Text-DiFuse: An Interactive Multi-Modal Image Fusion Framework based on Text-modulated Diffusion Model

*Hao Zhang, Lei Cao and Jiayi Ma**

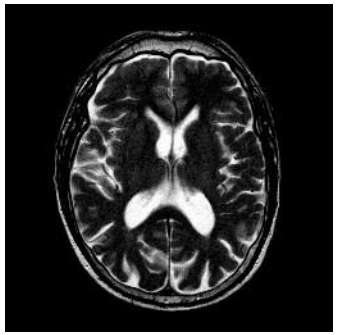
Electronic Information School, Wuhan University, Wuhan, China

➢ Introduction

Multi-modal Image Fusion

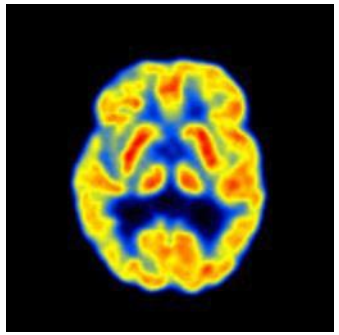
Combine the important information contained in multi-modal images of the same scene to generate a fused image that can **describe the scene content more comprehensively and accurately**, thereby helping people or machines better understand the scene and complete decisions.

Medical Image Fusion



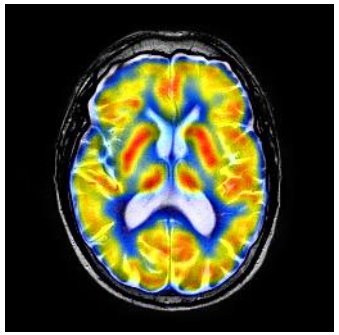
MRI

+



PET

=



$$I_f = F(I_1, I_2)$$

Infrared and Visible Image Fusion

IR

+

VIS

=

➢ Challenges & Motivations

☐ Composite Degradation Challenge

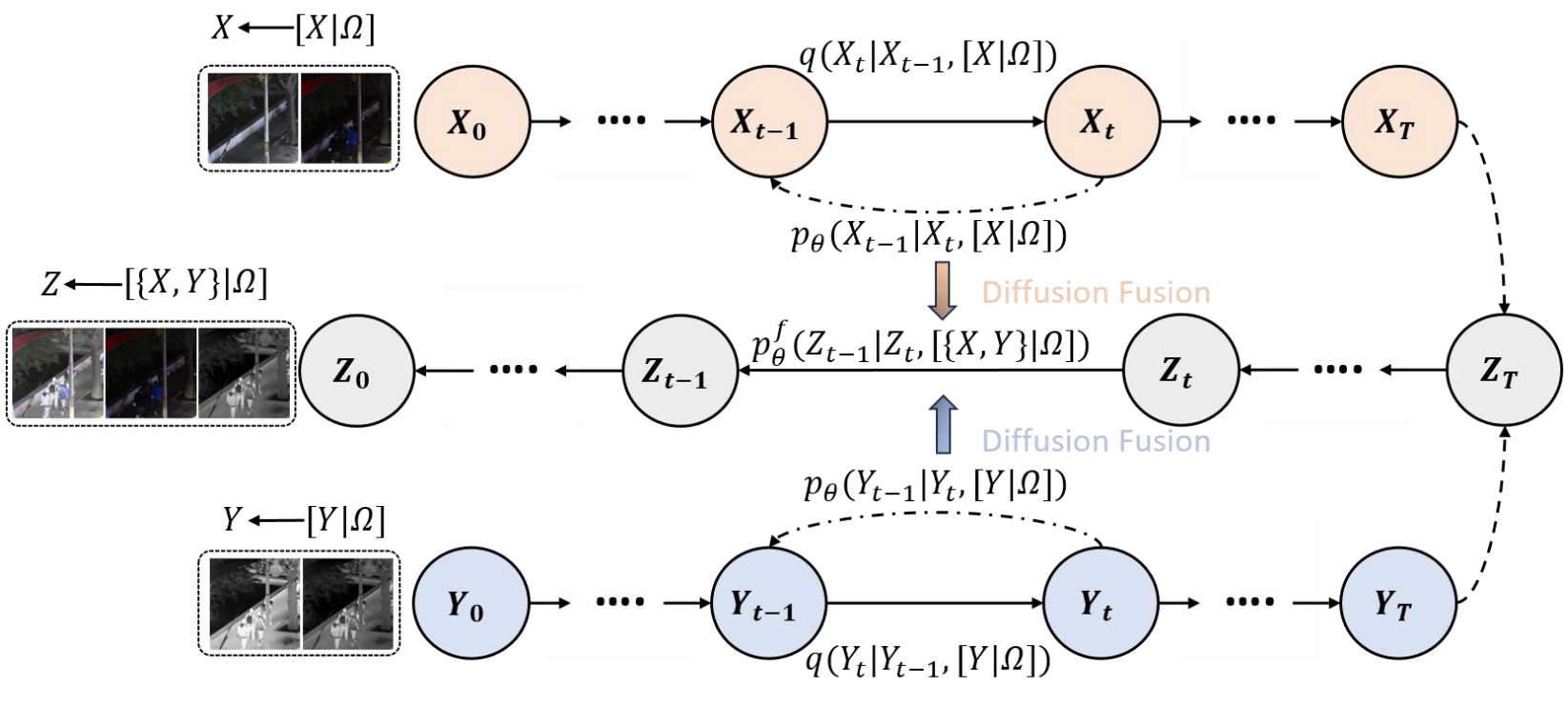
- Current methods falter in scenes with degradation, especially composite degradation, which we refer to as the **composite degradation challenge**. Essentially, current methods prioritize multi-modal information integration without considering effective information restoration from degradation. As a result, the fused images still show a considerable degree of degradation, even obscuring valuable scene details.

