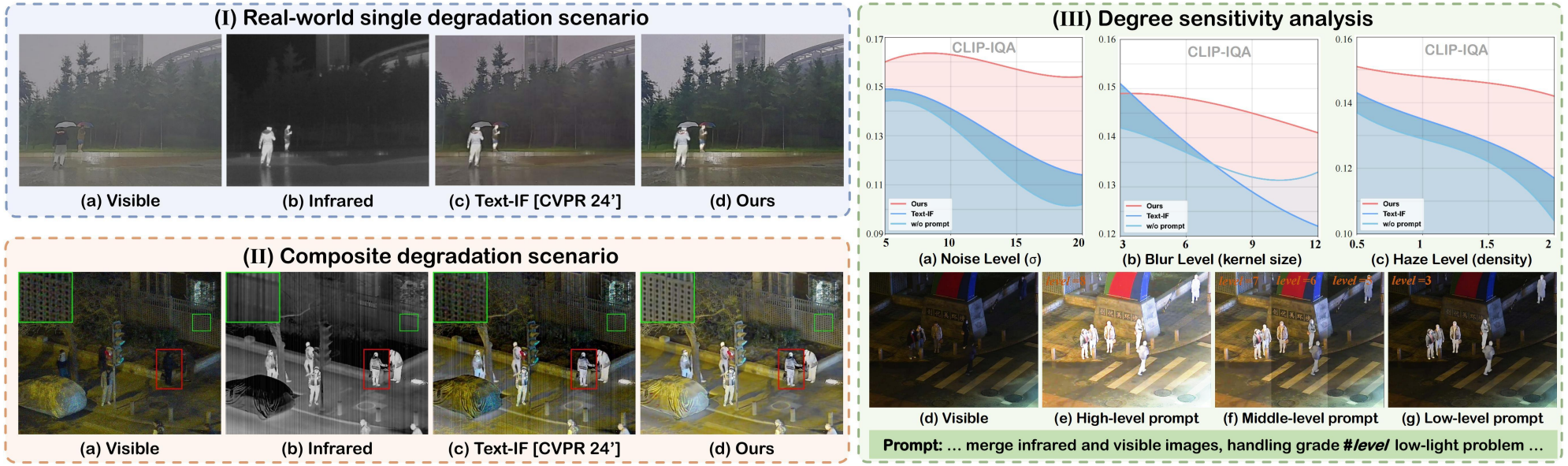
ControlFusion: A Controllable Image Fusion Network with Language-Vision Degradation Prompts

Linfeng Tang[1,†], **Yeda Wang**[1,†], **Zhanchuan Cai**[2], **Junjun Jiang**[3], **Jiayi Ma**[1,*]
[1]Electronic Information School, Wuhan University
[2]School of Computer Science and Engineering, Macau University of Science and Technology
[3]Faculty of Computing, Harbin Institute of Technology
linfeng0419@gmail.com, wangyeda@whu.edu.cn,
zccai@must.edu.mo, jiangjunjun@hit.edu.cn, jyma2010@gmail.com,

https://github.com/Linfeng-Tang/ControlFusion

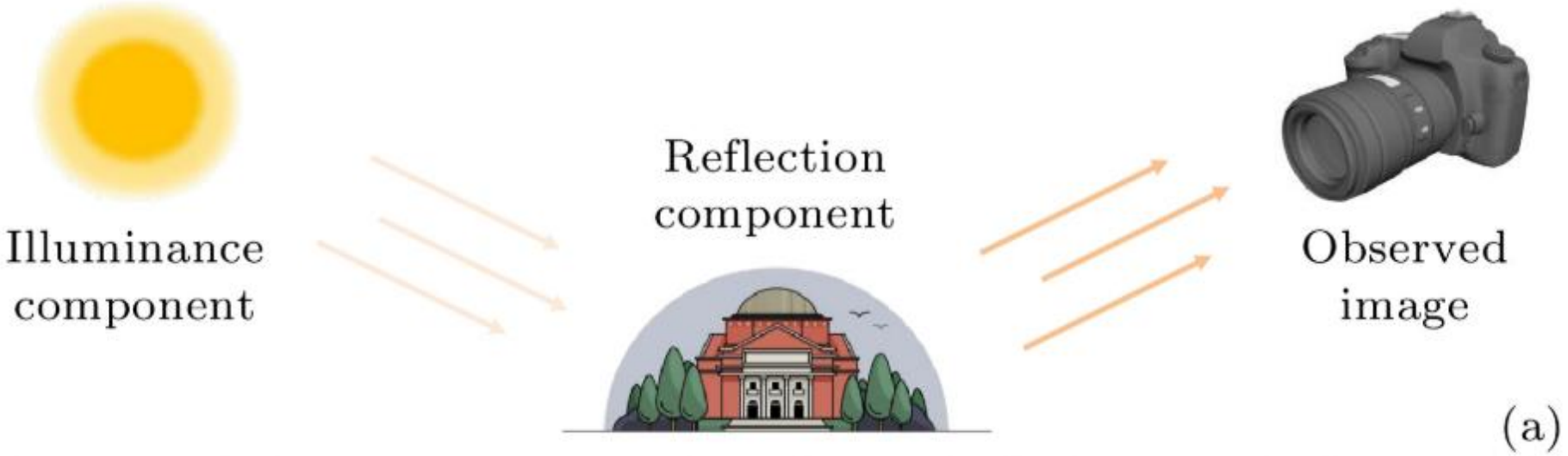
This work was supported by National Natural Science Foundation of China (**No. 62276192**).

