

Taming Flow Matching with Unbalanced Optimal Transport into Fast Pansharpening

Zihan Cao^{1,*} Yu Zhong^{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@gmail.com, yuuzhong1011@gmail.com, liangjian.deng@uestc.edu.cn

* Equal contribution, † Corresponding author

Abstract

Pansharpening, a pivotal task in remote sensing for fusing high-resolution panchromatic and multispectral imagery, has garnered significant research interest. Recent advancements employing diffusion models based on stochastic differential equations (SDEs) have demonstrated state-of-the-art performance. However, the inherent multi-step sampling process of SDEs imposes substantial computational overhead, hindering practical deployment. While existing methods adopt efficient samplers, knowledge distillation, or retraining to reduce sampling steps (e.g., from 1,000 to fewer steps), such approaches often compromise fusion quality. In this work, we propose the Optimal Transport Flow Matching (OTFM) framework, which integrates the dual formulation of unbalanced optimal transport (UOT) to achieve one-step, high-quality pansharpening. Unlike conventional OT formulations that enforce rigid distribution alignment, UOT relaxes marginal constraints to enhance modeling flexibility, accommodating the intrinsic spectral and spatial disparities in remote sensing data. Furthermore, we incorporate task-specific regularization into the UOT objective, enhancing the robustness of the flow model. The OTFM framework enables simulation-free training and single-step inference while maintaining strict adherence to pansharpening constraints. Experimental evaluations across multiple datasets demonstrate that OTFM matches or exceeds the performance of previous regression-based models and leading diffusion-based methods while only needing **one sampling step**. Project website: <https://github.com/294coder/PAN-OTFM>.

1. Introduction

Multispectral images, captured by collecting different wavelengths reflected by various surface objects, can reveal more comprehensive spectral and chromatic character-

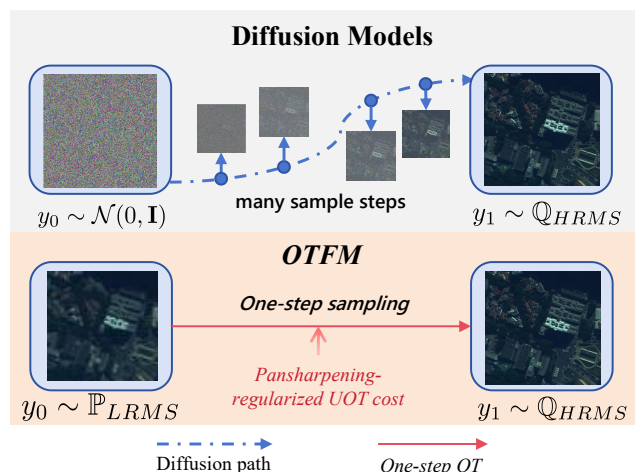


Figure 1. **Key distinction from previous diffusion models.** Traditional diffusion models typically sample from a Gaussian distribution, requiring numerous iterative steps (e.g., 1000) to achieve results. In contrast, our OTFM harnesses the power of unbalanced optimal transport, enabling high-quality pansharpening with just *one-step* sampling step.

istics of a region, thereby reflecting more complete properties of surface features [12]. However, constrained by physical imaging limitations, remote sensing satellites can only acquire low-resolution multispectral images (LRMS). Pansharpening generates high-resolution multispectral images (HRMS) which are unattainable under practical imaging conditions by fusing single-channel Panchromatic Images (PAN) with high spatial resolution and LRMS. This technique has garnered widespread attention from both the research community and commercial companies. Most satellites, such as World-View3, GaoFen-2, and QuickBird, are capable of simultaneously capturing PAN and LRMS data.

Existing pansharpening methods can be categorized into two groups: traditional methods (e.g., component substitution methods) [40, 54, 60] and deep learning approaches [14, 27, 39]. The first three conventional methods are of-

ten constrained by technical limitations, manifesting spectral and spatial distortion issues. Due to the development of deep learning, lots of deep models are proposed in the aspects of regression-based [14, 27, 51], unfolding-based [8, 63], and generative-based models [9, 39, 65]. Regression-based methods take LRMS and PAN as inputs and utilize residual learning to capture high-frequency information, achieving better performance than traditional methods. However, these approaches often suffer from poor interpretability and generalization. Unfolding methods leverage the concept of iterative optimization by unfolding the model, which can enhance model performance and provide interpretability. Nevertheless, unfolding with a naive for-loop cannot theoretically guarantee convergence approximation, and the unfolded model significantly expands the computational graph, leading to a substantial increase in inference overhead. Methods based on generative models, for example, GANs [37], employ a discriminator to distinguish between degraded and real PAN and LRMS, thereby improving the generator’s output quality and offering sufficient flexibility. However, these methods often encounter training instability and are prone to mode collapse.

Recent advances in generative modeling have witnessed diffusion models [24, 42, 45] emerge as powerful tools, with pioneering work [39, 65] successfully adapting these architectures to pansharpening tasks. These implementations typically employ LRMS and PAN images as conditional inputs to guide the diffusion process. Subsequent innovations like Flow Matching (FM) [34] and Rectify Flow [35] have streamlined the training paradigm through linear interpolation-based time scheduling, establishing theoretical equivalence to conventional diffusion models when initialized with Gaussian distributions. Nevertheless, both diffusion-based and FM-based approaches rely on stochastic or ordinary differential equations (SDEs/ODEs) for image synthesis, requiring computationally intensive sampling processes with numerous number of network evaluations (NFEs). While acceleration strategies including advanced samplers [29, 48] and knowledge distillation techniques [35, 38] have been developed to reduce NFEs, they face critical limitations: 1) either necessitate complete network retraining or 2) incur inevitable performance compromises.

In the spirit of directly reducing NFEs during training, we leverage the theory of unbalanced optimal transport (UOT) to derive its dual formulation, and further parameterizing it into a mapping network and a potential network. To further constrain the marginals of UOT, we formulate the pansharpening task as an integration of the UOT cost into the UOT framework, enhancing generalization performance. Subsequently, we adapt the UOT loss into a training diagram suitable for flow matching (FM), stabilizing the training process while achieving *one-step* generation with FM, eliminating the need for post-training or design-

ing complex samplers. To summarize, our contributions are as follows:

- We derive a dual formulation of UOT specifically tailored for the pansharpening task, which facilitates efficient one-step sampling;
- We integrate a regularization term customized for the pansharpening task into the UOT cost within the flow matching framework. This approach enhances both the stability of the training process and the generalization capability of the model;
- Our training framework, termed OTFM, enables the sampling of HRMS in *a single step* while achieving performance that matches or surpasses that of previous diffusion-based methods.

2. Preliminary

We present the preliminaries of pansharpening and diffusion models in Suppl. Sect. 1, and review the basic concepts of optimal transport and flow matching in this section.

