Residual-Conditioned Optimal Transport: Towards Structure-Preserving Unpaired and Paired Image Restoration

Xiaole Tang [1] **Xin Hu** [1] **Xiang Gu** [1] **Jian Sun** [1]

Abstract

Deep learning-based image restoration methods generally struggle with faithfully preserving the structures of the original image. In this work, we propose a novel Residual-Conditioned Optimal Transport (RCOT) approach, which models image restoration as an optimal transport (OT) problem for both unpaired and paired settings, introducing the transport residual as a unique degradation-specific cue for both the transport cost and the transport map. Specifically, we first formalize a Fourier residual-guided OT objective by incorporating the degradation-specific information of the residual into the transport cost. We further design the transport map as a two-pass RCOT map that comprises a base model and a refinement process, in which the transport residual is computed by the base model in the first pass and then encoded as a degradation-specific embedding to condition the second-pass restoration. By duality, the RCOT problem is transformed into a minimax optimization problem, which can be solved by adversarially training neural networks. Extensive experiments on multiple restoration tasks show that RCOT achieves competitive performance in terms of both distortion measures and perceptual quality, restoring images with more faithful structures as compared with state-of-the-art methods.

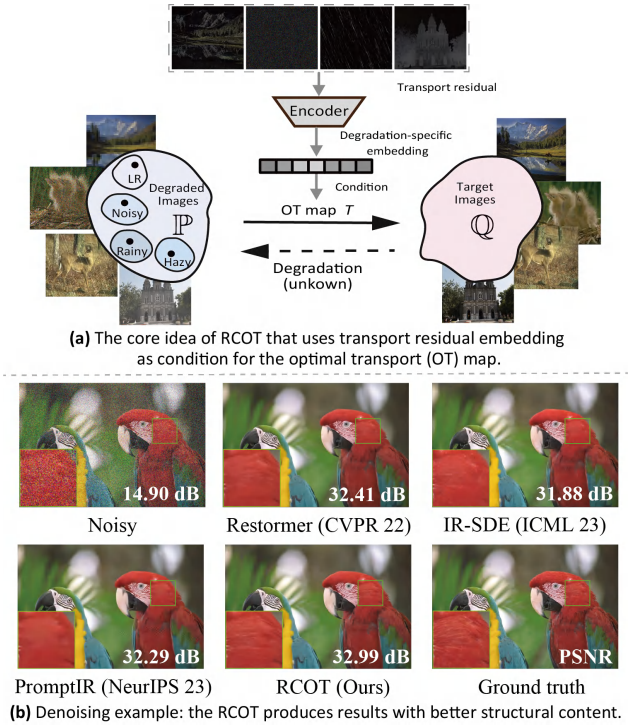


Figure 1: (a) The core idea of the RCOT framework that uses the transport residual embedding (representation of some degradation-specific knowledge) as a condition for the OT map. (b) A denoising demo under noise level $\sigma = 50$. RCOT produces a noise-free image with better structures.

1. Introduction

Image restoration is a fundamental low-level task of removing the degradation (e.g., noise, down-scaling, rain, haze, blur, etc.) from a degraded image. Traditional methods (He et al., 2010; Sun et al., 2008; Tang et al., 2022; 2023; Zheng et al., 2023) focus on designing an optimization problem that exploits suitable priors of the clear image. Recently, significant advancements in image restoration have been witnessed. They are primarily driven by the sophisticated network architectures (Chen et al., 2023; Cui et al., 2023a; Liang et al., 2021; Ren et al., 2016; Wang et al., 2023; 2022b; Zamir et al., 2021; 2022; Zhou et al., 2023) and efficient generative models, especially Generative Adversarial Networks (GANs) (Ledig et al., 2017; Li et al., 2018b; Pan et al., 2020; 2021; Zhao et al., 2020) and diffusion probabilistic models (DPMs) (Choi et al., 2023; Kawar et al., 2022; Luo et al., 2023; Murata et al., 2023; Saharia et al., 2022a;b; Zhu et al., 2023). The mainstream of the generation-based methods focuses on how to effectively condition the generator on degraded images for high-fidelity restoration.

[1]School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China. Correspondence to: Jian Sun <jiansun@xjtu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Classic restoration methods (Potlapalli et al., 2023; Zamir et al., 2021; 2022) utilize sophisticated networks to fit regression models that yield deterministic results, by minimizing distortion measures (e.g., MSE, SSIM) w.r.t. the ground-truth. However, due to the ill-posedness of image restoration, they often capture the "mean" of the posterior distribution of the high-quality data given the degraded data, sampling results with excessive smoothness and compromised structural details, which may deviate from human perception (e.g., Restormer (Zamir et al., 2022), PromptIR (Potlapalli et al., 2023) in Figure 1 (b)). Differently, generation-based methods (Gao et al., 2023; Luo et al., 2023; Saharia et al., 2022b) regard the restoration task as a conditional distribution modeling problem, which generally results in visually appealing results. However, they often use the degraded image as a conditional input without including specific information about the degradation, which can result in outputs with remaining distortions and inaccurate structural details (e.g., IR-SDE (Luo et al., 2023) in Figure 1 (b)). Therefore, how to faithfully preserve the structures while minimizing the distortion remains challenging.

To tackle the challenge, we propose to model image restoration as an optimal transport (OT) problem, in which we introduce a novel transport residual as a degradation-specific cue for both the transport cost and transport map. Specifically, we present a two-pass Residual-Conditioned Optimal Transport (RCOT) approach to realize a degradation-aware and structure-preserving OT map, which applies to both unpaired and paired data settings. The key idea is to incorporate the degradation-specific knowledge (from the residual or its embedding) into the transport cost, and more importantly, into the transport map via a two-pass process, in which the transport residual is computed by the base model in the first pass and then encoded as a degradation-specific embedding to condition the second-pass restoration. This conditioning mechanism enables the transport map to adjust its behaviors for multiple restoration tasks and restore images with better structural content (Figure 1 (b)).

In summary, our contributions mainly include:

- We model image restoration as an OT problem, in which we introduce a Fourier residual-guided OT objective, allowing us to incorporate degradation-specific knowledge into the transport cost. We further deduce a minimax dual formulation for the OT model.
- We propose a two-pass RCOT approach, which conditions the transport map on the residual embedding. This conditioning mechanism dynamically injects degradation-specific knowledge from the residual embedding into the RCOT map, enhancing its capability to preserve the image structure.
- Extensive experiments on multiple tasks, e.g., image denoising, super-resolution, deraining, and dehazing on both synthetic and real-world datasets show the effectiveness of our method in terms of both distortion measures and perceptual quality. In particular, our method restores images with more faithful structural details, compared with existing approaches.

2. Background

Image restoration. Classic state-of-the-art restoration methods focus on minimizing distortion (e.g., MSE). They are primarily driven by efficient network architectures (Liang et al., 2021; Wang et al., 2022b; Zamir et al., 2021; 2022; Zhou et al., 2023) and model the restoration as regression models that produce deterministic results. However, they rely on large amounts of paired training data and are prone to "mean" results with compromised perceptual quality.

Differently, deep generative models have made much progress that can generate perceptually realistic images. The early mainstream methods are built on the conditional generative adversarial network (CGAN) (Mirza & Osindero, 2014), in which the restoration problem is treated as a conditional generation problem. RoCGAN (Chrysos et al., 2020) utilizes high-quality data for supervised training of restoration from degraded inputs. AmbientGAN (Bora et al., 2018) generates clean images from noisy input, assuming the degradation satisfies certain conditions. Recently, a promising avenue is the adoption of diffusion models. Saharia et al. (2022b) utilize the degraded images as conditions to train diffusion models. Kawar et al. (2021; 2022) and Chung et al. (2023) operate under the assumption that the degradation and its parameters are known at test time. Gao et al. (2023) propose a scale-adaptive condition on the LR image for high-fidelity super-resolution. Luo et al. (2023) propose a maximum likelihood-based loss function to train a mean-reverting score-based model. However, these methods more or less omit the degradation-specific knowledge, which limits their ability to produce faithful structures while minimizing the distortion.

Intuitively, the generation-based methods are seeking an efficient map T to transform the distribution of the low-quality image distribution $\mathbb{P}$ into the high-quality image distribution $\mathbb{Q}$. The key issue lies in the ambiguity that there may exist infinite maps satisfying this constraint. Prior knowledge is required to determine which is the optimal map. In this paper, we model image restoration from an OT perspective, seeking the most efficient map that transports $\mathbb{P}$ to $\mathbb{Q}$ with a minimal transport cost.

Optimal transport and its applications in restoration. OT problem seeks the optimal *transport map* (Monge Problem (Monge, 1781), a.k.a., MP) or *transport plan* (Kantorovich problem (Kantorovich, 1942), a.k.a., KP) to transform a distribution $\mathbb{P} \in \mathcal{P}(Y)$ to another distribution $\mathbb{Q} \in \mathcal{P}(X)$ with

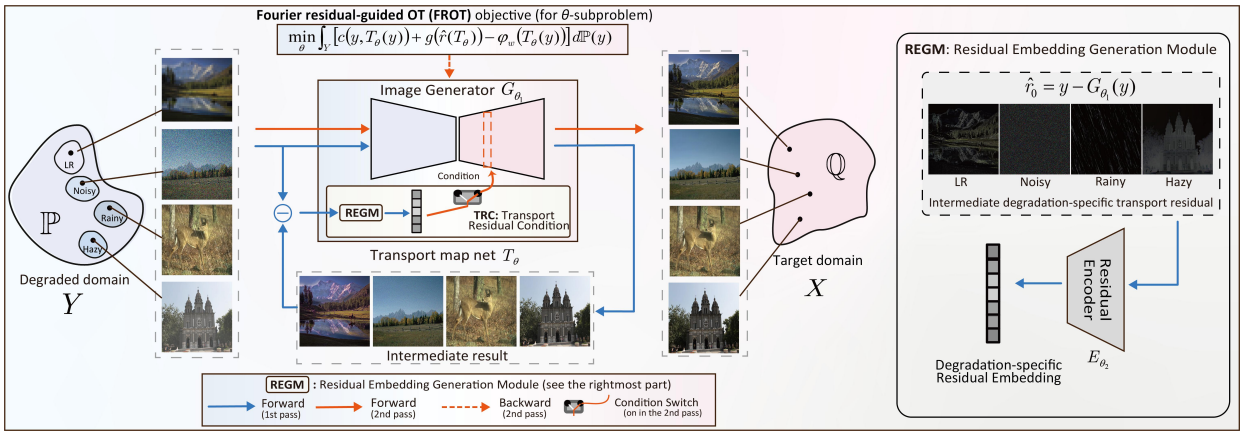


Figure 2: **Overview of the two-pass RCOT framework for structure-preserving restoration.** RCOT integrates the transport residual into the transport cost, leading to a FROT objective; and more crucially, into the transport map via a two-pass conditioning process. The first pass unconditionally generates an intermediate result along with the estimated residual $\hat{r}_0$. The second pass restores the refined result conditioned on the residual embedding $E_{\theta_2}(\hat{r}_0)$.

