Supplementary of "MMAIF: Multi-task and Multi-degradation All-in-One for Image Fusion with Language Guidance"

Zihan Cao[1] Yu Zhong[1] Ziqi Wang[1] Liang-Jian Deng[1, 2, ⋆]
[1] University of Electronic Science and Technology of China (UESTC)
[2] Multi-Hazard Early Warning Key Laboratory of Sichuan Province, UESTC
{iamzihan666,yuuzhong1011}@gmail.com, liangjian.deng@uestc.edu.cn
⋆ Corresponding author

Abstract

In this supplementary material, we provide some background on the mixture-of-experts (MoE) and works about all-in-one restoration. The derivatives of how to sample deterministic trained flow matching model in a stochastic path and the background of recent bridge models. Then, we include the configuration of the dataset collection stage. Finally, we provide more quantitative results of MEF and MFF datasets and visual comparisons with previous methods.

1. Backgrounds

In this section, we introduce the mixture-of-experts (MoE), and recent all-in-one methods for image restoration.

1.1. Mixture-of-Experts (MoE)

Shazeer *et al.* [21] introduced the MoE layer, which comprises a set of E experts $\{\mathcal{E}_i(x)\}$ (each typically a FeedForward network (FFN)) and a learnable router with weights $\mathbf{W}_r \in \mathbb{R}^{D\times E}$. For a given token representation $\mathbf{x} \in \mathbb{R}^{L\times D}$, the router selects the top-k experts based on the gating value computed as $S = \mathbf{x} \cdot \mathbf{W}^{router}$. The MoE layer's output is a weighted combination of the selected experts' computations, where the weights are normalized gating values derived from the softmax distribution. Formally, the MoE layer can be expressed as:

$$\mathbf{x}_i = \sum_{i=1}^{E} (g_{i,j} \mathrm{FFN}_i(\mathbf{x}_j)) + \mathbf{x}_j, \tag{1}$$

$$g_{i,j} = \begin{cases} s_{i,j}, & s_{i,j} \in \mathrm{TopK}(\{s_k | 1 \le k \le N\}, K), \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

$$s_{i,j} = \mathrm{Softmax}_i(\mathbf{x}_j \mathbf{W}_i^{router}). \tag{3}$$

Since tokens select their most suitable experts, this top-k routing mechanism is also referred to as *token-choice routing*.

Recently, the MoE technique has been proven to be effective in language modeling [4, 11, 23]. With large model capacity, the MoE-based model often reaches lower validation loss compared with those dense models. Deepseek v3 [11] has scaled the MoE model into near 685B parameters while only activating near 37B parameters in inference. Many efforts have been made to improve the MoE training [24, 26] and its effectiveness [16]. Nevertheless, the application of MoE in image fusion-related architectures remains scarce. To the best of our knowledge, the MoE-based DiT improved in MMAIF represents the first attempt in the context of the image fusion task.

1.2. All-in-one Image Restoration

Image restoration focuses on recovering a high-quality image from its corrupted version, serving as a fundamental and enduring problem in computer vision. It encompasses various tasks, including image denoising, de-hazing, deblurring, *etc*. Despite the traditional approach of training individual restoration networks for specific degradations, recent works have shifted towards all-in-one methods, which aim to train a *single* model capable of restoring multiple types of degradations simultaneously. Compared to degradation-specific approaches, all-in-one restoration offers superior model storage efficiency and practical applicability. The primary challenge lies in employing a single set of model parameters to address diverse degradation types while accurately restoring corresponding components. To tackle this, AirNet [6] introduced contrastive learning to capture discriminative degradation representations. PromptIR [19] enhanced multi-degradation handling through vision prompts. [18] proposed an adaptive discriminative filter-based model for specific degradations to restore images with unknown degradations. IDR adopted a two-stage ingredients-oriented restoration framework. More recently, CLIP-AWR [22], DA-CLIP [15], and DINO-IR [9] harnessed pre-trained large-scale vision models to achieve state-of-the-art performance in all-in-one restoration. Perceive-IR [29] proposed a two-stage learning strategy, learning semantic visual prompts and harnessing CLIP [20] visual-text alignment to fulfill the restoration task.

However, for degradations in image fusion, the need for both a restoration network and a fusion network to generate a clean image from degraded image pairs introduces significant complexity to the inference pipeline. Particularly, for CLIP-feature-based restoration methods, which additionally require loading the CLIP model during inference, at least three large models coexist to fuse a single image, further complicating the process.

2. Stochastic Path of Flow Matching and Bridge Models

In this section, we discuss the stochasticity in the sampling path for flow matching and other bridge models.

2.1. Stochastic Sampling of Flow Matching

Flow matching (FM) [10] and rectify flow [14] provides a deterministic sampling path, or says ODE, represented as:

$$dZ_t = v_\theta(Z_t, t)dt, \tag{4}$$

where Z_t is the flow hidden state at timestep t, starting from $Z_0 \sim \mathcal{N}(0, \mathbf{I}_d)$ to the data distribution $Z_1 \sim p_{data}$. This ODE path simplifies the generative process against diffusion models that rely on SDE. Some recent works show that the diffusion SDE can be converted to the marginal-preserved ODE without retraining the flow model. Intuitively, the flow ODE can also be converted to SDE that preserves the sample endpoints in order to obtain the *stochastic* sampler in inference.