2.1. Optimal Transport

In this section, we introduce the background of optimal transport, including the general formulations of the Monge [41] and Kantorovich OT problems [28]. We begin by briefly defining the notation. There are two distributions $\mathbb{P} \in \mathcal{P}(Y)$ and $\mathbb{Q} \in \mathcal{P}(X)$. $\mathcal{P}(X)$ and $\mathcal{P}(Y)$ represent the set of distributions on the compact metric space. Let T be a push-forward operator that works to push a probability mass to another, *i.e.*, $T_{\#}\mathbb{P} = \mathbb{Q}$. $\Pi(\mathbb{P}, \mathbb{Q})$ denotes the set of joint probability distributions on $X \times Y$ whose marginals are $\mathbb{P}$ and $\mathbb{Q}$, respectively. Let $\pi \in \mathcal{M}_+(X \times Y)$ denote the set of joint positive measures, with $\pi_0(x)$ and $\pi_1(y)$ representing the marginals with respect to X and Y , respectively. A cost function $c(x, y)$ is defined on $X \times Y$ and $x \in X, y \in Y$. In this paper, we consider $X = Y \subset \mathbb{R}^d$ with a quadratic cost $c(x, y) = \|x - y\|_2^2$, where d is the dimension of the data.

Traditional OT formulations have two formats, including Monge and Kantorovich formulations. The Monge formulation is defined as,

$$C_{Monge}(\mathbb{P}, \mathbb{Q}) = \inf_{T_{\#}\mathbb{P}=\mathbb{Q}} \int_Y c(x, T(x)) d\mathbb{P}(x). \quad (1)$$

The optimal map T^* of Monge formulation takes over all transport maps $T : Y \rightarrow X$. The constraint $T_{\#}\mathbb{P} = \mathbb{Q}$ is somehow too strict, then can be relaxed into Kantorovich formulation,

$$C_{Kant}(\mathbb{P}, \mathbb{Q}) = \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{X \times Y} c(x, y) d\pi(x, y). \quad (2)$$

The optimal coupling π^* is taken over all transport plans on $X \times Y$ whose marginals are $\mathbb{P}$ and $\mathbb{Q}$.

Recently, many works have introduced OT into image restoration or translation tasks. For instance, [31] formulated the Kantorovich formulation as a minimax problem for image translation; [21] computed OT maps within a dual framework for image super-resolution; [57] relaxed the Monge formulation into a Wasserstein-1 penalty to constrain the push-forward operator for image denoising; and LightSB [32] introduced entropy-regularized OT, linking it to the Schrödinger bridge [33] for fast bridge matching. However, these methods are often constrained by the OT formulations, leading to issues such as reduced network expressiveness [2] and unstable training [4, 31].

2.2. Flow Matching

Flow matching [34, 35], as a generative method, has proven highly effective in producing high-quality images [13, 34], videos [26], text [19], and molecular structures [16]. Compared to diffusion models [24, 42], flow matching offers a simpler implementation of the training and sampling processes. Specifically, during training, we sample y_0, y_1 from prior and data distributions $\mathbb{P}, \mathbb{Q}$, and then sample $t \in \mathcal{U}(0, 1)$ to construct y_t through a flow path interpolation,

$$y_t = ty_0 + (1 - t)y_1, \quad (3)$$

where y_0 is sampled from the standard Gaussian distribution. We can train a velocity network $s_{\theta,t}(y_t, t)$ using the following flow loss:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{y_0 \in \mathbb{P}, y_1 \in \mathbb{Q}, t \in \mathcal{N}(0,1)} \|s_{\theta,t}(y_t, t) - (y_1 - y_0)\|_2^2, \quad (4)$$

where $\mathbb{P}$ and $\mathbb{Q}$ are two interpolated distributions. After the network converges, we can efficiently perform ODE sampling using an ODE sampler (e.g., Euler sampler):

$$dy_t = s_{\theta,t}(y_t, t)dt. \quad (5)$$

When generating samples using flow matching with an ODE sampler, it typically requires 100 to 1000 sampling steps to achieve reasonable results. Even with the use of more efficient ODE samplers [48], this issue can only be slightly alleviated. Recently, distillation methods have proven highly effective in reducing the number of sampling steps, such as online distillation [6, 50], offline Re-Flow [30, 35]. However, these methods either load multiple teacher/student networks to optimize or require offline generation of a large number of (noise, generated image) data pairs. Therefore, a key question arises: *can we directly obtain a model capable of one-step sampling during a single training process?*

3. Methodology

In this section, we first introduce how to formulate pansharpening under the flow matching framework. Then, we

incorporate unbalanced optimal transport (UOT) into flow matching to enable one-step generation. Finally, we present network architectures of our proposed OTFM.

3.1. Pansharpening as Flow Matching

For previous deep regression-based methods [14, 27, 58], one can formulate the pansharpening problem as,

$$y = T_{\theta}(m, p), \quad (6)$$

where $p \in P$, $m \in M$ are PAN, LRMS and in distribution $\mathbb{C}$. $y \in \mathbb{Q}$ is the fused HRMS. They are coupled in a product distribution $\mathbb{Q} \times \mathbb{C}$. T_{θ} is implemented as a deep network that takes p and m as input. Under the flow matching framework, the pansharpening problem is broken down into many ODE sampling steps, formulated as,

$$dy_t = s_{\theta,t}(y_t, t, m, p)dt, \quad (7)$$

where $y_0 \in \mathcal{N}(0, \mathbf{I})$, $y_1 \in \mathbb{Q}$ is the HRMS distribution, and s_{θ} is the velocity network. Intuitively, we can turn the network T_{θ} to be timestep, PAN, and LRMS conditioned to learn the flow velocity s using the flow loss.

However, challenges persist, such as network training errors [29], discretization errors during sampling [49], and accumulated errors caused by varying flow schedulers [43, 47] that result in divergent sampling paths. These issues create a dilemma between sampling efficiency and generation quality: *achieving high-quality generation requires numerous sampling steps, while directly reducing the number of steps exacerbates errors and degrades the performance.*

3.2. Pansharpening Regularized Unbalanced OT

In this subsection, we introduce how to leverage the *dual formulation* of UOT to construct an optimization framework and further introduce the *spatial and spectral regularization* in the UOT cost function. This framework is then integrated into the flow matching training framework, culminating in a simulation-free OTFM training and sampling paradigm that supports one-step sampling while preserving task-specific regularization.

