SHORT TITLE

MODEL OPTIMIZATION FOR FEDERATED LEARNING AND FLOW-BASED IMAGE RESTORATION

By LIANGYAN LI, M.A.Sc

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Ph.D. of Electrical and Computer Engineering

McMaster University

PH.D. OF ELECTRICAL AND COMPUTER ENGINEERING (2025)

Hamilton, Ontario, Canada (Electrical and Computer Engineering)

TITLE:	Model Optimization for Federated Learning and Flow-Based Image Restoration
AUTHOR:	Liangyan Li M.A.Sc (Electronic Information Engineering), Tianjin University of Science and Technology, Tianjin, China
SUPERVISOR:	Dr. JunChen
NUMBER OF PAGES:	xx, 131

Lay Abstract

This thesis explores how we can improve artificial intelligence models in two key areas: Federated Learning (FL) and Image Restoration (IR).

FL allows multiple clients to train a shared model without revealing their private data. However, differences in data distributions across clients, known as data heterogeneity (non-i.i.d.), can negatively impact the model's learning process and accuracy. We develop an informed client selection strategy that prioritizes clients based on the diversity of their local gradients.

In IR, we focus on removing complex visual distortions while preserving perceptual quality and natural image structure. We develop a two-stage restoration framework to address the fundamental distortion–perception tradeoff. The first stage generates a coarse estimate optimized for distortion-oriented metrics, while the second stage refines this estimate using generative model-based methods that learn an efficient distribution mapping to enhance perceptual fidelity. Our approach achieves high-quality, visually realistic restorations even under challenging real-world degradations by combining distortion reduction with perceptual enhancement. These contributions advance flow-based optimization strategies for image restoration in the presence of distribution shifts.

Abstract

This thesis explores model optimization strategies for two fundamental areas in machine learning: Federated Learning (FL) and Image Restoration (IR), both of which must address challenges posed by data heterogeneity and distribution shifts. We present three contributions aimed at improving robustness, adaptability, and performance in these settings.

The first chapter introduces a gradient-based client selection method for FL. We propose a novel ℓ_4-norm cosine similarity metric that captures higher-order gradient structures, allowing the server to prioritize clients whose updates are more aligned and informative. This approach accelerates convergence and improves the final model quality compared to random or traditional ℓ_2-based selection strategies, especially under non-i.i.d. client distributions.

The second chapter presents MoiréXNet, a multi-scale image restoration network designed to remove complex visual distortions such as moiré patterns. Our framework integrates linear attention modules for efficient feature aggregation, test-time training for adaptation to unseen degradations, and a truncated flow matching prior to enforce structural consistency. MoiréXNet achieves state-of-the-art performance across several real-world benchmarks.

The third chapter addresses the "Last Mile" of image restoration through a rectified flow-based refinement process. We design a two-stage restoration framework: a coarse estimate is first optimized for distortion-oriented metrics, followed by refinement using generative model-based methods that learn efficient distribution mappings to enhance perceptual fidelity. This strategy balances distortion reduction and perceptual quality, producing visually realistic results even under severe degradation.

Collectively, this thesis advances gradient-based optimization for federated systems and flow-guided adaptive restoration methods, contributing to the development of AI models that are robust, efficient, and capable of handling messy, unpredictable real-world data.

To my beloved family

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Jun Chen, for his unwavering support and exceptional mentorship since 2017, when he served as my co-supervisor during my master's degree. His guidance, patience, and expertise have ignited my passion for research and profoundly shaped my academic and personal growth, making this thesis possible.

I would like to thank my committee members, Dr. Sorina Dumitrescu, Dr. Ratnasingham Tharmarasa and Dr. Narayanaswamy Balakrishnan, for giving me guidance and support over the years. I would also like to thank Dr. Herbert Yang for being my external examiner. My gratitude also goes to my friends and fellow collaborators, especially to Huan, Yangyi, Stefano, Xiangyu, Kevin, Matthew, Wei, Zijun, Xiaohong, and Yimo for generously sharing their advice and knowledge.

Lastly, I would like to thank my family for their unwavering support throughout my Ph.D. journey. To them, I dedicate this thesis.

Table of Contents

List of Figures

List of Tables

Notation, Definitions, and Abbreviations

Notation

$A \leq B$	A is less than or equal to B
$\mathbb{E}[\cdot]$	Expectation operator
$(\cdot)^T$	Transpose operator
$\text{tr}(\cdot)$	Trace operator
$\text{diag}(x_1, ..., x_L)$	$L \times L$ diagonal matrix with diagonal entries $x_1, ..., x_L$
$\lvert \cdot \rvert$	Determinant operator
$\mathbf{1}$	All ones vector
$\mathbf{0}$	All zeros vector
$\lVert \cdot \rVert_2$	2-norm
$\lVert \cdot \rVert_p$	p-norm

$\lvert\mathcal{X}\rvert$	Cardinality of a set $\mathcal{X}$
Σ_X	Covariance matrix of X
$I(X;Y)$	Mutual information between X and Y

Definitions

Challenge With respect to video games, a challenge is a set of goals presented to the player that they are tasks with completing; challenges can test a variety of player skills, including accuracy, logical reasoning, and creative problem solving

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
CS	Client Selection
DDPM	Denoising Diffusion Probabilistic Model
FM	Flow Matching
FL	Federated Learning
GAN	Generative Adversarial Network
GT	Ground Truth

HR	High Resolution
ISP	Image Signal Processor
IR	Image Restoration
KV	Key–Value
LFEF	Learnable Frequency Enhanced Filter
LR	Low Resolution
LSTM	Long Short-Term Memory
MAML	Model-Agnostic Meta-Learning
MSE	Mean Squared Error
MMSE	Minimum Mean Squared Error
NIQE	Natural Image Quality Evaluator
PS	Parameter Server
PSNR	Peak Signal-to-Noise Ratio
PnP	Plug-and-Play
RF	Rectified Flow
RNN	Recurrent Neural Network
SD3	Stable Diffusion 3
SGD	Stochastic Gradient Descent

SISR	Single Image Super-Resolution
SOTA	State of the Art
SSIM	Structural Similarity Index
TTT	Test-Time Training

Declaration of Academic Achievement

1. Liangyan Li, Kevin Le, Ruibin Li, Matthew Ferreira, Wei Dong, Jun Chen, Xiangyu Xu. Solving the Last Mile Problem of Image Restoration with Rectified Flow. Work-in-progress

2. Liangyan Li, Yimo Ning, Wei Dong, Kevin Le, Yunzhe Li, Xiaohong Liu, Jun Chen. MoireXNet: Adaptive Multi-Scale Demoiréing with Linear Attention Test-Time Training and Truncated Flow Matching Prior. arXiv preprint `https://arxiv.org/abs/2506.15929`.

3. Liangyan Li, Yangyi Liu, Yimo Ning, Stefano Rini, Jun Chen. A Gradient-Based Selection Scheme Leveraging Cos_4 Cosine Similarity for Federated Learning under Data Heterogeneity. arXiv preprint `https://arxiv.org/abs/2506.15923`.

4. Li Xie, Liangyan Li, Jun Chen, and Zhongshan Zhang. Output-constrained lossy source coding with application to rate-distortion-perception theory. IEEE Trans- actions on Communications, 2024.

5. Li Xie, Liangyan Li, Jun Chen, Lei Yu, and Zhongshan Zhang. Gaussian rate- distortion-perception coding and entropy-constrained scalar quantization. arXiv preprint arXiv:2409.02388, 2024.

6. Marcos V Conde, Florin-Alexandru Vasluianu, Radu Timofte, Jianxing Zhang, Jia Li, Fan Wang, Xiaopeng Li, Zikun Liu, Hyunhee Park, Sejun Song, et al. Deep raw image super-resolution. a ntire 2024 challenge survey. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6745–6759, 2024.

7. Yangyi Liu, Huan Liu, Liangyan Li, Zijun Wu, and Jun Chen. A data-centric solution to nonhomogeneous dehazing via vision transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1406–1415, 2023.

8. Ke Chen, Liangyan Li, Huan Liu, Yunzhe Li, Congling Tang, and Jun Chen. Swin- fsr: Stereo image super-resolution using swinir and frequency domain knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1764–1774, 2023.

9. Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, Han Zhou, Wei Dong, Yangyi Liu, Jun Chen, Huan Liu, Liangyan Li, et al. Ntire 2023 hr nonhomogeneous dehazing challenge report. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1808–1825, 2023

10. Huan Liu, Zijun Wu, Liangyan Li, Sadaf Salehkalaibar, Jun Chen, and Keyan Wang. Towards multi-domain single image dehazing via test-time training. In

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5831–5840, 2022.

11. Ren Yang. Ntire 2021 challenge on quality enhancement of compressed video: Methods and results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 647–666, 2021.

Chapter 1

Introduction

Artificial intelligence (AI) has achieved remarkable success across various applications [32, 41, 139, 30, 28], from personalized services and medical diagnosis to autonomous driving and computational photography. Despite these advances, deploying AI models in real-world settings often presents significant challenges, particularly in data heterogeneity, distribution shifts, and limited supervision. This thesis seeks to enhance model performance while improving the robustness and effectiveness of AI models in two pivotal areas: *Federated Learning (FL)* and *Image Restoration (IR)*.

1.0.1 Problem Formulation: Federated Learning

FL aims to collaboratively train a global model $w \in \mathbb{R}^d$ across a population of N decentralized clients, each holding a private local dataset $\mathcal{D}_i$ $(i = 1, \ldots, N)$. The central server seeks to solve the following optimization problem:

$$\min_{w} F(w) = \sum_{i=1}^{N} p_i F_i(w), \tag{1.0.1}$$

where $F_i(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}_i}[\ell(w; x, y)]$ is the local loss function on client i, ℓ denotes a supervised loss (e.g., cross-entropy), and p_i reflects the relative importance or size of client i's dataset.

At each communication round t, a subset of clients $\mathcal{S}_t \subseteq \{1, \ldots, N\}$ is selected to perform local updates. Each selected client $i \in \mathcal{S}_t$ updates its model parameters by minimizing $F_i(w)$ locally, typically via stochastic gradient descent (SGD), and sends the update Δw_i back to the server for aggregation [133].

Challenges under Data Heterogeneity. In practice, client datasets are often non-i.i.d. and highly heterogeneous [80, 230, 97, 143, 1], leading to divergent updates that slow convergence and degrade the performance of the aggregated global model. Random client sampling fails [133] to account for the variability and alignment of client updates, resulting in suboptimal optimization dynamics. To address these challenges, this thesis formulates client selection as an informed sampling process. In each round, the server aims to select clients whose updates are expected to be diverse, yet aligned with global optimization. Specifically, in chapter 2, we introduce a gradient-based selection strategy leveraging ℓ_4-norm cosine similarity (Cos_4), which measures high-order similarities between local gradients and the global model:

$$Cos_4(g_i, g_j) = \frac{\langle g_i, g_j \rangle}{\|g_i\|_4 \|g_j\|_4}, \tag{1.0.2}$$

where g_i and g_j are flattened gradients or model updates from clients i and j, and $\|\cdot\|_4$ denotes the ℓ_4 norm.

By prioritizing clients based on their Cos_4 similarity to the global direction, the server selects a subset $\mathcal{S}_t$ that promotes faster convergence and more consistent model

updates, effectively mitigating the impact of data heterogeneity.

1.0.2 Problem Formulation: Image Restoration

IR aims to recover a clean image $X \in \mathbb{R}^{H\times W\times C}$ from a degraded observation $Y \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote the image height, width, and channels, respectively. The degradation process is typically modeled as a conditional distribution $p(Y \mid X)$, which may be complex, non-invertible, and unknown in real-world scenarios [114, 25, 103].

Given paired training data $\{(Y_i, X_i)\}_{i=1}^{N}$, traditional supervised learning approaches aim to learn a deterministic mapping $f_w : Y \mapsto \hat{X}$ by minimizing a distortion-oriented loss:

$$w^* = \arg\min_{w} \mathbb{E}_{(Y,X)} \left[\mathcal{L}_{\text{distortion}} \left(f_w(Y), X\right)\right], \tag{1.0.3}$$

where $\mathcal{L}_{\text{distortion}}$ typically measures pixel-wise differences, such as ℓ_1 or ℓ_2 loss.

However, minimizing distortion metrics alone often leads to overly smooth reconstructions that lack perceptual realism, particularly when the degradation distribution at test time deviates from the one seen during training. This issue is known as the *distortion–perception tradeoff* [15].

To address these challenges, we adopt a two-stage restoration framework:

- **Stage 1: Coarse Estimation.** A mapping $f_w : Y \mapsto Z$ is trained to minimize distortion and produce an initial estimate Z of the clean image.
- **Stage 2: Perceptual Refinement.** A rectified flow model $v_\theta : Z \mapsto \hat{X}$ is used to refine Z toward a perceptually improved reconstruction $\hat{X}$ via deterministic

sample transport.

In Chapter 3, we implement the first-stage coarse estimator by designing MoiréXNet, an adaptive multi-scale restoration network designed for demoiréing. MoiréXNet integrates linear attention for efficient feature aggregation and employs test-time training (TTT) [178] for online adaptation. To support perceptual refinement, it also incorporates a truncated flow matching [100] prior, improving robustness to unseen degradations.

In Chapter 4, we extend this two-stage framework into a general-purpose solution for universal image restoration. We treat the outputs of existing MMSE-based models as coarse estimates and refine them using a rectified flow model [112]. This approach addresses the *Last Mile* problem in IR—improving perceptual quality without retraining the entire pipeline. We theoretically and empirically demonstrate that rectified flow offers an efficient refinement mechanism that enhances both distortion and perception metrics across diverse restoration benchmarks.

1.0.3 Contributions and Thesis Organization

The thesis consists of three articles (under review) addressing model optimization for client selection in Federated Learning and Image Restoration. Contributions are detailed in the sections of Chapters 2, 3 and 4. Summarized reference information follows:

- **Liangyan Li**, Yangyi Liu, Yimo Ning, Stefano Rini, Jun Chen. "A Gradient-Based Selection Scheme Leveraging L4 Cosine Similarity for Federated Learning under Data Heterogeneity". arXiv preprint `https://arxiv.org/abs/2506.15923`.