the minimal transport cost. $\mathcal{P}(X)$ and $\mathcal{P}(Y)$ respectively represent the sets of probability distributions on the Polish spaces X and Y. Formally, the Monge and Kantorovich problems can be stated as follows

$$C_{mp}(\mathbb{P},\mathbb{Q}) \triangleq \inf_{T_{\#}\mathbb{P}=\mathbb{Q}} \int_Y c\big(y, T(y)\big) d\mathbb{P}(y), \quad (1)$$

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

where $c : Y \times X \to \mathbb{R}_+$ measures the transport cost between two samples. In the MP (1), the map T^* attaining the infimum is called the *optimal transport map*, which is taken over all the transport maps $T : Y \to X$. The constraint $T_{\#}\mathbb{P} = \mathbb{Q}$ means that T pushes forward the probabilistic mass of $\mathbb{P}$ to $\mathbb{Q}$, where $T_{\#}$ is the push-forward operator. In the KP (2), the coupling π^* attaining the minimum is called the *optimal transport plan*, which is taken over all the transport plans π on $X \times Y$ whose marginals are $\mathbb{P}$ and $\mathbb{Q}$.

Recently, many attempts have been made to construct the translation/restoration as an OT problem. In this context, $\mathbb{P}, \mathbb{Q}$ represent the degraded distribution and target distribution respectively. Gu et al. (2023) propose to use optimal transport to guide the training of the conditional score-based diffusion model for super-resolution and translation. Korotin et al. (2023a;b) compute optimal transport maps and plans using neural networks under the duality framework and apply their method to unpaired image translation. They use ℓ_2 regularizer for the transport cost $c(x, y)$. Wang et al. (2022a) relax the transport constraint in Monge formulation with a Wasserstein-1 discrepancy penalty $W_1(\mathbb{Q}, T_{\#}\mathbb{P})$ between the target distribution $\mathbb{Q}$ and the push-forward distribution $T_{\#}\mathbb{P}$. Likewise, they empirically use ℓ_2 regularizer for $c(x, y)$ for the denoising task. These methods pioneer the way of modeling image translation/restoration problems as an OT problem, but their performances in multiple restoration problems are limited without proper prior knowledge about the correspondence between $\mathbb{P}$ and $\mathbb{Q}$.

Different from the aforementioned studies, our RCOT customizes image restoration as an OT problem, crafting the transport cost and map through the integration of transport residual. This innovation leads to a degradation-aware and structure-preserving transport map, i.e., the RCOT map.

3. Method

The key idea of RCOT is to introduce the transport residual as a degradation-specific cue for both the transport cost and transport map. We first model image restoration as an OT problem, exploiting the frequency knowledge of the residual, yielding the Fourier residual-guided OT (FROT) objective (section 3.1). Secondly, and most crucially, we integrate the degradation-specific knowledge from the residual embedding in the transport map via a two-pass process (section 3.2), in which the transport residual is computed by the base model in the first pass and then encoded as a residual embedding to condition the second-pass restoration (see Figure 2). In section 3.3, we present the learning algorithm for the two-pass RCOT map by adversarially training two neural networks to solve the minimax optimization problem in both unpaired and paired settings.

3.1. Residual-guided OT Formulation for Restoration

We first discuss formulizing image restoration as an OT problem, which can be applied to both unpaired and paired data. For the convenience of understanding, we elaborate our method for the unpaired data setting in sections 3.1 and 3.2. We will specify and extend to the paired case in section 3.3. We represent the domains of degraded images and target images by Y and X, whose distributions are $\mathbb{P}$ and $\mathbb{Q}$, respectively. Then the Kantorovich (Kantorovich,

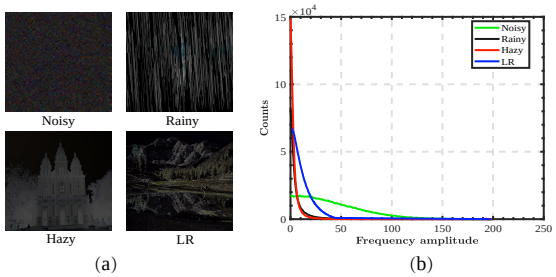


Figure 3: (a) Visual examples of the transport residual r. (b) Counts of the frequency amplitude of residuals for four types of degradation. For the draining, dehazing, and super-resolution tasks, the residuals are generally sparse in the frequency domain. The curves are averaged with 40 images.

1942) form of OT cost can be defined by (2). However, this formulation does not consider the prior knowledge of the degradation, limiting its applicability to multiple restoration tasks. To this end, we suggest the FROT objective by introducing a penalty term $g(\cdot)$ on the degradation domain gap (i.e., the transport residual $r = y - x$) in the transport cost, leading to $\tilde{c}(y, x) = c(x, y) + g(r)$. The FROT objective is then defined as

$$\text{FROT}(\mathbb{P}, \mathbb{Q}) \triangleq \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{X \times Y} \tilde{c}(y, x) d\pi(y, x). \quad (3)$$

Figure 3 presents the images of transport residuals for four types of degradation (noise, rain, haze, and low-resolution) along with their Fourier signal histograms. These histograms indicate that for degradations like rain, haze, and low-resolution, the residuals tend to be sparse in the frequency domain. In the case of noise, the histograms exhibit a smoother profile, resembling a Gaussian distribution. Based on the observation, we formalize the ℓ_1 regularizer on the Fourier residuals for the deraining, dehazing, and super-resolution tasks, i.e., $g(\cdot) = ||\mathcal{F}(\cdot)||_1$. For the denoising task, we specify the ℓ_2 regularizer on the Fourier residual. Our objective is to find the corresponding OT map T^* that attains the infimum of Monge's formulation (1) under cost $\tilde{c}$. The duality of (3) will lead to a more manageable approach. According to (Villani et al., 2009), (3) takes the following dual form:

$$\text{FROT}(\mathbb{P}, \mathbb{Q}) = \sup_{\varphi} \int_Y \varphi^{\tilde{c}}(y) d\mathbb{P}(y) + \int_X \varphi(x) d\mathbb{Q}(x),$$

where $\varphi^{\tilde{c}}(y) = \inf_{x \in X} [c(x, y) + g(r) - \varphi(x)]$ is the c-transform of φ. Replacing the optimization of the first term over target $x \in X$ with an equivalent optimization (Rockafellar interchange theorem (Rockafellar, 1976), Theorem 3A) over the map of interest $T : X \to Y$, we obtain the minimax reformulation of dual form:

$$\text{FROT}(\mathbb{P}, \mathbb{Q}) = \sup_{\varphi} \inf_{T} \Big\{ \mathcal{L}(T, \varphi) \triangleq \int_X \varphi(x) d\mathbb{Q}(x) + \int_Y [c(T(y), y) + g(\hat{r}(T)) - \varphi(T(y))] \, d\mathbb{P}(y) \Big\}, \quad (4)$$

where $\hat{r}(T) = y - T(y)$ represents the transport degradation domain gap (termed as transport residual). Now we show that tackling this minimax problem provides the OT map.

Proposition 3.1. *(Saddle points of FROT provide OT maps). For any optimal potential function $\varphi^* \in \arg\sup_{\varphi} \mathcal{L}(T, \varphi)$, it holds for the Monge OT map T^* that*

$$T^* \in \arg\min_{T} \mathcal{L}(T, \varphi^*). \quad (5)$$

The proof is given in Appendix A. Proposition 3.1 affirms the feasibility of solving the minimax problem (4) to acquire an optimal pair, constituting an OT map from $\mathbb{P}$ to $\mathbb{Q}$. For general $\mathbb{P}$ and $\mathbb{Q}$, given some optimal potential φ^*, the set $\arg\inf_T$ may encompass not only the OT map T^* but also other saddle points, which are capable of delivering decent performance as in experiments (section 4). To tackle the minimax problem (4), we can approximate the map T and potential φ with neural networks T_θ and φ_ω.

3.2. Two-pass Residual-Conditioned OT Map

After parameterizing the transport map with a neural network T_θ, to better preserve image structures for restoration under different degradations, we are motivated to incorporate degradation-specific and structural knowledge into the map T_θ. Intuitively, the transport residual may contain degradation-specific knowledge (e.g., degradation type and level) and structural information (as empirically shown in Appendix I). We then suggest utilizing the transport residual as an additional condition for the transport map T_θ. To achieve this, we develop a two-pass RCOT map (Figure 2) based on a transport residual condition (TRC) module, in which the transport residual is computed by the base model in the first pass and then encoded as a degradation-specific embedding to condition the second-pass restoration. Correspondingly, we introduce an image generator G_{θ_1} to generate the restored image, and a residual encoder E_{θ_2} to control the restoration with the residual embedding as a condition. G_{θ_1} and E_{θ_2} constitute the map T_θ.