3.2.1. UOT Generative Modeling

Under some mild assumption of $(X, \mathbb{P})$ and $(Y, \mathbb{Q})$ [53], the minimizer π^* of Kantorovich problem (2) always exists, and the dual form is given as,

$$C(\mathbb{P}, \mathbb{Q}) = \sup \left[\int_X u(x)d\mathbb{P}(x) + \int_Y v(y)d\mathbb{Q}(y) \right], \quad (8)$$

where the sup is taken under $\mathbb{P}(x) + \mathbb{Q}(y) \leq c(x, y)$. u and v are density of distribution $\mathbb{P}$ and $\mathbb{Q}$. Assuming $u = -v$ and u is 1-Lipschitz [53], the Eq. (8) can be reformulated using *Kantorovich-Rubinstein duality* [28] as,

$$C(\mathbb{P}, \mathbb{Q}) = \sup_{v \in L^1(\mathbb{Q})} \left[\int_X v^c(x)d\mathbb{P}(x) + \int_Y v(y)d\mathbb{Q}(y) \right], \quad (9)$$

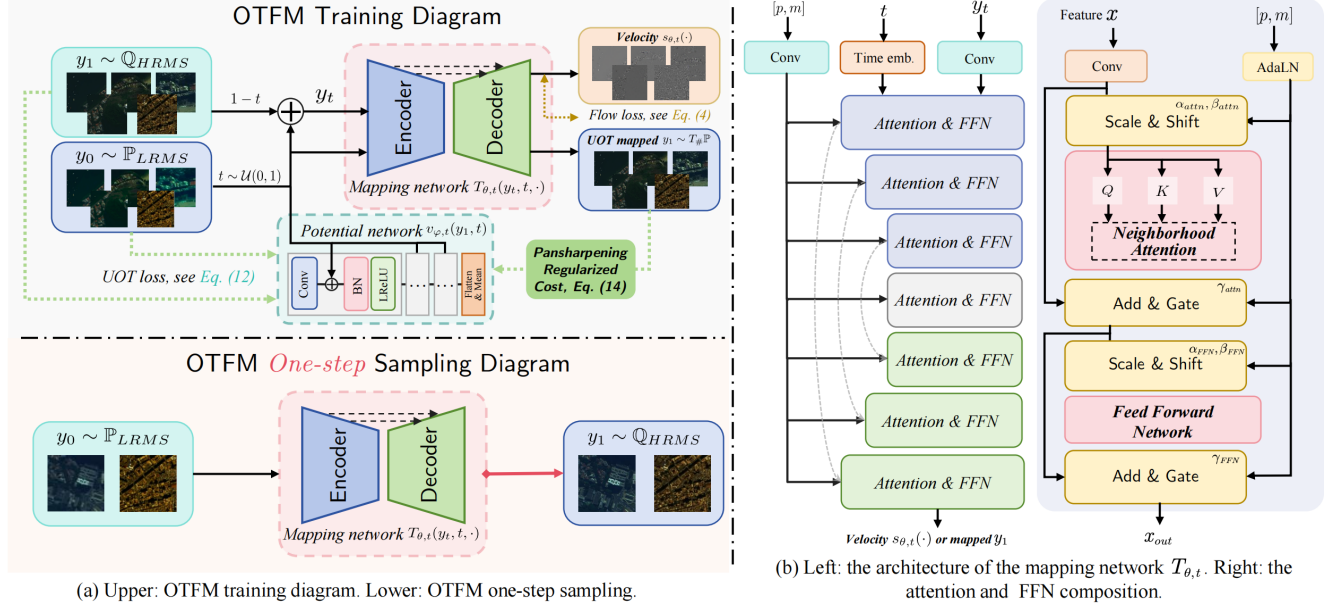


Figure 2. (a) **Training and one-step sampling diagrams of the proposed OTFM.** Due the flow matching velocity construction (see Eq. (4)), the UOT mapped $\hat{y}_1$ can be simply obtained from the predicted velocity $s_{\theta,t}(y_t, t, \cdot)$. A potential network $v_{\varphi}(\cdot)$ is parameterized from UOT dual formulation (see Prop. 3.1) and supports training a one-step mapping network. Note that conditions PAN and LRMS in the mapping network are omitted for simplicity. (b) **Mapping network designs for pansharpening.** The network is adopted as a U-net [46] architecture. To inject conditions (i.e., $[p, m]$), the zero-AdaLN is used to scale, shift, gate the feature x .

where v^c denotes the c -transform [5] of v , and meets $v^c(x) = \inf_{y \in Y} (c(x, y) - v(y))$.

To relax the hard marginal constraints of Eq. (1) which cause performance drop and optimization unstableness, UOT [11] can be introduced. Let an entropy function $f : [0, \infty) \rightarrow [0, \infty)$ be convex, non-negative, and lower semi-continuous and $D_f(\mu|\nu)$ be the f -divergence for the case that the measure μ is not absolutely continuous w.r.t measure ν . Thus, we can relax the marginal constraint by defining the UOT formulation $C_{UOT}(\mathbb{P}, \mathbb{Q})$,

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

We can summarize some properties of UOT and further, according to [17] and Eq. (2), the dual form of UOT can be proposed with the proof in Suppl. Sect. 2.1.

Remark 3.1 (Properties of UOT). 1) UOT can transport any measurable mass because it relaxes the marginal constraint; 2) UOT is not sensitive to out-of-distribution (OOD) samples, which limits traditional OT methods. If a sample is OOD, UOT can adaptively move the marginals.

Proposition 3.1 (Dual formulation of UOT). The UOT dual formulation, $C_{UOT}(\mathbb{P}, \mathbb{Q})$, can be obtained by using the c -transform:

$$\sup_{v \in C} \left[\int_X -f(-v^c(x)) d\mathbb{P} + \int_Y -f(-v(y)) d\mathbb{Q} \right], \quad (10)$$

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

Remark 3.2 (Possible choices of function f). UOT only constrains f to be convex, non-decreasing, lower semi-continuous, and differentiable functions. Since the conjugate of a function is convex and lower semi-continuous, we empirically choose $f := \exp(\cdot)$ to be a general case.

At this time, we can use the dual form to construct a generative framework. Then we introduce the mapping network T_{θ} to approximate the v^c as following,

$$T_{\theta}(x) = \arg \inf_{y \in Y} [c(x, T_{\theta}(x)) - v(T_{\theta}(x))]. \quad (11)$$

For v , the UOT dual-formulation objective $\mathcal{K}(v)$ can be derived from the supremum of Eq. (10) and Eq. (11),

$$\int_X f(-[c(x, T_{\theta}(x)) - v(T_{\theta}(x))]) d\mathbb{P} + \int_Y f(-v(y)) d\mathbb{Q}.$$

We can leverage a min-max training diagram [20] for training the T_{θ} for each v , since there is no closed-form for the optimal T . By parameterizing the v as a neural potential network v_{φ} , the networks' objective $\mathcal{L}_{T,v}$ is,

$$\inf_{v_{\varphi}} \left[\int_X f \left(\sup_{T_{\theta}} \{c(x, T_{\theta}(x)) - v_{\varphi}(T_{\theta}(x))\} \right) d\mathbb{P} + \int_Y f(-v_{\varphi}(y)) d\mathbb{Q} \right]. \quad (12)$$

We discuss differences with GANs [20] in Suppl. Sect. 3.