Let us review the flow matching forward path:

$$Z_t = \alpha_t Z_1 + \beta_t Z_0, \tag{5}$$

when given that noise and data coupling (Z_0, Z_1). In common works, α_t, β_t are often set to $t, 1-t$. The velocity field is given by:

$$v_t = \mathbb{E}[\dot{Z}_t | Z_1 - Z_0]. \tag{6}$$

Formally, we can train a flow model to minimize the real velocity:

$$\mathbb{E}_{t,Z_0,Z_1}[\|\dot{Z}_t - v_\theta(Z_t, t)\|^2]. \tag{7}$$

The marginal preserving law in FM demonstrates that: the distribution of Z_t on the ODE trajectory matches the distribution on the interpolation at each time t. Thus, the final output Z_1 of the ODE can follow the same distribution of p_{data}. However, in practice, when we solve the flow ODE, the sampling error may accumulate over time. These errors come from the model approximation and numerical discretization, causing the drift between the final sampled distribution and the true data distribution.

To solve the problem, we can introduce the Langevin dynamic to correct errors. Let ρ_t be the density of Z_t, representing the true distribution at timestep t. In principle, we can apply a short Langevin step to adjust the sampled trajectory's distribution towards ρ_t:

$$dZ_{t,\tau} = \sigma_t^2 \nabla \rho_t(Z_{t,\tau}) d\tau + \sqrt{2}\sigma_t dW_\tau, \tau \geq 0, \tag{8}$$

where τ is the auxiliary time scale for Langevin dynamics. In FM, the trajectory is already close to ρ_t at t, so, a single step of Langevin dynamics can be sufficient to mitigate the distributional drift, yielding a compositional SDE:

$$dZ_t = v_\theta(Z_t, t)dt + \sigma_t^2 \nabla \log \rho_t(Z_t)dt + \sqrt{2}\sigma_t dW_t, \tilde{Z}_0 = Z_0. \tag{9}$$

The Langevin component acts as a negative feedback loop, correcting distributional drift without bias when $\tilde{Z}_t$ and ρ_t are well aligned. In Eq. (9), we need to estimate the score $\nabla \log \rho_t$. In FM case, the score function can be directly accessed from velocity v_t, without training an additional score model. Specifically, when the coupling Z_0 and Z_1 is independent, by Tweedie's formula, we have:

$$\nabla \log \rho_t(z) = -\frac{1}{\beta_t}\mathbb{E}[Z_0 | Z_t = z]. \tag{10}$$

We have the estimated velocity in hand:

$$v_t(z) = \mathbb{E}[\dot{\alpha}_t Z_1 + \dot{\beta}_t Z_0 | Z_t = z]. \tag{11}$$

Using both equations, we can estimate the score function:

$$\nabla \log \rho_t(z) = \frac{\alpha_t v_t(z) - \dot{\alpha}_t z}{\lambda_t \beta_t}, \tag{12}$$

where $\lambda_t = \dot{\alpha}_t \beta_t - \alpha_t \dot{\beta}_t$. As a result, the flow SDE is:

$$dZ_t = v_t(Z_t, t)dt + \gamma(\alpha_t v_t(Z_t, t) - \dot{\alpha}_t Z_t)dt + \sqrt{2\lambda_t \beta_t \gamma_t} dW_t, \tag{13}$$

where $\sigma_t^2 = \lambda_t \beta_t \gamma_t$. In FM setting, $\alpha_t = t, \beta_t = 1-t$, we have the score function:

$$\nabla \log \rho_t(z) = \frac{t v_t(z) - z}{1-t}, \tag{14}$$

causing the SDE:

$$dZ_t = v_t(Z_t, t)dt + \gamma_t(t v_t(z) - z)dt + \sqrt{2\gamma_t(1-t)} dW_t. \tag{15}$$

By choosing different γ_t to introduce the different levels of stochasticity in the path, we can reach the sampling Algo. 1 in the main text.

2.2. Bridge Models

Recent works on image-to-image translation introduce another model, bridge models, which directly interpolate two distributions, rather than the one is pointed at the Gaussian distribution. Some of them are based on Brownian bridge [7] and Schödinger bridge (SB) [5]. We will briefly introduce this kind of model.

SB is an entropy optimal transport model following the forward and backward SDEs:

$$dX_t = [f_t(X_t)dt + \beta_t \nabla \log \Psi(X_t, t)]dt + \sqrt{\beta_t} dW_t,$$
$$dX_t = [f_t(X_t)dt - \beta_t \nabla \log \hat{\Psi}(X_t, t)]dt + \sqrt{\beta_t} d\bar{W}_t,$$

where $X_0 \sim p_A$ and $X_1 \sim p_B$, and p_A, p_B can be any two distinct distributions. The functions $\Psi, \hat{\Psi} \in C^{2,1}(\mathbb{R}^d, [0, 1])$ are energy potentials that solve the PDEs:

$$\begin{cases} \frac{\partial \Psi(x,t)}{\partial t} = -\nabla \Psi^\top f - \frac{1}{2}\beta \Delta \Psi, \\ \frac{\partial \hat{\Psi}(x,t)}{\partial t} = -\nabla \cdot (\hat{\Psi}^\top f) + \frac{1}{2}\beta \Delta \hat{\Psi}, \end{cases} \tag{16}$$
$$s.t.\ \Psi(x, 0)\hat{\Psi}(x, t) = p_A(x), \Psi(x, 1)\hat{\Psi}(x, 1) = p_B(x).$$

