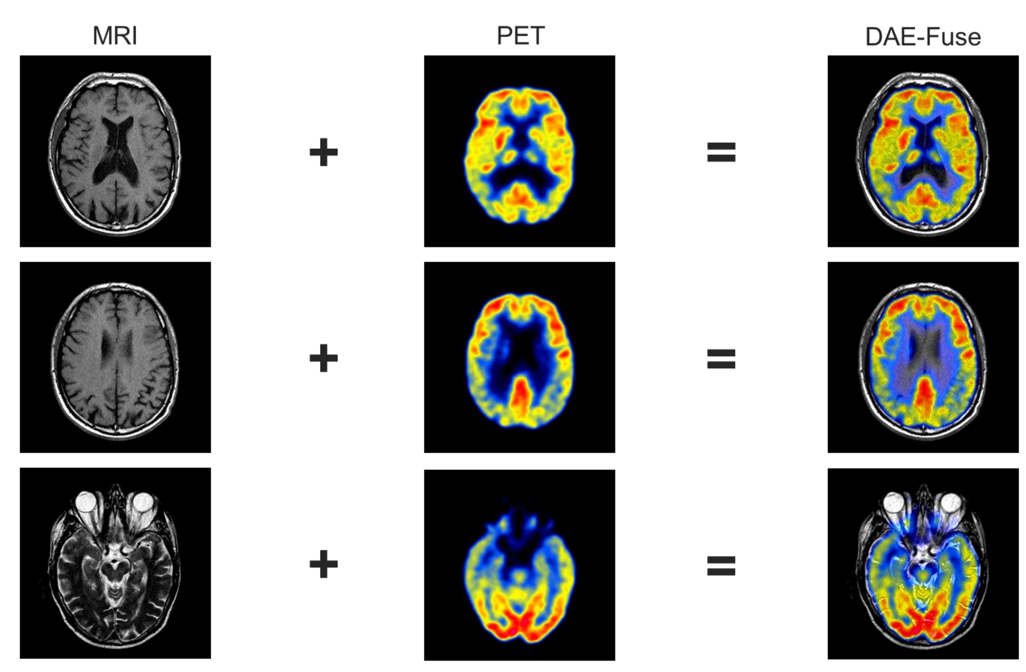


Quantitative:

via evaluation metrics for image fusion

Eight unsupervised metrics:
- entropy (**EN**)
- standard deviation (**SD**)
- spatial frequency (**SF**)
- visual information fidelity (**VIF**)
- sum of correlation of differences (**SCD**)
- mutual information (**MI**)
- a new quality metric[1] (**Qabf**)
- structural similarity index measure (**SSIM**)

Medical Image Fusion



Dataset: MRI-CT

	EN	SD	SF	MI	SCD	VIF	Qabf
DIDFuse	4.37	58.34	34.64	1.71	0.69	0.41	0.38
U2Fusion	4.21	61.98	32.54	2.08	0.75	0.37	0.46
RFN-Nest	4.97	70.36	33.42	1.98	0.68	0.43	0.52
DDcGAN	4.26	62.56	30.61	1.72	0.65	0.38	0.42
TarDAL	4.35	61.14	28.38	1.94	0.92	0.32	0.56
CDDFuse	4.49	71.36	34.02	2.16	1.18	0.44	0.56
DDFM	4.77	69.35	32.77	1.98	1.03	0.41	0.54
Ours	4.83	76.19	35.56	2.20	1.21	0.49	0.57

Dataset: MRI-SPECT

	EN	SD	SF	MI	SCD	VIF	Qabf
DIDFuse	2.97	55.12	13.69	1.51	0.42	0.46	0.51
U2Fusion	3.45	45.66	14.58	1.54	0.44	0.36	0.45
RFN-Nest	3.53	44.50	16.51	1.89	0.39	0.51	0.62
DDcGAN	3.33	49.33	13.48	1.72	0.35	0.46	0.36
TarDAL	3.02	65.63	15.54	1.62	0.32	0.48	0.58
CDDFuse	3.67	52.68	17.32	1.83	0.96	0.58	0.64
DDFM	3.57	56.42	16.81	1.66	0.49	0.53	0.55
Ours	4.17	62.61	22.56	1.87	1.58	0.68	0.70

Dataset: MRI-PET

	EN	SD	SF	MI	SCD	VIF	Qabf
DIDFuse	3.97	65.12	24.62	1.63	0.42	0.41	0.53
U2Fusion	3.83	59.66	24.58	1.61	0.44	0.32	0.43
RFN-Nest	3.84	64.50	25.51	1.77	0.39	0.52	0.62
DDcGAN	3.78	66.33	23.48	1.72	0.35	0.36	0.47
TarDAL	4.02	65.63	23.54	1.62	0.32	0.46	0.54
CDDFuse	4.02	74.15	25.48	1.78	1.42	0.57	0.65
DDFM	4.06	71.42	25.69	1.66	0.49	0.54	0.71
Ours	4.45	75.11	29.20	1.87	1.68	0.66	0.65