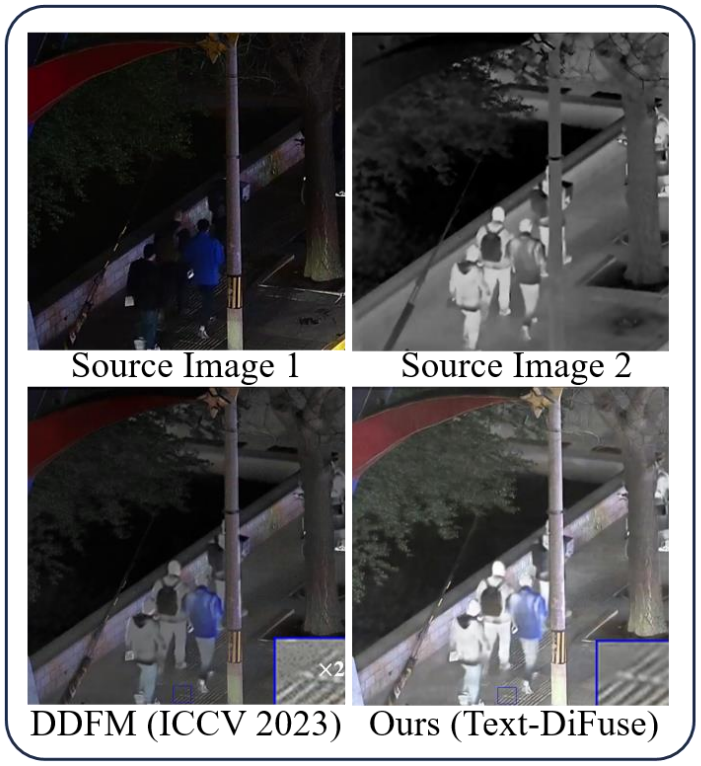
☐ Under-Customization Objects Limitation

- Existing fusion methods fail to account for the specificity of objects in the scene (*e.g.*, pedestrians, vehicles), applying the same fusion rules indiscriminately to both foreground and background. This lack of differentiation, termed the **under-customization objects limitation**, is unreasonable and may compromise the delineation of crucial objects.