Motivation



- Existing restoration-fusion methods **overlooking the domain gap** between simulated data and realistic images, which hampers their generalizability in practical scenarios.
- Existing methods are **tailored for specific or single types of degradation**, making them ineffective in handling more complex composite degradations.
- Existing methods **lack degradation level modeling**, causing a sharp decline in performance as degradation intensifies.

- **Methodology**



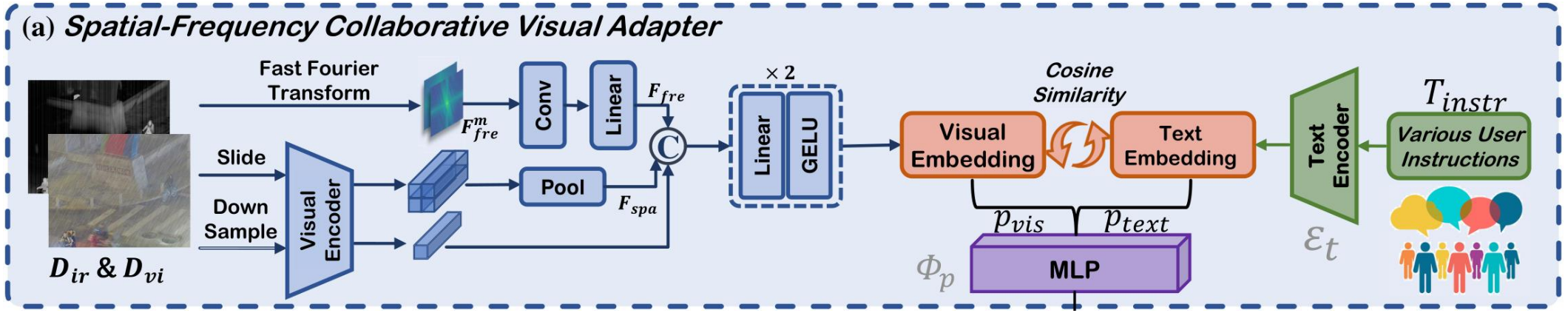
Stripe noise: $D_{ir}^{s} = \mathcal{P}_s(I_{ir}) = \alpha \cdot I_{ir} + \mathbf{1}_H \mathbf{n}^\top,$

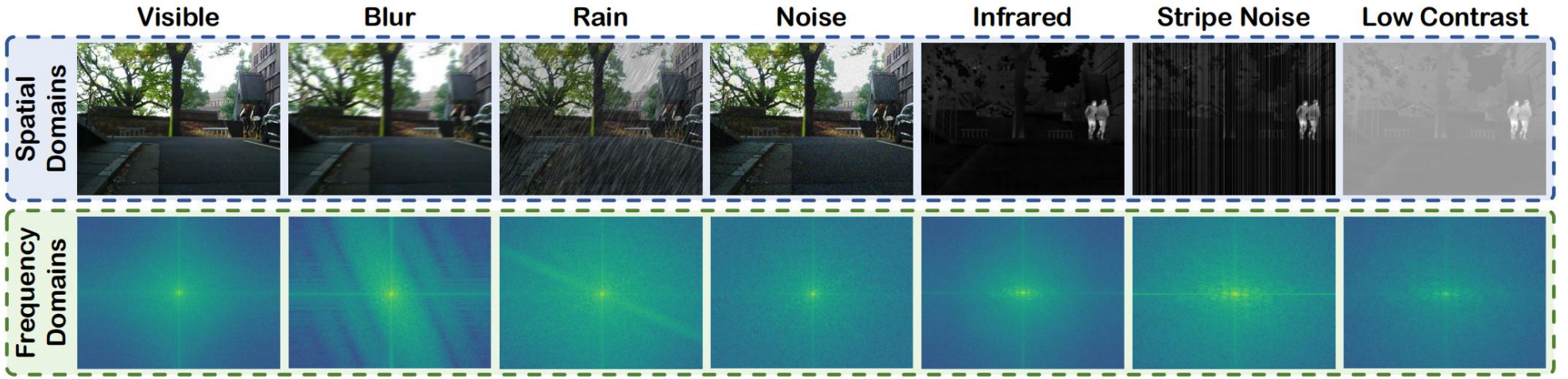
Blur&Noise: $D_{m}^{s} = \mathcal{P}_s(I_m) = I_m * K(N, \theta) + \mathcal{N}(0, \sigma^2),$