3.2.2. Pansharpening Regularized UOT Cost Function

However, the relaxed marginals do not admit the pansharpening regularizations, that is, the sampled (or fused) image may not meet the spatial or spectral consistency from PAN and LRMS. Hence, we introduce a novel regularization term and embed it into the UOT cost function. Following [7, 59], the HRMS y can be degraded back to PAN p and LRMS m in spatial and spectral aspects, respectively:

$$\begin{cases} \text{Spatial :} & \tilde{p} = yBS \\ \text{Spectral :} & \tilde{m} = y - yB \odot (\hat{p} \odot \hat{p}_L), \end{cases} \quad (13)$$

where $B \in \mathbb{R}^{HW \times HW}$ is matrixization blurring operator, $S \in \mathbb{R}^{HW \times hw}$ denotes sparse decimation matrix for LRMS bands weighting, $\hat{p}$ can be regarded as a function of LRMS bands and PAN and pre-processed by spectral matching [52], and $\hat{p}_L$ is obtained by exploiting MTF-filters [59]. $\odot$ and $\oslash$ are elementwise product and division. Based on this, we can modify the original UOT cost in Eq. (12) into *pansharpening-regularized cost*,

$$\tilde{c}(x, y) = c(x, y) + \|p - \tilde{p}\|_2^2 + \|m - \tilde{m}\|_2^2, \quad (14)$$

where $\tilde{p}$ and $\tilde{m}$ are obtained from Eq. (13).

Proposition 3.2 (Saddle points of pansharpening-regularized UOT provide the OT maps). *For any optimal potential function $v^* \in \arg \sup_v \mathcal{L}_{T,v}$ (i.e., Eq. (12)), it provides the OT map T^* which holds*

$$T^* \in \arg \min_T \mathcal{L}(T, v^*). \quad (15)$$

The proof is provided in Suppl. Sect. 2.2. Prop. 3.2 gives that the although pansharpening task introduces additional cost in the original OUT cost, there is an optimal pair that can be acquired by solving the min-max problem in Eq. (12). By using the optimal pair, an OT map from $\mathbb{P}$ to $\mathbb{Q}$ can be constituted.

3.3. Taming FM with UOT into One-step Generator

Recall that our initial goal is to turn the multistep flow matching into a one-step generator. In this subsection, we proposed a novel algorithm that embeds the UOT min-max problem into flow matching learning, thereby forming the *simulation-free* training and *one-step* sampling framework.

Note that in original flow matching, we need to learn a velocity network $s_{\theta,t}$ to approximate the true velocity $y - x$, that is, $y_0 := x \sim \mathcal{N}(0, \mathbf{I})$ and $y_1 := y \sim \mathbb{Q}_{HRMS}$. Note that we *change the start point of FM y_0 into LRMS (not Gaussian)* since it eases the UOT cost learning and sets an appropriate estimation to sample [10]. In flow matching, if the velocity s_t is obtained, we can access the predicted $y_1 = s_t + y_0$. Thus, we can convert the velocity network into a one-step generator: $T_{\theta,t} := s_{\theta,t} + y_0$, where

Algorithm 1: Training algorithm of OTFM.

Input: The coupled HRMS, PAN, and LRMS distribution $\mathbb{Q} \times \mathbb{C}$, convex function f , one-step flow mapping network $T_{\theta}(\cdot)$, potential network v_{φ} .

Output: Trained flow mapping network $T_{\theta^*}(\cdot)$

```

1 while Not Converged do
2    $Y_1, P, M \sim \mathbb{Q} \times \mathbb{C}$ . ▷ A batch of training data
3    $Y_0 \leftarrow M$  ▷  $y_0$  is LRMS, not a Gaussian
4    $t \leftarrow \mathcal{U}(0, 1)$ ,  $Y_t \leftarrow tY_0 + (1-t)Y_1$ . ▷ Flow state
   ▷ Flow matching loss
5    $\mathcal{L}_{flow} = \frac{1}{|Y_0|} \sum_{y_0 \in Y_0, y_1 \in Y_1} \|s_{\theta,t}(y_t, t, m, p) - (y_1 - y_0)\|_2^2$ .
6    $\hat{Y}_1 = s_{\theta,t} + Y_t$ . ▷ Get OT mapped  $y$ 
   ▷ UOT mapping network loss,  $\tilde{c}$  is in Eq. (14)
7    $\mathcal{L}_T = \frac{1}{|Y_0|} \sum_{y_0, y_1, \hat{y}_1} [\tilde{c}(y_0, \hat{y}_1) - v_{\varphi,t}(\hat{y}_1, t)]$ .
8   Update  $\theta$  using loss  $\mathcal{L}_{flow} + \mathcal{L}_T$ .
   ▷ Potential network loss
9    $\mathcal{L}_v = \frac{1}{|Y_0|} \sum_{y_0, \hat{y}_1} [f(-\tilde{c}(y_0, \hat{y}_1) + v_{\varphi,t}(\hat{y}_1, t))] + \frac{1}{|Y_1|} \sum_{y_1} [f(-v_{\varphi,t}(y_1, t))]$ .
10  Update  $\varphi$  using loss  $\mathcal{L}_v$ .
11 end
12  $T_{\theta^*} \leftarrow T_{\theta}$ . ▷ Trained mapping network  $T_{\theta^*}$ 

```

T is the UOT mapping network. Note that the flow condition is unchanged, i.e., PAN, and LRMS. Naturally, the flow learning objective is converted into y_1 -predicted under the UOT context. Note that we keep the flow matching loss (4), since flow loss constrains the flow path $\{y_t\}_{t=0}^1$ into a near-straight path [35], forming a *simulation-free* training and further stabilizes the UOT training. One question is whether converting the network's objective may affect the network's performance. As shown in EDM [29], re-weighting the output of the network into y_1 has proven to be effective in learning a better diffusion model. For the potential network, due to the timestep introduced into OTFM, we condition the potential network *w.r.t* to t , that is $v_{\varphi,t} := v_{\varphi}$.

Finally, we can summarize the training algorithm in Algo. 1. After the mapping network is well-trained, we can use the mapping network to directly sample in *one step*.

3.4. Deep Network for OTFM

Network designing is a vital part of OTFM, to suit pansharpening task that provides two images as conditions, we design an efficient network architecture for the mapping network T_{θ} and the potential network v_{φ} .

As shown in Fig. 2 (b), since the U-net [46] is widely used in diffusion methods, we adopt a UNet-like structure for the mapping network. Specifically, the network consists of three encoder layers and decoder layers. Each en-

coder layer is followed by a downsampling operation with a factor of 2, and each decoder layer is followed by an up-sampling operation to adjust spatial resolution and channel quantity during feature transmission. The encoder and decoder are each composed of two stacked basic blocks. Each basic block adopts attention and FFN blocks. For an efficient attention computation, we leverage the neighborhood attention [22] to obtain the global feature response. As for the FFN, it is implemented as three MLPs. Note that, the mapping network is additionally conditioned on PAN and LRMS. To fulfill the condition injection, we utilize the DiT [44] AdaLN-zero conditioning,

$$\begin{cases} \alpha, \beta, \gamma = MLP(LN([m, p])), \\ x_{op} = (1 + \alpha) \cdot op(x) + \beta, \\ x_{out} = x + \gamma \cdot x_{op}, \end{cases} \quad (16)$$

where x is the feature, $op \in \{\text{attention}, \text{FFN}\}$, $[\cdot]$ is the channel concatenation, and LN denotes the layer-norm. The encoder layer encodes input features, downsamples them, and forwards the output to the corresponding decoder layer through skip connections. The decoder layer decodes features, integrates them with skip-connected features, and upsamples the result. Convolutions are employed at the network’s front and back to ensure feature dimensions align with the network or target image’s channel specifications.

