Supplementary Materials: "Taming Flow Matching with Unbalanced Optimal Transport into Fast Pansharpening"

Abstract

In this supplementary, we provide more background on pansharpening and diffusion models in Sect. 1. The proof of Prop. 3.2 is detailed in Sect. 2 that proves the additional pansharpening-based regularization in the unbalanced optimal transport cost can provide an OT map in saddle points. More discussions of proposed OTFM are detailed in Sect. 3. Additional quantitative results and visualizations are provided in Sect. 4 to verify the effectiveness of the proposed OTFM.

1. Background

In this section, we provide recent works of deep-learning-based models for pansharpening, basic preliminary of diffusion models, and diffusion applications for pansharpening.

1.1. Deep Models for Pansharpening

Recently, deep learning methods are widely used for pansharpening. They are more flexible and effective than traditional methods, leading to better reconstruction quality and increased popularity in the field. PNN [10] is the earliest deep model applied to the Pansharpening task, achieving results far superior to traditional methods using a three-layer CNN architecture. Recognizing the spatial and spectral properties in Pansharpening, PanNet [17] injects high-frequency information from PAN into the upsampled LRMS to produce HRMS. Inspired by traditional methods, FusionNet [3] designs a neural network that takes the difference between upsampled LRMS and PAN as input, yielding finer panchromatic sharpened images. The global receptive field of the Self-attention mechanism addresses the local receptive field issue in CNNs, significantly advancing deep model development. ViTPAN [12] inputs cropped LRMS and PAN into a three-layer self-attention-based encoder, resulting in high-quality HRMS. PMACNet [9] utilizes a parallel dual-branch network structure to process spatial and spectral features, extending the Self-attention mechanism to pixel-level fusion for more refined results.

1.2. Diffusion Models and Applications on Pansharpening

The diffusion model, as a generative model, has demonstrated remarkable performance in image processing tasks such as image super-resolution [5] and image restoration [19, 8]. Recently, several outstanding works have also emerged in Pansharpening. PanDiff [11] decomposes the fusion process into multiple Markov processes and utilizes U-Net to reconstruct HRMS from random Gaussian noise. DDIF [2] injects coarse-grained style information and fine-grained high and low-frequency details of PAN and LRMS into the diffusion model to reconstruct high-quality images. SSDiff [18] approaches the Pansharpening task from a subspace decomposition perspective using a diffusion model, employing an alternating projection method to fuse discriminative spatial and spectral features, achieving superior reconstruction quality.

The diffusion model consists of forward and reverse processes. The forward process gradually adds noise to the prior distribution x_0 over T steps of a Markov chain, transforming it into an approximate standard normal distribution $x_T \sim \mathcal{N}(0, \mathbf{I})$. Through the reparameterization trick, x_t of any timestep t can be directly obtained from x_0 using the following formula.

$$q(\mathbf{x}_t|\mathbf{x}_0) = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{17}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=0}^{t} \alpha_i$, β_t is a pre-defined variance schedule.

The reverse process aims to eliminate the degradation introduced in the forward process, gradually denoising from x_t to recover x_0. To achieve this, a neural network can be utilized to learn the distribution of $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, followed by iterative denoising as follows:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)), \tag{18}$$

where μ_θ and Σ_θ are the mean and variance of $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, respectively, and θ is the parameters of model. The mean and variance can be computed as follows:

$$\mu_\theta = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)), \tag{19}$$

$$\Sigma_\theta(\mathbf{x}_t, t) = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t. \tag{20}$$

After performing the aforementioned T-step sampling, a high-quality reconstructed image can be obtained, but this comes at the cost of significantly higher time overhead compared to single-step sampling models.

2. Proof

In this section, we provide the proof of Props. 3.1 and 3.2, which is about the dual formulation of the unbalanced optimal transport (UOT) and the saddle points for pansharpening-regularized UOT cost.