I^{2}SB [12] reformulates the SB drift as the score function:

$$dX_t = f_t(X_t)dt + \sqrt{\beta_t} dW_t, X_0 \sim \hat{\Psi}(\cdot, 0), \tag{17}$$
$$dX_t = f_t(X_t)dt + \sqrt{\beta_t} d\bar{W}_t, X_1 \sim \Psi(\cdot, 1). \tag{18}$$

By assuming the coupling is available (*e.g.*, in image fusion tasks), *i.e.*, $p(X_0, X_1) = p_A(x_0)p_B(X_1|X_0)$, and set the SDE drift $f_t := 0$, one can obtain the DDPM-like [2] forward process:

$$q(X_t|X_0, X_1) = \mathcal{N}(X_t, \mu_t(X_0, X_1), \Sigma_t). \tag{19}$$

I^{2}SB set the mean $\mu_t = \frac{\bar{\sigma}_t}{\bar{\sigma}_t + \sigma_t} X_0 + \frac{\sigma_t}{\sigma_t + \bar{\sigma}_t} X_0$ and variance $\Sigma_t = \frac{\sigma_t^2 \bar{\sigma}_t^2}{\bar{\sigma}_t^2 + \sigma_t^2} \cdot \mathbf{I}$, where $\sigma_t^2 := \int_0^t \beta_\tau d\tau$ and $\bar{\sigma}_t^2 := \int_t^1 \beta_\tau d\tau$. Akin to flow matching, we can still learn a velocity-based model:

$$\mathcal{L} = \mathbb{E}_{t, X_0, X_1} \left[\left\| v_\theta(X_t, t) - \frac{X_t - X_0}{\sigma_t} \right\|_2^2 \right]. \tag{20}$$

When sampling, the posterior $p(X_{t-\tau}|X_0^v, X_t)$ can be obtained using Eq. (19). Thus, stochasticity is naturally introduced into the sampling process and relaxing the p_A to be any distribution at the same time.

3. Dataset Collection Details

Table 1. Training and testing numbers of pairs on each dataset.

Datasets	LLVIP	M3FD	MSRS	RealMFF	MFF-WHU	SICE	MEFB
Train	12025	3900	1083	639	92	288	60
Test	3463	300	361	71	30	72	40

Table 2. Configurations of heavy and mild degradations.

Heavy Degradation	Configurations
Low Exposure	Brightness: [0.3, 0.6]
Gaussian Blur	Kernel Size: [5, 21], Sigma: [1.0, 2.0]
Motion Blur	Kernel Size: [5, 21], Angle: 35.0, Direction: 0.5
Gaussian Noise	Mean: 0.0, Std: 0.0588 (15/255)
Rain	Rain mode (light and heavy), Apply Blur (Ksize: [5, 17], Sigma: [1.0, 2.0]), Low-light: [0.8, 1.0]
Haze	Haze Distance: [3, 15], IR Blur (Ksize: [5, 17], Sigma: [1.0, 2.0])
Snow	Snow Mode: ['Combine Snow', 'Small Snow', 'Mid Snow']
JPEG Compression	JPEG Quality: [10, 30]
Low Contrast	Contrast: [0.3, 0.8]
Downsample	Down Scale: [2.0, 3.5]
Strip IR	Direction: vertical and horizontal
Mild Degradation	**Configurations**
Low Exposure	Brightness: [0.5, 0.8]
Gaussian Blur	Kernel Size: [5, 9, Sigma: [1.0, 1.5]
Motion Blur	Kernel Size: [5, 9], Angle: 35.0, Direction: 0.5
Gaussian Noise	Mean: 0.0, Std: 0.0392 (10/255)
Rain	Rain mode: Light, Blur (Ksize: [5, 9], Sigma: [1.0, 1.5]), Low-light[0.85, 1.0]
Haze	Haze Distance: [5, 20], IR Blur (Ksize: [5, 9], Sigma: [1.0, 1.5])
Snow	Snow Mode: ['Small Snow', 'Middle Snow']
JPEG Compression	JPEG Quality: [30, 60]
Low Contrast	Contrast: [0.5, 0.8]
Downsample	Down Scale: [1.4, 2.0]
Strip IR	Direction: vertical and horizontal

In the dataset collection stage, SwinFusion [17] and DeFuse [8] are used to generate ground truth (GT) for each clean image pair, and the datasets are split into training and testing sets, with the train/test split ratios detailed in Tab. 1. Degraded image pairs are generated separately for the training and testing sets. The validation set is then derived from the generated degraded training set, using a fixed proportion. To create degraded image pairs, we introduce $n = 1, 2, 3$ levels of degradation based on the description in Sect. 3.1 of the main text. We apply a heavy degradation when only one type of degradation is present. For $n = 2$ or 3, we opt for milder degradations; severe compositional degradations can result in unrealistic information loss. The configurations of degradation are provided in Tab. 2.

Furthermore, due to the paired nature of our images, careful consideration must be given to degradation consistency and modality-specific characteristics. Specifically, for the VIF task, the infrared modality may exhibit blurring artifacts under rainy, hazy, or snowy weather conditions. For MEF and MFF tasks, motion blur should maintain consistent direction and intensity across the image pair. Rainy weather may also induce low-light conditions and additional blurring. The application of haze should simulate depth-dependent effects, with haze density increasing with distance from the camera lens.