To leverage the potential network, we adopt a patch GAN discriminator [25] into a t -conditioned network. Specifically, as illustrated in Fig. 2 (a), the potential network comprises three convolutional blocks, each consisting of a “Conv+BN+LReLU” layer sequence. The timestep t is encoded using a cosine function and integrated with the feature map. Finally, the resulting feature is flattened and averaged to yield the potential value v_φ .

4. Experiments

In this section, we will provide a comprehensive overview of the experimentation settings including the implementation details and datasets. After that, we will present the benchmarks employed to evaluate the performance of our approach to pansharpening. Finally, the main results and ablation studies will be provided to quantitatively illustrate the effectiveness of the proposed method.

4.1. Implementation Details

The mapping network and the potential network are optimized with AdamW optimizer [36] with learning rates set to $2e^{-4}$ and $1e^{-4}$, respectively. We use exponential moving average for both models with the decay ratio set to 0.99. The maximum training steps are set to 100k, and the batch size is 52. The model was trained for one day with two NVIDIA GeForce RTX 4090 GPUs.

4.2. Datasets

To better validate the effectiveness of the proposed OTFM, we conducted training and testing on a public dataset, Pan-collection¹, which includes data from three types of remote sensing satellites: WV3 (8 bands), GF2 (4 bands), and QB (4 bands). More details are provided in Suppl. Sect. 4.1.

4.3. Benchmark and Metrics

To evaluate the performance of our OTFM, we compare it with various state-of-the-art (SOTA) methods of pansharpening (on WV3, GF2, and QB datasets). We choose two traditional methods: MTF-GLP-FS [55], BT-H [1], as well as eight deep learning-based methods: DiCNN [23], FusionNet [15], LAGConv [27], DCFNet [58], MMNet [62], Pandiff [39], HFIN [51], and DDIF [9]. We utilize four common metrics to evaluate the results of these methods on the pansharpening reduced-resolution datasets, including SAM [64], ERGAS [56], Q2n [18], and SCC [66]. Meanwhile, the D_λ , D_s , and HQNR [3] indices are used to evaluate the full-resolution datasets.

4.4. Main Results

Results on WV3. We evaluated the performance of OTFM on 20 test images from both the reduced-resolution dataset and the full-resolution dataset of WV3, with the results presented in Tab. 1. Compared to two traditional methods and eight SOTA DL-based methods, the proposed OTFM performs the best overall performance on the reduced-resolution dataset. To illustrate the performance differences among each method clearly, we present the fusion result images and the error maps for some methods in Fig. 3 with a specific location zoomed in. Additionally, our OTFM achieves SOTA on the real full-resolution dataset, achieving an HQNR of 0.954. With only one sampling step, OTFM can generate high-quality HRMS better than DDIF.

Results on GF2. Tab. 2 presents the test results for the reduced-resolution 20 test images GF2 dataset. Our OTFM achieves SOTA performance, and as shown in Fig. 3, the fusion results from OTFM exhibit reduced spectral distortion while retaining more spatial details compared to other methods, which is comparable with 25-sampling-step DDIF. Furthermore, OTFM outperforms previous methods on the real full-resolution dataset, with visual comparisons available in the Suppl. Sect. 4.2. This demonstrates that the proposed OTFM possesses excellent fusion performance, as well as effectiveness on real datasets.

Results on QB. We also conduct experiments on the QB dataset to evaluate performance at reduced resolutions. Similarly, Tab. 3 reports the quality metrics obtained from 20 randomly selected test samples from the QB dataset. Our

¹<https://liangjiandeng.github.io/PanCollection.html>.

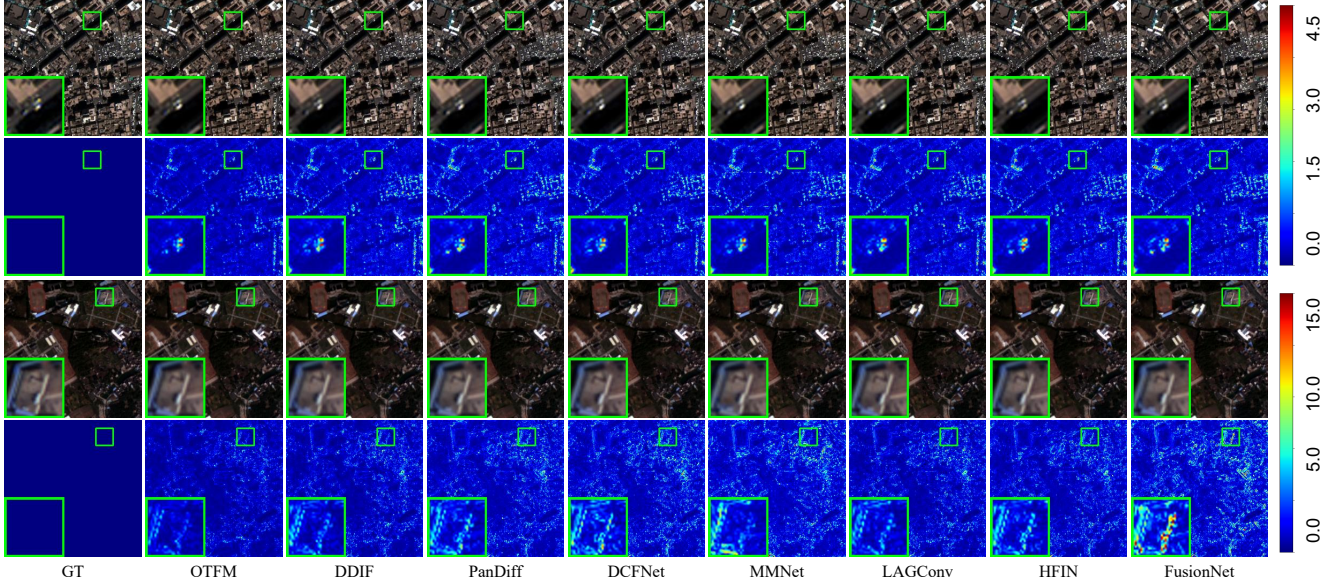


Figure 3. Visual comparisons on WV3 (1-2 rows) and GaoFen-2 (3-4 rows) cases. The second the fourth rows are error maps.