- **Liangyan Li**, Yimo Ning, Wei Dong, Kevin Le, Yunzhe Li, Xiaohong Liu, Jun Chen, J. (2025). "MoireXNet: Adaptive Multi-Scale Demoiréing with Linear Attention Test-Time Training and Truncated Flow Matching Prior". arXiv preprint `https://arxiv.org/abs/2506.15929`.

- **Liangyan Li**, Kevin Le, Ruibin Li, Matthew Ferreira, Wei Dong, Jun Chen, Xiangyu Xu, (2025). "Solving the Last Mile Problem of Image Restoration with Rectified Flow". Manuscript submitted to TIP 2025 (under review).

The rest of the thesis is organized as follows:

1. **Federated Learning Client Selection.** We introduce a gradient-based client selection strategy using a novel $\mathcal{L}_4$ cosine similarity metric, denoted Cos_4, which more sensitively captures client gradient differences under data heterogeneity.

2. **MoiréXNet: Adaptive Demoiréing.** We present *MoiréXNet*, an adaptive, multi-scale framework that combines linear attention, test-time training, and a truncated flow-matching prior to robustly remove complex moiré artifacts while preserving fine detail.

3. **Rectified Flow for Last-Mile Restoration.** We propose a two-stage, rectified-flow–based refinement for the last-mile problem in image restoration, balancing distortion-minimization objectives with perceptual fidelity via efficient deterministic transport.

4. **Conclusions and Future Directions.** We synthesize our key findings and outline promising directions for future work in both federated learning and flow-based image restoration.

The following chapter is reproduced from a submitted IEEE paper:

Liangyan Li, Yangyi Liu, Yimo Ning, Stefano Rini, Jun Chen. "A Gradient-Based Selection Scheme Leveraging L4 Cosine Similarity for Federated Learning under Data Heterogeneity". arXiv preprint `https://arxiv.org/abs/2506.15923`.

Contribution Declaration: Liangyan Li is the first author and primary contributor: she conceived the algorithm, implemented the code, conducted all experiments, and drafted the manuscript. Yangyi Liu performed preliminary coding and contributed to discussions, while Yimo Ning assisted in organizing the experimental data. Stefano Rini contributed to discussions and the final manuscript writing. Prof. Jun Chen supervised the research.

Chapter 2

A Gradient-Based Selection Scheme Leveraging Ł$_4$ Cosine Similarity for Federated Learning under Data Heterogeneity

2.1 Abstract

Federated Learning (FL) has gained increasing popularity for its ability to utilize diverse datasets from multiple sources without requiring data centralization. However, existing works often overlook the nuanced gradient correlations among remote clients, which can be particularly detrimental in the face of data heterogeneity. In this paper, we propose $\mathsf{cos}_4-\mathsf{select}$, a novel FL framework that exploits an ℓ_4-norm-based cosine similarity measure of the model updates to identify and select clients for

model aggregation. By emphasizing higher-order gradient moments, $\cos_4-\mathsf{select}$ effectively mitigates the adverse impacts of non-IID data distributions, improving both the convergence speed and the final accuracy. In addition, we incorporate a simple yet effective algorithm that promotes diversified selections by tracking a queue of previously selected clients. We validate our approach using a VGG16 model across various data partition schemes and initialization settings, demonstrating consistent gains over state-of-the-art selection strategies.

2.2 Introduction

Across the many distributed deep learning frameworks proposed in the literature, *Federated Learning* (FL) has gained significant attention for its ability to train models on decentralized data without requiring raw data exchange [133, 95]. In an FL system, a central *parameter server* (PS) coordinates the training of a global model by collecting model updates from multiple remote *clients*, each holding its local dataset. In many practical scenarios, the convergence of FL is adversely affected by *local data heterogeneity*, i.e., when the data distributions among clients differ significantly. Simply averaging the local updates in such a setting can severely degrade performance. A second challenge in FL involves *communication efficiency*: because FL relies on frequent exchanges of model updates between clients and the PS [89, 95], the communication overhead often becomes the main bottleneck. This paper addresses the intersection of these two issues. In particular, we seek to answer the question:

"What is the best strategy for selecting a subset of clients for training under local data heterogeneity?"

Our core insight differs from existing client-selection literature in two key respects.

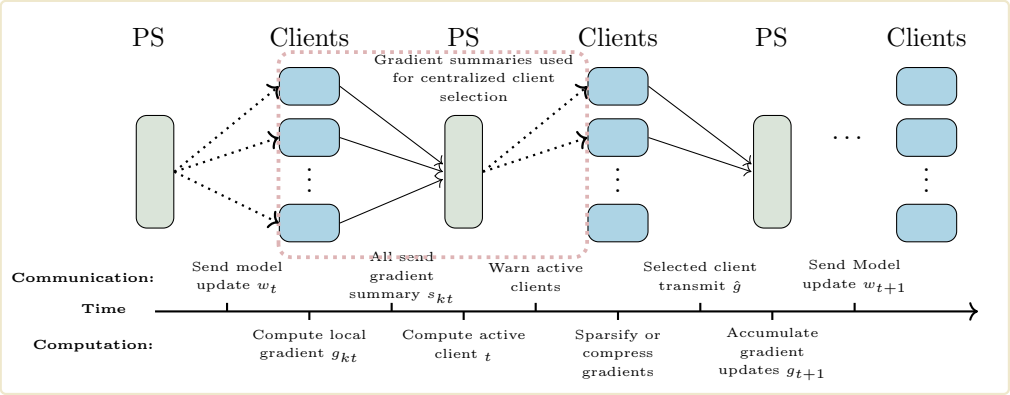


Figure 2.1: A conceptual representation of the *Gradient Summaries for Centralized Client Selection* (GSCCS) setting.

First, we highlight that *diversity* among the selected clients, rather than strict gradient alignment, is often the more effective criterion, as divergent gradients can provide complementary updates for the global model. Second, we propose the cosine similarity induced by the ℓ_4 norm—denoted $\cos_4-\mathsf{select}$—as a particularly robust selection metric under various data heterogeneity levels.

Building on these insights, we design a $\cos_4$-based centralized client-selection mechanism for FL, which we refer to as $\cos_4-\mathsf{select}$. The method is simple, yet can be implemented with minimal communication overhead, since each client only transmits a small *sketch* of its local gradient to the server. Empirically, $\cos_4-\mathsf{select}$ accelerates convergence and improves model accuracy compared to existing strategies, especially in highly heterogeneous environments. We believe our findings offer a practical and scalable approach for enhancing both the communication efficiency and the overall performance of federated learning.

Our contributions can be outlined as follows:

- **Novel Problem Formulation for Client Selection:** We formulate the client

selection problem in federated learning as a logistic regression model in which the input features capture the pairwise gradient diversity between any two clients, and the output is the probability of how likely the pair may be the best choice for model update. This notion naturally extend to the multi-client case by averaging the pairwise selection performance.

- **Feature Exploration for Gradient Alignment:** We thoroughly evaluate a wide range of features for two-client selection in a single-layer FL training scenario, including both statistical and geometrical criteria. Our analysis–conducted across varying number levels of data heterogeneity and number of selected clients– reveals that the cosine similarity induced by the ℓ_K norm (with $K \in \{1, 2, 4\}$) performs robustly across all scenarios.

- **Identification of $\mathsf{cos}_4\mathsf{-select}$ as the Best Single Feature:** Among the different ℓ_K-based cosine similarity measures, we identify $\mathsf{cos}_4\mathsf{-select}$ as the most effective single feature for client selection, as it better captures gradient alignment under diverse training conditions.

- **Empirical Validation and Performance Gains:** We conduct extensive numerical evaluations demonstrating that leveraging $\cos_F$ for client selection significantly improves convergence speed and final model accuracy compared to existing selection strategies. Our results suggest that a simple yet powerful metric such as $\cos_F$ can be effectively integrated into FL protocols with minimal overhead.

Our proposed approach builds upon prior works that explore higher-order statistics of gradient updates. For instance, [31] demonstrates that the kurtosis of the gradient can serve as a relevant statistical measure for training performance, while

[115] explores gradient distortions involving higher-order norms. These insights collectively motivate our focus on ℓ_4-based cosine similarities in the present work.

2.3 Related Work

Client selection in Federated Learning (FL) is motivated by various local dataset dynamics such as data imbalance, heterogeneity, and computational constraints, as established in foundational works [134, 127, 94]. Research has shown that strategic client selection can significantly improve accuracy [92], ensure fairness [176], enhance robustness [10], and accelerate convergence [144, 176]. Several studies tackle the multi-objective nature of this task by applying classical optimization tools to balance fairness, resource constraints, and model performance. For instance, Lyapunov optimization frameworks [73, 162, 12, 240] dynamically manage system stability across these dimensions, while greedy algorithms [135, 224, 118] and Hungarian matching methods [141, 26] offer more computationally efficient solutions. Reinforcement learning approaches [231, 3, 8] also allow adaptive selection policies to be learned from environmental interactions.

On the other hand, single-factor optimizations focus on isolated aspects like computational capacity [189, 243] or client reputation [200, 11, 199, 180], potentially overlooking the comprehensive trade-offs required for optimal global performance. Closer to the scope of this work are approaches that select clients based on their local training losses. The Active FL (AFL) method [56] assigns selection probabilities via differential privacy applied to local losses, and POWER-OF-CHOICE (POWD) [79] prioritizes clients with high local losses. While these loss-based methods can speed up convergence, they may degrade overall performance in non-IID settings by ignoring

correlations among clients.

Rather than relying on local losses, another class of methods selects clients by inspecting their gradients [131, 166, 214]. In particular, [131] selects clients whose gradients have the largest norms, and [166] employs a gradient-based approach that leverages Shapley values to identify clients most representative of the global dataset, improving efficiency and robustness under non-IID conditions. Similarly, [214] jointly optimizes client selection and gradient compression, aiming to reduce communication costs while maintaining overall performance. However, these gradient-based methods typically treat each client's contribution independently, risking biased updates when chosen clients fail to represent the global distribution.

To address this limitation, more recent strategies factor in the relationships among clients. For example, FedCor [181] models correlations between clients' losses via a Gaussian Process, allowing it to iteratively select clients based on predicted performance gains. By focusing on client-to-client relationships, such an approach can prioritize the updates most likely to benefit the global model over successive training rounds.

Existing approaches often rely on local losses, resource metrics, or static client correlations, which can introduce biases or fail to adapt to evolving data conditions. In contrast, our method emphasizes gradient alignment and diversity among clients by directly analyzing inter-client gradient relationships rather than relying solely on losses or reputations. This allows us to avoid redundancy from correlated updates, mitigate biases arising from non-representative client sampling, and enhance robustness under data heterogeneity by prioritizing gradient diversity. Unlike purely

gradient-norm or loss-based selection strategies, this framework explicitly models interdependencies between clients, thereby yielding more coherent and globally beneficial updates.

2.4 Preliminaries

2.4.1 Notations

We use lowercase boldface letters (e.g., $\mathbf{z}$) to denote vectors and calligraphic uppercase letters (e.g., $\mathcal{A}$) to denote sets. For any set $\mathcal{A}$, let $|\mathcal{A}|$ be its cardinality. We adopt the shorthand $[m:n] \triangleq \{m, \ldots, n\}$ and $[n] \triangleq \{1, \ldots, n\}$. Throughout the paper, (i) $k \in [K]$ indicates a client index – we use k' needing two client indexes, (ii) $t \in [T]$ an iteration index, (iii) $s \in \mathcal{S} \subseteq [K]$ a selected client index with $|\mathcal{S}| = S$, (iv) $j \in [J]$ denotes a shard (or partition) index, and $r \in [R]$ is the random seed.

2.4.2 Federated Learning

Consider the FL setting with K clients, each possessing a local dataset $\mathcal{D}_k \in \mathcal{D}$, for $k \in [K]$ wishing to minimize the *loss function* $\mathcal{L}$ as evaluated across all the clients and over the model weights $\mathbf{w} \in \mathbb{R}^m$, where m denotes the dimensionality of the model parameter. This minimization is coordinated by the PS as follows: in round $t \in [T]$, the clients transmit local gradients to the PS; the PS generates a model update, and sends the updated model back to the clients. The above steps are repeated for T times: the model obtained at time T is declared as the converged model. Mathematically,

the loss function $\mathcal{L}$ is defined as

$$\mathcal{L}(\mathbf{w}) = \frac{1}{|\mathcal{D}|} \sum_{k \in [K]} \mathcal{L}_k(\mathcal{D}_k, \mathbf{w}), \tag{2.4.1}$$

where $\mathcal{L}_k(\mathcal{D}_k, \mathbf{w})$ is the local loss function quantifying the prediction error of the k-th client's model. A common approach for numerically finding the optimal value of $\mathbf{w}$ is through the iterative application of (synchronous) stochastic gradient descent (SGD). We define the local gradients calculated at communication round t as

$$\mathbb{E}[\mathrm{g}_{kt}] = \mathbb{E}[\nabla \mathcal{L}_k(\mathcal{D}_k, \mathbf{w}_t)], \tag{2.4.2}$$

where $\nabla \mathcal{L}_k(\mathcal{D}_k, \mathbf{w}_t)$ denotes the local gradients of the model evaluated at the local dataset of the k-th client by minimizing the local loss function. Note that the expectation in (2.4.2) is taken over the randomness in evaluating the gradients, e.g., mini-batch effects. The PS aggregates all the local gradients and forms the new global weights

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{g}_t, \text{ for } t \in [T], \text{ where} \tag{2.4.3}$$

for

$$\mathbf{g}_t = \frac{1}{K} \sum_{k \in [K]} \mathbf{g}_{kt}, \tag{2.4.4}$$

where $\mathbf{w}_0$ is a random initialization.

2.4.3 Client Selection

Communication overhead is a major bottleneck in Federated Learning (FL), since each training round typically involves transmitting updates between multiple clients and a

central PS. To address this, a common strategy is to select only a subset of clients in each round, reducing communication while still gathering sufficiently diverse updates. Formally, let $\mathcal{S}_t \subseteq [K]$ denote the set of active clients at round $t \in [T]$. This selection process aims to balance two objectives: keeping $|\mathcal{S}_t|$ small to mitigate communication costs, and ensuring that the selected gradients are sufficiently diverse to enhance the global model's generalization.

Client selection can follow two main paradigms. In *decentralized* approaches, each client independently decides whether to participate based on local metrics. In contrast, *centralized* approaches place this responsibility on the PS, which can use aggregated information (e.g., partial gradient statistics) for client selection. In this work, we focus on centralized selection under heterogeneous data distributions, leveraging gradient-based measures to identify and activate clients whose updates offer the greatest potential for improving model convergence.