➢ Method

☐ Explicitly Couples Information Fusion and Diffusion



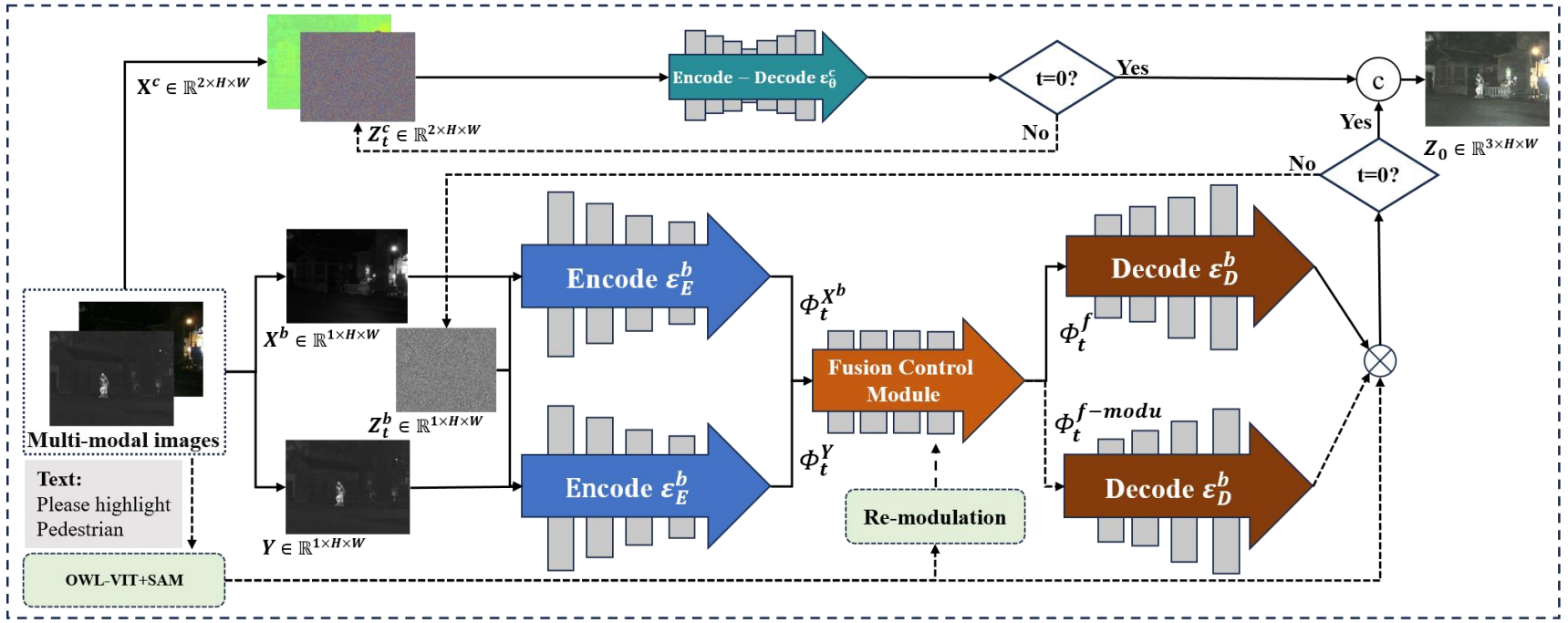


$$\nabla_\theta \| \epsilon_t - \epsilon_\theta(s_t, \Omega(s), t) \|^2 + \lambda D_{KL}(q(s_{t-1}|s_t, s_0, \Omega(s)) \| p_\theta(s_{t-1}|s_t, \Omega(s))), t > 1$$

$$\nabla_\theta \| \epsilon_t - \epsilon_\theta(s_t, \Omega(s), t) \|^2 - \lambda log p_\theta(s_0|s_1, \Omega(s)), t = 1$$