Two-pass transport map. Given a degraded image y, the first pass unconditionally generates an intermediate restored result $G_{\theta_1}(y)$ via the image generator and calculates its corresponding intermediate transport residual $\hat{r}_0 = y - G_{\theta_1}(y)$. The second pass extracts the residual embedding $E_{\theta_2}(\hat{r}_0)$ via the residual encoder and uses it as a condition for the image generator, which then restores a refined result

$G_{\theta_1}(y|E_{\theta_2}(\hat{r}_0))$. This transport process is summarized as

$$\hat{r}_0 = y - G_{\theta_1}(y), \quad T_\theta(y) = G_{\theta_1}\left(y|E_{\theta_2}(\hat{r}_0)\right). \quad (6)$$

Transport residual condition (TRC). The TRC module is intended to compensate for the degradation-specific information that can be lost through the degradation. Specifically, the TRC module consists of two key components: a residual embedding generation module (REGM) and a condition integration module. In REGM, the estimated transport residual in the first pass is projected by the residual encoder to a residual embedding $E_{\theta_2}(\hat{r}_0)$, which is then utilized as a degradation-specific condition for the second-pass restoration. In the condition integration module, we employ the cross-stage feature fusion (Zamir et al., 2021) to integrate the features from the image generator G_{θ_1} with the degradation-specific embedding $E_{\theta_2}(\hat{r}_0)$ for structure-preserving restoration.

3.3. Overall Training

With the parameterization of T_θ and φ_ω, the optimization objective function of (4) can be written as

$$\mathcal{L}_{\mathrm{FROT}}(\omega,\theta) = \mathbb{E}_{x\sim\mathbb{Q}}\left[\varphi_\omega(x)\right] + \mathbb{E}_{y\sim\mathbb{P}}\left[c(T_\theta(y),y) + g(\hat{r}(T_\theta)) - \varphi_\omega(T_\theta(y))\right]. \quad (7)$$

For unpaired setting, we train the networks T_θ and φ_ω by respectively minimizing and maximizing $\mathcal{L}_{\mathrm{FROT}}(\omega,\theta)$, i.e., $\max_\omega \min_\theta \mathcal{L}_{\mathrm{FROT}}(\omega,\theta)$. This can be achieved by adversarially training T_θ and φ_ω, in which we estimate the expectation using mini-batch data in each training step.

(Partially) Paired Setting. In practice, most datasets may include a fraction of available paired samples. For this partially paired setting, we can leverage the paired samples to enforce $T_\theta(y)$ to approximate the target x for any pair $(y,x) \in P$ with a squared ℓ_2 loss (where P denotes the paired subsets of $X \times Y$):

$$\mathcal{L}_{\text{paired}}(\theta) = \frac{\gamma}{|P|}\sum_{(y,x)\in P}\|T_\theta(y) - x\|^2. \quad (8)$$

Consequently, the training objective for the partially paired setting is $\max_\omega \min_\theta\{\mathcal{L}_{\mathrm{FROT}}(\omega,\theta)+\mathcal{L}_{\text{paired}}(\theta)\}$. The overall algorithm is detailed in Algorithm 1 in the Appendix B.

4. Experiments

We evaluate the proposed RCOT on benchmark datasets on four representative image restoration tasks: image denoising, deraining, dehazing, and super-resolution (SR). In tables, the best and second-best quality metrics (PSNR/SSIM for measuring pixel/structure similarity, and LPIPS (Zhang et al., 2018a)/FID (Heusel et al., 2017) for perceptual deviation measuring) are **highlighted** and underlined. The **implementation details and selected compared methods** are introduced in Appendix C. Extra evaluations (e.g., parameter quantity, cross-dataset and preliminary multiple-in-one comparisons) are included in Appendix G.

4.1. Results

Gaussian Image Denoising. For Gaussian image denoising, we train the model on a combination of BSD400 (Arbelaez et al., 2010) and WED (Ma et al., 2016) datasets. The BSD400 dataset comprises 400 training images, while the WED dataset consists of 4,744 images. Gaussian noise with level $\sigma \in \{25, 50\}$ is separately added to generate noisy images for training. We evaluate RCOT on the Kodak24 (Franzen, 1999) and CBSD68 (Martin et al., 2001) datasets under noise levels $\sigma \in \{25, 50\}$. Table 1 reports the PSNR/SSIM/LPIPS/FID scores of the compared methods. The RCOT achieves the best performance under all metrics. Particularly, the LPIPS and FID values of RCOT's results are remarkably better than those of other methods. Figure 4 and Figure 13 in the Appendix H.3 display visual examples of $\sigma = 50$. MPRNet, Restormer, and PromptIR (Potlapalli et al., 2023; Zamir et al., 2021; 2022), directly minimizing the ℓ_1 loss, can perform well in distortion measures (PSNR and SSIM), but they produce results with smoothed structures. The generation-based methods IR-SDE (Luo et al., 2023) and OTUR (Wang et al., 2022a) seem to produce realistic structures but are limited in removing the heavy noise completely. As a comparison, our RCOT reproduces a clean and sharp image with faithful textures.

Image Deraining. We evaluate RCOT on both **synthetic** dataset Rain100L (Yang et al., 2017) and **real-world** dataset SPANet (Wang et al., 2019). For Rain100L (Yang et al., 2017), we train the model with 13,712 paired clean-rain images collected from multiple datasets (Fu et al., 2017; Li et al., 2016; Yang et al., 2017; Zhang & Patel, 2018; Zhang et al., 2019a). For real-world SPANet (Wang et al., 2019), it contains 27.5K paired rainy and rain-free images for training, and 1, 000 paired images for testing.

Table 2 reports the performance of the evaluated methods. The proposed RCOT achieves the best performance over both distortion measures and perceptual quality measures. The underlying reason should be that the RCOT benefits from the FROT objective and residual embedding that exploits the information of rain streaks. From the results shown in Figures 5 and 14, we can observe that results of NOT (Korotin et al., 2023b) are discolored. Restormer (Zamir et al., 2022) and PromptIR (Potlapalli et al., 2023) effectively remove the rain streaks, but the structural details are oversmoothed. IR-SDE (Luo et al., 2023) produces a restored result that still exhibits a slight presence of rain streaks. RCOT restores rain-free images with better structural content.

Table 1: Denoising results (PSNR/SSIM/LPIPS/FID) on Kodak24 (Franzen, 1999) and CBSD68 (Martin et al., 2001) datasets. RCOT achieves competitive qualitative performance. (*) indicates the method in an unpaired setting.

Method	Kodak24 (Franzen, 1999)		CBSD68 (Martin et al., 2001)	
	$\sigma = 25$	$\sigma = 50$	$\sigma = 25$	$\sigma = 50$
NOT* (Korotin et al., 2023b)	29.13/0.786/0.131/70.17	27.12/0.725/0.227/104.5	29.76/0.802/0.119/73.68	26.82/0.723/0.212/113.3
OTUR* (Wang et al., 2022a)	31.05/0.848/0.104/55.74	28.03/0.744/0.178/83.27	30.27/0.833/0.095/67.74	27.36/0.733/0.169/93.17
MPRNet (Zamir et al., 2021)	31.96/0.868/0.112/43.98	28.36/0.785/0.185/73.26	30.89/0.880/0.103/59.23	27.56/0.779/0.163/86.42
Restormer(Zamir et al., 2022)	32.13/0.880/0.097/40.22	29.25/0.799/0.156/64.26	31.20/0.887/0.090/55.28	27.90/0.794/0.149/66.12
IR-SDE (Luo et al., 2023)	31.40/0.842/0.080/45.56	28.03/0.721/0.134/83.66	30.46/0.856/0.075/57.30	26.98/0.737/0.138/96.40
RCD(Zhang et al., 2023)	32.18/0.880/0.089/39.85	29.22/0.795/0.147/63.26	31.28/0.886/0.089/50.96	28.01/0.796/0.149/69.78
PromptIR (Potlapalli et al., 2023)	32.25/0.883/0.091/30.91	29.19/0.799/0.154/60.42	31.31/0.888/0.085/45.45	28.03/0.797/0.143/63.98
RCOT*	31.84/0.860/0.085/35.28	28.64/0.792/0.152/63.27	30.77/0.849/0.079/47.21	27.69/0.775/0.145/73.43
RCOT	**32.64/0.885/0.070/19.33**	**29.53/0.828/0.121/33.13**	**31.52/0.889/0.062/34.24**	**28.25/0.799/0.118/56.60**

14.60/0.256 Noisy | 26.60/0.822 NOT | 29.29/0.921 OTUR | 29.77/0.940 MPRNet | **31.32**/0.955 Restormer | 29.52/0.921 IR-SDE | 30.98/0.951 PromptIR | 31.26/**0.961** RCOT

Figure 4: Visual comparison of denoising on Kodak24 (Franzen, 1999) with $\sigma = 50$. Our RCOT produces a noise-free image with clear textural details.

Table 2: Deraining results on synetheic dataset Rain100L (Fan et al., 2019) and real-world dataset SPANet (Wang et al., 2019). $(*)$ indicates the method in an unpaired setting.

Method	Synthetic Rain100L (Fan et al., 2019)				Real-world SPANet (Wang et al., 2019)			
	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)
NOT* (Korotin et al., 2023b)	29.29	0.911	0.030	54.12	32.55	0.901	0.028	53.28
OTUR* (Wang et al., 2022a)	33.71	0.954	0.027	36.64	39.23	0.961	0.017	29.64
MPRNet (Zamir et al., 2021)	34.95	0.964	0.039	21.61	39.52	0.967	0.021	28.13
Restormer (Zamir et al., 2022)	36.74	0.978	0.026	13.29	41.39	0.981	0.013	19.67
SFNet (Cui et al., 2023b)	36.56	0.974	0.023	13.12	41.02	0.980	0.015	21.52
IR-SDE (Luo et al., 2023)	36.94	0.978	0.014	9.52	42.56	0.987	0.009	16.25
PromptIR (Potlapalli et al., 2023)	37.09	0.979	0.025	10.21	39.17	0.965	0.015	28.93
RCOT*	36.22	0.972	0.019	12.59	41.05	0.977	0.014	18.93
RCOT	**37.27**	**0.980**	**0.015**	**7.97**	**43.77**	**0.993**	**0.008**	**9.52**

Image Dehazing. We evaluate RCOT on the **synthetic** SOTS (Li et al., 2018a) dataset, which contains 72,135 images for training and 500 images for testing, and **real hazy** O-HAZE (Ancuti et al., 2018) dataset, which contains 45 paired hazy and haze-free images, which are collected at the same scene under the same illumination conditions. 40 images are used for training and the other 5 images are used for testing.