2.4.4 Data Heterogeneity

In many practical scenarios, the local dataset at each client is intrinsically heterogeneous. One way to capture this heterogeneity is by assuming that each client's data are sampled from a mixture of several prior distributions. Specifically, let $\mathbf{m}_{ik}$ denote the i-th data point (features and labels) at client k. We write

$$\mathbf{m}_{ik} \sim P_{\mathbf{m}}^{k} = \sum_{j \in [J]} \lambda_{jk} P_{\mathbf{m}}^{(j)}, \tag{2.4.5}$$

where $\{P_{\mathbf{m}}^{(j)}\}_{j \in [J]}$ is a kernel of J distributions, and the mixing coefficients $\lambda_k =$

$\{\lambda_{jk}\}_{j\in[J]}$ specify the proportion of data at client k coming from each kernel distribution.

In the following, we implement the mixture model of (2.4.5), we partition the full CIFAR dataset into 10 shards–one per label– and assign to each client J *shards*. Every client k then receives exactly J shards *chosen uniformly at random* from the available $J \cdot K$ total shards. This produces non-i.i.d. local data distributions reflecting the mixture structure. By varying J or the manner in which shards are allocated, one can control the degree of heterogeneity across clients.

2.5 Problem Formulation

In many practical FL scenarios, it is advantageous for the parameter server (PS) to decide which clients will actively participate in a training round, based on partial information about each client's local gradients. We refer to this approach as the *Gradient Summaries for Centralized Client Selection* (GSCCS) setting – see Fig. 2.1. Specifically, the selection process proceeds in two phases:

1. **Summary Transmission:** Each client $k \in [K]$ sends a *gradient summary*—a compressed or otherwise optimized representation of its local gradient $\mathbf{g}_{kt}$ —to the PS.

2. **Centralized Client Selection:** Upon receiving these summaries, the PS examines their contents and decides on the set $\mathcal{S}_t \subseteq [K]$ of active clients for round t, with $|\mathcal{S}_t| = S$.

By introducing this extra summary-transmission phase, the PS can make a more informed selection of active clients. Compared to one-shot aggregation schemes that

skip client selection, GSCCS adds one extra transmission step prior to the main gradient exchange. However, these summaries typically have much lower dimension or precision than the full gradient vectors, allowing the PS to *exclude* clients whose updates do not significantly benefit the global model.

Mathematically, let the *gradient summary* of client k at time t be defined as

$$\mathbf{s}_{kt} = \phi(\mathbf{g}_{kt}), \tag{2.5.1}$$

where $\phi : \mathbb{R}^m \to \mathbb{R}^p$ outputs a lower-dimensional or compressed vector. Each client $k \in [K]$ transmits $\mathbf{s}_{kt}$ to the PS, which then uses these summaries to compute pairwise gradient affinities and ultimately select the active set $\mathcal{S}_t$ of cardinality J.

Once $\mathcal{S}_t$ is determined, the gradient update at round t is given by

$$\widehat{\mathbf{g}}_t = \frac{1}{J} \sum_{s \in \mathcal{S}_t} \mathbf{g}_{st}, \tag{2.5.2}$$

and the corresponding sequence of model weights $\{\widehat{\mathbf{w}}_t\}_{t\in[T]}$ follows from standard gradient-based updates. We denote the client selection policy as

$$\mathcal{S}_t = \pi\big(\{\mathbf{s}_{kt}\}_{k\in[K]}\big), \tag{2.5.3}$$

where among the $\binom{K}{J}$ possible subsets of clients, the *optimal* choice $\mathcal{S}_t^*$ minimizes the loss $\mathcal{L}(\cdot)$ at time t. Thus, the client selection problem under the GSCCS setting can be formulated as

$$\mathsf{P}: \quad \min_{\phi,\pi} \quad \frac{1}{T} \sum_{t\in[T]} \left(\mathcal{L}\big(\widehat{\mathbf{g}}_t^*\big) \; - \; \mathcal{L}(\widehat{\mathbf{g}}) \right), \tag{2.5.4}$$

where we seek to minimize the average additional loss incurred by selecting J out of

K clients based on summaries of dimension p.

2.5.1 Assumptions, Observations, and Comments

Sketch dimension. We do not delve deeply into how the sketch size r affects training performance, leaving this investigation for future work. Our current framework assumes r is large enough that each sketch accurately captures the relevant gradient information needed for client selection.

Reliability of summaries. We assume the transmitted summaries are sufficiently accurate for the parameter server to compute meaningful gradient-affinity measures. In practice, one can optimize the sketching or sampling procedure (e.g., choosing an appropriate matrix M) to balance communication overhead against selection accuracy.

Randomness in the training process. When referring to "optimal" client selection, we do so in an *average* sense—averaging over the stochasticity introduced by random shard allocations, parameter initialization, and mini-batch sampling. Analyzing the full dynamics of this random process is beyond our current scope.

Final-loss selection criterion. Selecting clients to directly minimize the *final* loss after T rounds can be impractical, as it requires predicting the model's future evolution across multiple rounds. Instead, our approach focuses on a per-round (or greedy) selection strategy, striking a practical balance between simplicity and performance.

2.6 Proposed Method: $\mathsf{cos}_4-\mathsf{select}$

In this section, we present the main contribution of our work: the $\mathsf{cos}_4-\mathsf{select}$ algorithm for centralized client selection under the *Gradient Summaries with Centralized Client Selection* (GSCCS) setting (Fig. 2.1). We organize our discussion as follows. First, Section 2.6.1 introduces a logistic regression formulation for selecting *exactly two* clients. Section 2.6.2 provides numerical insights into the most predictive gradient-based features, and Section 2.6.3 describes the full $\mathsf{cos}_4-\mathsf{select}$ algorithm.

2.6.1 Logistic Regression Formulation for Pairwise Client Selection

To begin, we focus on selecting exactly two clients out of K, treating the client selection task as a logistic regression problem. Concretely, the *input features* capture how two clients' updates interact (e.g., pairwise cosine similarity), while the *labels* reflect normalized validation accuracies after training on those two clients. The resulting *dataset* consists of feature-label pairs gathered under different rounds, heterogeneity levels, and random seeds.

Mathematically, we define the *client selection* (CS) loss:

$$\mathcal{L}_{CS} = \min_{\mathbf{w}} \frac{1}{J} \sum_{\substack{k,k' \\ k \neq k'}} \mathsf{BCE}\Big(\sigma(\mathbf{w}^\top \mathbf{x}_{k,k'}),\, y_{k,k'}\Big), \tag{2.6.1}$$

where BCE is the binary cross-entropy loss, $\sigma(\cdot)$ is the sigmoid function, $\mathbf{x}_{k,k'}$ is a d-dimensional feature vector (e.g., pairwise gradient metrics) for clients (k, k'), and $y_{k,k'} \in [0, 1]$ reflects how well the pair (k, k') performs if selected. Intuitively,

$\mathcal{L}_{CS}$ measures how accurately the logistic model (parameterized by $\mathbf{w}$) predicts the performance of any two-client combination under the specified training conditions.

Table 2.1: Comparison of $\cos_1$, $\cos_2$, $\cos_3$, and $\cos_4$ under various experimental settings.

Charac.	Type	$\cos_1$		$\cos_2$		$\cos_3$		$\cos_4$	
		Rank	Rel.A	Rank	Rel.A	Rank	Rel.A	Rank	Rel.A
Shard	1	2.85	0.1790	2.85	0.1771	2.65	0.1761	1.65	**0.1747**
	2	3.05	0.0993	2.55	0.1017	2.55	0.1002	1.85	**0.0985**
	5	3.00	**0.0334**	2.50	0.0337	2.40	0.0393	2.10	0.0415
Iteration	5	3.33	0.0833	2.33	0.0766	2.67	0.0677	1.67	**0.0672**
	10	3.33	0.0762	2.00	0.0645	3.33	0.0657	1.33	**0.0634**
	15	4.00	**0.1073**	2.00	0.1189	2.33	0.1352	1.67	0.1227
All	Avg	2.97	**0.1039**	2.63	0.1042	2.53	0.1052	1.87	0.1049
	Std	0.59	0.073	0.60	0.072	0.51	0.069	0.51	**0.067**

Feature Vector Construction. In our experiments, we collect various gradient-based metrics into the feature vector $\mathbf{x}_{k,k'}$ (see Table 2.1). One key candidate is the *cosine similarity* under the L_p norm:

$$\cos_p(\mathbf{g}_k, \mathbf{g}_{k'}) \;=\; \frac{\langle \mathbf{g}_k, \mathbf{g}_{k'} \rangle_p}{\|\mathbf{g}_k\|_p \, \|\mathbf{g}_{k'}\|_p} \tag{2.6.2}$$

with

$$\langle u, v \rangle_p = \frac{\|u+v\|_p - \|u-v\|_p}{4}. \tag{2.6.3}$$

This generalization encompasses the traditional L_2 cosine similarity as a special case and allows for higher-order norms (e.g., L_4) that can emphasize dominant gradient coordinates.

Label Definition. To generate the label $y_{k,k}$ in (2.6.1) for a particular pair of clients

k, k' we:

- Let $y'_{k',k}(n, j, r)$ represent the vector of accuracies (or losses) for all possible $K(K-1)$ pairs under the three training conditions: (i) iteration n, (ii) heterogeneity level j, (iii) and random seed r.
- Map these raw accuracies to $[0, 1]$ considering for instance, via a softmax or min-max scaling to obtain y_j. Other scaling functions can be considered, such as linear or entropy scaling.

This yields a binary or continuous label indicating the relative performance of each pair k, k' across the three training conditions.

Once the logistic model is trained, the learned weight vector $\mathbf{w}$ reveals which pairwise features (or combinations thereof) best predict successful client selection. Of course, care must be taken to avoid issues like collinearity, overfitting, and data imbalance; yet, this framework offers a systematic way to distill the most informative gradient-based metrics for further analysis or for building practical selection heuristics.

2.6.2 Numerical Findings

To gain further insight on the client selection formulation in Sec. 2.6.1 we consider the training of the last layer of VGG for $K = 10$ clients, $T = 15$ iterations, $J = \{1, 2, 5\}$ shards, and $R = 10$ seeds, For each of the training settings, i.e. (t, j, r) we consider the selection of all possible $K(K-1)$ pair of users k, k' and record (i) a set of pairwise gradient features $\mathbf{x}_{k,k,}(t, j, r)$, and the accuracy of the model under this selection $y'_{k,k'}(t, j, r)$ as discussed above. The features we consider are quite extensive, such as

- *per-user, geometric features* such as $(\|\mathbf{g}\|_p + \|\mathbf{g}'\|_p)$,
- *pairwise geometric features* e.g. $\|\mathbf{g} - \mathbf{g}'\|_p$, $\cos_p(\mathbf{g}, \mathbf{g}')$
- *per-user, statistical features* e.g. $(\mathsf{Kurt}[\mathbf{g}] + \mathsf{Kurt}[\mathbf{g}'])$,
- *pairwise statistical features* e.g. $\mathsf{Cov}(\mathbf{g}, \mathbf{g}')$

After careful experimentation, we conclude that the $\cos_p(\mathbf{g}, \mathbf{g}')$ for $p \in [4]$ provides the most accurate and robust user selection performance. A summary of the training for this set of features is provided in Table 2.1: here we report the performance in terms of the (i) rank of the features in the feature important analysis and (ii) the relative loss, obtained, as normalized over the loss of the worst user selection minus the loss of the user selection and across (i) shards and (ii) iterations. Note that the results in each row are averaged across the other training conditions. That is the results for shard j are averaged over the iterator t and random seed r.

From the above, we glean the two following main insights:

- **Single-Feature Accuracy:** When restricting to a single feature for simplicity, $\cos_4$ consistently emerges as the most predictive and robust metric for successful client-pair selection. While $\cos_1$ achieves a smaller relative accuracy overall, it has a higher variance.
- **Negative Alignment:** selecting the pair with the *most negative* $\cos_4$ improves performance, suggesting that selecting complementary gradients–as opposed to overly similar ones– is beneficial in most settings.

Algorithm 1 $\cos_4$-select: FL Training under GSCCS

Require: Rounds T, number of clients K, selection size J, sketch function $\phi(\cdot)$, queue length L

Initialize global model w_0, queue $\mathcal{Q} = \emptyset$
for $t = 1$ to T **do**
  PS broadcasts w_t to all clients
  for each client $k \in [K]$ in parallel **do**
    $\mathbf{g}_{kt} \leftarrow$ **ClientUpdate**$(\mathcal{D}_k, w_t)$
    $\mathbf{s}_{kt} \leftarrow \phi(\mathbf{g}_{kt})$ ▷ Gradient summary (Eq. (2.4.2))
    Send $\mathbf{s}_{kt}$ to the PS
  end for
  PS computes $\cos_4(\mathbf{S}_{it}, \mathbf{S}_{jt})$ for all $i, j \notin \mathcal{Q}$ ▷ Eq. (2.6.4)
  Select $\mathcal{S}_t = \text{Top-}J(-\cos_4(\mathbf{s}_{it}, \mathbf{s}_{jt}))$ ▷ Negative alignment
  for each $x \in \mathcal{S}_t$ **do**
    enqueue x into $\mathcal{Q}$ ▷ Track selected clients
  end for
  Selected clients $\mathcal{S}_t$ send $\mathbf{g}_{kt}$ to PS
  $\widehat{\mathbf{g}}_t = \frac{1}{J} \sum_{k \in \mathcal{S}_t} \mathbf{g}_{kt}$ ▷ Eq. (2.5.2)
  $w_{t+1} \leftarrow$ **Optimizer**$(w_t, \widehat{\mathbf{g}}_t)$ ▷ Eq. (2.4.3)
  PS updates queue $\mathcal{Q}$ by removing clients if their dwell time $> L$
end for

2.6.3 Proposed Algorithm: $\cos_4-$select

Having gleaned the insights from Section 2.6.2, we now present our proposed solution for the GSCCS setting, which we refer to as $\cos_4-$select. Before describing the full procedure, we introduce three additional elements that generalize and stabilize the selection process.