Table 1. Quantitative results on WV3 dataset. The best results are in red and the second best results are in blue.

Methods	Reduced-Resolution (RR): Avg $\pm$ std				Full-Resolution (FR): Avg $\pm$ std			NFEs
	SAM ($\downarrow$)	ERGAS ($\downarrow$)	Q2n ($\uparrow$)	SCC ($\uparrow$)	D_λ ($\downarrow$)	D_s ($\downarrow$)	HQNR ($\uparrow$)	
MTF-GLP-FS [55]	5.32 $\pm$ 1.65	4.65 $\pm$ 1.44	0.818 $\pm$ 0.101	0.898 $\pm$ 0.047	0.021 $\pm$ 0.008	0.063 $\pm$ 0.028	0.918 $\pm$ 0.035	–
BT-H [1]	4.90 $\pm$ 1.30	4.52 $\pm$ 1.33	0.818 $\pm$ 0.102	0.924 $\pm$ 0.024	0.057 $\pm$ 0.023	0.081 $\pm$ 0.037	0.867 $\pm$ 0.054	–
DiCNN [23]	3.59 $\pm$ 0.76	2.67 $\pm$ 0.66	0.900 $\pm$ 0.087	0.976 $\pm$ 0.007	0.036 $\pm$ 0.011	0.046 $\pm$ 0.018	0.920 $\pm$ 0.026	1
FusionNet [15]	3.33 $\pm$ 0.70	2.47 $\pm$ 0.64	0.904 $\pm$ 0.090	0.981 $\pm$ 0.007	0.024$\pm$0.009	0.036 $\pm$ 0.014	0.941 $\pm$ 0.020	1
LAGConv [27]	3.10 $\pm$ 0.56	2.30 $\pm$ 0.61	0.910 $\pm$ 0.091	0.984 $\pm$ 0.007	0.037 $\pm$ 0.015	0.042 $\pm$ 0.015	0.923 $\pm$ 0.025	1
DCFNet [58]	3.03 $\pm$ 0.74	2.16 $\pm$ 0.46	0.905 $\pm$ 0.088	0.986$\pm$0.004	0.078 $\pm$ 0.081	0.051 $\pm$ 0.034	0.877 $\pm$ 0.101	1
MMNet [62]	3.08 $\pm$ 0.64	2.34 $\pm$ 0.63	0.916 $\pm$ 0.086	0.983 $\pm$ 0.006	0.054 $\pm$ 0.023	0.034 $\pm$ 0.012	0.914 $\pm$ 0.028	1
HFIN [51]	3.08 $\pm$ 0.63	2.31 $\pm$ 0.58	0.912 $\pm$ 0.089	0.984 $\pm$ 0.006	0.025 $\pm$ 0.008	0.043 $\pm$ 0.017	0.934 $\pm$ 0.024	1
PanDiff [39]	3.30 $\pm$ 0.60	2.47 $\pm$ 0.58	0.898 $\pm$ 0.088	0.980 $\pm$ 0.006	0.027 $\pm$ 0.012	0.054 $\pm$ 0.026	0.920 $\pm$ 0.036	1000
DDIF [9]	2.74$\pm$0.51	2.01$\pm$0.45	0.920$\pm$0.082	0.988$\pm$0.003	0.026 $\pm$ 0.008	0.023$\pm$0.008	0.952$\pm$0.017	25
OTFM (ours)	2.76$\pm$0.52	2.03$\pm$0.44	0.919$\pm$0.083	0.988$\pm$0.003	0.018$\pm$0.007	0.029$\pm$0.006	0.954$\pm$0.012	1

OTFM method achieves the best overall performance across the four metrics. Additional results analysis and visual comparisons for QB dataset can be found in Suppl. Sect. 4.3.

5. Ablation Study

In this section, we ablate the proposed UOT training loss, pansharpening-regularized OT cost, and compare it with previous OT and GAN methods on the WV3 dataset.

5.1. Compared with Reflow and Consistency Model

Recently, Reflow [35] has been proposed to distill a multi-step sampling FM into a one-step model, while the Consistency Training (CT) [50] aims to train a model supporting one-step sampling by leveraging the ODE endpoint consistency law. For Reflow, we adhere to its distillation algorithm: first, we train an FM until convergence, then sample (noise, HRMS) pairs, and finally perform distillation on these pairs. For CT, we employ the Euler discretization

approach and directly train the one-step CT. Their performances are presented in Tab. 4 (a-b). Although both methods perform sampling in a single step, their results still fall behind. Particularly for Reflow, due to the accumulation of distillation errors, it exhibits the worst performance.

5.2. UOT makes a One-step Generator

We claim that UOT enables a one-step generator, which is demonstrated in this subsection by ablating the loss function associated with the proposed UOT loss. Specifically, we remove Lines 6-10 in Algo. 1, effectively training a flow matching model directly. The results are presented in Tab. 4 (c). This model requires 100 ODE sampling steps to achieve the same pansharpening performance as our OTFM. Notably, replacing flow matching with diffusion falls back to the setting of PanDiff. We also provide the performance of diffusion within our designed deep architecture (see Tab. 4 (d)); this variant still necessitates more sampling steps and exhibits inferior performance.

Table 2. Quantitative results on the GaoFen-2 datasets. The best results are in red and the second best results are in blue.

Method	Reduced-Resolution (RR): Avg±std				Full-Resolution (FR): Avg±std				NFEs
	SAM (↓)	ERGAS (↓)	Q2n (↑)	SCC (↑)	D_λ (↓)	D_s (↓)	HQNR (↑)		
MTF-GLP-FS [55]	1.68±0.35	1.60±0.35	0.891±0.026	0.939±0.020	0.035±0.014	0.143±0.028	0.823±0.035	–	
BT-H [1]	1.68±0.32	1.55±0.36	0.909±0.029	0.951±0.015	0.060±0.025	0.131±0.019	0.817±0.031	–	
DiCNN [23]	1.05±0.23	1.08±0.25	0.959±0.010	0.977±0.006	0.041±0.012	0.099±0.013	0.864±0.017	1	
FusionNet [15]	0.97±0.21	0.99±0.22	0.964±0.009	0.981±0.005	0.040±0.013	0.101±0.013	0.863±0.018	1	
LAGConv [27]	0.78±0.15	0.69±0.11	0.980±0.009	0.991±0.002	0.032±0.013	0.079±0.014	0.891±0.020	1	
DCFNet [58]	0.89±0.16	0.81±0.14	0.973±0.010	0.985±0.002	0.023±0.012	0.066±0.010	0.912±0.012	1	
MMNet [62]	0.99±0.14	0.81±0.12	0.969±0.020	0.986±0.002	0.043±0.030	0.103±0.013	0.858±0.027	1	
HFIN [51]	0.84±0.15	0.73±0.13	0.977±0.011	0.990±0.002	0.027±0.020	0.062±0.010	0.912±0.018	1	
PanDiff [39]	0.89±0.12	0.75±0.10	0.979±0.010	0.989±0.002	0.027±0.020	0.073±0.010	0.903±0.021	1000	
DDIF [9]	0.64±0.12	0.57±0.10	0.985±0.008	0.986±0.003	0.020±0.011	0.041±0.010	0.939±0.014	25	
Proposed	0.64±0.12	0.57±0.11	0.985±0.007	0.993±0.001	0.019±0.009	0.042±0.008	0.939±0.010	1	

Table 3. Results on the QuickBird reduced-resolution datasets. The best results are in red and the second best results are in blue.