2.1. Proof of Proposition 3.1

Proposition 2.1 (Dual formulation of UOT)**.** *The UOT dual formulation, $C_{UOT}(\mathbb{P}, \mathbb{Q})$, can be obtained by using the c-transform:*

$$\sup_{v\in C}\left[\int_X -f(-v^c(x))d\mathbb{P} + \int_Y -f(-v(y))d\mathbb{Q}\right], \tag{21}$$

where f is the entropy function and v^c is the c-transformation of v in Eq. (9).

Proof. Recall that the UOT problem is,

$$\inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})}\int_{X\times Y}(x,y)d\pi(x,y) + D_f(\pi_0|\mathbb{P}) + D_f(\pi_1|\mathbb{Q}).$$

Using the f-divergence's definition $D_f(\mu,\nu) = \int f(\frac{d\mu}{d\nu})d\nu$, it can be rewritten as,

$$\begin{aligned}\inf_{\pi}\int_{X\times Y} c(x,y)d\pi(x,y) + \int_X f\left(\frac{d\pi_0(x)}{d\mathbb{P}(x)}\right)d\mathbb{P} \\ + \int_Y f\left(\frac{d\pi_1(y)}{d\mathbb{Q}(y)}\right)d\mathbb{Q}.\end{aligned} \tag{22}$$

The Lagrangian formulation of Eq. (22) is,

$$\begin{aligned}\mathcal{L}(\pi,\eta,\kappa,u,v) = &\int_{X\times Y} c(x,y)d\pi(x,y) \\ &+ \int_X f\left(\frac{d\eta(x)}{d\mathbb{P}(x)}\right)d\mathbb{P}(x) + \int_Y f\left(\frac{d\kappa(y)}{d\mathbb{Q}(y)}\right)d\mathbb{Q}(y) \\ &+ \int_Y u(x)\left(d\eta(x) - \int_Y d\pi(x,y)\right) \\ &+ \int_Y v(y)\left(d\kappa(y) - \int_X d\pi(x,y)\right),\end{aligned} \tag{23}$$

where η, κ are slack variable of π_0 and π_1. u, v are the Lagrange multipliers associated with the constraints. One can form the dual Lagrangian function, which is,

$$g(u,v) = \inf_{\pi,\eta,\kappa}\mathcal{L}(\pi,\eta,\kappa,u,v). \tag{24}$$

Then the optimization are seperated *w.r.t* each varibles,

$$\begin{aligned}g(u,v) = &\inf_\eta\left[\int_X u(x)d\eta(x) + f\left(\frac{d\eta(x)}{d\mathbb{P}(x)}\right)d\mathbb{P}(x)\right] \\ &+ \inf_\kappa\left[\int_X v(y)d\kappa(y) + f\left(\frac{d\kappa(y)}{d\mathbb{Q}(y)}\right)d\mathbb{Q}(y)\right] \\ &+ \inf_\pi\left[\int_{X\times Y}(c(x,y) - u(x) - v(y))d\pi(x,y)\right].\end{aligned} \tag{25}$$

The first term is,

$$-\sup_\eta\left[\int_X\left[-u(x)\frac{d\eta(x)}{d\mathbb{P}(x)} - f\left(\frac{d\eta(x)}{d\mathbb{P}(x)}\right)\right]d\mathbb{P}(x)\right] \tag{26}$$

$$= -\int_X f(-u(x))d\mathbb{P}(x) \tag{27}$$

Then, we can write Eq. (25) into,

$$\begin{aligned}g(u,v) = &\inf_\pi\left[\int_{X\times Y}(c(x,y) - u(x) - v(y))d\pi(x,y)\right] \\ &- \int_X f(-u(x))d\mathbb{P}(x) - \int_Y f(-v(x))d\mathbb{Q}(x).\end{aligned} \tag{28}$$