Table 3 reports the qualitative results. The proposed RCOT achieves the best performance overall. Notably, the unpaired RCOT also achieves a promising performance, especially in terms of the LPIPS and FID values. From the results shown in Figure 6 and 12, we can observe that NOT (Korotin et al., 2023b) and IR-SDE (Luo et al., 2023) produce sharp results but are not faithful in color and removing haze. Restormer, MPRNet, and PromptIR can remove the haze but produce results with distorted color. As a comparison, the RCOT can remove the haze while faithfully preserving the color.

Image Super-resolution. For image super-resolution, we evaluate the performance on the challenging DIV2K (Agustsson & Timofte, 2017) dataset, which consists of 800 (4x) LR and HR image pairs for training and 100 pairs for testing.

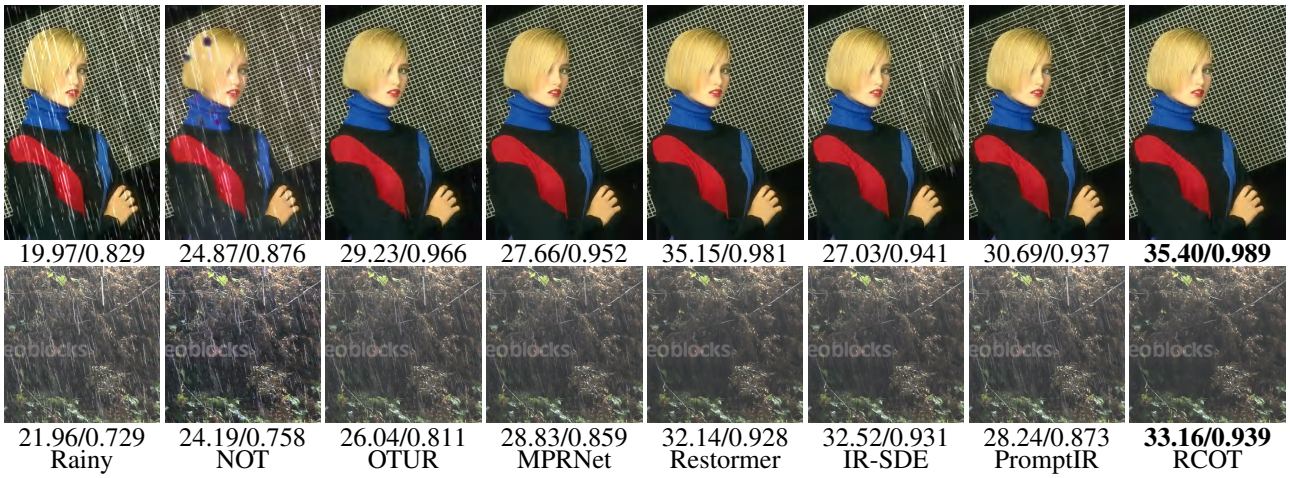


Figure 5: Visual comparison of deraining results on synthetic Rain100L (Fan et al., 2019) (row 1) and real-world SPANet (Wang et al., 2019) (row 2). The RCOT reproduces rain-free images with realistic structural details.

Table 3: Dehazing results on synetheic dataset SOTS (Li et al., 2018a) and real-world dataset O-HAZE (Ancuti et al., 2018). ($*$) indicates the method in an unpaired setting.

Method	Synthetic SOTS (outside) (Fan et al., 2019)				Real-world O-HAZE (Ancuti et al., 2018)			
	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)
NOT* (Korotin et al., 2023b)	24.21	0.900	0.046	21.35	15.13	0.673	0.271	258.68
OTUR* (Wang et al., 2022a)	26.36	0.953	0.024	18.96	16.43	0.719	0.209	248.74
MPRNet (Zamir et al., 2021)	28.31	0.954	0.029	17.79	21.55	0.778	0.256	223.86
Restormer (Zamir et al., 2022)	30.87	0.969	0.026	13.29	25.20	0.804	0.221	198.07
Dehazeformer (Song et al., 2023)	31.45	**0.978**	0.021	15.54	25.56	0.812	0.209	199.35
IR-SDE (Luo et al., 2023)	30.55	0.968	0.018	12.76	22.13	0.776	0.160	179.23
PromptIR (Potlapalli et al., 2023)	31.31	0.973	0.021	16.28	25.27	0.813	0.216	217.66
RCOT*	30.34	0.965	0.020	12.97	21.01	0.773	0.186	203.72
RCOT	**31.66**	0.976	**0.015**	**10.21**	**27.16**	**0.839**	**0.145**	**169.38**

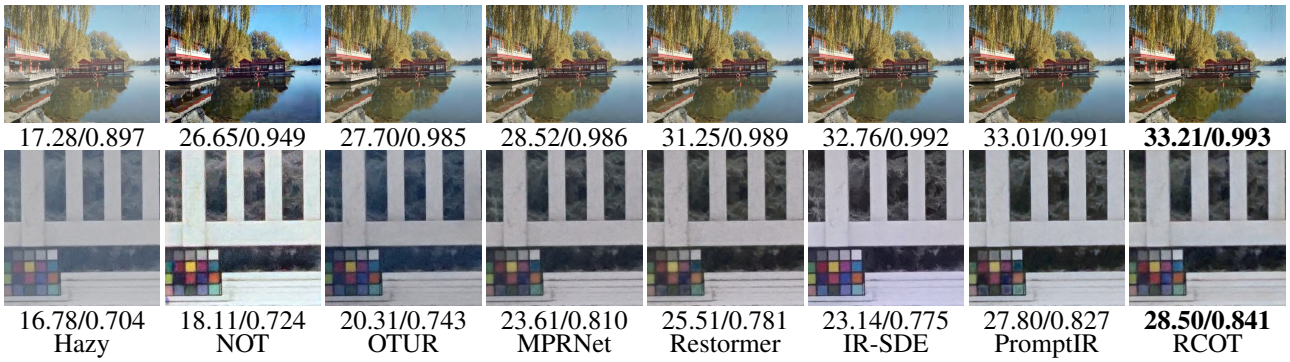
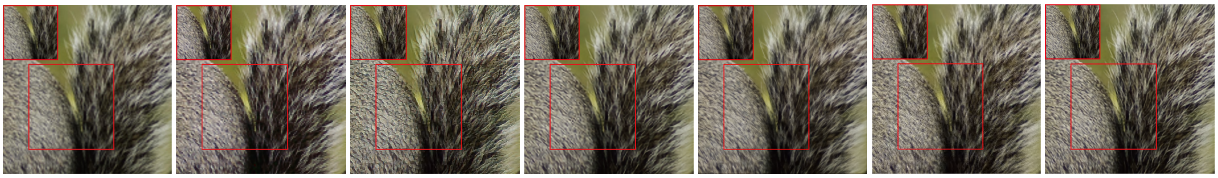


Figure 6: Visual comparison of dehazing results on synethtic SOTS (Li et al., 2018a) (row 1) and real-world O-HAZE (Ancuti et al., 2018) (row 2). RCOT produces haze-free images with faithful color.

The quantitative results in Table 4 show that the proposed RCOT outperforms the most recent generative methods. The qualitative gains of both distortion measures and perceptual quality are significant. Notably, even in unpaired setting the RCOT achieves the third-best performance in terms of the FID score. The visual examples in Figure 7 and 15 show that the DPMs-based methods (Gao et al., 2023; Luo et al., 2023) produce unnatural structures. The OT-based methods (Korotin et al., 2023b; Wang et al., 2022a) exhibit severe distortion. As a comparison, our RCOT restores a sharp image with more realistic details.

Table 4: SR results on DIV2K (Agustsson & Timofte, 2017). (*) indicates the method in an unpaired setting.

Method	Bicubic	NOT* (Korotin et al., 2023b)	OTUR* (Wang et al., 2022a)	Restomer (Zamir et al., 2022)	IR-SDE (Luo et al., 2023)	IDM (Gao et al., 2023)	LINF (Yao et al., 2023)	RCOT*	RCOT
PSNR↑	26.70	25.73	24.88	27.90	26.89	27.59	27.33	26.78	**28.41**
SSIM↑	0.771	0.718	0.679	0.796	0.775	0.785	0.769	0.758	**0.804**
LPIPS↓	0.186	0.136	0.128	0.136	0.118	0.121	0.116	0.125	**0.114**
FID↓	15.88	19.27	24.65	6.14	8.56	5.73	3.26	4.51	**1.22**



25.22/0.782 Bicubic | 24.76/0.768 NOT | 24.34/0.744 OTUR | 25.54/0.772 Restormer | 25.78/0.799 IDM | 25.36/0.785 IR-SDE | **26.12/0.814** RCOT

Figure 7: Visual comparison of 4x SR on the DIV2K (Agustsson & Timofte, 2017) dataset. Please zoom in for better visualization. RCOT produces the sharpest HR image with more realistic structures.

Table 5: Results w/ and w/o TRC. TRC produces a significant boost of PSNR/SSIM/FID values.

Method	SR on DIV2K	Deraining on Rain100L	Denoising on Kodak24	Average
w/o TRC	27.24/0.785/6.72	34.20/0.955/28.63	28.45/0.790/64.20	29.97/0.843/32.47
w/ TRC	**28.41/0.804/1.22**	**37.27/0.978/8.67**	**29.53/0.828/33.13**	**31.72/0.870/14.34**

Table 6: Effect of the loss components. The model is trained under different losses for denoising and tested on CBSD68 (Martin et al., 2001) with noise level $\sigma = 50$. FROT shows decent performance, working alone and with the ℓ_2 loss.