We defined several aspect ratios for images: 1:1, 3:4, 4:3, 16:9, and 9:16, to facilitate model adaptation across image pairs with varying resolutions. For the LLVIP, M3FD, MSRS, SICE, and RealMFF datasets, for the training stage,

Table 3. Quantitative metrics of MEFB and MFF-WHU task in degraded MEF and MFF tasks. ERN denotes the existing restoration network. Reg and FM mean regression and flow matching, respectively. The best and second-best results are colored in red and blue.

Methods	MEFB MEF Dataset					Methods	MFF-WHU MFF Dataset				
	Q_{cv}	$Q^{A/BF}$	BRISQUE	MUSIQ	CLIPIQA		Q_{cv}	$Q^{A/BF}$	BRISQUE	MUSIQ	CLIPIQA
ERN+U2Fusion [27]	515.189	0.164	69.806	23.216	0.174	ERN+U2Fusion[27]	146.070	0.398	34.325	48.858	0.323
ERN+DeFuse [8]	347.178	0.351	36.510	46.131	0.315	ERN+DeFuse [8]	165.911	0.381	38.286	44.923	0.309
ERN+HoLoCo [13]	347.178	0.351	36.510	46.131	0.315	ERN+ZMFF [3]	254.337	0.389	42.414	49.859	0.364
ERN+TC-MoA [30]	403.580	0.315	42.038	35.792	0.288	ERN+TC-MoA [30]	306.029	0.264	55.497	22.367	0.156
ERN+PSLPT [25]	438.174	0.409	31.211	51.945	0.341	ERN+PSLPT [25]	175.625	0.426	28.780	57.716	0.403
Ours (Reg)	368.647	**0.412**	29.313	**56.908**	**0.362**	Ours (Reg)	**105.342**	**0.415**	**24.483**	**64.669**	**0.411**
Ours (FM)	**366.802**	**0.385**	**21.753**	**62.971**	**0.362**	Ours (FM)	**107.638**	**0.416**	**25.401**	**64.947**	**0.446**

Table 4. Results of total n degradations in LLVIP dataset. "R+F" denotes degradation-level all-in-one DA-CLIP + EMMA pipeline. "w/" and "w/o" mean "with" and "without", respectively. "lang." is the language guidance.

Config	LLVIP Dataset				
	VIF	$Q^{A/BF}$	BRISQUE	MUSIQ	CLIPIQA
$n = 1$ R+F	0.496	0.370	28.647	52.015	0.286
$n = 1$ w/o lang.	0.589	0.379	27.826	52.978	0.321
$n = 1$ w/ lang.	**0.604**	**0.384**	**27.535**	**53.277**	**0.330**
$n = 2$ R+F	0.450	0.346	33.256	48.589	0.264
$n = 2$ w/o lang.	0.584	0.370	29.037	50.226	0.318
$n = 2$ w/ lang.	**0.601**	**0.380**	**28.346**	**51.807**	**0.326**
$n = 3$ R+F	0.431	0.325	35.761	44.285	0.247
$n = 3$ w/o lang.	0.562	0.351	32.079	46.792	0.304
$n = 3$ w/ lang.	**0.584**	**0.366**	**30.432**	**48.691**	**0.317**

we synthesized 2000 clean/degradation/prompt/GT pairs under each aspect ratio and degradation degree. For MFF-WHU and MEFB datasets, we generated 500 and 300 pairs, respectively. For the validation stage, we synthesized 300 pairs for each dataset under each aspect ratio and degradation degree. In total, we synthesized 157500 pairs of data for training, 31500 pairs for validation, and 6300 pairs for testing.

4. Degraded Dataset Comparisons with TextIF EMS Dataset

In TextIF [28], a degraded/clean/prompt dataset named EMS was proposed. In comparison, the dataset we propose differs in the following aspects:

1) Our dataset incorporates a greater variety of degradations, including noise, low lighting, JPEG compression artifacts, motion blur, rain, snow, haze, IR strips, downsampling, low contrast, and Gaussian blur. In contrast, the EMS dataset is *limited to only a few degradation strategies and does not align well with real-world scenarios*. For instance, the haze effect applied in EMS is centered, which results in models trained on it having poor robustness when handling real-world degradations for fusion tasks;
2) Our dataset not only includes data for the VIF task but also encompasses fusion degradation data for MEF and MFF tasks. In contrast, the EMS dataset *only contains data for the VIF task*, making it challenging to train an all-in-one model that generalizes across multiple fusion tasks;
3) Our dataset is *significantly larger* than the EMS dataset. While the EMS dataset contains only thousands of degradation pairs, our dataset includes nearly 160,000 pairs with *diverse degradation types, varying degradation levels, and different image aspect ratios*. This scale enables the training of larger and more complex models;
4) Additionally, our dataset *provides image latents encoded by the Cosmos tokenizer* [1] to support the training of Transformer networks that operate in the latent space, enabling faster inference. It also supports the training of models in the pixel space.

5. Additional Results

In this section, we provide more quantitative and visual comparisons to demonstrate the effectiveness of MMAIF: it supports multiple degradations, various fusion tasks, and fusion tasks involving multiple simultaneous degradations within a single image pair. Moreover, it outperforms previous SOTA restoration+fusion pipelines as well as recent all-in-one methods in terms of degradation fusion performance.