Generalizing Beyond Pairs. Although the earlier analysis focused on selecting exactly two clients, we naturally extend this to subsets of size $J > 2$. Rather than computing pairwise $\cos_4$ for two users, we consider the average similarity across all

pairs in the subset $\mathcal{S}$:

$$\overline{\cos_4}(\mathcal{S}) \;=\; \frac{1}{\binom{J}{2}} \sum_{\substack{k,k' \in \mathcal{S} \\ k>k'}} \cos_4\big(\mathbf{g}_k, \mathbf{g}_{k'}\big). \tag{2.6.4}$$

A lower value of $\overline{\cos_4}(\mathcal{S})$ indicates a higher degree of gradient diversity, which can be beneficial in non-i.i.d. settings. We will demonstrate in subsequent sections that this multi-client criterion maintains the advantages of the pairwise approach while scaling to larger subsets.

Generalizing Beyond a Single Layer. Next, while Section 2.6.2 considered only one layer of the model, Equation (2.6.4) extends naturally to deeper networks. We simply compute each $\cos_4(\mathbf{g}_k, \mathbf{g}_{k'})$ over multiple layers (or an appropriately weighted sum of per-layer similarities) and then aggregate those into an overall $\overline{\cos_4}$ score. This allows our method to capture gradient diversity spanning the entire neural architecture, rather than focusing on a single layer.

AoU-Queue for Client Rotation. Finally, to avoid selecting the same clients repeatedly—especially in highly heterogeneous scenarios—we introduce an *Age-of-Update Queue*, or *AoU-Queue*. Whenever a client is chosen for transmission, we place it in the AoU-Queue for a fixed number of rounds, preventing its immediate re-selection. Mathematically, let L be the length of the queue, then if a user is selected for transmission at time t, it will be available for selection at time t' with $t' > t + L/S$. Conceptually, this provides a "cool-down period," ensuring that we periodically sample less frequently chosen clients.

Equipped with these three ingredients—multi-user selection, multi-layer gradients, and an AoU-Queue—we now proceed to present the full $\mathsf{cos_4{-}select}$ algorithm in

detail. Algorithm 1 outlines the full procedure.

Each client k first compresses its local gradient $\mathbf{g}_{kt}$ into a lower-dimensional sketch $\mathbf{s}_{kt} = \phi(\mathbf{g}_{kt})$ and sends it to the PS. Based on these summaries, the PS approximates the $\cos_4$ similarity for every pair of clients not currently in the queue. It then selects a subset $\mathcal{S}_t$ of size J by *maximizing negative alignment* (i.e., minimizing $\overline{\cos_4}$). Those J clients are enqueued for ℓ rounds to avoid repeated selection, while they transmit their *full* gradients to the PS. Finally, the server aggregates these gradients, updates the global model, and broadcasts it back to all clients. As illustrated in Fig. 2.6 , introducing a nonzero queue length significantly improves overall accuracy for different shard configurations, highlighting the benefit of balancing gradient-based selection with controlled client rotation.

By mixing a negative $\cos_4$-based alignment criterion with a simple queue mechanism to ensure rotation, $\mathsf{cos_4{-}select}$ balances gradient diversity with fair client participation. Our experiments confirm that this approach achieves both faster convergence and higher accuracy in strongly heterogeneous FL environments.

2.7 Experiments

2.7.1 Settings

Experiments are conducted on the CIFAR-10 [90] and Fashion-MNIST datasets [204] with an **ImageNet pre-trained** VGG16 network [165], where the feature extraction layers are frozen and only the classifier layers are fine-tuned. Each experiment is repeated with 10 random seeds to ensure reliability. To evaluate the performance of our client selection strategy, we compare it against three baselines: FedCor [181],

AFL [56] and POWER-OF-CHOICE [79].

To compare the $\cos_4$ client selection method with relevant baselines in heterogeneous federated learning (FL) settings, the datasets are partitioned using the Partition by Shards (PS) method. Specifically, the dataset is divided into $K \times S = 10 \times S$ shards, where $K = 10$ represents the number of clients and S is a hyperparameter controlling the level of heterogeneity (with lower S indicating higher non-IIDness). Within each shard, data points share identical labels. Our method was evaluated under three heterogeneity levels (S=1, 2, 5) and two selection configurations: performing selection on only the layer 6 (one-layer) and on the layers 3 and 6 (two-layer) of the VGG classifier.

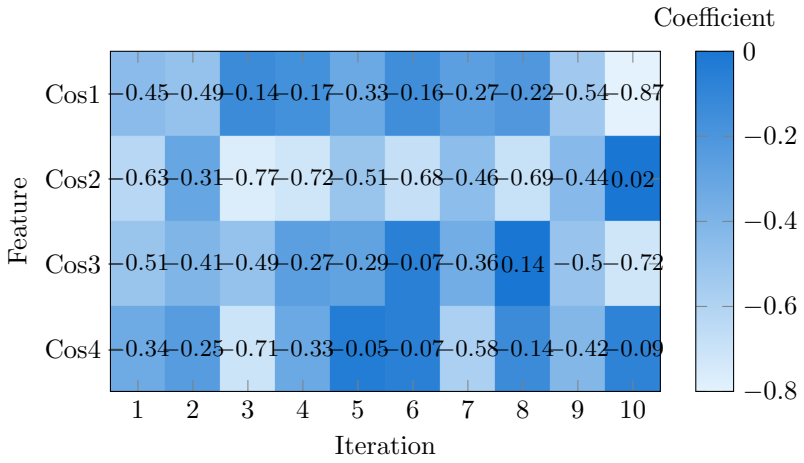


Figure 2.2: Comparation of cosine similarities for shard $j = 1$ when choose 2 clients from 10 on CIFAR10 dataset.

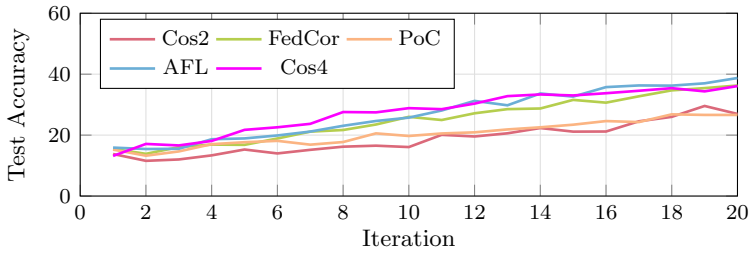


Figure 2.3: Comparing our Cos4 for shard $j = 1$ against Cos2 and other SOTA methods.

Table 2.2: Comparison of $\cos_4$ with the baselines under various experimental settings.

Charac.	Type	$\cos_4$	AFL	FedCor	PoC	$\cos_4$	AFL	FedCor	PoC
		One Layer				Two Layers			
Shard	1	**29.29**	28.24	25.47	27.95	**28.94**	28.68	23.69	27.27
	2	**42.68**	38.99	39.91	38.83	**43.53**	39.03	39.22	39.26
	5	**56.46**	53.79	54.62	53.56	**56.60**	52.16	52.83	52.97
Iteration	3	**34.82**	31.65	32.42	32.53	**36.07**	31.30	32.84	33.25
	5	**40.88**	37.77	36.89	36.17	**40.42**	35.66	35.54	35.68
	7	**44.22**	40.35	40.41	41.41	**44.78**	41.01	38.16	38.89
	9	**46.51**	44.70	42.61	43.67	**47.13**	45.13	42.01	43.88
All	Avg	**42.81**	40.34	40.00	40.11	**43.02**	39.96	38.58	39.83
	Std	7.91	8.73	7.97	8.30	8.08	8.75	7.47	7.85

2.7.2 Compare to SOTA

CIFAR-10 dataset with 2 clients selected

Table 2.2 demonstrates that our $\cos_4$ method consistently outperforms the baseline methods on CIFAR-10 under different data heterogeneity (Shard=1,2,5). At the top half of the table, we display the converged test accuracy of each method. The results show that $\cos_4$ converges to the best test accuracy compared to the baseline methods across all three heterogeneity settings. At the bottom half of the table, the test accuracy (averaged across three heterogeneity settings) at selected iterations

is illustrated. Also, Figure 2.3 shows a heatmap of the coefficients $\mathbf{w}$ for different iterations $t \in [10]$ for shard $j = 1$, choosing $J = 2$ users over $K = 10$ averaged across $R = 10$ seeds. Our $\cos_4$ method demonstrates robust performance across the entire converging process. Extending to the two-layer configuration, while the baseline methods generally result in a loss of performance, our $\cos_4$ method may achieve higher test accuracy.

CIFAR-10 dataset with 4 clients selected

In Figure 2.4, we compare our method with the baseline methods when 4 out of 10 total clients need to be selected. Although all four methods converge to similar final test accuracies after 20 iterations, our method excels the other in terms of converging speed. While the baseline methods all take more than 5 iterations to reach 40% test accuracy, our method achieves that at the third iteration. This advantage in convergence speed persists until our method is the first to achieve the final converged test accuracy.

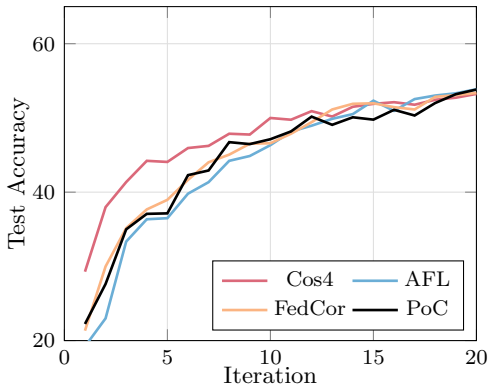


Figure 2.4: Comparison with baselines on the CIFAR-10 dataset for shard number $J = 2$, $S = 4$, $K = 10$, in the one-layer setting.

Fashion-MNIST dataset with 2 clients selected

In order to validate the effectiveness of our method on other datasets, we deploy all three baselines and $\cos_4-\mathsf{select}$ on **Fashion-MNIST** dataset. In Figure 2.5, we present the experiment results under the setting, where local dataset is highly heterogeneous (Shard =1) and 2 out of 10 clients need to be selected for the one-layer setting. By analyzing the training curves, we observe that our $\cos_4$ method starts to outperform the baseline methods starting from the fourth iteration, and our method converges to the best final test accuracy among all comparisons.

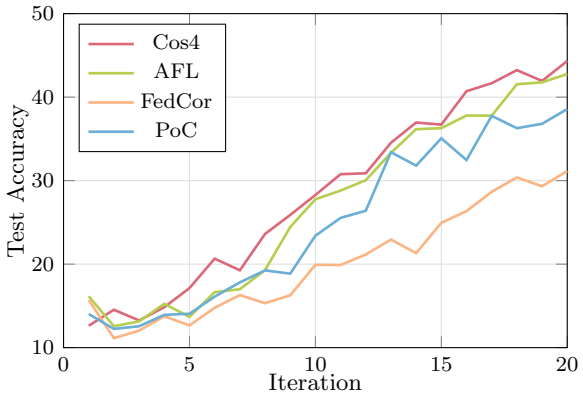


Figure 2.5: Comparison with baselines on the Fashion-MNIST dataset (Shard = 1, selecting 2 clients from 10) with one-layer setting.

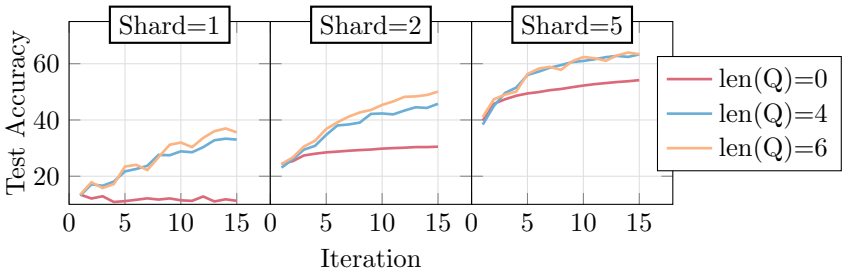


Figure 2.6: Comparison with different queue length using dataset CIFAR-10 for Shard $= 1, 2$ and 5 under GSCCS settings.

Figure 2.6 shows that the choice of queue length significantly influences learning efficiency: Q=4 optimally balances stability and adaptability across shard sizes. For smaller shards (1-2) with higher data heterogeneity, Q=4 maintains competitive convergence accuracy by 10 iterations vs Q=0/Q=6 stagnation) by preserving critical historical updates without over-saturation. In larger shards (5), homogeneous data naturalizes queue length impact, though Q=4 still enables slightly faster early convergence. Excessively long queues (Q=6) consistently underperform, demonstrating diminishing returns in data retention efficiency.

2.8 Conclusion

This paper introduces a novel framework that leverages the cosine similarity between clients' local gradients to select clients, thereby reducing communication costs and improving convergence in heterogeneous federated learning (FL). Specifically, we propose a simple yet highly efficient method for client selection. Extensive experimental results validate the effectiveness of our approach across various FL scenarios. In future work, we will demonstrate that our method remains effective even when using limited summaries of local gradients to determine which clients are selected for global model updates.

The following chapter is reproduced from a submitted IEEE paper:

Liangyan Li, Yimo Ning, Kevin Le, Wei Dong, Yunzhe Li, Jun Chen, and Xiaohong Liu (2025). "MoiréXNet: Adaptive Multi-Scale Demoiréing with Linear Attention, Test-Time Training, and Truncated Flow Matching Prior." arXiv preprint `https://arxiv.org/abs/2506.15929`.

Contribution Declaration:

Liangyan Li is the first author and primary contributor: she conceived the algorithm, implemented the code, conducted all experiments, and drafted the manuscript. Yimo Ning summarized the results. Yizhe reproduced the state-of-the-art (SOTA) methods in the earlier stage of the project. Wei Dong and Xiaohong Liu contributed to discussions. Prof. Jun Chen supervised the research.

Chapter 3

MoiréXNet: Adaptive Multi-Scale Demoiréing with Linear Attention Test-Time Training and Truncated Flow Matching Prior

3.1 Abstract

This paper introduces a novel framework for image and video demoiréing by integrating Maximum A Posteriori (MAP) estimation with advanced deep learning techniques. Demoiréing addresses inherently nonlinear degradation processes, which pose significant challenges for existing methods. Traditional supervised learning approaches either fail to remove moiré patterns completely or produce overly smooth results. This stems from constrained model capacity and scarce training data, which

inadequately represent clean image distribution and hinder accurate reconstruction of ground-truth images. While generative models excel in image restoration for linear degradations, they struggle with nonlinear cases such as demoiréing and often introduce artifacts.