Loss	PSNR ↑	SSIM ↑	FID ↓
$\mathcal{L}_{\text{FROT}}$	27.60	0.772	69.43
supervised ℓ_2	27.69	0.779	78.69
$\mathcal{L}_{\text{FROT}}$+ ℓ_2	**28.25**	**0.799**	**56.60**

Degraded | w/o TRC | w/ TRC | Ground truth

Figure 8: Visual results to show the importance of TRC. The restored images with TRC in three tasks (SR, Deraining, Denoising) contain better structures.

4.2. Ablation Study

We primarily showcase the ablation studies on the TRC module and training loss components in the main body. **Four additional ablation experierments** are reported in Appendix D.

Importance of the TRC module. We validate the importance of the transport residual condition module on three tasks (SR on DIV2K (Agustsson & Timofte, 2017), Deraining on Rain100L (Fan et al., 2019), and Denoising on Kodak24 (Franzen, 1999)). We report the quantitative results with and without TRC in Table 5 and display visual comparisons in Figure 8. The proposed TRC module yields an average gain of 1.75 dB of PSNR value over the basic model. With the TRC, the model restores images with better structures.

Effect of the loss components. To investigate the effect of different loss components (FROT and supervised ℓ_2 loss), we train the transport map T_θ under different loss functions and evaluate its denoising performance. Table 6 reports the results on CBSD68 (Martin et al., 2001) with noise level $\sigma = 50$. When training under $\mathcal{L}_{\text{FROT}}$, the performance is close to that under supervised ℓ_2 loss. As a comparison, the training objective for the paired case, which integrates $\mathcal{L}_{\text{FROT}}$ and supervised ℓ_2, provides a significant gain to our model. These results validate the effectiveness of the proposed FROT under unpaired setting and show the importance of integrating supervised ℓ_2 loss with $\mathcal{L}_{\text{FROT}}$when the targets are available.

5. Conclusion

This paper proposed a novel Residual-Conditioned Optimal Transport (RCOT) approach to preserve perceptual structures while effectively removing the distortion, which treated image restoration as an OT problem and introduced the unique transport residual as a degradation-specific cue for both the transport cost and transport map. We first customized the Fourier residual-guided OT objective that exploits Fourier statistics of the degradation domain gap (represented by the residual). Based on FROT, we deduced a minimax problem that can be tackled to provide OT maps. Then we developed a two-pass RCOT map, in which the second pass generates refined results conditioned on the intermediate transport residual computed by the first pass for structure preservation. Extensive experiments demonstrated the effectiveness of RCOT for achieving competitive restoration performance, especially in terms of the ability to preserve structural content. In the future, we are also interested in developing an all-in-one RCOT framework for restoration by properly integrating domain knowledge.

Limitation and Future Research

In this work, based on the empirical observation of the degradation domain gap in the frequency domain, we utilize handcrafted priors in the frequency domain to characterize the Fourier residual, which has been shown to be effective in the experiments. However, it may not be optimal for different degradations. Automatic, adaptive, and optimal learning for it will be explored in our future work. We hope the method can be applied to all-in-one image restoration.

Impact Statement

This work aims to advance the field of machine learning with applications in image restoration. It may be valuable to the research of optimal transport designed as a deep generative model working with low-quality data and has no ethical concerns as far as we know.

Acknowledgement

This work was supported by National Key R&D Program 2021YFA1003002, and NSFC (12125104, U20B2075, 623B2084, 12326615).

References

Agustsson, E. and Timofte, R. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 126–135, 2017.

Ancuti, C., Ancuti, C., Timofte, R., and De Vleeschouwer, C. O-haze: a dehazing benchmark with real hazy and haze-free outdoor images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 754–762, 2018.

Arbelaez, P., Maire, M., Fowlkes, C., and Malik, J. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(5):898–916, 2010.

Bora, A., Price, E., and Dimakis, A. G. Ambientgan: Generative models from lossy measurements. In *International Conference on Machine Learning (ICML)*, 2018.

Chen, X., Li, H., Li, M., and Pan, J. Learning a sparse transformer network for effective image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5896–5905, 2023.

Choi, J., Park, Y., and Kang, M. Restoration based generative models. In *International Conference on Machine Learning (ICML)*, pp. 5787–5816, 2023.

Chrysos, G. G., Kossaifi, J., and Zafeiriou, S. Rocgan: Robust conditional gan. *International Journal of Computer Vision (IJCV)*, 128:2665–2683, 2020.

Chu, X., Chen, L., and Yu, W. Nafssr: Stereo image super-resolution using nafnet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1239–1248, June 2022.

Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye, J. C. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023.

Cui, Y., Ren, W., Yang, S., Cao, X., and Knoll, A. Irnext: Rethinking convolutional network design for image restoration. In *International Conference on Machine Learning (ICML)*, 2023a.

Cui, Y., Tao, Y., Bing, Z., Ren, W., Gao, X., Cao, X., Huang, K., and Knoll, A. Selective frequency network for image restoration. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023b.

Fan, Q., Chen, D., Yuan, L., Hua, G., Yu, N., and Chen, B. A general decoupled learning framework for parameterized image operators. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(1):33–47, 2019.

Franzen, R. Kodak lossless true color image suite. *source: http://r0k. us/graphics/kodak*, 4(2):9, 1999.

Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., and Paisley, J. Removing rain from single images via a deep detail network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3855–3863, 2017.

Gao, S., Liu, X., Zeng, B., Xu, S., Li, Y., Luo, X., Liu, J., Zhen, X., and Zhang, B. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10021–10030, 2023.

Gu, X., Yang, L., Sun, J., and Xu, Z. Optimal transport-guided conditional score-based diffusion model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

He, K., Sun, J., and Tang, X. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(12):2341–2353, 2010.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

Kantorovich, L. V. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pp. 199–201, 1942.

Kawar, B., Vaksman, G., and Elad, M. Snips: Solving noisy inverse problems stochastically. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:21757–21769, 2021.

Kawar, B., Elad, M., Ermon, S., and Song, J. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Korotin, A., Selikhanovych, D., and Burnaev, E. Kernel neural optimal transport. In *International Conference on Learning Representations (ICLR)*, 2023a. URL https://openreview.net/forum?id=Zuc_MHtUma4.

Korotin, A., Selikhanovych, D., and Burnaev, E. Neural optimal transport. In *International Conference on Learning Representations (ICLR)*, 2023b. URL https://openreview.net/forum?id=d8CBRlWNkqH.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4681–4690, 2017.

Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., and Wang, Z. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing (TIP)*, 28 (1):492–505, 2018a.

Li, R., Pan, J., Li, Z., and Tang, J. Single image dehazing via conditional generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8202–8211, 2018b.

Li, Y., Tan, R. T., Guo, X., Lu, J., and Brown, M. S. Rain streak removal using layer priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2736–2744, 2016.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., and Timofte, R. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1833–1844, 2021.

Luo, Z., Gustafsson, F. K., Zhao, Z., Sjölund, J., and Schön, T. B. Image restoration with mean-reverting stochastic differential equations. *International Conference on Machine Learning (ICML)*, 2023.

Ma, K., Duanmu, Z., Wu, Q., Wang, Z., Yong, H., Li, H., and Zhang, L. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing (TIP)*, 26(2):1004–1016, 2016.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pp. 416–423, 2001.

Mirza, M. and Osindero, S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Monge, G. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pp. 666–704, 1781.

Murata, N., Saito, K., Lai, C.-H., Takida, Y., Uesaka, T., Mitsufuji, Y., and Ermon, S. GibbsDDRM: A partially collapsed gibbs sampler for solving blind inverse problems with denoising diffusion restoration. In *International Conference on Machine Learning (ICML)*, 2023.

Pan, J., Dong, J., Liu, Y., Zhang, J., Ren, J., Tang, J., Tai, Y.-W., and Yang, M.-H. Physics-based generative adversarial models for image restoration and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(7):2449–2462, 2020.

Pan, X., Zhan, X., Dai, B., Lin, D., Loy, C. C., and Luo, P. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(11): 7474–7489, 2021.

Potlapalli, V., Zamir, S. W., Khan, S., and Khan, F. S. Promptir: Prompting for all-in-one blind image restoration. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ren, W., Liu, S., Zhang, H., Pan, J., Cao, X., and Yang, M. H. Single image dehazing via multi-scale convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 154–169, 2016.

Rockafellar, R. Integral functionals, normal integrands and measurable selections. *Nonlinear Operators and the Calculus of Variations*, pp. 157–207, 1976.

Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., and Norouzi, M. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10, 2022a.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(4):4713–4726, 2022b.

Song, Y., He, Z., Qian, H., and Du, X. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing (TIP)*, 32:1927–1941, 2023.

Sun, J., Xu, Z., and Shum, H.-Y. Image super-resolution using gradient profile prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8. IEEE, 2008.

Tang, X., Zhao, X., Liu, J., Wang, J., Miao, Y., and Zeng, T. Uncertainty-aware unsupervised image deblurring with deep priors guided by domain knowledge. *arXiv e-prints*, pp. arXiv–2210, 2022.

Tang, X., Zhao, X., Liu, J., Wang, J., Miao, Y., and Zeng, T. Uncertainty-aware unsupervised image deblurring with deep residual prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9883–9892, 2023.

Villani, C. et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Wang, C., Pan, J., Wang, W., Dong, J., Wang, M., Ju, Y., and Chen, J. Promptrestorer: A prompting image restoration method with degradation perception. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Wang, T., Yang, X., Xu, K., Chen, S., Zhang, Q., and Lau, R. W. Spatial attentive single-image deraining with a high quality real rain dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Wang, W., Wen, F., Yan, Z., and Liu, P. Optimal transport for unsupervised denoising learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(2):2104–2118, 2022a.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., and Change Loy, C. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., and Li, H. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17683–17693, 2022b.

Wei, Y., Zhang, Z., Wang, Y., Xu, M., Yang, Y., Yan, S., and Wang, M. Deraincyclegan: Rain attentive cyclegan for single image deraining and rainmaking. *IEEE Transactions on Image Processing*, 30:4788–4801, 2021.