If $c(x,y) - u(x) - v(y) < 0$ and all mass of the plan π are concentrated into one coupling. The problem (28) is $-\infty$. Under this trivial solution, we can set $c(x,y) - u(x) - v(y) \geq 0$. Finally, the UOT dual formulation is,

$$\begin{aligned}\sup_{u(x)+v(y)\leq c(x,y)}\Bigg[&\int_X -f(-u(x))d\mathbb{P}(x) \\ &+ \int_Y -f(-v(y)d\mathbb{Q}(y)\Bigg],\end{aligned} \tag{29}$$

which concludes this proposition. □

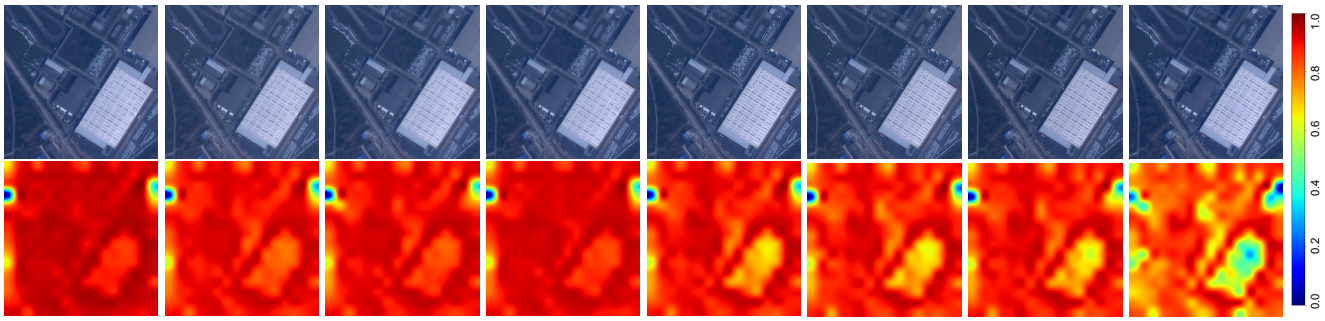
Note that $c(x,y) = u(x) + v(y)$ is taken over π-almost everywhere, because f is non-decreasing, which means,

$$\begin{aligned}\sup_{(u,v)\in C(X)\times C(Y)}\Bigg[&\int_X -f(-u(x))d\mathbb{P}(x) \\ &+ \int_Y -f(-v(y)d\mathbb{Q}(y) - \mathcal{I}(u+v\leq c)\Bigg],\end{aligned} \tag{30}$$

where $\mathcal{I}(\cdot)$ is the indicator function. In remark 3.2, we set f to be convex, non-decreasing, and differentiable. Thus, all terms in Eq. (30) are finite by letting $u = -1$ and $v = -1$. Moreover, the strong duality is still held by using Fechel-Rockafellar's theorem. Finally, the dual form with constraint $u(x) + v(y) \leq c(x,y)$ is concluded as,

$$\begin{aligned}\sup_{v\in C}\Bigg[&\int_X -f\Big(-\inf_{y\in Y}[c(x,y) - v(y)]\Big)d\mathbb{P}(x) \\ &+ \int_Y -f(-v(y))d\mathbb{Q}(y)\Bigg].\end{aligned} \tag{31}$$

This leads to the UOT objective $\mathcal{L}_{T_\theta, v_\varphi}$ in Eq. (12).



OTFM PanDiff DCFNet DDIF LAGConv FusionNet DiCNN BT-H

Figure 1. Fused GF2 full-resolution data and their corresponding HQNR map. The high value in the HQNR map means better full-resolution fusion performance.

2.2. Proof of Proposition 3.2

Proposition 2.2 (Saddle points of pansharpening-regularized UOT provide the OT maps)**.** *For any optimal potential function* $v^* \in \arg\sup_v \mathcal{L}_{T,v}$*, it provides the OT map* T^* *which holds*

$$T^* \in \arg\min_T \mathcal{L}_{T,v}(T, v^*). \tag{32}$$

Proof. Denote the pansharpening-regularization cost as $g(y)$ (see Eq. (14)). The full UOT cost is: $\tilde{c}(x,y) = c(x,y) + g(y)$. From Eq. (31), when v^* is obtained, we have the UOT loss *w.r.t* T_θ,

$$\int_X -f[-c(x,T_\theta(x)) + g(T_\theta(x)) + v^*_\varphi(T_\theta(x))]d\mathbb{P}(x) + \int_Y -f(-v^*_\varphi(y))d\mathbb{Q}(y). \tag{33}$$