Weather related degradation: $D_{vi}^{w} = \mathcal{P}_w(I_{vi}) = I_{vi} \cdot t + A(1-t) + R,$

Illumination degradation: $D_{vi}^{i} = \mathcal{P}_i(I_{vi}) = \frac{I_{vi}}{L} \cdot L^{\gamma},$

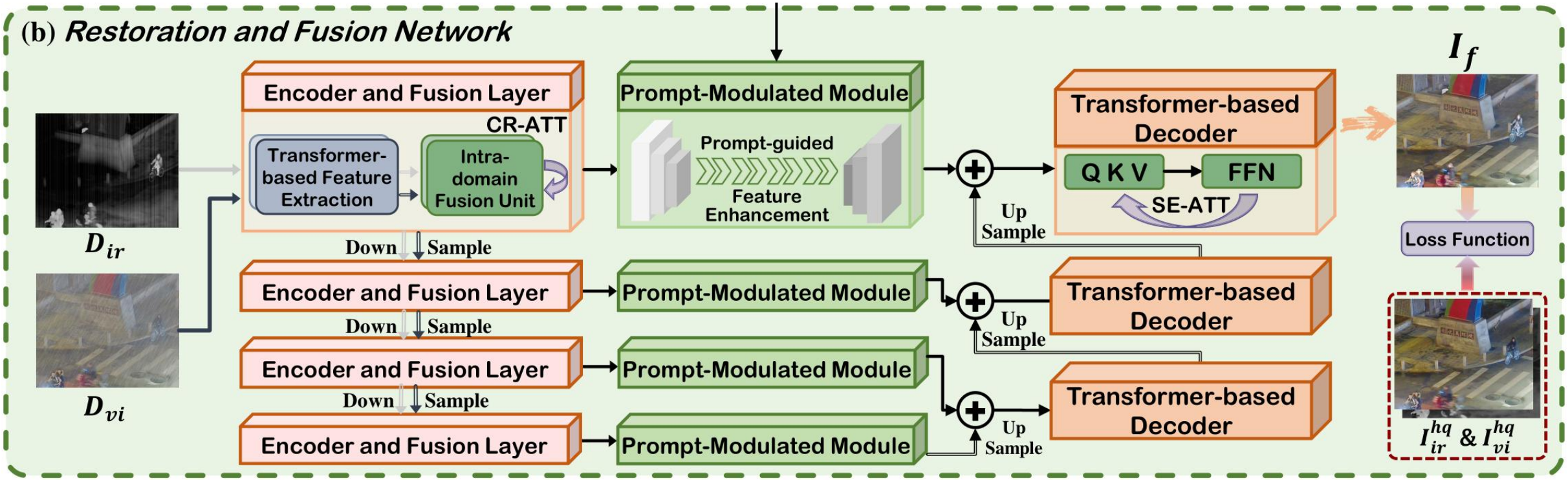
- **Methodology**





$$F^m_{fre} = \sum_{x=0}^{W-1}\sum_{y=0}^{H-1} D_m(x,y)e^{-j2\pi\left(\frac{ux}{W}+\frac{vy}{H}\right)}, F_{fre} = \text{Linear}\left(\text{Conv}\left([F^{ir}_{fre}, F^{vi}_{fre}]\right)\right)$$

- **Methodology**



$$\{Q_{ir}, K_{ir}, V_{ir}\} = \mathcal{F}_{ir}^{qkv}(F_{ir}), \{Q_{vi}, K_{vi}, V_{vi}\} \quad = \mathcal{F}_{vi}^{qkv}(F_{vi}).$$

$$F_f^{ir} = \text{softmax}\left(\frac{Q_{vi}K_{ir}}{\sqrt{d_k}}\right)V_{ir}, F_f^{vi} = \text{softmax}\left(\frac{Q_{ir}K_{vi}}{\sqrt{d_k}}\right)V_{vi},$$

Methodology

- **Stage I loss function**

$$\mathcal{L}_I = \lambda_1 \underbrace{\|p_{vis} - p_{text}\|^2}_{\mathcal{L}_{\text{mse}}} + \lambda_2 \underbrace{(1 - \frac{p_{vis} \cdot p_{text}}{\|p_{vis}\|\|p_{text}\|})}_{\mathcal{L}_{\cos}},$$