➢ Method

☐ Text-Controlled Fusion Re-Modulation Strategy



- With the degradation removing has already been encompassed in ' ε_E^b ' and ' ε_D^b '.

$$\left[\left\{\Phi_t^{X^b}, \Phi_t^Y\right\}|\Omega\right] = \varepsilon_E^b\left(Z_t^b, \left[\{X^b, Y\}|\Omega\right], t\right), t \in \{T, \cdots 0\}.$$

$$\left[\Phi_t^f|\Omega\right] = \left[\left\{\Phi_t^{X^b}, \Phi_t^Y\right\}|\Omega\right] \odot \{\omega_t^{X^b}, \omega_t^Y\}$$

$$\epsilon_\theta(t), \nu_\theta(t) = \varepsilon_D^b\left(\left[\Phi_t^f|\Omega\right], t\right)$$

$$\hat{Z}_0^b = \frac{Z_t^b - \sqrt{1-\bar{\alpha}}\epsilon_\theta(t)}{\sqrt{\bar{\alpha}}}$$

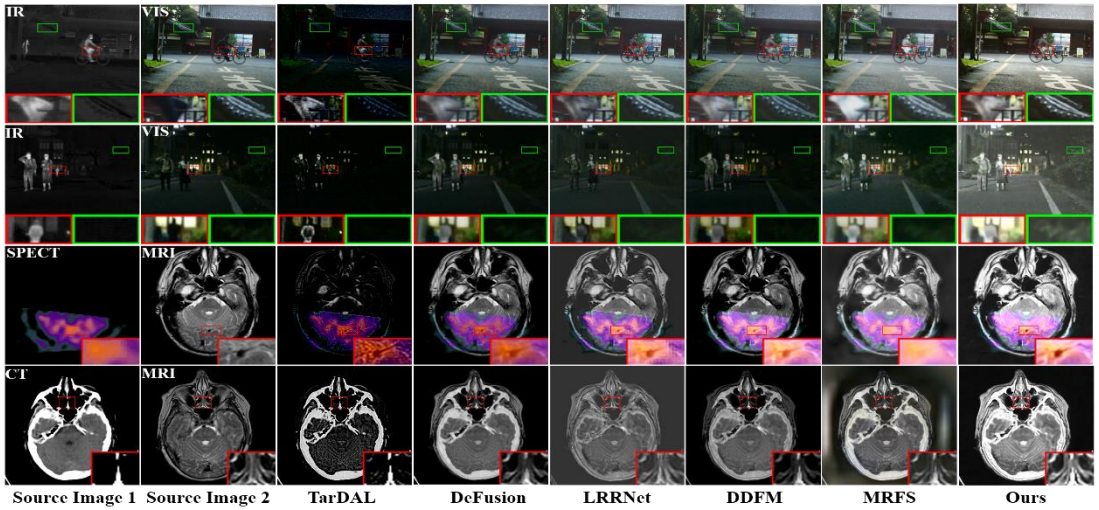
$$\mathcal{L}_{\text{int}} = \| \left|\hat{Z}_0^b\left(Z_t^b, \epsilon_\theta(t)\right)\right| - \max\{|X^b|, |Y|\} \|$$

$$\mathcal{L}_{grad} = \| \nabla\hat{Z}_0^b\left(Z_t^b, \epsilon_\theta(t)\right) - \max\{\nabla X^b, \nabla Y\} \|$$

- Potential modulation: $\left[\Phi_t^{f-modu}|\Omega\right] = \left[\left\{\Phi_t^{X^b}, \Phi_t^Y\right\}|\Omega\right] \odot \{\omega_t^{X^b}, \omega_t^Y\} \odot \{\kappa^{X^b}, \kappa^Y\}$