Using $T^*_\#\mathbb{P} = \mathbb{Q}$ and the change of variable $T^*_\theta(x) = y$, we have,

$$\int_X -f(-v^*_\varphi(T^*_\theta(x)))d\mathbb{P}(x) = \int_Y -f(-v^*_\varphi(y))d\mathbb{Q}(y). \tag{34}$$

Due to f is non-decreasing, we have

$$\mathcal{L}_{T,v}(T^*, v^*) = \int_X [c(x,T^*(x)) + g(T^*(x))]d\mathbb{P}(x) \tag{35}$$

$$= \inf_{T_\#\mathbb{P}=\mathbb{Q}} \int_Y [c(x,T^*_\theta(x)) + g(T^*_\theta(x))]d\mathbb{P}(x) \tag{36}$$

$$= \text{Monge}(\mathbb{P},\mathbb{Q}), \tag{37}$$

which proves this proposition. □

3. Differences with GANs

Our UOT learning objective shares a similar min-max learning paradigm with GANs, as both involve optimizing competing objectives. However, unlike GANs, our UOT approach is not grounded in the adversarial game-theoretic framework. Instead, it relies on a well-defined OT cost function. Consequently, our method avoids the inherent instability issues commonly encountered in GAN training, such as mode collapse or oscillatory convergence, as the optimization process is guided by a stable and mathematically grounded objective rather than the dynamic balance between a generator and a discriminator. Furthermore, the use of UOT allows for greater control over the marginal constraints, enabling more flexible and robust distribution matching compared to the implicit density estimation performed by GANs.

In OTFM, the integration with Flow Matching (FM) ensures that UOT and FM losses operate simultaneously. The UOT loss enforces the mapping network T_θ to act as a one-step generator, while the FM loss guides the velocity network s_θ to learn the correct flow along the interpolation path. Notably, s_θ and T_θ are the same network, as the velocity $y_1 - y_0$ and the mapped HRMS y_1 can be derived from one another. In contrast, GANs rely on an adversarial training framework, where a generator and a discriminator compete in a minimax game. While GANs can achieve impressive results, they are prone to training instability and mode collapse due to the lack of explicit constraints on the intermediate dynamics or flow, which are inherently addressed in OTFM through FM.

4. Additional experiments

4.1. Datasets

To evaluate the performance of OTFM against other state-of-the-art (SOTA) methods, calculating both reference and non-reference metrics on reduced resolution and full resolution datasets, respectively. Specifically, in Gaofen-2 and QuickBird, the spatial resolutions of the PAN images are 0.8 meters and 0.6 meters, respectively, while the corresponding MS images have four bands (red, green, blue, and near-infrared) with spatial resolutions of 3.2 meters and 2.4