- **Stage II loss function**

$$\mathcal{L}_{int} = \frac{1}{HW}\|I_f - \max(I_{ir}^{hq}, I_{vi}^{hq})\|_1,$$

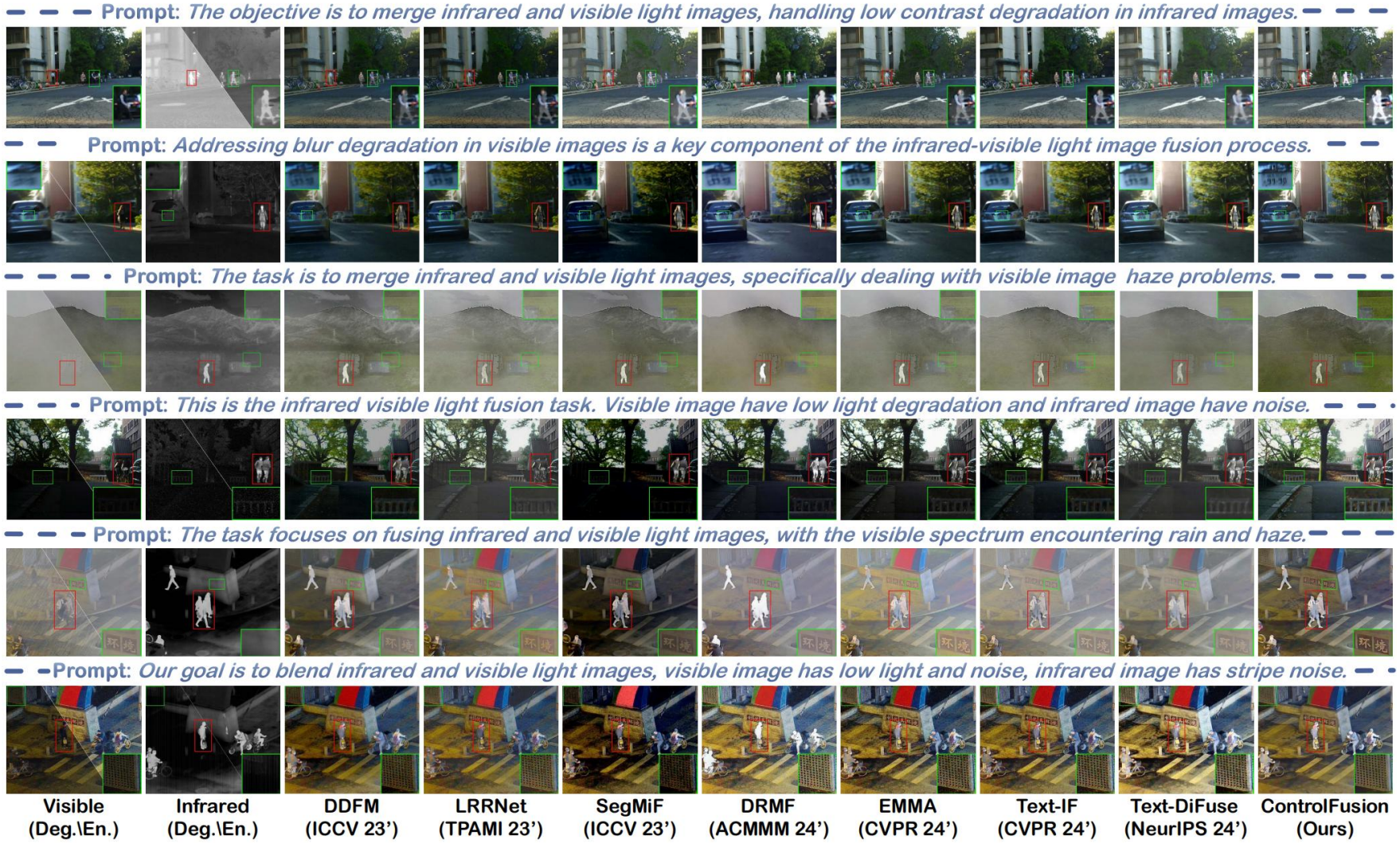
$$\mathcal{L}_{ssim} = 2 - (\text{SSIM}(I_f, I_{ir}^{hq}) + \text{SSIM}(I_f, I_{vi}^{hq})).$$

$$\mathcal{L}_{grad} = \frac{1}{HW}\|\nabla I_f - \max(\nabla I_{ir}^{hq}, \nabla I_{vi}^{hq})\|_1,$$

$$\mathcal{L}_{color} = \frac{1}{HW}\|\mathcal{F}_{CbCr}(I_f) - \mathcal{F}_{CbCr}(I_{vi}^{hq})\|_1,$$

$$\mathcal{L}_{II} = \alpha_{int} \cdot \mathcal{L}_{int} + \alpha_{ssim} \cdot \mathcal{L}_{ssim} + \alpha_{grad} \cdot \mathcal{L}_{grad} + \alpha_{color} \cdot \mathcal{L}_{color}$$

- Experimental results(Fusion results on challenging scenarios)