To address these limitations, we propose a hybrid MAP-based framework that integrates two complementary components. The first is a supervised learning model enhanced with efficient linear attention Test-Time Training (TTT) modules, which directly learn nonlinear mappings for RAW-to-sRGB demoiréing. The second is a Truncated Flow Matching Prior (TFMP) that further refines the outputs by aligning them with the clean image distribution, effectively restoring high-frequency details and suppressing artifacts. These two components combine the computational efficiency of linear attention with the refinement abilities of generative models, resulting in improved restoration performance.

3.2 Introduction

Moiré patterns, caused by interference between grid-like structures such as camera sensors and LED screens [188], are visually disruptive artifacts characterized by wavy lines, ripples, or colorful distortions [179]. These patterns degrade image quality and are challenging to remove due to their complex, content-dependent variations in thickness, frequency, and color, which often blend with fine image details [238].

Conventional demoiréing methods, relying on classical filters or signal decomposition models [163, 177, 217, 86, 203], struggle to handle the non-linear and intricate nature of moiré artifacts. The introduction of paired real [179, 64, 222, 219, 237,

Figure 3.7: Visual comparison of demoiréing methods. (a) Clean, (b) Moiré, (c) PnP Flow, (d) MoiréXNet, (e) MoiréXNet + TFMP.

148, 42, 223] and synthetic datasets [239] has enabled supervised learning methods [179, 102, 220, 34, 62, 54, 64, 105, 237, 109, 238, 190, 219, 145, 188] to achieve notable success in recovering clean sRGB images from corrupted inputs. However, these methods are limited by their reliance on finite training datasets, which fail to capture the true distribution of clean images. This limitation, compounded by the non-linear transformations in the Image Signal Processor (ISP) pipeline, often results in oversmoothed outputs with missing high-frequency details. While some efforts have been made to incorporate frequency-domain information [64, 105, 43, 237, 188] or utilize RAW domain data [222, 223, 208], they still fall short of accurately recovering fine textures and edges, leading to suboptimal restoration quality.

Figure 3.8: Visual comparison of moiré artifact removal and detail preservation: (a) Clean Images, (b) Moiré Images, (c) PnP Flow Matching with Moiré sRGB as inputs, (d) MoiréXNet results (ours), and (e) MoiréXNet results enhanced refinement via TFMP. Using pretrained PnP Flow Matching with a linear kernel on moiré sRGB inputs (d) leads to artifacts like bullring effects due to the nonlinear nature of the moiré pattern. Our framework (d) achieves superior structural fidelity and artifact suppression while preserving high-frequency details. Refining the results of MoiréXNet with TFMP (e) achieves further performance enhancements.

Generative models [58, 88, 71, 130] have shown strong performance in image restoration tasks by leveraging learned priors to recover missing details. For tasks such as denoising, deblurring, and super resolution, plug-and-play (PnP) denoisers [16, 9, 202, 5, 13, 234, 149, 38, 170, 75, 104, 132] are widely adopted for reconstruction. However, these approaches predominantly assume *linear* degradation processes (e.g., additive noise, known blur kernels, uniform downsampling). Moiré patterns, in contrast, pose a fundamentally different and more complex challenge.

They arise from nonlinear interactions between scene textures and sensor sampling grids, resulting in spatially varying aliasing effects that resist closed-form characterization. This inherent nonlinearity limits generative models' ability to disentangle artifacts from true image content, often leading to residual artifacts or hallucinated details in restored images, as illustrated in Figure 3.8, column (c). MRGAN [221], an unsupervised approach based on CycleGAN [244], demonstrates progress in moiré removal by training generators with self-supervised techniques. However, it fails to fully exploit the benefits of supervised learning and clean image priors, which restricts its effectiveness.

In this paper, we propose a generic approach to tackling image/video restoration and demonstrate its effectiveness, particularly in the challenging task of demoiréing. Our contributions can be summarized as follows:

1. **Hybrid MAP-based framework**: We introduce a novel framework that combines supervised learning with generative priors to address the nonlinear and non-stationary nature of moiré degradation.

2. **Efficient Test-Time Training (TTT) modules**: We incorporate linear attention TTT modules into a supervised model, enabling efficient and robust RAW-to-sRGB demoiréing through direct nonlinear mappings.

3. **Flow Matching generative prior**: We leverage a TFPM to refine restoration outputs, effectively aligning them with the clean image distribution to recover high-frequency details and suppress artifacts.

4. **State-of-the-art performance**: Our approach demonstrates superior results on benchmark datasets, achieving significant improvements in quantitative metrics (e.g., PSNR) and visual quality over prior methods.

3.3 Related Works

3.3.1 Moiré Pattern Removal

Moiré patterns result from the interference of similar frequencies, degrading the quality of screen captures. A moiré pattern remover restores clean images or videos by eliminating these patterns and correcting color deviations. Conventional methods [159, 163] primarily focus on specific types of moiré patterns. In contrast, supervised learning-based image demoiréing approaches excel at learning diverse moiré patterns. This progress has been driven by the availability of high-quality moiré datasets [222, 237, 219, 179, 65, 42, 223] and advancements in deep learning backbones [68, 183, 116, 178].

For image demoiréing, most approaches utilize Convolutional Neural Network (CNN)-based architectures, integrating multiscale features [102, 179, 34, 62, 237, 33, 108, 187], attention mechanisms [85, 206] (e.g., channel, spatial, and color) and frequency-domain techniques [33, 237, 64, 105, 187, 122, 238] to tackle the complex patterns of moiré artifacts. 3DNet [190] leverages both spatial- and frequency-domain knowledge through a dual-domain distillation network. DDA [235] focuses on efficient image demoiréing for real-time applications. The aforementioned methods predominantly operate in the sRGB domain, where the ISP discards much of the original sensor information through processes such as tone mapping, white balance, and compression. In contrast, RAW-domain images retain unprocessed sensor data, preserving richer details and naturally exhibiting reduced moiré patterns. RDNet [222] introduces the first RAW-domain demoiréing dataset, leveraging the richer information available in RAW images and incorporating a multi-scale encoder with multi-level

feature fusion. However, despite employing a class-specific learning strategy to handle different types of screen content, it lacks flexibility when applied to diverse scenes, ultimately limiting its generalization capability. Studies such as [222, 210, 208] explore image demoiréing in both RAW and sRGB domains. However, these methods struggle to handle diverse and complex scenarios effectively.

For video demoiréing, VDmoiré [42] introduces the first dedicated dataset and a baseline model, while RawVDemoiré [223] proposes a temporal alignment method specifically for RAW video demoiréing. Compared to their image demoiréing counterparts, which primarily focus on feature extraction and fusion, video demoiréing methods [209, 146, 106, 36, 150, 42] emphasize leveraging temporal information from neighboring frames and aggregating multi-frame features to enhance the quality of the restored video frames.

A key challenge in both image and video reconstruction lies in extracting rich features that can effectively capture spatial and temporal dependencies. Transformers [183, 50] have consistently outperformed CNN-based models [164, 67] across a wide variety of tasks [193, 124, 126, 24, 111, 110, 125, 113, 192–194], thanks to their ability to effectively capture global dependencies through self-attention mechanisms. Such mechanisms rely on the Key-Value (KV) cache to store historical context, with the attention output at time t given by:

$$z_t = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{3.3.1}$$

where the softmax operation computes the attention weights. Self-attention explicitly stores all historical context, resulting in memory requirements that grow linearly with the sequence length $O(t)$ and computational complexity of $O(t^2)$ due to pairwise

interactions between all tokens. In contrast, Mamba RNNs [207, 59] are an effective state space model with linear computation complexity. Mamba RNNs leverage selective state spaces(input-dependent gating) to achieve linear-time complexity ($O(t)$) while maintaining global receptive fields. Mamba dynamically adjusts its state transition parameters based on the input, enabling both hardware-aware efficiency (via parallel scan operations) and data-dependent context compression.

More recently, TTT blocks [178, 236] have emerged as a method to bridge the gap between Transformers and RNNs by employing efficient parametric updates. TTT avoids maintaining a growing KV cache by using a parametric hidden state s_t, which is updated as follows:

$$s_t = f(s_{t-1}, x_t; W), \tag{3.3.2}$$

where s_{t-1} is the previous hidden state, x_t is the current token input, and W represents the learned parameters. The output is generated as:

$$z_t = g(s_t; W). \tag{3.3.3}$$

This hidden state is iteratively updated at each time step, representing a compressed summary of all previous tokens. TTT does not maintain an explicit KV cache, resulting in a fixed-size representation with $O(1)$ memory requirements. While TTT is highly efficient for processing long sequences, it is typically less expressive than full self-attention mechanisms.

In this work, we adopt TTT linear attention layers as our primary building blocks, leveraging their long-range attention capabilities while maintaining computational efficiency. Frequency-domain features play a crucial role in computer vision

tasks [23, 215, 225, 151, 37]. To harness the global representational power of the frequency domain and mitigate the attention layers' inherent bias toward low-frequency features, we introduce a Learnable Frequency Enhanced Filter (LFEF) block before the TTT linear attention layer. LFEF adaptively enhances both high- and low-frequency components, ensuring a balanced and enriched feature representation for improved downstream processing. We further integrate Invertible Neural Networks (INNs) [45, 87, 44] during preprocessing to improve the feature extraction ability of the TTT backbone. INNs are invertible models that excel in image processing tasks [246, 78, 57, 6] by enabling efficient, lossless transformations and high-fidelity reconstructions. Comprehensive architectural details are provided in Section 3.4.1, with quantitative and qualitative performance evaluations discussed in Section 4.5. The efficiency of the TTT linear attention layers is reflected in lower computational complexity and faster inference times as shown in Table 3.3, while still achieving high restoration performance as evidenced by competitive PSNR and SSIM scores.

3.3.2 Pretrained Image Priors for Image Restoration

Although traditional supervised image demoiréing methods achieve high PSNR [201], their results often suffer from oversmoothing artifacts [77]. This limitation stems from two key factors: (1) Euclidean distance-based loss functions in neural networks prioritize pixel-wise fidelity at the expense of perceptual quality, and (2) constrained dataset diversity restricts the model's capacity to learn accurate image mappings.

To address these challenges, recent approaches have integrated prior knowledge of natural image statistics into restoration pipelines. Image restoration is typically formulated as an inverse problem, where the goal is to recover a clean image x from

noisy observations y based on the forward model: $y = Hx + n$, where H is the forward operator, which is typically a *linear* operation (e.g., a blurring matrix or downsampling operator), and n represents additive noise, often assumed to follow a Gaussian distribution. One common approach to solving this inverse problem is to optimize a regularized objective function:

$$\hat{x} = \arg\min_x \frac{1}{2}\|Hx - y\|_2^2 + \lambda\Phi(x). \tag{3.3.4}$$

The first term, $\frac{1}{2}\|Hx - y\|_2^2$, enforces consistency with the observed data (data fidelity term), while the second term, $\lambda\Phi(x)$, encodes prior knowledge about the image (regularization term), such as smoothness or sparsity. Traditional methods [153, 245, 55, 66, 155] relied on explicit mathematical models of natural image statistics, such as Fourier spectrum [154], total variation (TV) [155, 20, 48, 49, 142], sparsity priors [128, 22, 19, 155], and patch-based Gaussian mixtures priors [248, 104] to guide the restoration process. The Plug-and-Play (PnP) [185, 175] framework revolutionized image restoration by decoupling the prior from the forward model. Instead of explicitly defining a regularization term $\Phi(x)$, PnP leverages powerful image denoising algorithms as implicit priors. For example, advanced nonlearned denoisers like BM3D [21] and CNN-based denoiser [227, 7, 29, 121, 136, 228] have been used for this purpose. Many modern learned priors exploit the capabilities of generative models, which excel at capturing complex natural image distributions. Generative models such as GANs [58], VAEs [88], diffusion models [172, 71, 167], and normalizing flows [45, 87] have shown immense potential in this regard, making them valuable tools for modeling priors [16, 9, 202, 5, 13, 234, 149, 38, 170]. A recent extension of

normalizing flows, known as flow matching [100, 112], optimizes transport paths between distributions and has been explored for various image restoration tasks within PnP frameworks [132].

Even though PnP denoisers excel in addressing linear degradation tasks such as denoising, deblurring, and super-resolution, effectively balancing fidelity and perceptual quality, their effectiveness is significantly constrained in handling nonlinear degradation tasks like demoiréing or JPEG artifact removal. As shown in Fig. 3.8, when applied to image demoiréing, PnP-Flow tends to introduce artifacts due to the inherent complexity of the degradation process. To overcome this limitation, we propose using TFMP as a refinement step to enhance the demoiréed images produced by a supervised model.

3.4 Methodology

Moiré removal involves recovering a clean image x from a degraded observation $y = M(x) + n$, where $M(\cdot)$ represents a *nonlinear*, scene-dependent degradation caused by interference between high-frequency textures and sampling grids. In this section, we present our hybrid approach: A supervised learning model in Section 3.4.1, enhanced with efficient linear attention TTT modules, directly learns nonlinear mappings for RAW-to-sRGB demoiréing. This stage incorporates INN and LFEF modules to refine features in both the spatial and frequency domains, effectively removing coarse patterns while preserving structural content. A truncated flow matching model in Section 3.4.2 further refines the outputs by aligning them with the clean image distribution. This step restores high-frequency details and suppresses residual artifacts through distribution matching, enabling photorealistic texture recovery.