Method	Reduced-Resolution (RR): Avg±std			
	SAM (↓)	ERGAS (↓)	Q2n (↑)	SCC (↑)
MTF-GLP-FS [55]	8.11±1.95	7.51±0.79	0.829±0.090	0.899±0.019
BT-H [1]	7.19±1.55	7.40±0.84	0.833±0.088	0.916±0.015
DiCNN [23]	5.38±1.03	5.14±0.49	0.904±0.094	0.962±0.013
FusionNet [15]	4.92±0.91	4.16±0.32	0.925±0.090	0.976±0.010
LAGConv [27]	4.55±0.83	3.83±0.42	0.933±0.088	0.981±0.009
DCFNet [58]	4.54±0.74	3.83±0.29	0.933±0.090	0.974±0.010
MMNet [62]	4.56±0.73	3.67±0.30	0.934±0.094	0.983±0.007
HFIN [51]	4.54±0.81	3.81±0.32	0.934±0.085	0.980±0.010
PanDiff [39]	4.58±0.74	3.74±0.31	0.935±0.090	0.982±0.090
DDIF [9]	4.35±0.73	3.52±0.27	0.938±0.090	0.984±0.079
Proposed	4.39±0.76	3.54±0.28	0.938±0.088	0.985±0.070

Table 4. Ablation studies on different one-step generators, the proposed pansharpening-regularization, and comparisons with previous OT and GAN methods.

Methods	SAM (↓)	ERGAS (↓)	Q2n (↑)	SCC (↑)	HQNR (↑)
(a) Reflow [35]	3.18	2.32	0.915	0.983	0.920
(b) CT [50]	2.88	2.12	0.916	0.985	0.936
(c) w/o UOT (FM)	2.83	2.10	0.917	0.986	0.941
(d) w/o UOT (Diff)	2.87	2.12	0.915	0.986	0.942
(e) w/o Pan-Reg cost	2.78	2.08	0.918	0.986	0.940
(f) FM+NOT [31]	2.89	2.11	0.915	0.985	0.934
(g) FM+GAN [61]	2.85	2.12	0.916	0.986	0.938

Table 5. Efficiency results on the 256×256 resolution image.

Method	Proposed	DDIF	PanDiff	DCFNet	MMNet
latency (s)	0.029	2.602	261.410	0.257	0.348

5.3. Pansharpening-regularized Cost Enhances Generalization

In Sect. 3.2.2, we introduce a task-level regularization on the UOT cost for the pansharpening task. To validate the effectiveness of this pansharpening-regularized cost, we remove the regularization term and retrain the OTFM. The results are presented in Tab. 4 (e). It can be seen that, with the regularization term, OTFM can reach higher reduced and full resolution metrics. This regularization brings full-

resolution HQNR from 0.940 to 0.954. Thus, we can conclude that the proposed pansharpening OT regularization enhances the model’s generalization capability.

5.4. UOT beats previous OT and GAN

Recent OT approaches, such as NOT [31] and the GAN-based method [61], have demonstrated strong performance in image translation tasks. We integrate these methods with FM to verify the effectiveness of the proposed OTFM. Both approaches follow a one-step generation process, and their results are presented in Tab. 4 (f-g). It can be observed that our OTFM outperforms them on both reduced and full resolution metrics. Additionally, we found that the training loss of GAN methods is less stable compared to UOT. We conjecture that this may be due to the GAN discriminative instability. Although the NOT+FM approach exhibits stable training, its performance is inferior.

5.5. Inference Latency

Inference latency is a vital aspect to examine the efficiency of pansharpening methods. In Tab. 5, we include the comparisons with previous SOTA multi-step diffusion methods: DDIF and PanDiff, as well as previous regression-based methods: DCFNet and MMNet. By testing them and ours under 256×256 resolution, OTFM can fuse the HRMS using only 0.029 seconds due to the efficient one-step sampling and designs, much faster than previous diffusion-based methods.

6. Conclusion

This work introduces unbalanced optimal transport (UOT) into flow matching to address multi-step sampling limitations. By integrating pansharpening-specific constraints into UOT’s cost function, we mitigate restrictive marginal conditions and enable one-step generation without teacher-student networks or extensive (noise, HRMS) pair sampling. Our method matches the performance of multi-step diffusion approaches on pansharpening datasets while achieving $\sim 10\times$ faster inference.

7. Acknowledgement

This work is supported by the Project of the Department of Science and Technology of Sichuan Province (Grant No. 2025YFNH0001) and NSFC (Grant No.12271083). The computational resource is provided by Center for HPC, University of Electronic Science and Technology of China.