Yang, W., Tan, R. T., Feng, J., Liu, J., Guo, Z., and Yan, S. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1357–1366, 2017.

Yao, J.-E., Tsao, L.-Y., Lo, Y.-C., Tseng, R., Chang, C.-C., and Lee, C.-Y. Local implicit normalizing flow for arbitrary-scale image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M.-H., and Shao, L. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., and Yang, M.-H. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Zhang, H. and Patel, V. M. Density-aware single image deraining using a multi-stream dense network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 695–704, 2018.

Zhang, H., Sindagi, V., and Patel, V. M. Image de-raining using a conditional generative adversarial network. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 30(11):3943–3956, 2019a.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018a.

Zhang, W., Liu, Y., Dong, C., and Qiao, Y. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3096–3105, 2019b.

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., and Fu, Y. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 286–301, 2018b.

Zhang, Z., Jiang, Y., Shao, W., Wang, X., Luo, P., Lin, K., and Gu, J. Real-time controllable denoising for image and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14028–14038, June 2023.

Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Eric, I., Chang, C., and Xu, Y. Large scale image completion via co-modulated generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Zheng, H., Pang, T., and Ji, H. Unsupervised deep video denoising with untrained network. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, pp. 3651–3659, 2023.

Zhou, M., Huang, J., Guo, C.-L., and Li, C. Fourmer: an efficient global modeling paradigm for image restoration. In *International Conference on Machine Learning (ICML)*, 2023.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2223–2232, 2017.

Zhu, Y., Zhang, K., Liang, J., Cao, J., Wen, B., Timofte, R., and Van Gool, L. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1219–1229, 2023.

Appendix

A. Proof

Proposition A.1. *(Saddle points provide OT maps). For any optimal potential function $\varphi^* \in \arg\sup_\varphi \mathcal{L}(T, \varphi)$, it holds for the OT map T^* (i.e., the transport map attaining the infimum of Monge's formulation (1) under $\tilde{c}$) that*

$$T^* \in \arg\min_T \mathcal{L}(T, \varphi^*) \tag{9}$$

Proof. First, we give Monge's primal form of the FROT objective

$$\mathbb{M}\text{-FROT}(\mathbb{P}, \mathbb{Q}) = \inf_{T\#\mathbb{P}=\mathbb{Q}} \int_Y \left[c(T(y), y) + g(\hat{r}(T))\right] d\mathbb{P}(y). \tag{10}$$

For the OT map T^*, it holds $\hat{r}(T^*) = y - T^*(y) = r$, then we find

$$\mathcal{L}(T^*, \varphi^*) = \int_X \varphi^*(x) d\mathbb{Q}(x) + \int_Y \left[c(T^*(y), y) + g(r) - \varphi^*(T^*(y))\right] d\mathbb{P}(y). \tag{11}$$

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

$$\int_Y \varphi^*(T^*(y)) d\mathbb{P}(y) = \int_X \varphi^*(x) d\mathbb{Q}(x).$$

Substituting this equality into (11), we obtain

$$\mathcal{L}(T^*, \varphi^*) = \int_Y \left[c(T^*(y), y) + g(r)\right] d\mathbb{P}(y) = \inf_{T\#\mathbb{P}=\mathbb{Q}} \int_Y \left[c(T(y), y) + g(\hat{r}(T))\right] d\mathbb{P}(y) = \mathbb{M}\text{-FROT}(\mathbb{P}, \mathbb{Q}).$$

□

B. Algorithm

Algorithm 1 RCOT Solver to compute the OT map.

Input: degraded dataset $Y \sim \mathbb{P}$; high-quality dataset $X \sim \mathbb{Q}$; transport network: T_θ;
potential network: φ_w; the number of iterations of θ per iteration of ω: n_T;

while θ has not converged **do**
  Sample batches $\mathcal{Y}$ from Y, $\mathcal{X}$ from X;
  $\mathcal{L}(\omega) \leftarrow \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} \varphi_w(T_\theta(y)) - \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} \varphi_w(x)$;
  Update w by using $\frac{\partial \mathcal{L}_\varphi}{\partial w}$;
  for $t = 0, \cdots, n_T$ **do**
    Compute $T_\theta(y)$ via (6);
    if there exist paired samples **then**
      $\mathcal{L}(\theta) \leftarrow \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} \left[c(y, T_\theta(y)) + g(\hat{r}(T_\theta)) - \varphi_w(T_\theta(y))\right] + \frac{\gamma}{|P|}\sum_{(y,x)\in P} \|T_\theta(y) - x\|^2$
    else
      $\mathcal{L}(\theta) \leftarrow \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} \left[c(y, T_\theta(y)) + g(\hat{r}(T_\theta)) - \varphi_w(T_\theta(y))\right]$;
    end if
    Update θ by using $\frac{\partial \mathcal{L}(\theta)}{\partial \theta}$;
  end for
end while

C. Implementation Details and Baselines

This section introduces the compared methods and detailed settings in our experiments.

Implementation details. We train separate models for different tasks using the RMSProp optimizer with a learning rate of 1×10^{-4} for the transport network T_θ and 0.5×10^{-4} for the potential network φ_w. The inner iteration number n_T is set to be 1. The learning rate is decayed by a factor of 10 after 100 epochs. In all experiments, the transport network T_θ uses the backbone in MPRNet (Zamir et al., 2021). The residual encoder consists of two CNN down-sampling layers with residual channel attention block (RCAB) (Zhang et al., 2018b). In the FROT objective, $c(x, y)$ is suggested as ℓ_2-norm. During training, we crop patches of size 256x256 as input and use a batch size of 4. All the experiments are conducted on the Pytorch framework with an NVIDIA 4090 GPU. For super-resolution, there is an extra preprocessing step. The LR images undergo bicubic rescaling to match the dimensions of their respective high-resolution counterparts. The source code will be released after the possible publication of our work.

For the unpaired setting, although datasets that contain paired labels are utilized for training, we randomly shuffle the target x and degraded input y to ensure the loss are isolated from paired information, which is a common strategy (Korotin et al., 2023a;b; Wang et al., 2022a) of unpaired training for restoration problems.

Representative Compared methods. For image denoising, deraining, and dehazing, we choose four most recent representative methods with state-of-the-art performance and two recent OT-based generative methods as competitors. They include MPRNet (Zamir et al., 2021), Restormer (Zamir et al., 2022), IR-SDE (Luo et al., 2023), OTUR (Wang et al., 2022a), NOT (Korotin et al., 2023b), and PromptIR (Potlapalli et al., 2023).

MPRNet (Zamir et al., 2021) and Restormer (Zamir et al., 2022) respectively specialize two backbones for restoration. Notably, Restormer (Zamir et al., 2022) designs an efficient Transformer model by making several key designs in the building blocks (multi-head attention and feed-forward network) such that it can capture long-range pixel interactions, while still remaining applicable to large images. It achieves state-of-the-art results on several image restoration tasks.

PromptIR (Potlapalli et al., 2023) utilizes learnable task prompts to incorporate the degradation-specific knowledge in the Transformer model to achieve task-aware restoration. We train the PromptIR model **in single task setting**.

For image super-resolution, we choose the most recent state-of-the-art generative diffusion-based (Gao et al., 2023; Luo et al., 2023), sophisticated transformer-based method Restormer (Zamir et al., 2022), and OT-based methods (Korotin et al., 2023b; Wang et al., 2022a) as competitors for comparison.

D. Additional Ablation Studies

D.1. Versatility and generalizability beyond specific network designs

The RCOT can be easily applied to different network architectures or frameworks. The two-pass TRC module is a plug-in module (Figure 2), allowing us to use any architecture as a base model to generate the restored result, and then use this result to calculate the residual. We have now included a comparison on Rain100L dataset (Fan et al., 2019) between the MPRNet (Zamir et al., 2021), NAFNet (Chu et al., 2022), Restormer (Zamir et al., 2022) methods and the corresponding methods with the proposed TRC module.

Table 7: The influence of the transport residual condition (TRC) module being integrated into different network backbones. The metrics are presented as PSNR (↑)/SSIM (↑)/LPIPS (↓)/FID (↓) values.

Method	MPRNet (Zamir et al., 2021)	NAFNet (Chu et al., 2022)	Restormer (Zamir et al., 2022)
w/o TRC	34.95/0.964/0.0387/21.61	35.58/0.969/0.0355/18.35	36.74/0.978/0.0226/13.29
w/ TRC	**36.78/0.972/0.0145/12.58**	**37.10/0.976/0.0134/11.86**	**38.22/0.984/0.0102/7.01**

The results in Table 7 show that the TRC module brings a meaningful boost to three SOTA network architectures, which validates its versatility and generalizability beyond specific network designs.

D.2. Effect of different conditions

In this paper, we treat the transport residual as a degradation-specific condition by encoding its embedding to adaptively enhance the representation of the restoration. To better demonstrate the benefits, we also try conditioning directly on the output restored image $\hat{x}_0$ of the base model. The average qualitative results on (SR, Deraining, and Denoising) are reported

in Table 8 which sustains our claim.

Table 8: The qualitative results of the model with different conditions for the second pass restoration.

Method	w/o condition	conditioned on $\hat{x}_0$	conditioned on residual $\hat{r}_0$
PSNR	29.97	30.39	**31.72**
SSIM	0.843	0.848	**0.870**
FID	32.47	26.52	**14.34**

D.3. Effect of the Fourier residual penalty in FROT objective

We investigate the effect of the proposed FROT cost on four restoration tasks (Denoising on Kodak24 (Franzen, 1999) with noise level $\sigma = 50$, SR on DIV2K (Agustsson & Timofte, 2017), Deraining on Rain100L (Fan et al., 2019), and Dehazing on SOTS (Li et al., 2018a)). In Table 9, we compare the performance of the models trained under the unpaired FROT cost (w/ $g(\hat{r})$) and regular OT cost (w/o $g(\hat{r})$). The results show that the Fourier residual penalty, integrating degradation-specific knowledge into the transport cost, brings meaningful gains.