5.1. Additional MEF and MFF Results

In this section, we provide the degraded fusion performance of MEF MEFB and MFF MFF-WHU datasets in Tab. 3. As evidenced by the results, our MMAIF significantly outperforms restoration+fusion pipelines across nearly every metric, demonstrating the effectiveness of our proposed method.

5.2. Performance without Language Guidance

Language guidance plays a prominent role in our MMAIF. To verify its effectiveness, during training, we randomly drop the encoded prompt feature and replace it with a zero tensor, which allows us to avoid manually estimating the degradation type during testing. This approach results in a certain degree of performance loss but enables MMAIF to

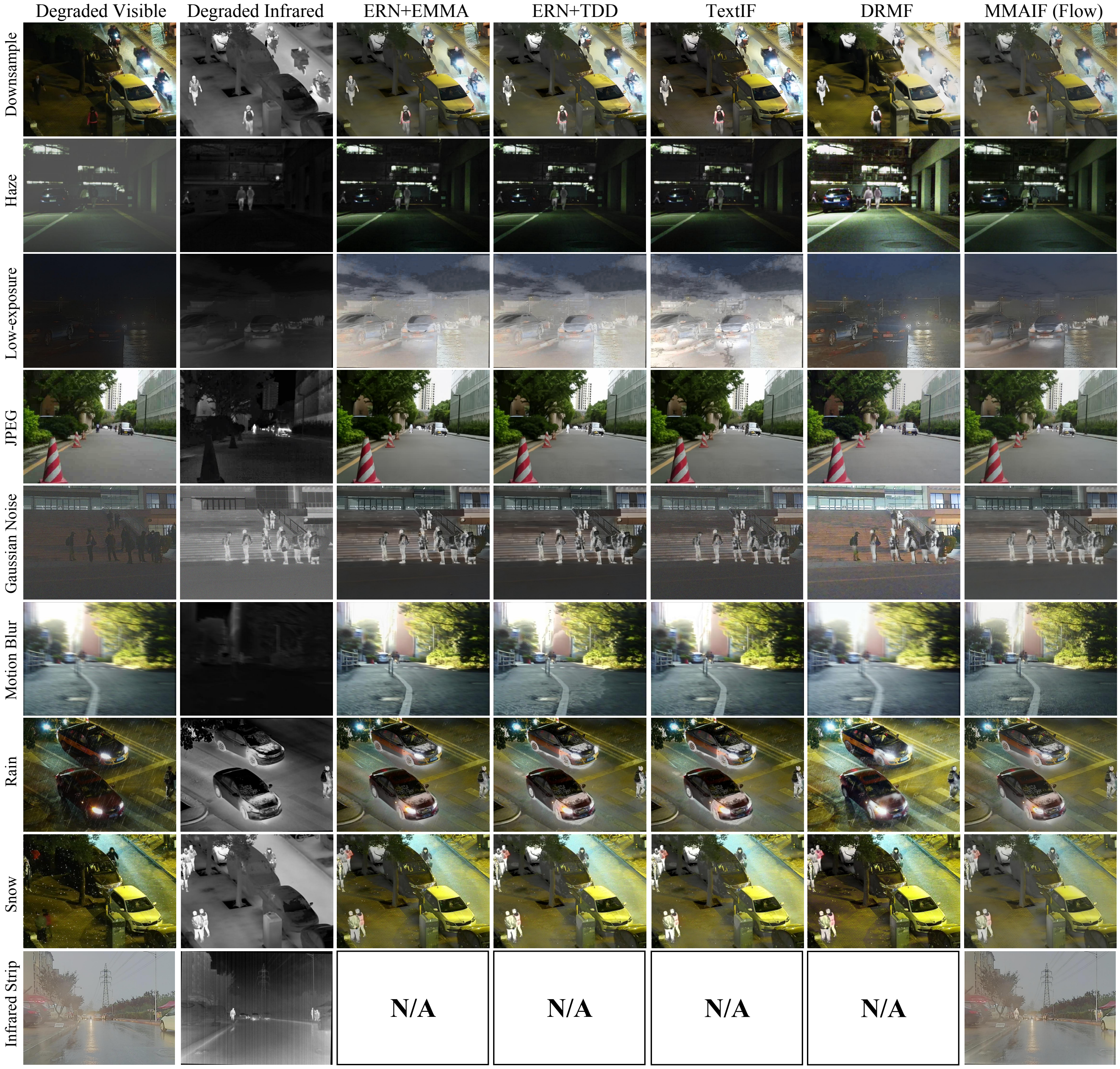

Figure 1. More visual results of the proposed MMAIF and recent SOTA fusion methods incorporated with existing restoration methods ("ERN" in this figure) on the VIF degraded image fusion task. "N/A" means that no degradation-oriented restoration method is involved.

perform degradation image fusion more flexibly. The performance without language guidance is shown in Tab. 4. Note that, even without language guidance, MMAIF still performs better than previous methods, demonstrating the effectiveness of our framework.