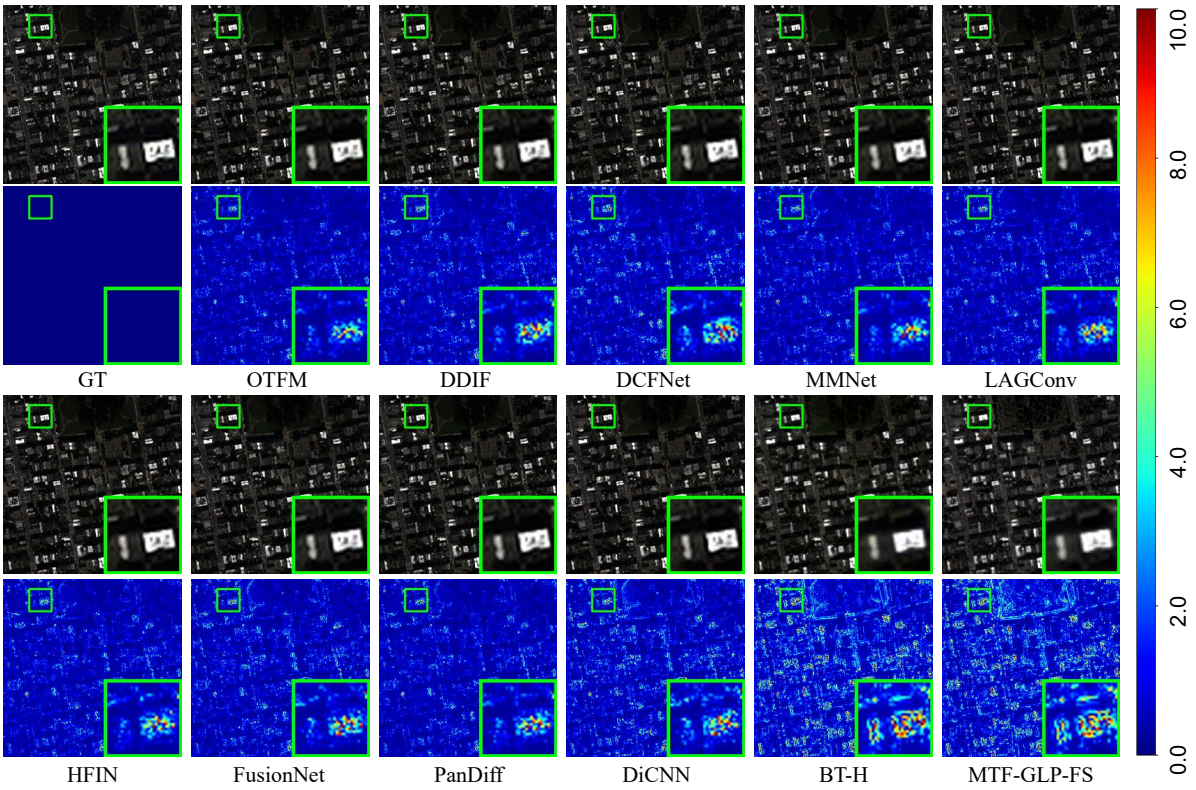


Figure 2. Visual comparisons on QuickBird reduced-resolution dataset. The second and fourth rows are error maps.

Table 1. Results on the QuickBird full-resolution datasets. The best results are in red, and the second best results are in blue.

Method	Full-Resolution (FR): Avg±std		
	D_λ (↓)	D_s (↓)	HQNR (↑)
MTF-GLP-FS [14]	**0.049±0.015**	0.138±0.024	0.820±0.034
BT-H [1]	0.230±0.072	0.165±0.016	0.643±0.065
DiCNN [6]	0.092±0.014	0.107±0.021	0.811±0.031
FusionNet [4]	0.059±0.019	0.052±0.009	0.892±0.022
LAGConv [7]	0.084±0.024	0.068±0.013	0.854±0.018
DCFNet [15]	**0.045±0.015**	0.124±0.027	0.836±0.016
MMNet [16]	0.089±0.051	0.097±0.038	0.823±0.032
HFIN [13]	0.065±0.025	0.078±0.019	0.862±0.019
PanDiff [11]	0.059±0.022	0.064±0.025	0.881±0.042
DDIF [2]	0.058±0.013	**0.049±0.010**	**0.895±0.021**
Proposed	0.053±0.017	**0.025±0.013**	**0.922±0.020**

meters. WorldView-3 offers even higher spatial resolutions (PAN: 0.3 m and MS: 1.2 m) and provides spectral information across eight bands, which additionally includes coastal, yellow, red edge, and near-infrared-2 bands.

4.2. Results on GaoFen-2

Fig. 1 presents a visual comparison of the real full-resolution GF2 dataset and the corresponding HQNR map. The closer the HQNR map approaches red, the better the reconstruction quality of the real image. It can be observed that OTFM, PanDiff, DCFNet, and DDIF all exhibit strong performance, with distortions mainly occurring at the image edges. OTFM, in particular, handles architectural details in the central region of the image more effectively, demonstrating its efficacy on real datasets.