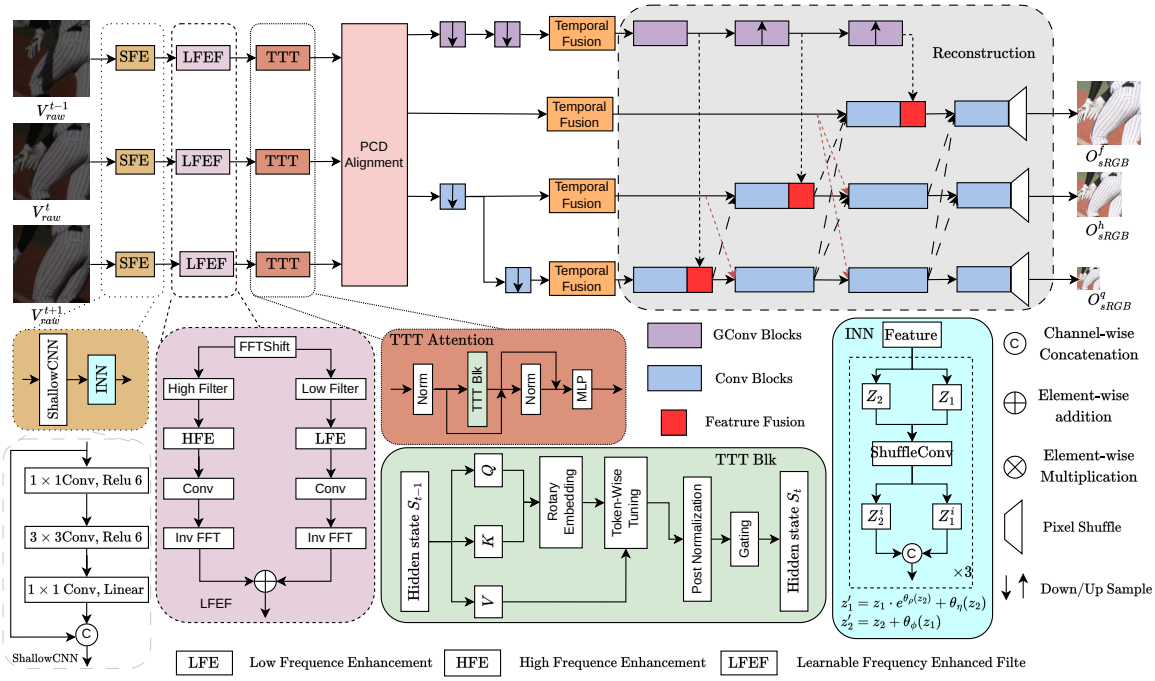


Figure 3.9: An overview of the proposed method.

3.4.1 Demoiréing Network Architecture

Our demoiréing framework builds upon the VDRaw framework [36], replacing the convolutional layers in the preprocessing phase with a combination of ShallowCNN and an INN module to enhance high-frequency detail preservation. Furthermore, we adaptively combine the frequency domain features by learning a weighted filter before applying TTT linear attention for multi-scale feature extraction. In our model, we selected the TTT 1B version and adjusted the hidden size to 256.

Fig. 3.9 provides an overview of our proposed **MoiréXNet** model for video demoiréing. More specifically, the key stages include Shallow Feature Extraction (SFE), Deep Feature Extraction (DFE), Auxiliary Frames Alignment and Blend (AFAB) and Hierarchical Reconstruction (HR). The model takes three neighboring

RAW images, $\mathbf{V}_{raw}^{i} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times 4}$, where $i \in \{t-1, t, t+1\}$, in the 4-channel RGGB format as input and directly reconstructs the corresponding central frame sRGB image $\mathbf{O}_{sRGB}^{t}$. For clarity and consistency, we adopt the same notation as VDRaw.

The SFE block is designed to extract shallow features, denoted as $\mathbf{F}_{\text{raw}}$, from the RAW input data $\mathbf{V}_{\text{raw}}$. The SFE block uses lightweight convolutional layers to embed raw input frames (reference + auxiliary frames) into an initial feature space. Then, an INN module, as proposed in [236], further ensures lossless information preservation by enabling the mutual reconstruction of input and output features. This unique property allows the INN to serve as a lossless feature extraction mechanism, ensuring that all critical information is retained for the subsequent DFE process. The DFE block generates multi-scale features by downsampling shallow features using bilinear interpolation with scaling factors of 0.5 and 0.25. At each level, the combined LFEF and TTT Linear Attention blocks extract deep features from each frame. The LFEF addresses the limitations of attention layers, which tend to lose high-frequency information, by extracting richer features and enhancing overall model performance. Detailed architecture is shown in Figure 3.9.

At each scale, TTT linear attention blocks capture long-range dependencies to enhance global context understanding while maintaining fine-grained details. The multi-scale features are finally fused using the VDRaw strategy, ensuring structural consistency and effective feature integration.

We have enhanced the feature extraction blocks compared to the VDRaw framework, while keeping the feature alignment and reconstruction blocks unchanged. Following the approach in [42, 196, 208], we employ pyramid cascading deformable (PCD)

alignment to align the features of $\mathbf{V}_{raw}^{t-1}$ and $\mathbf{V}_{raw}^{t+1}$ with those of $\mathbf{V}_{raw}^{t}$. For reconstruction, we adhere to the original VDRaw framework, where the final output consists of three different resolutions of sRGB images: full resolution, half resolution, and quarter resolution $\mathbf{O}_{sRGB}^{f}$, $\mathbf{O}_{sRGB}^{h}$, $\mathbf{O}_{sRGB}^{q}$, which are used to calculate the multiscale loss.

For image demoiréing tasks, features are extracted from a single input frame, bypassing the need for the PCD module. Instead, multi-scale features are fused using TTT linear attention blocks before being passed to the reconstruction backbone for image restoration.

3.4.2 Flow Matching for Iterative Refinement of Degraded Images

We propose an iterative refinement process leveraging a TFMP to enhance the outputs of the MoiréXNet model. The denoiser learns a velocity field that maps degraded images to clean ones, enabling a stepwise recovery that progressively brings images closer to the ground truth. We initialize the iterative refinement process with a degraded image $\tilde{x}$, which serves as the initial estimate of the clean image x, as $\tilde{x}$ is significantly closer to x compared to the moiréimage y. Thus, we set $x_t = \tilde{x}$, with t starting from a higher value (e.g., $t = 0.95$) rather than 0. Here, $t \in [0.95, 1]$ represents the progression through the refinement process, with five samples drawn at each step. A refined version of the MoiréXNet model's output is obtained by applying a few iterations of the denoiser.

The flow matching model learns a velocity field $\frac{\partial x_t}{\partial t} = v(x_t, t)$, which defines the gradient direction guiding the degraded image toward the clean image at timestep

t. In our approach, this velocity field is iteratively applied to refine the MoiréXNet model's output by updating x_t as follows:

$$x_{t-1} = x_t + \Delta t \cdot v(x_t, t), \tag{3.4.1}$$

where Δt is the step size.

The flow matching model is pretrained to learn the transformation dynamics from degraded images to clean images, ensuring robust refinement. The iterative process progressively improves the image, making it cleaner and closer to the ground truth. The method works for a wide variety of image degradation types, such as noise, blur, and compression artifacts.

3.5 Experiments

Table 3.3: A comparison of state-of-the-art methods for image and video demoiréing in the context of Raw Video Demoiréing, evaluated using average PSNR, SSIM, LPIPS, and computational complexity. The best results are bolded in red, while the second-best results are bolded in black. This table highlights the state-of-the-art performance of our model, MoiréXNet, in both image and video demoiréing tasks.

	Method	Input type	PSNR↑	SSIM↑	LPIPS↓	Inference time (s)
Image	RDNet [222]	RAW	25.892	0.8939	0.1508	2.514
	RRID [210]	sRGB+RAW	**27.283**	**0.9029**	**0.1168**	**0.501**
	MoiréXNet	RAW	**29.590**	**0.9170**	**0.0936**	**0.070**
Video	VDMoiré [42]	sRGB+RAW	27.277	0.9071	0.1044	1.057
	VDMoiré∗	sRGB+RAW	27.747	0.9116	0.0995	1.125
	DTNet [209]	sRGB	27.363	0.8963	0.1425	0.972
	DTNet∗	sRGB	27.892	0.9055	0.1135	1.050
	VDRaw [36]	sRGB+RAW	28.706	0.9201	0.0904	1.247
	DemMamba [207]	RAW	30.004	0.9169	**0.0901**	**0.446**
	MoiréXNet	RAW	**30.127**	**0.9258**	**0.0847**	**0.070**
	TFMP	RAW	**30.214**	**0.9281**	0.0973	**-**

We compare the proposed MoiréXNet and its refined version, TFMP, with state-of-the-art methods and evaluate their performance on both video and image demoiréing tasks.

3.5.1 Experimental Setup

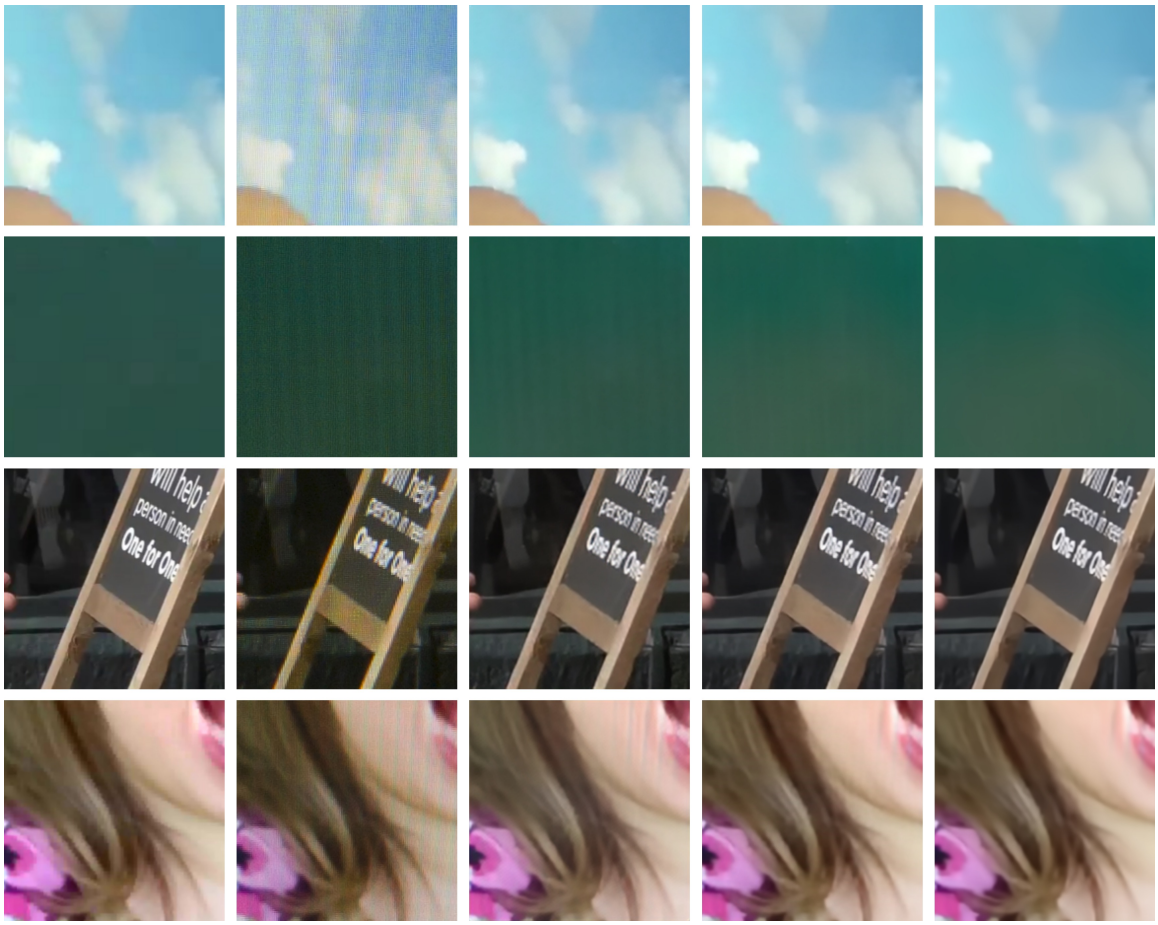
Training Details. Our experiments are conducted on a machine equipped with two NVIDIA A100 GPUs. We train our methods using the AdamW optimizer with an initial learning rate of 3×10^{-4}, betas (0.9, 0.999) for momentum and variance smoothing, and weight decay to improve generalization. The learning rate is adjusted using the ReduceLROnPlateau scheduler, which reduces it by a factor of 0.8 if validation loss does not improve for 3 consecutive epochs, with a minimum learning rate of 5×10^{-6}. This setup ensures efficient and stable training with adaptive learning rate adjustments. We begin training with L1 VGG loss for 175 epochs, followed by fine-tuning with wavelet loss for an additional 41 epochs to further refine the results.

3.5.2 Datasets

For the RAW-domain image and video demoiréing task, we conduct experiments using the RawVDemoiré dataset [36] and the TMM22 dataset [222]. The VDMoiré dataset includes 300 training videos and 50 testing videos, each with 60 frames at a resolution of 1080×720 (720p). For image demoiréing, the TMM22 dataset provides 540 RAW and sRGB image pairs for training and 408 pairs for testing, with image patches cropped to 256×256 for training and 512×512 for testing. In both datasets, RAW moiré inputs are compared against sRGB ground truth images to evaluate the performance of the proposed method.

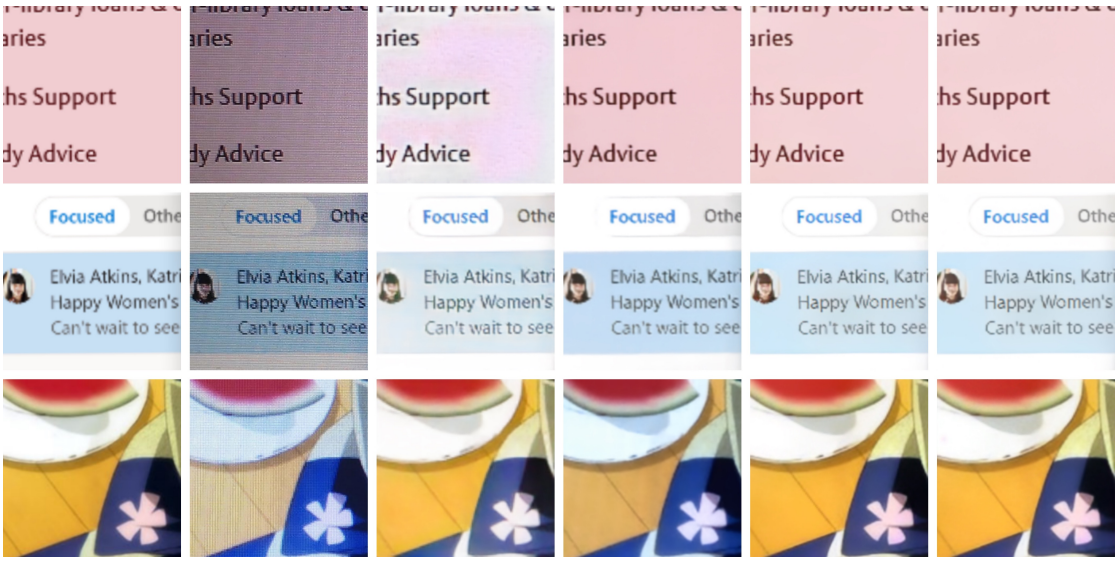
Table 3.4: Quantitative comparison with the state-of-the-art demoiréing approaches and RAW image restoration methods on TMM22 dataset [222] in terms of average PSNR, SSIM, LPIPS and computational complexity. The best results are bolded in red, while the second-best results are bolded in black.