5.3. More Visual Results

We provide additional visual comparisons in Figs. 1, 2, and 3 for VIF, MEF, and MFF tasks, as well as more examples of combined degradation (*i.e.*, $n \in \{2, 3\}$) in Fig. 4 of MMAIF's degraded image fusion results. The results demonstrate that MMAIF outperforms other restoration+fusion pipelines and recent all-in-one methods in degraded image fusion, producing clearer and more realistic fused images with superior adaptability and robustness when handling multiple co-existing degradations. It is still worth noting that all samples are processed by *a single, all-in-one model across various image fusion tasks and differ-*

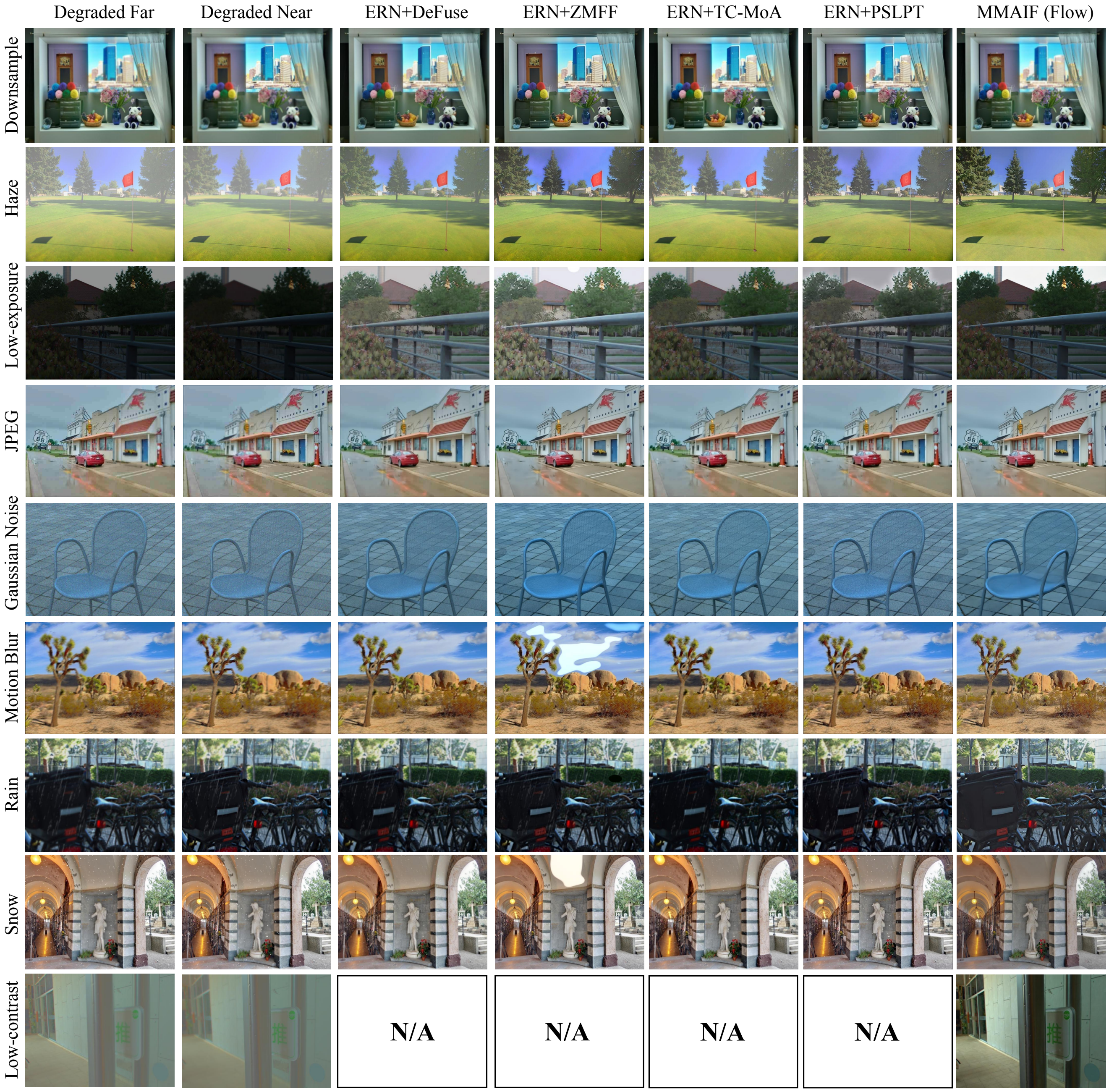

Figure 2. More visual results of the proposed MMAIF and recent SOTA fusion methods incorporated with existing restoration methods ("ERN" in this figure) on the MFF degraded image fusion task. "N/A" means that no degradation-oriented restoration method is involved.

ent types of degradation.

References

[1] NVIDIA et. al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 4

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[3] Xingyu Hu, Junjun Jiang, Xianming Liu, and Jiayi Ma. Zmff: Zero-shot multi-focus image fusion. *Information Fusion*, 92:127–138, 2023. 4