Table 9: Results of the OT cost w/ $g(\hat{r})$ (FROT) and w/o $g(\hat{r})$. The models are trained by only using unpaired FROT, in which ℓ_2-norm is chosen as $c(x, y)$ and $g(\hat{r}) = \|\mathcal{F}(\hat{r})\|_1$ for SR, draining, and dehazing while utilizing $g(\hat{r}) = \|\mathcal{F}(\hat{r})\|_2$ for denoising.

Task	Transport cost	PSNR ↑	SSIM ↓	FID ↓
Denoising	w/o $g(\hat{r})$	28.37	0.780	69.12
	w/ $g(\hat{r})$	**28.64**	**0.792**	**63.27**
SR	w/o $g(\hat{r})$	25.96	0.723	11.41
	w/ $g(\hat{r})$	**26.78**	**0.758**	**4.51**
Deraining	w/o $g(\hat{r})$	35.69	0.945	20.13
	w/ $g(\hat{r})$	**36.22**	**0.972**	**12.59**
Dehazing	w/o $g(\hat{r})$	29.72	0.953	15.56
	w/ $g(\hat{r})$	**30.34**	**0.965**	**12.57**

D.4. Effect of different regularizers for the Fourier residual

We conduct an extra ablation study on the regularizer for the Fourier residual. The results are reported in Table 10.

Table 10: Comparison of different regularizers for the Fourier residual.

Regularizer	$L_{0.5}$	ℓ_2	L_1
SR	**26.82**/0.756	26.49/0.747	26.78/**0.758**
Denoising	28.52/0.782	**28.79/0.795**	28.64/0.792
Deraining	36.01/0.968	35.80/0.948	**36.22/0.972**
Dehazing	30.18/0.958	29.80/0.954	**30.34/0.965**

For the denoising task, the ℓ_2 regularizer is more suitable for the Fourier residual since the Gaussian noise is basically from a Gaussian distribution, which is equivalent to the ℓ_2 regularization. The ℓ_1 sparsity regularizer applies to other degradations.

E. Sensitivity to the Supervised Trade-off Parameter γ

We test the sensitivity of RCOT to the trade-off parameter γ on SR tasks. Figure 9 shows that when $\gamma = 1 \times 10^4$, RCOT's performance might approach its limit. Interestingly, the distortion measures, i.e., PSNR and SSIM, go steady when γ

increases after 1×10^3. However, the FID score seems to turn bad if γ keeps increasing. Based on this result, we fix $\gamma = 10^4$ in our experiments.

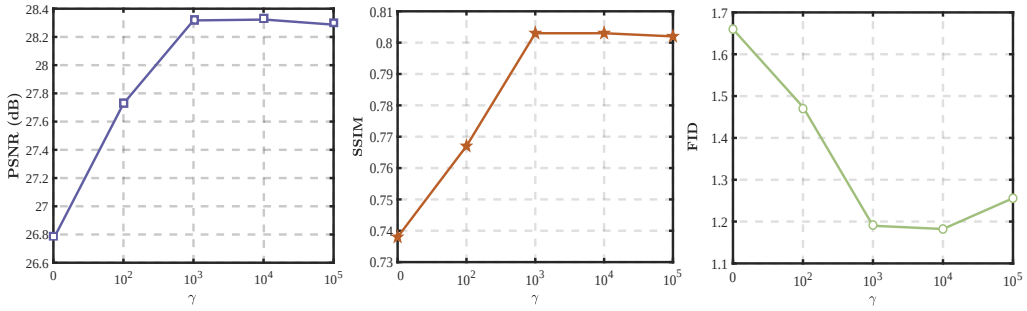


Figure 9: Test of the sensitivity of RCOT to the parameter γ on the DIV2K (Agustsson & Timofte, 2017) dataset for SR task.

F. Training Cost Curves for Three tasks

In Figure 10, we display the cost curves over three tasks (i.e., SR, Deraining, and Denoising) of T_θ and φ_w in the training process. The T_θ cost curve is normalized to $[0, 1]$. φ_w cost is scaled to $[0, 1]$ and then take the negative.

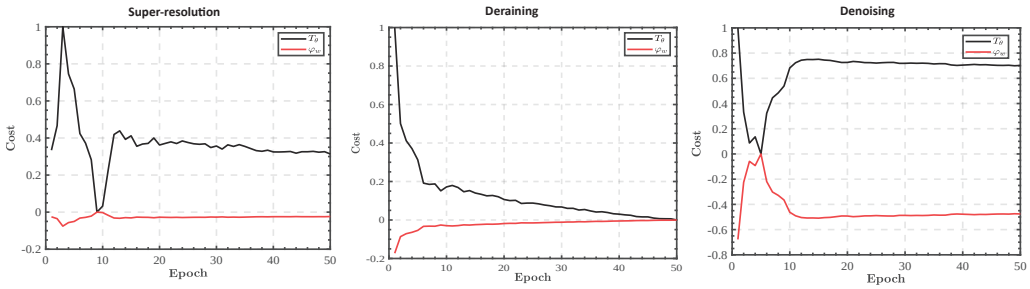


Figure 10: The cost curves for three tasks. The cost of T_θ is scaled to $[0, 1]$. The cost of φ_w is scaled to $[0, 1]$ and then takes the negative.

G. Additional Evaluations and Comparisons

Parameter quantity and computational complexity. We report the parameter quantity and computational complexity in Table 11. We compare with the two most recent SOTA methods IRSDE and PromptIR. The results show that our approach is based on a relatively lightweight model, which sustains its effectiveness. **Cross-domain evaluations.** We conduct two representative cross-domain evaluations to compare the generalizability of RCOT and other methods. Specifically, we train the deraining models on the synthetic Rain13K dataset (Fu et al., 2017; Li et al., 2016; Yang et al., 2017; Zhang & Patel, 2018; Zhang et al., 2019a) and then test their performances on the real rainy dataset SPANet (Wang et al., 2019). Then we train the dehazing models on the synthetic SOTS dataset and then test their performances on the real hazy dataset O-HAZE (Ancuti et al., 2018). Table 12 reports the quantitative results. The results show that the compared methods generally exhibit a declined performance in the presence of a domain gap. However, RCOT still achieves notable performance under all metrics, which sustains the generalizability of RCOT over real-world datasets.

Method	IR-SDE	PromptIR	RCOT
#Param	36.2M	33M	14.2M
Flops	117G	158G	142G
Inference time	9.09s	3.25s	1.06s

Table 11: Comparison of the number of parameters, model computational efficiency, and inference time. The metrics computed on Rain100L (Fan et al., 2019) with a single NVIDIA 4090 GPU.

Evaluation of a single model for multiple degradations. We are interested in extending RCOT to realize all-in-one

Table 12: Cross-domain comparisons. The models are trained on synthetic datasets and tested on real-world datasets (real rainy SPANet (Wang et al., 2019) and real hazy O-HAZE (Ancuti et al., 2018)). (∗) indicates an unpaired setting.

Method	Real rainy SPANet (Wang et al., 2019)				Real hazy O-HAZE (Ancuti et al., 2018)			
	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)
NOT* (Korotin et al., 2023b)	31.22	0.882	0.040	50.23	17.28	0.698	0.215	198.56
OTUR* (Wang et al., 2022a)	36.52	0.932	0.025	33.21	17.45	0.712	0.202	199.78
MPRNet (Zamir et al., 2021)	37.38	0.943	0.031	36.37	21.28	0.765	0.244	228.67
Restormer (Zamir et al., 2022)	37.86	0.951	0.022	33.65	24.56	0.788	0.233	196.33
IR-SDE (Luo et al., 2023)	38.29	0.966	0.013	19.52	24.53	0.796	0.169	186.44
PromptIR (Potlapalli et al., 2023)	37.23	0.947	0.019	34.29	24.32	0.776	0.240	205.66
RCOT*	37.37	0.948	0.015	21.20	22.35	0.775	0.178	189.72
RCOT	**40.02**	**0.972**	**0.009**	**15.66**	**26.59**	**0.827**	**0.148**	**165.12**

restoration in the future, which is to train a single model for multiple degradations. Since the Fourier residual regularizer in RCOT is task-specific, there is still a gap towards the all-in-one target. Nevertheless, the regularizers for deraining and dehazing are set to be ℓ_1 sparsity. Hence we train our model with a combination of rainy data (Rain100L training set) and noisy data (BSD400 and WED with $\sigma = 25$). Then we test on Kodak24 with noise level $\sigma = 25$ and Rain100L test set. The quantitative results are reported in Table 14, in which PromptIR (Potlapalli et al., 2023), the most recent representative all-in-one method is chosen as a competitor. The results show that our RCOT exhibits a better capability of handling multiple degradations. The reason behind the reason should be our residual embedding contains richer degradation-specific knowledge (e.g., degradation type and level) as compared to the learnable visual prompt in PromptIR (Potlapalli et al., 2023).

Table 13: Evaluation of a single model for both denoising and deraining.

Method	Rain100L	Kodak24
PromptIR (Potlapalli et al., 2023)	36.79/0.974	31.25/0.872
RCOT (ours)	**38.02/0.984**	**31.82/0.879**

G.1. Additional comparison with GAN-based methods

Given that the potential network φ can be considered as one discriminator, a comparison with GAN-based approaches could provide valuable insights into the effectiveness and uniqueness of the proposed method in terms of image restoration. We compare RCOT with the most recent GAN-based restoration methods for deraining (on synthetic Rain100L and real SPANet) and SR (on DIV2K).

Table 14: Comparison with GAN-based methods.

Method	Rain100L	Kodak24
CycleGAN (Zhu et al., 2017)	32.03/0.889/0.052/37.26	28.79/0.923/0.036/35.25
DeCyGAN (Wei et al., 2021)	32.75/0.893/0.047/32.15	34.78/0.929/0.042/32.67
RCOT	**37.27/0.980/0.012/7.97**	**43.77/0.993/0.008/9.52**

Method	DIV2K
ESRGAN (Wang et al., 2018)	26.22/0.752/0.112/6.59
RankSRGAN (Zhang et al., 2019b)	26.55/0.75/0.117/9.52
RCOT	**28.41/0.804/0.114/1.22**

H. Additional Visual Results

H.1. Additional Comparison on Real Rainy Dataset

We evaluate the performance of the compared methods on the real rainy dataset SPANet (Wang et al., 2019). This dataset contains 27.5K paired rainy and rain-free images for training, and 1, 000 paired images for testing.