Index	Methods										
	DMCNN [179]	WDNet [105]	ESDNet [219]	RDNet [222]	VDRaw [36]	CR3Net [168]	RRID [210]	DemMamba [207]	MoiréXNet Ours	TFMP Ours
# Input type	sRGB	sRGB	sRGB	RAW	RAW	sRGB+RAW	sRGB+RAW	RAW	RAW	RAW
PSNR↑	23.54	22.33	26.77	26.16	27.26	23.75	27.88	**28.14**	28.03	**28.08**
SSIM↑	0.885	0.802	0.927	0.921	0.935	0.934	**0.938**	0.936	**0.937**	**0.938**
LPIPS↓	0.154	0.166	0.089	0.091	0.075	0.102	0.079	**0.067**	**0.066**	**0.067**



(a) Clean Images (b) Moiré Images (c) VDRaw (d) MoiréXNet (e) TFMP

Figure 3.10: Qualitative comparison on RAW video demoiréing RawVDemoiré [36].



(a) Clean Images (b) Moiré Images (c) RDNet (d) RRID (e) MoiréXNet (f) TFMP

Figure 3.11: Qualitative comparison on RAW image demoiréing TMM22 dataset [222].

3.5.3 Loss Function

Relying solely on pixel-wise losses in the sRGB domain, such as L_1 or L_2, is often insufficient. We combine L_1 loss and VGG-based perceptual loss as follows:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{vgg}}\mathcal{L}_{\text{vgg}}(I_{\text{pred}}, I_{\text{gt}}) + \lambda_{\ell_1}\mathcal{L}_{\ell_1}(I_{\text{pred}}, I_{\text{gt}}), \tag{3.5.1}$$

where $\lambda_{\text{vgg}} = 0.3$ and $\lambda_{\ell_1} = 0.7$. The L_1 loss is:

$$\mathcal{L}_{\ell_1} = \|I_{\text{pred}} - I_{\text{gt}}\|_1. \tag{3.5.2}$$

The perceptual loss is computed using a pre-trained VGG-16 [164] network:

$$\mathcal{L}_{\text{vgg}} = \sum_{l \in \mathcal{F}} \|\phi_l(I_{\text{pred}}) - \phi_l(I_{\text{gt}})\|_1, \tag{3.5.3}$$

where ϕ_l represents the activation from layer l, and $\mathcal{F}$ is the set of feature layers.

3.5.4 Evaluation

For quantitative comparison, we use PSNR [201], SSIM[201], and LPIPS [233] to evaluate image quality. PSNR assesses pixel-level fidelity but overlooks structural and perceptual aspects. SSIM incorporates luminance, contrast, and structure, offering better alignment with human perception while remaining pixel-based. LPIPS leverages deep features to assess semantic distortions, enabling robust perceptual evaluation and broad use in demoiréing tasks [64, 233, 219, 205]. Additionally, we evaluate model efficiency by reporting inference time for a comprehensive analysis.

Table 3.3 compares the performance of various methods for image and video demoiréing based on PSNR, SSIM, LPIPS and inference time on RawVDemoiré dataset [223]. The results demonstrate that MoiréXNet outperforms all other methods, excelling in both reconstruction quality and computational efficiency, and establishes a new benchmark for RAW image and video demoiréing tasks. For image demoiréing, MoiréXNet achieves a PSNR of 29.590 dB, SSIM of 0.9170, and LPIPS of 0.0936, significantly outperforming RDNet [222] and RRID [210]. Specifically, MoiréXNet's PSNR is +3.698 dB higher than RDNet (25.892 dB) and +2.307 dB higher than RRID (27.283 dB), while its SSIM is +0.0231 dB higher than RDNet (0.8939) and +0.0141 dB higher than RRID (0.9029). the LPIPS is 0.0572 lower than RDNet and 0.0232 lower than RRID. For video demoiréing, MoiréXNet achieves a PSNR of 30.127 dB, an SSIM of 0.9258, and an LPIPS of 0.0847 while maintaining an efficient inference time of 0.070 seconds. When refined with PnP flow matching, MoiréXNet achieves state-of-the-art results with a PSNR of 30.214 dB, surpassing

DeMMamba [207] (30.004 dB) by +0.21 dB. It also achieves the highest SSIM of 0.9281, outperforming VDRaw[36] (0.9201) by +0.008, and achieves the lowest LPIPS score of 0.0795, which is 0.0054 lower than DeMMamba (0.0901).

These results demonstrate that the refined model leverages PnP flow matching to further enhance reconstruction quality. Both MoiréXNet variants excel in balancing computational efficiency and performance, with inference times competitive against lighter models. However, the increase in LPIPS alongside improved PSNR and SSIM after PnP flow matching refinement suggests that while the refinement enhances pixel-wise accuracy and structural fidelity, it may also introduce visually unnatural artifacts. Figure 3.10 illustrates the effectiveness of our model in removing moiré patterns directly from RAW moiré images, without relying on the sRGB color space for color correction. The results demonstrate that our model not only eliminates moiré artifacts but also preserves accurate colors, avoiding any visible color distortions or artifacts.

For the TMM22 dataset, which focuses on image demoiréing, we compare MoiréXNet with state-of-the-art methods, as presented in the table 3.4. As mentioned, the TTMM22 dataset only contains 540 images for training, which limits the model's ability to generalize effectively. Nevertheless, MoiréXNet achieves competitive performance against state-of-the-art methods, as illustrated in Figure 3.11.

3.5.5 Ablation Study

1) Ablation study on the model architecture. Table 3.5 highlights the significance of each block in the MoiréXNet architecture. The full model, which integrates INN, LFEF, and TFMP, achieves the highest reconstruction quality, as reflected in

Figure 3.12: Denoiser iterations vs PSNR.

the best PSNR and SSIM results. This underscores the necessity of combining these components to attain state-of-the-art performance on the VDraw dataset.

Specifically, including the INN block improves performance with a PSNR increase of +0.99 and an SSIM increase of +0.004. Adding the LFEF block on top of INN further enhances results, contributing a PSNR increase of +0.09 and an SSIM increase of +0.015. Finally, incorporating the TFMP block alongside INN and LFEF results in an additional PSNR increase of +0.09 and an SSIM increase of +0.003.

Table 3.5: Ablation study on Model Architecture.

Models	PSNR↑	SSIM↑
MoiréXNet (w/o INN, LFEF and TFMP)	29.04	0.906
MoiréXNet (w INN and w/o LFEF and TFMP)	29.36	0.910
MoiréXNet (w INN and LFEF w/o TFMP)	30.12	0.925
MoiréXNet (w INN LFEF and TFMP)	**30.21**	**0.928**

2) Optimal t for Flow-Matching Denoising. In flow-matching denoising, the parameter t determines the progression of the algorithm toward the clean image. Since $\tilde{x}$ is already close to the clean image, fewer iterations are required. The Figure 3.12 demonstrates that the PSNR peaks around iteration 15, where the algorithm achieves

optimal performance. This indicates that $\tilde{x}$ is approximately at $t = 0.98$. To avoid overshooting the peak, we set $t = 0.95$ for our method.

3.6 Conclusion

We proposed a hybrid approach for nonlinear moiré removal by combining an efficient supervised model with a denoising-based generative procedure, improving restoration quality and offering insights for handling linear and nonlinear degradations. Future work will focus on adaptive techniques to refine the data fidelity gradient, further enhancing the supervised model's performance.

The following chapter is reproduced from a submitted IEEE paper:

Liangyan Li, Kevin Le, Ruibin Li, Matthew Ferreira, Wei Dong, Jun Chen, and Xiangyu Xu. "Solving the Last Mile Problem of Image Restoration with Rectified Flow," submitted to IEEE Transactions on Image Processing and currently under review.

Contribution Declaration:

Liangyan Li is the first author and primary contributor: she conceived the algorithm, implemented the code, conducted all experiments, and drafted the manuscript. Kevin Le evaluated and summarized the results. Ruibin Li and Matthew Ferreira assisted in the later stages by preparing the results. Wei Dong contributed to discussions in the early stages, and Xiaohong Liu provided guidance and contributed to discussions throughout the experiments and the draft of the paper. Prof. Jun Chen supervised the research.

Chapter 4

Solving the Last Mile Problem of Image Restoration with Rectified Flow

4.1 Abstract

Image restoration seeks to reconstruct a clean image X from a degraded observation Y. Classical supervised models minimize distortion-based losses, yielding outputs that remain faithful to Y but usually lack perceptual quality. In contrast, generative approaches produce visually plausible samples aligned with the natural image distribution, yet these samples may contradict the observed evidence Y. This trade-off highlights a method that bridges the gap between distortion minimization and distributional realism. We frame this as a last mile problem, where the goal is to learn a deterministic rectified flow that transports MMSE estimates $Z \sim p_Z$ to clean samples $X \sim p_X$. This approach simultaneously minimizes the expected squared error and

aligns the output distribution with the natural image manifold. We provide theoretical proof and empirical validation of our methods on image super-resolution and image demoireíng tasks.

4.2 Introduction

Image restoration, *i.e.*, recovering a clean image from a corrupted/partial observation, is a fundamental problem in computer vision and signal processing. It lies at the core of numerous applications, from medical diagnostics [123, 242] and satellite imaging [161, 191] to photography [216, 211, 212] and digital forensics [53, 186]. The challenge is deeply rooted in the ill-posed nature of the task: given a degraded image, there are infinitely many plausible clean counterparts, making it difficult to produce results that are both accurate and perceptually convincing.

Most existing methods attempt to balance two competing goals: minimizing distortion (*e.g.*, via mean squared error, MSE) [47, 227] and maximizing perceptual realism [93, 212, 195]. However, these objectives are often at odds. Approaches that minimize MSE tend to produce blurry outputs that average over multiple plausible reconstructions. Conversely, methods that prioritize realism — through adversarial losses [93, 212] or powerful diffusion priors [157, 52, 247] — often generate visually pleasing results that diverge from the actual ground truth. Achieving both goals in a unified, principled manner has remained an open challenge.

In this paper, we revisit image restoration from a probabilistic and information-theoretic perspective. We ask: *How accurately can we reconstruct the original image while ensuring that the predicted images are indistinguishable from real, clean images?* Inspired by the Universal Rate-Distortion-Perception Theorem [226], we formalize this

as a perfect-perception-constrained optimization problem:

$$\min_{p_{\hat{X}|Y}: p_{\hat{X}} = p_X} \mathbb{E}[\|X - \hat{X}\|^2], \tag{4.2.1}$$

where X, Y, and $\hat{X}$ denote the clean image, the degraded observation, and the restored output, respectively. The constraint $p_{\hat{X}} = p_X$ enforces that the restored outputs lie on the natural image manifold, ensuring perceptual realism.

Our key insight is that this problem admits a theoretically optimal decomposition into two distinct stages. We prove that the optimal solution is obtained by (i) computing the minimum mean square error (MMSE) estimate $\mathbb{E}[X|Y]$, followed by (ii) applying an optimal transport map to align the distribution of MMSE outputs with the true data distribution. We refer to this second step — transforming the distortion-minimizing MMSE output into a realistic image — as the last mile of image restoration.

To address this last mile, we introduce a novel use of rectified flow [112, 101, 4], a flow-based generative model that learns continuous transport maps between arbitrary distributions, as illustrated in Figure 4.13. As proved in [107], rectified flow provides an effective mechanism for aligning distributions, making it particularly well-suited to our problem. Crucially, this establishes, for the first time, a complete, theoretically grounded solution to the perception-constrained image restoration problem.

Our main contributions are:

- We provide a foundational framework and a novel solution for **balancing the distortion-perception tradeoff** in image restoration, which minimizes distortion under a distribution-matching constraint, ensuring both fidelity and realism.

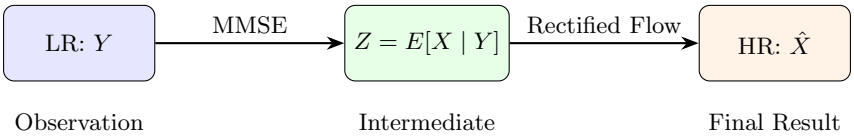


Figure 4.13: Overview of the proposed framework for IR, formulated as a composition of MMSE estimation followed by an optimal transport map, implemented using Rectified Flow.

- We prove that this problem admits **a unique and optimal solution**: a two-stage process combining MMSE estimation with an optimal transport map.
- We propose a **rectified flow-based** method to solve the last mile of image restoration and demonstrate its effectiveness both quantitatively and qualitatively.

4.3 Related Work

4.3.1 Image Restoration

Classical image restoration methods [84, 74, 184] frame the task as a least-squares or MAP estimation problem under Gaussian noise assumptions, minimizing pixel-wise reconstruction losses such as mean squared error (MSE). While these approaches guarantee stability and data consistency through convex optimization, they tend to average over all plausible solutions, producing oversmoothed outputs that lack fine-grained detail [227, 119]. This reflects the intrinsic perception–distortion tradeoff [15], where optimizing distortion metrics often degrades perceptual quality.

Modern generative models aim to overcome this tradeoff by introducing learned image priors. GAN-based methods [197, 229, 58] synthesize sharper textures via

Table 4.6: Summary of rectified flow experiments with different source and conditioning strategies. Here, `N` represents noise, and `NC` denotes no condition. Our proposed method, $Z2X \mid NC$, utilizes VAE(Z) as the source and operates without conditioning.