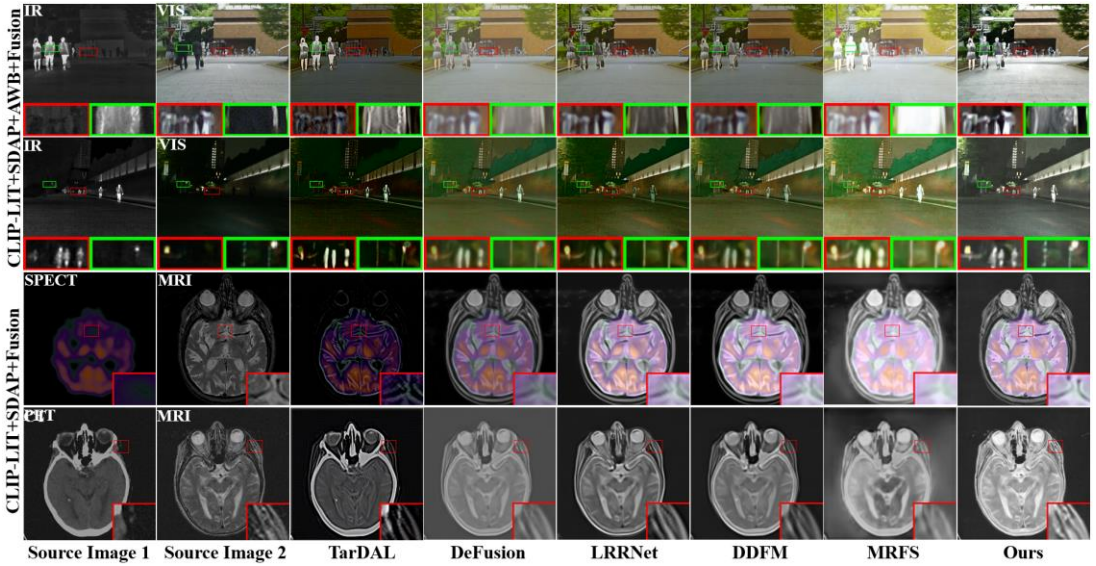
➢ Results

- **Multi-modal Image Fusion in the Scenario with Composite Degradation**



Methods	MSRS DataSet					Havard Medicine Dataset				
	EN ↑	AG ↑	SD ↑	SCD ↑	VIF ↑	EN ↑	AG ↑	SD ↑	SCD ↑	VIF ↑
RFN-Nest (InF'21)	5.89	1.84	26.03	1.41	0.63	5.34	4.07	63.65	1.58	0.43
GANMcC (TIM'21)	6.03	1.91	25.58	1.37	0.65	5.38	5.13	55.00	1.11	0.43
SDNet (IJCV'21)	4.90	2.33	16.35	0.91	0.49	5.56	6.22	46.64	0.48	0.38
U2Fusion (TPAMI'22)	5.19	2.46	24.82	1.21	0.52	5.22	6.08	53.20	1.03	0.41
TarDAL (CVPR'22)	3.30	2.04	18.52	0.63	0.15	5.66	5.13	41.94	1.12	0.18
DeFusion (ECCV'22)	6.22	2.31	32.34	1.36	0.75	4.96	4.47	55.45	0.93	0.48
LRRNet (TPAMI'23)	5.89	2.19	26.64	0.75	0.52	5.34	5.39	45.89	0.59	0.40
DDFM (ICCV'23)	5.81	2.65	24.98	1.37	0.62	5.00	5.04	63.53	1.59	0.48
MRFS (CVPR'24)	6.91	2.67	40.95	1.23	0.75	**7.24**	4.41	70.75	1.53	0.41
Ours (Text-DiFuse)	**7.08**	**3.31**	**47.44**	**1.44**	**0.76**	6.44	**7.31**	**80.19**	**1.69**	**0.49**