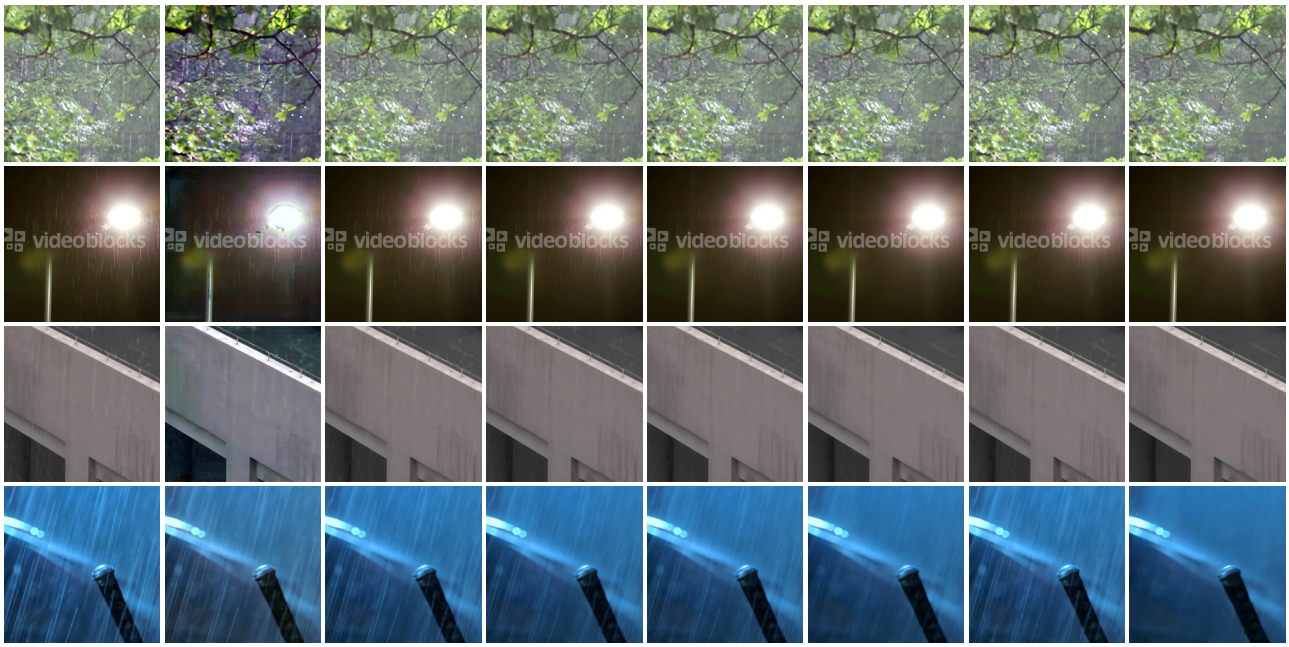


Figure 11: Visual comparison of deraining on the real dataset SPANet (Wang et al., 2019). The RCOT reproduces rain-free images with better structural details.

H.2. Additional Visual Comparison on Real Hazy Dataset

Figure 12: Visual comparison of dehazing on the real dataset O-HAZE (Ancuti et al., 2018). The RCOT reproduces haze-free images with more faithful color.

H.3. More visual examples

Figures 13, 14, and 15 display more visual results of the compared methods.

Figure 13: Visual comparison of denoising on noise level of $\sigma = 50$. Our RCOT produces results with better textural details.

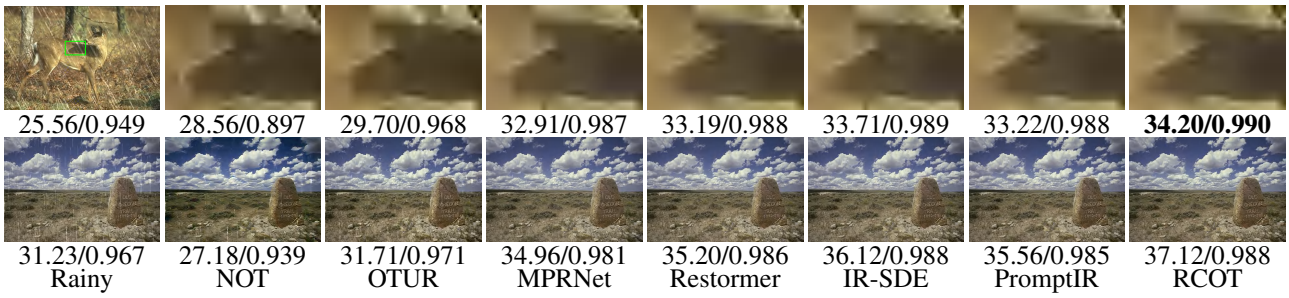


Figure 14: Visual comparison of deraining. Our RCOT produces results with better structural details.

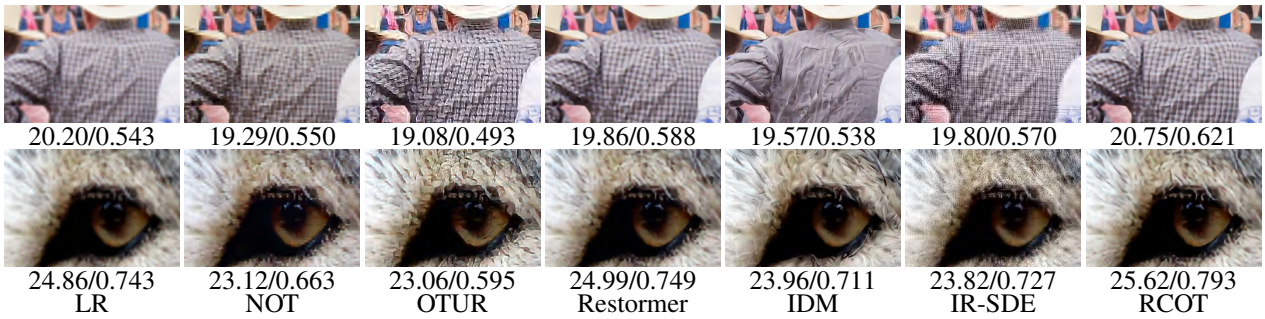


Figure 15: Visual comparison of 4x SR on the DIV2K (Agustsson & Timofte, 2017) dataset. RCOT produces the sharpest HR images with more realistic structures.

I. Visual examples of Transport Residuals

This section exhibits some visual examples of the transport residual $\hat{r}(T_\theta)$. We can observe that different degradations have unique degradation-specific transport residuals, which coincides with our motivation to utilize the residual embedding to encode the degradation-specific knowledge (e.g., degradation type and level).

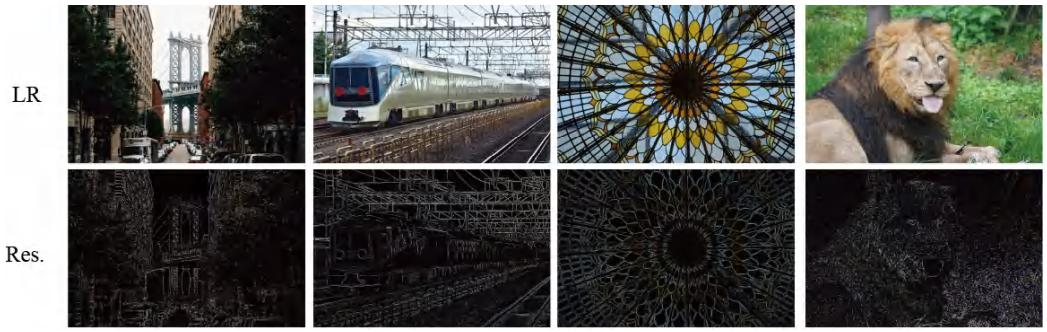


Figure 16: Visual examples of the estimated transport residuals $\hat{r}(T_\theta)$ for **SR** task.

Figure 17: Visual examples of the estimated transport residuals $\hat{r}(T_\theta)$ for **dehazing** task.

Figure 18: Visual examples of the estimated transport residuals $\hat{r}(T_\theta)$ for **denoising** task.

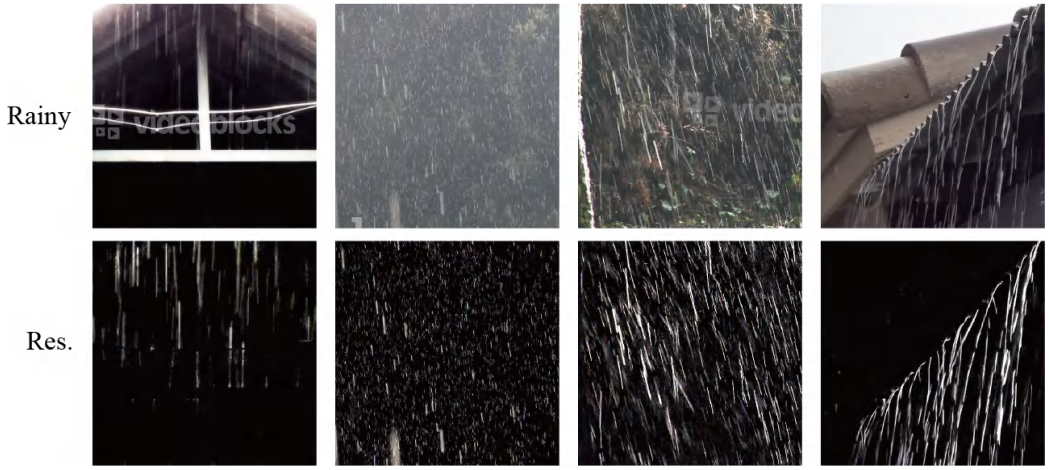


Figure 19: Visual examples of the estimated transport residuals $\hat{r}(T_\theta)$ for **draining** task.