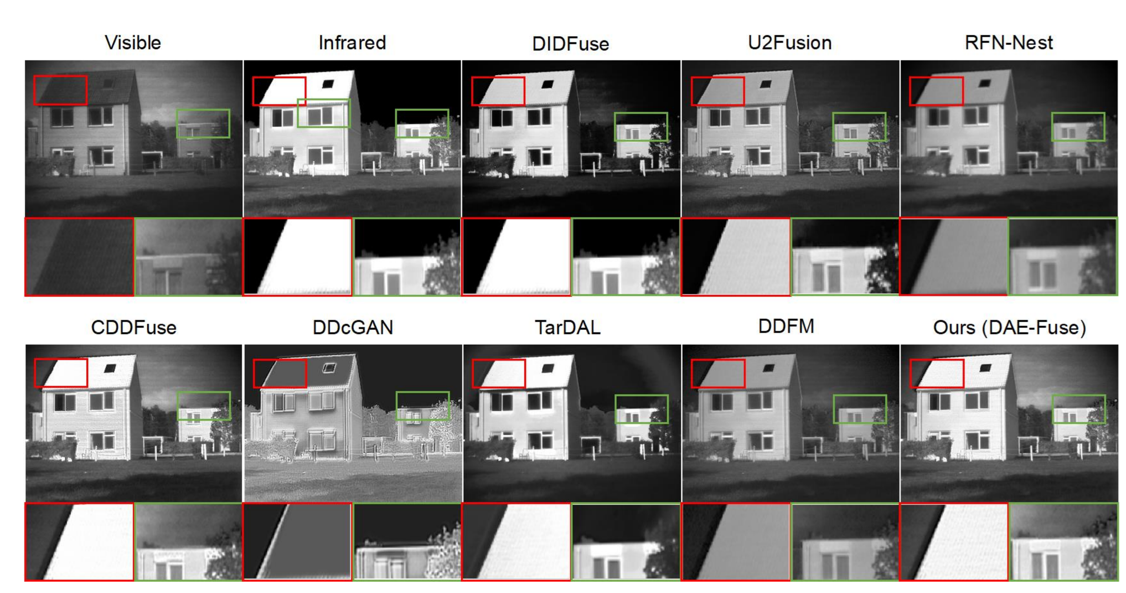
Infrared-Visible Image Fusion

Dataset: TNO

	EN	SD	SF	MI	SCD	VIF	Qabf
DIDFuse	6.97	45.12	12.59	1.63	1.71	0.58	0.42
U2Fusion	6.83	35.66	11.52	1.35	1.71	0.61	0.44
RFN-Nest	6.84	34.50	13.23	1.76	1.67	0.55	0.39
DDcGAN	6.78	46.33	11.68	1.78	1.72	0.48	0.35
TarDAL	7.02	45.63	10.68	2.17	1.62	0.57	0.32
CDDFuse	7.12	45.89	13.15	2.11	1.76	0.76	0.54
DDFM	7.06	48.42	13.03	2.06	1.66	0.83	0.49
Ours	7.17	46.63	13.31	2.21	1.89	0.75	0.55

Dataset: RoadScene

	EN	SD	SF	MI	SCD	VIF	Qabf
DIDFuse	7.07	51.12	14.59	2.11	1.69	0.60	0.39
U2Fusion	7.12	41.66	10.63	2.01	1.65	0.58	0.41
RFN-Nest	7.36	47.50	15.68	2.41	1.64	0.51	0.46
DDcGAN	7.43	38.32	11.68	2.25	1.56	0.48	0.31
TarDAL	7.22	45.63	11.68	2.31	1.71	0.54	0.47
CDDFuse	7.48	54.42	14.32	2.26	1.81	0.74	0.52
DDFM	7.52	53.74	13.57	2.21	1.66	0.81	0.51
Ours	7.57	56.05	14.36	1.86	1.85	0.76	0.57



Downstream: Multi-Modality Object Detection

Qualitative and quantitative results of MMOD task

	Peo	Car	Lam	Bus	Mot	Tru	mAP@50%
Ir	0.804	0.886	0.712	0.802	0.725	0.743	0.779
Vis	0.721	0.865	0.848	0.811	0.794	0.791	0.805
DID	0.791	0.924	0.857	0.833	0.787	0.788	0.830
U2F	0.802	0.922	0.870	0.839	0.783	0.786	0.833
RFN	0.813	0.915	0.851	0.829	0.813	0.875	0.849
DDc	0.797	0.908	0.832	0.895	0.805	0.872	0.851
TarD	0.835	0.947	0.854	0.928	0.811	0.874	0.874
CDD	0.846	0.928	0.864	0.931	0.813	0.891	0.878
DDFM	0.837	0.926	0.869	0.927	0.809	0.882	0.875
Ours	0.855	0.931	0.874	0.949	0.822	0.890	0.887

Conclusion

In conclusion, DAE-Fuse overcomes limitations in image fusion, producing sharp and natural images through adversarial feature extraction and attention-guided fusion. Discriminative blocks in both phases enhance feature extraction and structural fidelity. Public dataset experiments demonstrate DAE-Fuse's superiority over existing methods.

Publication

Yuchen Guo, Ruoxiang Xu, Rongcheng Li, Zhenghao Wu, Weifeng Su. DAE-Fuse: A Discriminative Autoencoder for Multi-Modality Image Fusion. 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2024) under review.

Arxiv Link: https://arxiv.org/pdf/2409.10080v1