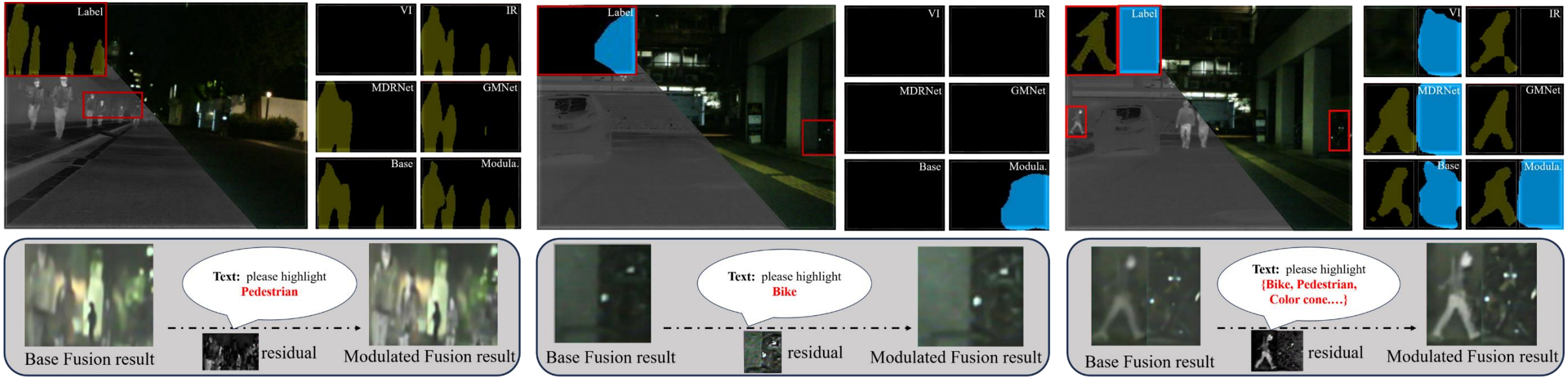
- **Enhancement Plus Multi-modal Image Fusion**



Methods		MSRS Dataset					Havard Medicine Dataset				
		EN ↑	AG ↑	SD ↑	SCD ↑	VIF ↑	EN ↑	AG ↑	SD ↑	SCD ↑	VIF ↑
CLIP-LIT SDAP AWB	RFN-Nest	6.43	2.23	27.17	1.38	0.60	5.72	4.11	77.46	1.64	0.35
	GANMcC	6.25	2.06	24.55	1.31	0.57	5.80	5.28	66.37	1.19	0.31
	SDNet	5.84	2.99	20.26	1.08	0.52	5.91	6.00	60.83	1.15	0.30
	U2Fusion	6.55	3.55	29.08	1.32	0.58	5.68	6.09	71.59	1.56	0.32
	TarDAL	5.29	**4.42**	25.22	1.00	0.35	6.11	4.81	36.54	0.69	0.23
	DeFusion	6.31	2.07	25.52	1.16	0.59	6.08	4.27	67.77	1.38	0.35
	LRRNet	6.55	2.68	31.19	1.13	0.54	5.86	5.23	62.91	1.34	0.21
	DDFM	6.39	2.43	26.40	1.16	0.60	5.70	4.48	77.40	1.64	0.35
	MRFS	6.84	2.86	32.28	1.28	0.58	**7.18**	4.19	**87.53**	1.50	0.31
Ours (Text-DiFuse)		**7.08**	3.31	**47.44**	**1.44**	**0.76**	6.44	**7.13**	80.19	**1.69**	**0.49**