4.3. Results on QuickBird

Fig. 2 displays the visual comparison results from the reduced-resolution dataset of QB, where OTFM reconstructs more details of the building roofs, which is particularly evident in the zoomed-in local images. Additionally, we conducted experiments on the full-resolution QuickBird dataset and evaluated the performance of OTFM. Similarly, the no-reference metrics were obtained from 20 randomly selected test images in the QB dataset. The performance comparison is reported in Tab. 1. Our OTFM achieves the best overall quality, with an HQNR score of 0.922. Surprisingly, the traditional method MTF-GLP-FS ranks second only to DCFNet in the D_λ index, indicating its superior capability in spectral reconstruction.

References

[1] Bruno Aiazzi, L Alparone, Stefano Baronti, Andrea Garzelli, and Massimo Selva. Mtf-tailored multiscale fusion of high-

resolution ms and pan imagery. *Photogrammetric Engineering & Remote Sensing*, 72(5):591–596, 2006. 4

[2] Zihan Cao, Shiqi Cao, Liang-Jian Deng, Xiao Wu, Junming Hou, and Gemine Vivone. Diffusion model with disentangled modulations for sharpening multispectral and hyperspectral images. *Information Fusion*, 104:102158, 2024. 1, 4

[3] Liang-Jian Deng, Minyu Feng, and Xue-Cheng Tai. The fusion of panchromatic and multispectral remote sensing images via tensor-based sparse modeling and hyper-laplacian prior. *Information Fusion*, 52:76–89, 2019. 1

[4] Liang-Jian Deng, Gemine Vivone, Cheng Jin, and Jocelyn Chanussot. Detail injection-based deep convolutional neural networks for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 59(8):6995–7010, 2020. 4

[5] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10021–10030, 2023. 1

[6] Lin He, Yizhou Rao, Jun Li, Jocelyn Chanussot, Antonio Plaza, Jiawei Zhu, and Bo Li. Pansharpening via detail injection based convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(4):1188–1204, 2019. 4

[7] Zi-Rong Jin, Tian-Jing Zhang, Tai-Xiang Jiang, Gemine Vivone, and Liang-Jian Deng. Lagconv: Local-context adaptive convolution kernels with global harmonic bias for pansharpening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1113–1121, 2022. 4

[8] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022. 1

[9] Yixun Liang, Ping Zhang, Yang Mei, and Tingqi Wang. Pmacnet: Parallel multiscale attention constraint network for pan-sharpening. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 1

[10] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7):594, 2016. 1

[11] Qingyan Meng, Wenxu Shi, Sijia Li, and Linlin Zhang. Pandiff: A novel pansharpening method based on denoising diffusion probabilistic model. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 1, 4

[12] Xiangchao Meng, Nan Wang, Feng Shao, and Shutao Li. Vision transformer for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. 1

[13] Jiangtong Tan, Jie Huang, Naishan Zheng, Man Zhou, Keyu Yan, Danfeng Hong, and Feng Zhao. Revisiting spatial-frequency information integration from a hierarchical perspective for panchromatic and multi-spectral image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25922–25931, 2024. 4

[14] Gemine Vivone, Rocco Restaino, and Jocelyn Chanussot. Full scale regression-based injection coefficients for panchromatic sharpening. *IEEE Transactions on Image Processing*, 27(7):3418–3431, 2018. 4

[15] Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, and Tian-Jing Zhang. Dynamic cross feature fusion for remote sensing pansharpening. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14687–14696, 2021. 4

[16] Keyu Yan, Man Zhou, Li Zhang, and Chengjun Xie. Memory-augmented model-driven network for pansharpening. In *European Conference on Computer Vision*, pages 306–322. Springer, 2022. 4

[17] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pan-sharpening. In *Proceedings of the IEEE international conference on computer vision*, pages 5449–5457, 2017. 1

[18] Yu Zhong, Xiao Wu, Liang-Jian Deng, Zihan Cao, and Hong-Xia Dou. Ssdiff: Spatial-spectral integrated diffusion model for remote sensing pansharpening. *Advances in Neural Information Processing Systems*, 37:77962–77986, 2025. 1

[19] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1219–1229, 2023. 1