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. 6, 7, 8
- [2] Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International conference on machine learning*, pages 146–155. PMLR, 2017. 3
- [3] Alberto Arienzo, Gemine Vivone, Andrea Garzelli, Luciano Alparone, and Jocelyn Chanussot. Full-resolution quality assessment of pansharpening: Theoretical and hands-on approaches. *IEEE Geoscience and Remote Sensing Magazine*, 10(3):168–201, 2022. 6
- [4] Arip Asadulaev, Alexander Korotin, Vage Egiazarian, Petr Mokrov, and Evgeny Burnaev. Neural optimal transport with general cost functionals. *arXiv preprint arXiv:2205.15403*, 2022. 3
- [5] Julio Backhoff-Veraguas, Mathias Beiglbock, and Gudmun Pammer. Existence, duality, and cyclical monotonicity for weak transport costs. *Calculus of Variations and Partial Differential Equations*, 58(6):203, 2019. 4
- [6] David Berthelot, Arnaud Autef, Jierui Lin, Dian Ang Yap, Shuangfei Zhai, Siyuan Hu, Daniel Zheng, Walter Talbott, and Eric Gu. Tract: Denoising diffusion models with transitive closure time-distillation. *arXiv preprint arXiv:2303.04248*, 2023. 3
- [7] Qi Cao, Liang-Jian Deng, Wu Wang, Junming Hou, and Gemine Vivone. Zero-shot semi-supervised learning for pansharpening. *Information Fusion*, 101:102001, 2024. 5
- [8] Xiangyong Cao, Xueyang Fu, Danfeng Hong, Zongben Xu, and Deyu Meng. Pancsc-net: A model-driven deep unfolding method for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. 2
- [9] 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. 2, 6, 7, 8
- [10] Zihan Cao, Xiao Wu, and Liang-Jian Deng. Neural schrödinger bridge matching for pansharpening. *arXiv preprint arXiv:2404.11416*, 2024. 5
- [11] Lenaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018. 4
- [12] Matteo Ciotola, Giuseppe Guarino, Gemine Vivone, Giovanni Poggi, Jocelyn Chanussot, Antonio Plaza, and Giuseppe Scarpa. Hyperspectral pansharpening: Critical review, tools, and future perspectives. *IEEE Geoscience and Remote Sensing Magazine*, 2024. 1
- [13] Quan Dao, Hao Phung, Binh Nguyen, and Anh Tran. Flow matching in latent space. *arXiv preprint arXiv:2307.08698*, 2023. 3
- [14] 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, 2, 3
- [15] 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. 6, 7, 8
- [16] Ian Dunn and David Ryan Koes. Mixed continuous and categorical flow matching for 3d de novo molecule generation. *ArXiv*, pages arXiv–2404, 2024. 3
- [17] Thomas Gallouët, Roberta Ghezzi, and François-Xavier Vialard. Regularity theory and geometry of unbalanced optimal transport. *arXiv preprint arXiv:2112.11056*, 2021. 4
- [18] Andrea Garzelli and Filippo Nencini. Hypercomplex quality assessment of multi/hyperspectral images. *IEEE Geoscience and Remote Sensing Letters*, 6(4):662–665, 2009. 6
- [19] Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37:133345–133385, 2025. 3
- [20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 4
- [21] Xiang Gu, Liwei Yang, Jian Sun, and Zongben Xu. Optimal transport-guided conditional score-based diffusion model. *Advances in Neural Information Processing Systems*, 36:36540–36552, 2023. 3
- [22] Ali Hassani and Humphrey Shi. Dilated neighborhood attention transformer. *arXiv preprint arXiv:2209.15001*, 2022. 6
- [23] 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. 6, 7, 8
- [24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3
- [25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 6
- [26] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024. 3

- [27] 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*, pages 1113–1121, 2022. 1, 2, 3, 6, 7, 8
- [28] Leonid Kantorovitch. On the translocation of masses. *Management science*, 5(1):1–4, 1958. 2, 3
- [29] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 2, 3, 5
- [30] Beomsu Kim, Yu-Guan Hsieh, Michal Klein, Marco Cuturi, Jong Chul Ye, Bahjat Kawar, and James Thornton. Simple reflow: Improved techniques for fast flow models. *arXiv preprint arXiv:2410.07815*, 2024. 3
- [31] Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. *arXiv preprint arXiv:2201.12220*, 2022. 3, 8
- [32] Alexander Korotin, Nikita Gushchin, and Evgeny Burnaev. Light schrödinger bridge. *arXiv preprint arXiv:2310.01174*, 2023. 3
- [33] Christian Léonard. A survey of the schrödinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*, 2013. 3
- [34] 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, 3
- [35] 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, 3, 5, 7, 8
- [36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [37] Jiayi Ma, Wei Yu, Chen Chen, Pengwei Liang, Xiaojie Guo, and Junjun Jiang. Pan-gan: An unsupervised pan-sharpening method for remote sensing image fusion. *Information Fusion*, 62:110–120, 2020. 2
- [38] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023. 2
- [39] 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, 2, 6, 7, 8
- [40] Xiangchao Meng, Jie Li, Huanfeng Shen, Liangpei Zhang, and Hongyan Zhang. Pansharpening with a guided filter based on three-layer decomposition. *Sensors*, 16(7):1068, 2016. 1
- [41] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781. 2
- [42] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2, 3
- [43] Yong-Hyun Park, Chieh-Hsin Lai, Satoshi Hayakawa, Yuhta Takida, and Yuki Mitsufuji. Jump your steps: Optimizing sampling schedule of discrete diffusion models. *arXiv preprint arXiv:2410.07761*, 2024. 3
- [44] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 6
- [45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2
- [46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4, 5
- [47] Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your steps: Optimizing sampling schedules in diffusion models. *arXiv preprint arXiv:2404.14507*, 2024. 3
- [48] Neta Shaul, Juan Perez, Ricky TQ Chen, Ali Thabet, Albert Pumarola, and Yaron Lipman. Bespoke solvers for generative flow models. *arXiv preprint arXiv:2310.19075*, 2023. 2, 3
- [49] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3
- [50] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023. 3, 7, 8
- [51] 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. 2, 6, 7, 8
- [52] Tanya S Unger Holtz. Introductory digital image processing: A remote sensing perspective, 2007. 5
- [53] Cédric Villani et al. *Optimal transport: old and new*. Springer, 2008. 3
- [54] Gemine Vivone, Rocco Restaino, and Jocelyn Chanussot. A regression-based high-pass modulation pansharpening approach. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):984–996, 2017. 1
- [55] 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. 6, 7, 8
- [56] Lucien Wald. *Data fusion: definitions and architectures: fusion of images of different spatial resolutions*. Presses des MINES, 2002. 6
- [57] Wei Wang, Fei Wen, Zeyu Yan, and Peilin Liu. Optimal transport for unsupervised denoising learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 2104–2118, 2022. 3

- [58] 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. [3](#), [6](#), [7](#), [8](#)
- [59] Zhong-Cheng Wu, Ting-Zhu Huang, Liang-Jian Deng, Jin-Fan Hu, and Gemine Vivone. Vo+net: An adaptive approach using variational optimization and deep learning for panchromatic sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021. [5](#)
- [60] Zhong-Cheng Wu, Ting-Zhu Huang, Liang-Jian Deng, Jie Huang, Jocelyn Chanussot, and Gemine Vivone. Lrtcfpan: Low-rank tensor completion based framework for pansharpening. *IEEE Transactions on Image Processing*, 32:1640–1655, 2023. [1](#)
- [61] Yanwu Xu, Yang Zhao, Zhisheng Xiao, and Tingbo Hou. Ufogen: You forward once large scale text-to-image generation via diffusion gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8196–8206, 2024. [8](#)
- [62] 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. [6](#), [7](#), [8](#)
- [63] Gang Yang, Man Zhou, Keyu Yan, Aiping Liu, Xueyang Fu, and Fan Wang. Memory-augmented deep conditional unfolding network for pansharpening. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1787, 2022. [2](#)
- [64] Roberta H Yuhas, Alexander FH Goetz, and Joe W Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop*, 1992. [6](#)
- [65] Yu Zhong, Xiao Wu, Liang-Jian Deng, Zihan Cao, and Hong-Xia Dou. Ssdif: Spatial-spectral integrated diffusion model for remote sensing pansharpening. *Advances in Neural Information Processing Systems*, 37:77962–77986, 2025. [2](#)
- [66] Jie Zhou, Daniel L Civco, and John A Silander. A wavelet transform method to merge landsat tm and spot panchromatic data. *International Journal of Remote Sensing*, 19(4):743–757, 1998. [6](#)