➢ Results

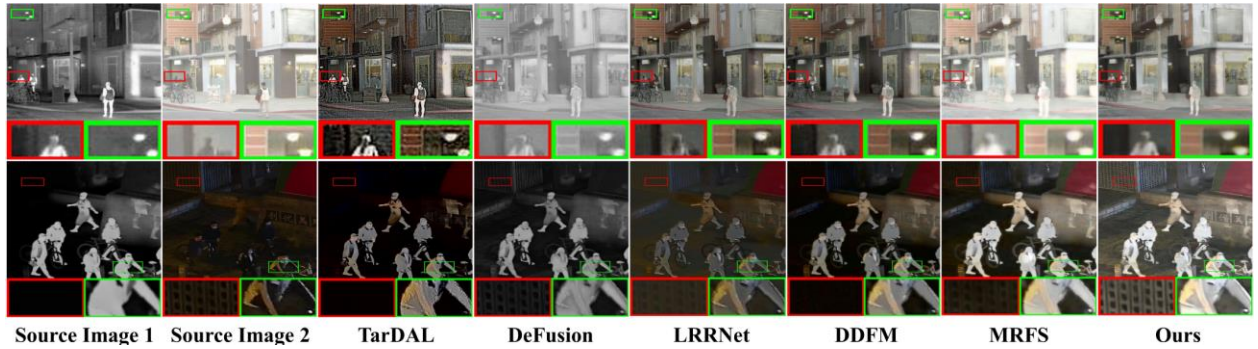
- **Text Re-Modulation Verification**



Segmentation	Source	Background	Car	Person	Bike	Curve	Car Stop	Cuardrail	Color cone	Bump	mIoU
MFNet	RGB-T	96.26	60.95	53.44	43.14	22.94	9.44	0.00	18.80	23.47	36.49
FEANet	RGB-T	98.00	87.41	70.30	62.74	45.33	29.80	0.00	29.07	48.95	55.28
EGFNet	RGB-T	98.01	87.84	71.12	61.08	46.48	22.10	6.64	55.35	47.12	54.76
CMX-B2	RGB-T	97.39	84.23	67.12	56.93	41.11	39.56	18.94	48.84	54.42	58.31
GMNet	RGB-T	98.00	86.46	73.05	61.72	43.96	42.25	14.52	48.70	47.72	57.34
MDRNet	RGB-T	97.90	87.07	69.81	60.87	47.80	34.18	8.21	50.18	54.98	56.78
SegNext-Base	IR	97.79	84.89	70.73	56.29	41.94	24.15	7.60	35.91	48.64	51.99
	VI	97.93	88.29	62.42	63.67	35.34	36.95	5.77	51.20	47.74	54.37
	Our basis	98.11	88.66	70.00	64.30	43.07	30.25	11.95	55.14	56.27	57.53
	Our modulatable	98.18	88.32	72.23	65.02	44.79	33.11	13.76	56.32	55.97	**58.63**