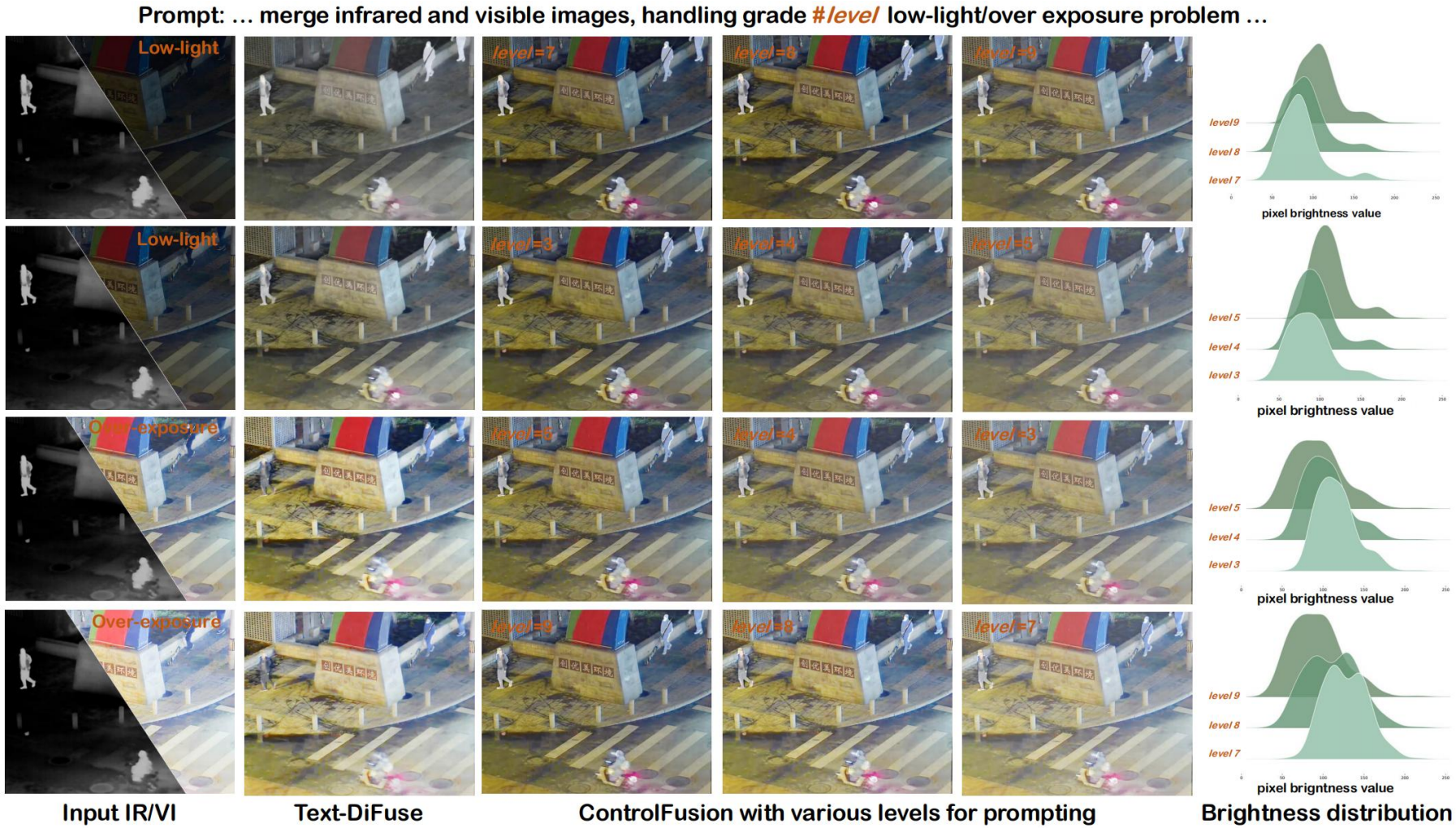


WUHAN UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

- Experimental results (Generalization results)

- Experimental results (Fusion results with various level prompts)



- Experimental results (Quantitative results on practical scenarios)