[4] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. 1

[5] Christian Léonard. A survey of the schrödinger problem

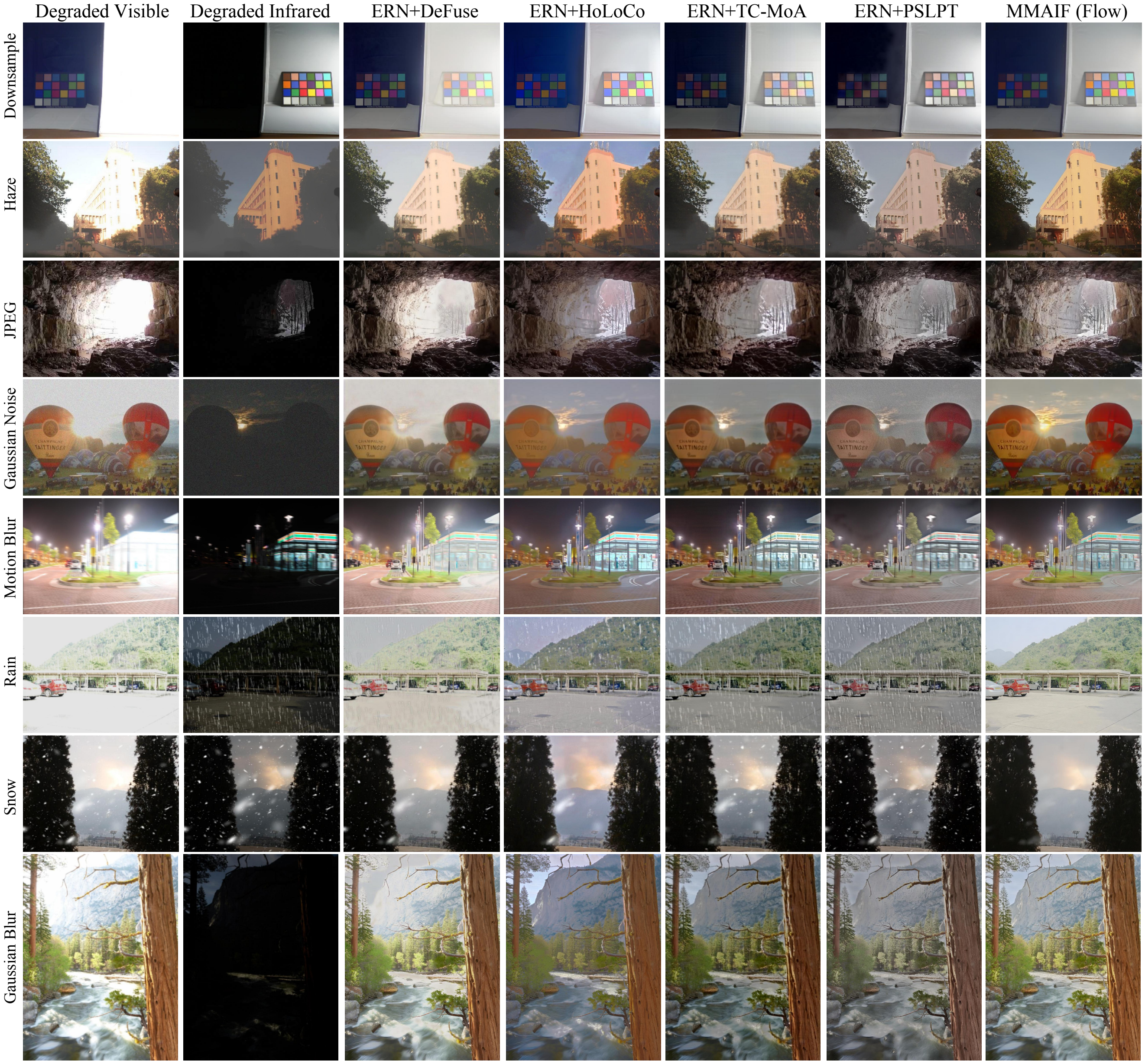

Figure 3. More visual results of the proposed MMAIF and recent SOTA fusion methods incorporated with existing restoration methods ("ERN" in this figure) on the MEF degraded image fusion task. "N/A" means that no degradation-oriented restoration method is involved.

and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*, 2013. 3

[6] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17452–17462, 2022. 1

[7] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 1952–1961, 2023. 3

[8] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *European Conference on Computer Vision (ECCV)*, 2022. 3, 4

[9] Xin Lin, Jingtong Yue, Kelvin CK Chan, Lu Qi, Chao Ren, Jinshan Pan, and Ming-Hsuan Yang. Multi-task image restoration guided by robust dino features. *arXiv preprint arXiv:2312.01677*, 2023. 2