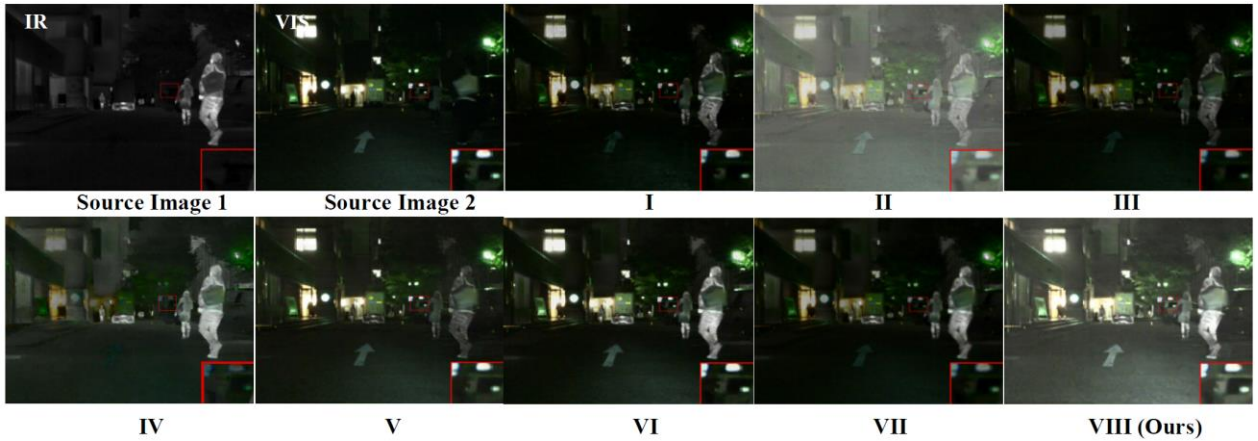
➢ Results

- **Generalization Evaluation**



Methods	LLVIP Dataset					RoadScene Dataset				
	EN ↑	AG ↑	SD ↑	SCD ↑	VIF ↑	EN ↑	AG ↑	SD ↑	SCD ↑	VIF ↑
RFN-Nest	6.37	2.24	26.66	1.63	0.73	7.37	2.62	46.77	1.66	0.58
GANMcC	6.24	2.09	27.02	1.59	0.65	7.24	3.58	43.68	1.39	0.57
SDNet	6.00	2.74	23.05	1.24	0.62	7.18	4.86	40.63	1.16	0.66
U2Fusion	5.52	2.69	21.12	1.32	0.61	7.32	4.92	43.99	1.49	0.66
TarDAL	3.85	2.59	23.05	0.92	0.22	7.35	**11.84**	52.30	0.97	0.47
DeFusion	6.46	2.36	29.48	1.48	0.82	6.97	2.85	35.96	0.98	0.59
LRRNet	5.67	2.28	19.49	1.06	0.57	7.19	3.55	44.01	1.47	0.58
DDFM	6.46	3.51	30.64	1.72	0.70	7.30	3.63	44.19	1.57	0.65
MRFS	7.00	2.34	40.90	1.67	0.86	7.18	2.70	46.57	1.20	0.52
Ours (Text-DiFuse)	**7.08**	**3.99**	**41.78**	**1.73**	**0.87**	**7.46**	2.96	**52.84**	**1.67**	**0.66**

- **Ablation Studies**



Index	Diff.	$\mathcal{L}_{int}$	$\mathcal{L}_{grad}$	FCM	EN	AG↑	SD↑	SCD↑	VIF↑
I	✓	✓	✓	✗/max	5.71	1.90	25.66	1.30	0.49
II	✓	✓	✓	✗/add	6.08	1.99	20.14	1.11	0.58
III	✓	✓	✓	✗/mean	5.60	1.49	20.71	1.15	0.56
IV	✓	✓	✓	✗/variance	5.91	1.90	27.59	0.91	0.46
V	✓	✓	✗	✓	6.20	2.75	33.78	1.42	0.73
VI	✓	✗	✓	✓	6.67	3.25	45.60	1.43	0.76
VII	✗/AE	✓	✓	✓	6.37	2.26	37.48	1.42	0.69
VIII	✓	✓	✓	✓	**7.08**	**3.31**	**47.44**	**1.44**	**0.76**

- ✓ **Model I**: removing FCM with using maximum rule
- ✓ **Model II**: removing FCM with using addition rule
- ✓ **Model III**: removing FCM with using mean rule
- ✓ **Model IV**: removing FCM with using variance-based rule
- ✓ **Model V**: removing $\mathcal{L}_{grad}$
- ✓ **Model VI**: removing $\mathcal{L}_{int}$
- ✓ **Model VII**: removing diffusion with using AE route
- ✓ **Model VIII**: our full model

➢ Conclusion & Contribution

- We propose a novel **explicit coupling paradigm of information fusion and diffusion**, solving the composite degradation challenge in the task of multi-modal image fusion.
- A **text-controlled fusion re-modulation strategy** is designed, allowing users to customize fusion rules with language to enhance the salience of objects of interest. This interactively improves the visual quality and semantic attributes of fused images.
- Our Text-DiFuse demonstrates the advantages over state-of-the-art methods in terms of **degradation robustness**, **generalization ability**, and **semantic properties**.