Methods	MSRS				LLVIP				RoadScene				FMB			
	EN	SD	VIF	Qabf	EN	SD	VIF	Qabf	EN	SD	VIF	Qabf	EN	SD	VIF	Qabf
DDFM	6.431	47.815	0.844	0.643	6.914	48.556	0.693	0.517	6.994	47.094	0.775	0.595	6.426	40.597	0.495	0.442
DRMF	6.268	45.117	0.669	0.550	6.901	50.736	0.786	0.626	6.231	44.221	0.728	0.527	6.842	41.816	0.578	0.372
EMMA	6.747	52.753	0.886	0.605	6.366	47.065	0.743	0.547	6.959	46.749	0.698	0.664	6.788	38.174	0.542	0.436
LRRNet	6.761	49.574	0.713	0.667	6.191	48.336	0.864	0.575	7.185	46.400	0.756	0.658	6.432	48.154	0.501	0.368
SegMiF	7.006	57.073	0.764	0.586	7.260	45.892	0.539	0.459	6.736	48.975	0.629	0.584	6.363	47.398	0.539	0.482
Text-IF	6.619	55.881	0.753	0.656	6.364	49.868	0.859	0.566	6.836	47.596	0.634	0.609	**7.397**	47.726	0.568	0.528
Text-DiFuse	6.990	56.698	0.850	0.603	**7.546**	55.725	0.883	0.659	6.826	50.230	0.683	0.662	6.888	49.558	0.793	0.653
ControlFusion	**7.340**	**60.360**	**0.927**	**0.718**	7.354	**56.631**	**0.968**	**0.738**	**7.421**	**51.759**	**0.817**	**0.711**	7.036	**50.905**	**0.872**	**0.730**

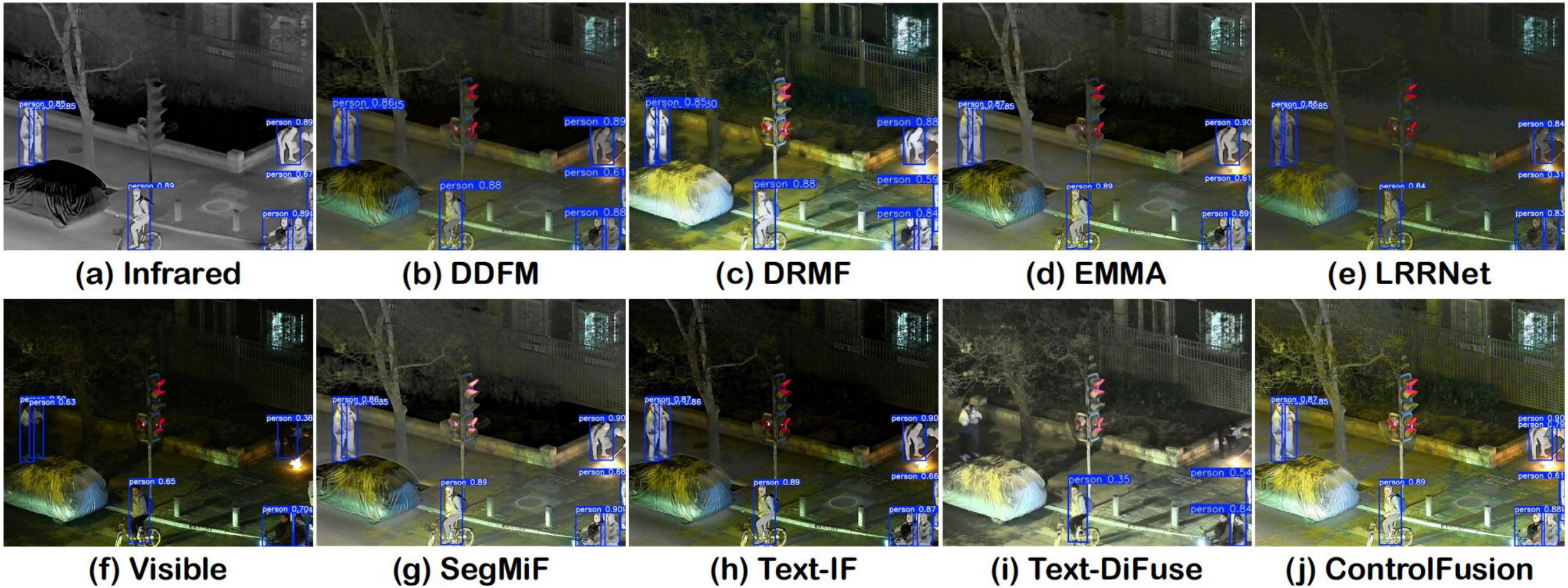
- Experimental results (Quantitative results on challenging scenarios)