[10] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 2

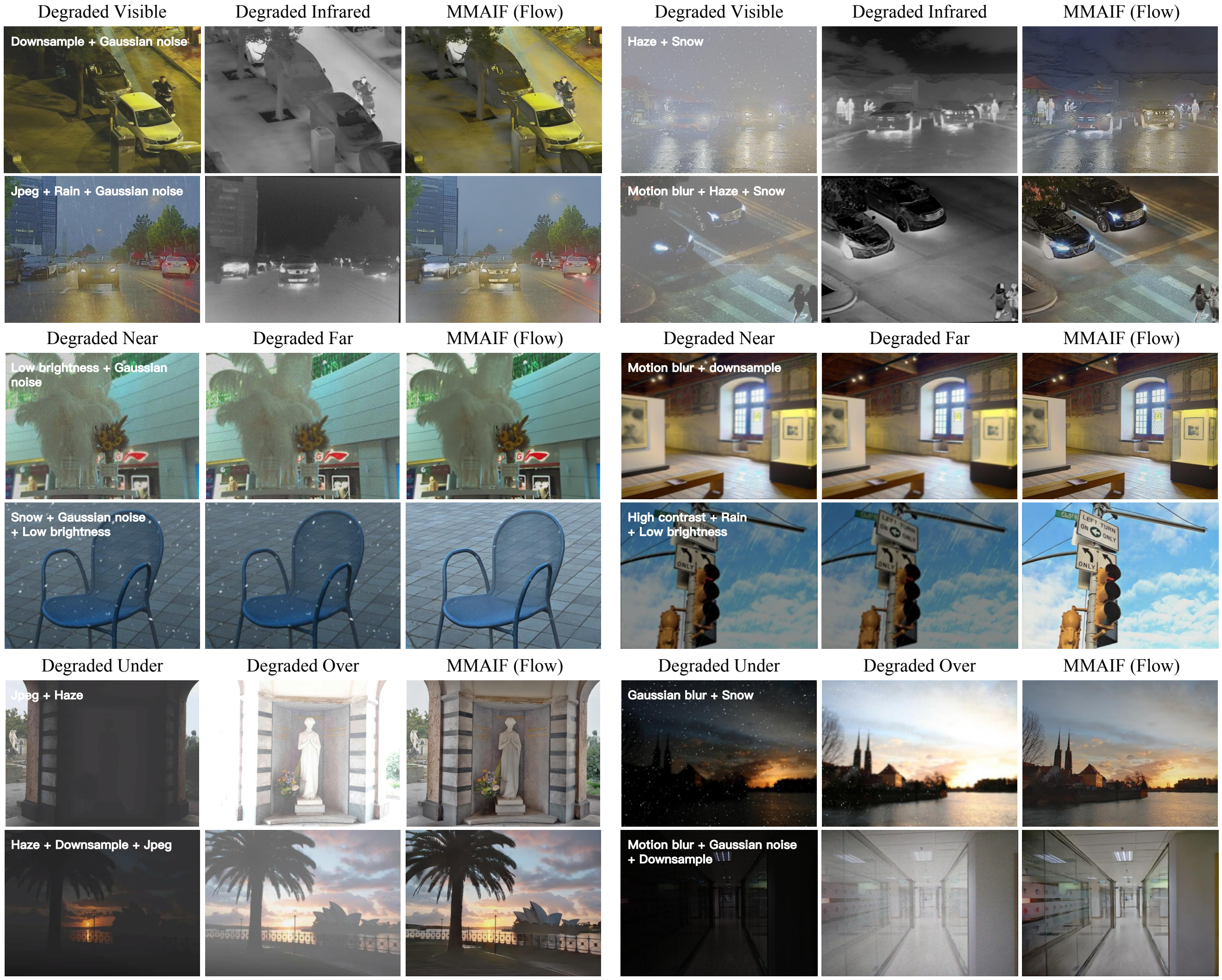

Figure 4. Fused images of combined degradations in one image pair ($n = 2, 3$). Our MMAIF can handle severe degradation and generate high-quality fused images. Zoom in for more details.

[11] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 1

[12] Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. I^{2}SB: Image-to-image schrödinger bridge. *arXiv preprint arXiv:2302.05872*, 2023. 3

[13] Jinyuan Liu, Guanyao Wu, Junsheng Luan, Zhiying Jiang, Risheng Liu, and Xin Fan. Holoco: Holistic and local contrastive learning network for multi-exposure image fusion. *Information Fusion*, 2023. 4

[14] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 2

[15] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Controlling vision-language models for multi-task image restoration. *arXiv preprint arXiv:2310.01018*, 2023. 2

[16] Ang Lv, Ruobing Xie, Yining Qian, Songhao Wu, Xingwu Sun, Zhanhui Kang, Di Wang, and Rui Yan. Autonomy-of-experts models. *arXiv preprint arXiv:2501.13074*, 2025. 1

[17] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022. 3

[18] Dongwon Park, Byung Hyun Lee, and Se Young Chun. All-in-one image restoration for unknown degradations using adaptive discriminative filters for specific degradations. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5815–5824. IEEE, 2023. 1

[19] V Potlapalli, SW Zamir, S Khan, and FS Khan. Promptir:

Prompting for all-in-one blind image restoration. arxiv 2023. *arXiv preprint arXiv:2306.13090*. 1

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[21] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 1

[22] Zhentao Tan, Yue Wu, Qiankun Liu, Qi Chu, Le Lu, Jieping Ye, and Nenghai Yu. Exploring the application of large-scale pre-trained models on adverse weather removal. *IEEE Transactions on Image Processing*, 2024. 2

[23] Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 1

[24] Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. Auxiliary-loss-free load balancing strategy for mixture-of-experts. *arXiv preprint arXiv:2408.15664*, 2024. 1

[25] Wu Wang, Liang-Jian Deng, and Gemine Vivone. A general image fusion framework using multi-task semi-supervised learning. *Information Fusion*, 108:102414, 2024. 4

[26] Ziteng Wang, Jianfei Chen, and Jun Zhu. ReMoE: Fully differentiable mixture-of-experts with relu routing. *arXiv preprint arXiv:2412.14711*, 2024. 1

[27] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2020. 4

[28] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27026–27035, 2024. 4

[29] Xu Zhang, Jiaqi Ma, Guoli Wang, Qian Zhang, Huan Zhang, and Lefei Zhang. Perceive-ir: Learning to perceive degradation better for all-in-one image restoration. *arXiv preprint arXiv:2408.15994*, 2024. 2

[30] Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for general image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7099–7108, 2024. 4