Methods	VI (Blur)				VI (Rain)				VI (Low light, LL)				VI (Over-exposure, OE)			
	CLIP-IQA	MUSIQ	TReS	SD	CLIP-IQA	MUSIQ	TReS	SD	CLIP-IQA	MUSIQ	TReS	SD	CLIP-IQA	MUSIQ	TReS	SD
DDFM	0.141	39.421	39.047	35.411	0.191	38.836	46.285	36.376	0.156	39.495	41.782	31.759	0.143	43.167	43.440	32.099
DRMF	0.128	40.739	40.968	40.722	0.174	48.164	48.565	41.174	0.143	41.428	37.947	38.287	0.190	48.334	42.582	44.256
EMMA	0.131	43.472	41.744	42.553	0.138	45.824	44.916	43.378	0.158	39.674	44.827	40.857	0.180	46.731	47.616	40.242
LRRNet	0.163	42.981	37.268	45.389	0.185	43.291	41.891	46.285	0.164	40.486	34.836	41.639	0.160	42.548	48.414	42.190
SegMiF	0.152	43.005	43.516	44.000	0.195	40.528	49.094	44.274	0.177	44.073	48.376	44.829	0.166	49.132	38.019	38.484
Text-IF	0.164	44.801	46.542	48.401	0.164	41.287	47.380	49.298	0.163	41.096	49.174	47.257	0.172	40.298	45.599	47.330
Text-DiFuse	0.172	44.958	47.699	46.376	0.173	39.243	50.017	47.297	**0.192**	46.734	50.126	49.883	0.183	39.095	49.596	50.279
ControlFusion	**0.184**	**47.849**	**50.240**	**50.287**	**0.196**	**52.319**	**52.465**	**50.901**	0.183	**48.420**	**51.072**	**53.787**	**0.191**	**50.301**	**52.961**	**54.218**

Methods	VI (Rain and Haze, RH)				IR (Low-contrast, LC)				IR (Random noise, RN)				IR (Stripe noise, SN)			
	CLIP-IQA	MUSIQ	TReS	EN	CLIP-IQA	MUSIQ	TReS	SD	CLIP-IQA	MUSIQ	TReS	EN	CLIP-IQA	MUSIQ	TReS	EN
DDFM	0.158	43.119	44.095	6.897	0.172	44.490	44.545	37.452	0.186	34.059	32.807	6.029	**0.209**	48.479	32.377	6.399
DRMF	0.171	45.524	43.693	6.230	**0.208**	43.982	45.561	40.315	0.192	44.617	44.085	5.744	0.187	44.714	43.650	6.354
EMMA	0.169	39.092	46.046	6.517	0.154	48.724	43.113	53.861	0.172	39.666	44.466	6.184	0.153	42.802	44.239	6.382
LRRNet	0.170	48.571	48.973	7.363	0.151	48.718	45.266	51.605	0.158	48.274	37.090	7.295	0.137	46.511	36.610	**7.702**
SegMiF	0.147	46.139	45.019	7.281	0.160	44.659	51.427	44.448	0.180	42.664	39.496	6.669	0.169	49.887	39.247	6.855
Text-IF	0.178	50.568	50.271	6.956	0.187	49.299	49.266	47.032	0.169	46.647	48.491	6.256	0.161	49.019	47.755	6.085
Text-DiFuse	0.175	52.788	53.073	7.470	0.165	50.092	51.715	39.429	**0.203**	49.278	49.479	6.982	0.194	**51.762**	49.288	7.089
ControlFusion	**0.189**	**54.287**	**54.465**	**7.891**	0.196	**51.986**	**52.846**	**57.827**	0.189	**50.711**	**51.668**	**7.724**	0.200	50.097	**50.264**	7.619

Methods	VI (OE) and IR (LC)				VI(Low light and Noise, LN)				VI (RH) and IR (RN)				VI (LL) and IR (SN)			
	CLIP-IQA	MUSIQ	TReS	SD	CLIP-IQA	MUSIQ	TReS	EN	CLIP-IQA	MUSIQ	TReS	SD	CLIP-IQA	MUSIQ	TReS	EN
DDFM	0.168	43.814	41.894	36.095	0.172	48.293	31.791	6.298	0.151	33.440	32.134	37.342	**0.189**	36.433	42.630	5.776
DRMF	0.184	42.399	39.374	40.847	0.201	44.363	43.063	5.875	0.174	43.663	43.858	37.997	0.142	38.241	41.049	5.280
EMMA	0.130	39.892	42.076	43.362	0.174	42.201	43.382	5.838	0.165	39.146	44.458	51.205	0.130	37.367	43.888	6.318
LRRNet	0.136	47.209	42.636	46.684	0.144	46.386	35.779	**7.306**	0.128	47.954	36.831	49.917	0.154	38.426	35.970	7.007
SegMiF	0.114	44.021	42.256	33.647	0.136	49.178	38.570	5.819	0.147	42.354	39.156	31.717	0.151	41.287	37.079	6.767
Text-IF	0.174	48.808	47.998	48.848	0.217	48.100	47.510	5.204	0.158	45.821	47.626	46.543	0.140	41.429	46.220	5.525
Text-DiFuse	0.131	49.021	**50.980**	47.640	0.185	**50.775**	48.610	6.440	**0.181**	48.645	48.937	38.808	0.161	47.734	48.448	6.738
ControlFusion	**0.187**	**50.479**	50.298	**50.955**	**0.225**	49.333	**49.513**	7.111	0.179	**50.107**	**51.091**	**55.417**	0.167	**50.632**	**48.971**	**7.055**

- Experimental results (Object detection)



- Experimental results (Object detection)

Methods	Prec.	Recall	AP@0.50	AP@0.75	mAP@0.5:0.95
DDFM	0.947	0.848	0.911	0.655	0.592
DRMF	0.958	0.851	0.937	0.672	0.607
EMMA	0.942	0.872	0.927	0.647	0.598
LRRNet	0.939	0.878	0.933	0.672	0.608
SegMiF	0.965	**0.896**	0.931	**0.690**	0.603
Text-IF	0.959	0.892	0.939	0.655	0.601
Text-DiFuse	0.961	0.885	0.941	0.656	0.606
ControlFusion	**0.971**	0.889	**0.949**	0.685	**0.609**

- **Experimental results (Ablation)**

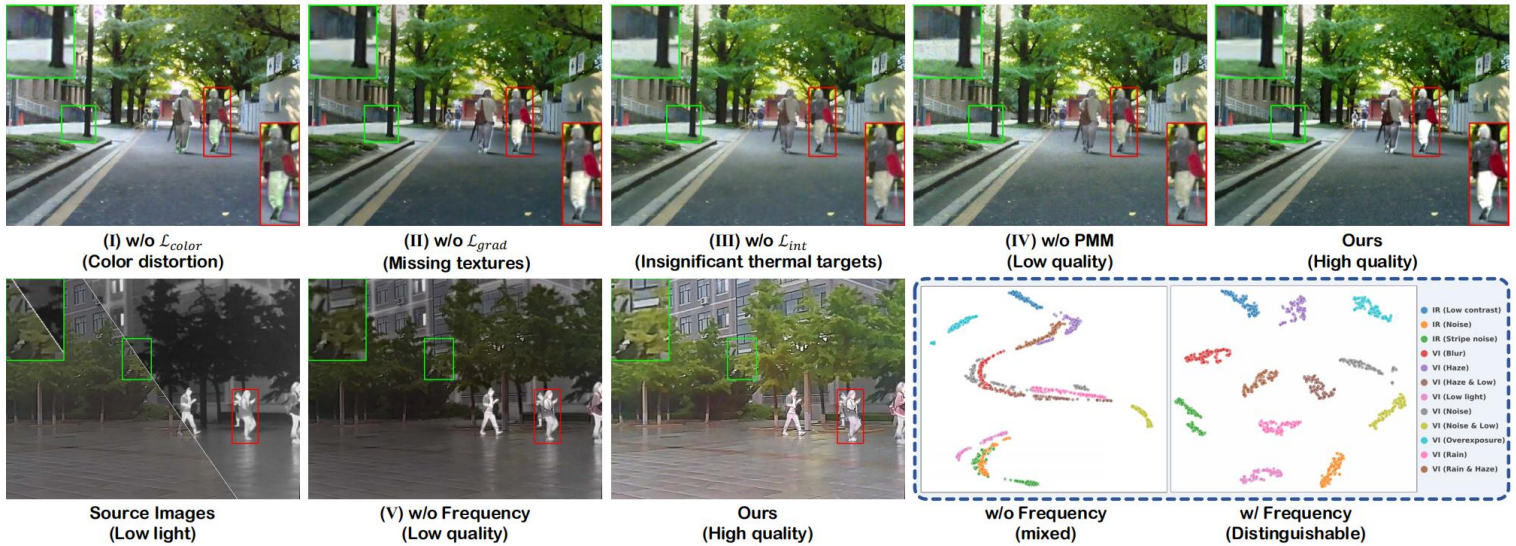


Table 6: Quantitative results of the ablation studies.

Configs	VI(LL & Noise)				VI(OE) and IR(LC)				VI (RH) and IR(Noise)			
	CLIP-IQA	MUSIQ	TReS	EN	CLIP-IQA	MUSIQ	TReS	SD	CLIP-IQA	MUSIQ	TReS	SD
I	0.132	42.424	42.839	4.788	0.166	41.983	42.841	39.917	0.147	43.619	47.932	48.935
II	0.152	45.582	45.358	5.855	0.151	43.646	42.002	39.208	0.167	47.862	45.007	44.347
III	0.154	46.571	44.495	5.013	0.155	44.286	44.561	42.068	0.156	43.007	43.816	41.544
IV	0.129	38.960	41.310	5.414	0.172	41.743	39.125	38.748	0.118	48.882	46.245	45.910
V	0.173	45.281	46.291	6.279	0.181	45.386	47.519	46.860	0.149	46.714	48.094	46.950
Ours	**0.225**	**49.333**	**49.513**	**7.111**	**0.187**	**50.479**	**50.298**	**50.955**	**0.179**	**50.107**	**51.091**	**55.417**

Thank You for Watching!

ControlFusion: A Controllable Image Fusion Network with Language-Vision Degradation Prompts

https://github.com/Linfeng-Tang/ControlFusion