Exp	**Source** x_0	**Condition**	**Target** x_1
$Z2X \mid NC$	$x_0 = \text{VAE}(Z)$	None	$x_1 = \text{VAE}(X)$
$Z2X \mid Y$	$x_0 = \text{VAE}(Z)$	VAE(Y)	$x_1 = \text{VAE}(X)$
$N2X \mid Y$	$x_0 \sim \mathcal{N}(0, I)$	VAE(Y)	$x_1 = \text{VAE}(X)$
$Y2X \mid NC$	$x_0 = \text{VAE}(Y)$	None	$x_1 = \text{VAE}(X)$
$N2X \mid Z$	$x_0 \sim \mathcal{N}(0, I)$	VAE(Z)	$x_1 = \text{VAE}(X)$

adversarial training, but suffer from instability, mode collapse, and hallucinated artifacts. PLUSE [138] explores the latent space of a pretrained StyleGAN [81] to produce realistic super-resolved outputs, employing a downscaling loss to ensure that the generated high-resolution image, once downscaled, aligns with the low-resolution observation. Likelihood-based models, including VAEs [88, 182] and normalizing flows [46, 87], enable stable optimization via variational or exact likelihoods but typically yield blurry outputs due to limitations in decoder expressivity. Denoising Diffusion Probabilistic Models (DDPM) [70, 157] have recently set new benchmarks in image generation by denoising Gaussian noise through a stochastic process. However, DDPM are computationally expensive at inference time and often exhibit weak fidelity to the input in restoration settings, making them less practical when strict data consistency is needed. Latent Diffusion Models (LDMs)[152] mitigate this cost by learning and operating in a lower-dimensional latent space. To further accelerate inference, methods such as Denoising Diffusion Implicit Models (DDIM) [169, 18, 39], consistency models (CMs)[173, 171], and diffusion distillation[137, 158, 241, 120] have been proposed. These models improve determinism and reduce sampling steps, but

still depend on stochastic iterative dynamics [82], making precise control over reconstruction fidelity and consistency challenging in IR tasks.

Moreover, the accumulation of inference errors during stochastic denoising can lead to uncontrolled artifacts and degraded fidelity [27, 99, 35]. To address this, DDNM[198] confines denoising to the null space of degradation operators, ensuring strict data consistency. SNIPS[83] blends diffusion sampling with iterative projections to jointly enforce fidelity and perceptual quality. Other approaches attempt to improve robustness by decoupling the reverse process[96], adapting noise schedules[40, 2, 156, 98, 174], or introducing regularization mechanisms[129]. Additionally, HyperDiffusion[72] proposes a hypernetwork to fuse multiscale information across diffusion stages, further enhancing restoration accuracy. Diff-Plugin [117] equips a single pretrained diffusion model with a lightweight dual-branch Task-Plugin that injects task-specific priors and guides the diffusion process for high-fidelity results.

4.3.2 Flow-Based Models for Image Restoration

Flow-based models offer a deterministic and efficient alternative to stochastic generation. Flow matching (FM) [101] learns a time-dependent velocity field $v(x, t)$ that continuously transports a simple initial distribution (e.g. Gaussian noise) into a target data distribution via an ODE-driven flow, which is widely used in vision tasks, such as [60, 160] . Unlike DDIM, which follows fixed, curvilinear diffusion trajectories (e.g., linear noise schedules) requiring numerous iterative steps for high-fidelity generation, FM theoretically enables straighter probability paths. However, in high-dimensional settings, FM's ODE dynamics can exhibit numerical stiffness, necessitating small

integration time steps that may compromise sampling efficiency.

Rectified flow [112] simplifies FM by enforcing approximately straight-line transport paths from source to target. It avoids unnecessary curvature, crossing, or intersection in the velocity field, resulting in faster and more stable generation. Rectified flow has demonstrated strong performance in high-resolution image generation and has been adopted in recent large-scale models, such as Stable Diffusion 3 (SD3) [51], FlowIE [247], FLUX [91] and PMRF [147].

Our contribution to this paper: (1) We provide new theoretical justification that the rectified flow from Z to X—formulated as an optimal transport problem—can provably approach the constrained optimum of the distortion-perception trade-off, despite the lack of a deterministic relationship between X and Z in typical restoration scenarios. (2) We conduct extensive experiments with SD3 to learn rectified flows in latent space and evaluate performance across multiple tasks, demonstrating that our method consistently achieves robust and stable refinement under diverse degradation conditions, including complex demoiréing scenarios.

4.4 Method

4.4.1 Preliminary: Universal Rate-Distortion-Perception

In image restoration, it is crucial to jointly balance distortion (*e.g.*, mean squared error) and perceptual quality (how realistic the output appears). A foundational result in this direction is the Universal Rate-Distortion-Perception Theorem [226], which characterizes the achievable distortion–perception tradeoffs for any learned representation.

Let X denote a clean image and Y be a distorted version of X. Define $Z = \mathbb{E}[X|Y]$, the optimal reconstruction of X under MMSE. Then the set of achievable distortion–perception pairs, denoted $\Omega(p_{Y|X})$, satisfies the following inclusion:

$$\Omega(p_{Y|X}) \subseteq \left\{(D, P) : D \geq \mathbb{E}[\|X - Z\|^2] + \inf_{p_{\hat{X}}: d(p_X, p_{\hat{X}}) \leq P} W_2^2(p_Z, p_{\hat{X}})\right\} \subseteq \mathrm{cl}(\Omega(p_{Y|X})), \tag{4.4.1}$$

where $d(\cdot, \cdot)$ is any divergence-based perception metric and W_2^2 denotes the squared 2-Wasserstein distance. This result implies that the minimal distortion achievable by any representation is lower-bounded by the MMSE estimation Z, plus an additional cost for making the reconstructed distribution perceptually match the data distribution.

In particular, the point $(D, P) = (\mathbb{E}[\|X - Z\|^2] + W_2^2(p_Z, p_X),\ 0)$ represents the optimal achievable performance under a perfect perception constraint, where the reconstructed samples are indistinguishable from the true data distribution (*i.e.*, $p_X = p_{\hat{X}}$). However, achieving this point in practice remains a significant challenge.

4.4.2 The Last Mile Problem

A generic image restoration problem can be formulated as (4.2.1). It is worth noting that enforcing $p_{\hat{X}} = p_X$ helps preserve the perceptual quality of the reconstruction [14]. A potential solution to (4.2.1) is given by posterior sampling with $p_{\hat{X}|Y}$ chosen to coincide with $p_{X|Y}$. This choice automatically ensures $p_{\hat{X}} = p_X$. A score-based diffusion posterior sampling method was recently proposed in [213]. We will demonstrate that posterior sampling is generally suboptimal for (4.2.1) and characterize the architectural principles underlying the optimal solution.

Let $Z := \mathbb{E}[X|Y]$. Due to the conditional independence of X and $\hat{X}$ given Y, we have

$$\mathbb{E}[\|X - \hat{X}\|^2] = \mathbb{E}[\|X - Z\|^2] + \mathbb{E}[\|Z - \hat{X}\|^2]. \tag{4.4.2}$$

If posterior sampling is used, then $p_{XZ} = p_{\hat{X}Z}$ and consequently

$$\mathbb{E}[\|X - \hat{X}\|^2] = 2\mathbb{E}[\|X - Z\|^2]. \tag{4.4.3}$$

In other words, the end-to-end distortion achieved by posterior sampling is twice the distortion achieved by the Minimum Mean Squared Error (MMSE) estimate of the clean image based on the degraded version. This result aligns with the findings in Ohayon et al. [147].

On the other hand, since $p_{\hat{X}} = p_X$, it follows that

$$\mathbb{E}[\|Z - \hat{X}\|^2] \geq W_2^2(p_Z, p_X), \tag{4.4.4}$$

where $W_2(p_X, p_Z)$ denotes the Wasserstein-2 distance between p_Z and p_X. Therefore, we have

$$\mathbb{E}[\|X - \hat{X}\|^2] \geq \mathbb{E}[\|X - Z\|^2] + W_2^2(p_Z, p_X). \tag{4.4.5}$$

This lower bound is in fact the minimum achieveable distortion for the image restoration problem in (4.2.1) and can be attained by first computing the MMSE estimate Z of X based on Y and then converting Z to $\hat{X}$ using the optimal transport plan associated with $W_2(p_Z, p_X)$.

The suboptimality of posterior sampling can be inferred from the fact that

$$\mathbb{E}[\|X - Z\|^2] > W_2^2(p_Z, p_X) \tag{4.4.6}$$

in general. Indeed, under certain regularity conditions [17], the transport plan that attains $W_2(p_Z, p_X)$ is determinstic whereas in most image restoration problems, due to the stochastic nature of the degradation kernel, X and Y (and consequently, X and Z) are not determinstically related. Moreover, we show in Appendix A.1 that even in scearios where Y is a low-dimensional projection of X—a situation commonly encountered in super-resolution and image inpainting—the inequality in (4.4.6) remains typically strict.

Given (4.4.5), the optimal image restortion scheme consists of two steps: MMSE estimation followed by optimal transport. The MMSE estimation step can be accomplished using conventoinal supervised training. However, the subsequent step, which converts the MMSE estimate Z into a perceptually perfect reconstruction $\hat{X}$, involves solving a challenging optimal transport problem. Here, we observe that with a well-chosen training set, the learned MMSE estimate Z is already quite close to X, implying that $W_2(p_Z, p_X)$ is also small. As a result, only a short-distance transportation is required, which we refer to as the "last mile problem".

4.4.3 Solving Last Mile with Rectified Flow

In this paper, we will address the "last mile problem" using the rectified flow approach. Specifically, we train a rectified flow model using paired (X, Z), where X is a clean image from the training set and Z is the corresponding MMSE estimate obtained from the first step. Note that X can be viewed as a sample from p_X while Z can be viewed

as a sample from p_Z. Here X and Z are not independent, but are jointly distributed according to p_{XZ}. By training a rectified flow, we can obtain a new coupling of p_X and p_Z, denoted by $\tilde{p}_{XZ}$, such that

$$\mathbb{E}_{\tilde{p}}[\|X - Z\|^2] \leq \mathbb{E}_p[\|X - Z\|^2], \tag{4.4.7}$$

where $\mathbb{E}_{\tilde{p}}[\|X - Z\|^2]$ is the incurred distortion when converting Z to $\hat{X}$ using the trained rectified flow. Note that the inequality (4.4.7) implies that the end-to-end distortion achieved by the rectified flow approach, which is $\mathbb{E}_p[\|X - Z\|^2] + \mathbb{E}_{\tilde{p}}[\|X - Z\|^2]$, is at least as small as that achieved by posterior sampling, which is $2\mathbb{E}_p[\|X - Z\|^2]$, although there is no guarantee that $\mathbb{E}_{\tilde{p}}[\|X - Z\|^2]$ can reach $W_2^2(p_Z, p_X)$.

Rectified flow evolves from the MMSE estimate Z to the clean target X along a straight-line path over time $t \in [0, 1]$. Unlike stochastic score-based methods, rectified flow learns a globally consistent velocity field that aligns source and target distributions via linear supervision. Specifically, we generate interpolated samples $Z_t = (1 - t)Z + tX$, and train the model to match the constant displacement vector $X - Z$:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(Z,X),\, t\sim\mathcal{U}[0,1]} \left[\|v_\theta(Z_t, t) - (X - Z)\|^2\right]. \tag{4.4.8}$$

This deterministic rectified flow efficiently transports MMSE estimates Z to clean images X, suitable for the short-range transport required in "last-mile" refinement.

4.5 Experiments

4.5.1 Model Architecture and Training

We utilize Stable Diffusion 3 (SD3) [51], adapting its rectified flow-based latent transformer architecture for image restoration tasks. Prompt conditioning is disabled during finetuning by setting encoder contexts to zero. Training and inference are conducted entirely in the latent space using the pretrained SD3 VAE. During training, we randomly crop images to 256×256. At test time, all evaluations are performed on full-resolution images. We use the AdamW optimizer with a batch size of 2 and a fixed learning rate of 5×10^{-5} to train the fusion module. All experiments are conducted on a single NVIDIA A100 GPU with 80GB of memory. Specifically, we finetune the following components:

- **ControlNet:** Encodes either the degraded observation Y or a null condition as auxiliary guidance.
- **Transformer Backbone:** Processes either the MMSE estimate Z or a noise input to predict velocity fields $v_\theta(x_t, t)$ for latent-space rectified transport.

Our proposed approach employs a null condition (NC) in the ControlNet module and utilizes the MMSE estimate Z in the Transformer backbone. We compare our method against several alternative experimental setups to evaluate its performance, as summarized in Table 4.6. As the performance of the `N2x-Z` variant is notably inferior to that of other methods, it will be excluded from subsequent comparisons.

4.5.2 Evaluation Metrics

We evaluate the quality of reconstructions using both distortion (PSNR, SSIM [201]) and perception-based metrics (LPIPS [232], FID [69], NIQE [140]). In evaluating our image restoration model $\hat{X} = T(Z)$, we consider metrics that reflect both distortion and perceptual quality. Below, we formalize each metric and highlight its theoretical motivation.

Mean Squared Error (MSE) and PSNR. We define the pixel-wise distortion between a prediction $\hat{x}$ and ground truth x as

$$\text{MSE}(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} \|x_i - \hat{x}_i\|^2, \tag{4.5.1}$$

$$\text{PSNR}(x, \hat{x}) = 10 \cdot \log_{10} \left(\frac{L^2}{\text{MSE}(x, \hat{x})} \right), \tag{4.5.2}$$

where L is the maximum pixel intensity (e.g., 1.0 or 255). While PSNR captures fidelity, it fails to reflect perceptual realism.

LPIPS: Learned Perceptual Image Patch Similarity. We adopt LPIPS [232] to quantify perceptual similarity using deep features:

$$\text{LPIPS}(x, \hat{x}) = \sum_{l} \frac{1}{H_l W_l} \|w_l \odot (\phi_l(x) - \phi_l(\hat{x}))\|_2^2, \tag{4.5.3}$$

where $\phi_l(\cdot)$ denotes the activation of a pre-trained network (e.g., VGG) at layer l, and w_l are learned channel-wise weights.

Frechet Inception Distance (FID). FID evaluates distributional similarity in the feature space of an Inception network. Assuming activations of real and generated images follow Gaussian distributions (μ_r, Σ_r) and (μ_g, Σ_g), the FID is given by:

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right). \tag{4.5.4}$$

It captures both the mean shift and covariance mismatch between the distributions.

Natural Image Quality Evaluator (NIQE). NIQE [140] measures the perceptual quality of images without requiring reference images. It computes the distance between the NSS (natural scene statistics) features of an image and a model trained on pristine natural images. Lower NIQE scores indicate higher perceptual quality:

$$\text{NIQE}(x) = d\big(\text{NSS features of } x, \text{NSS model}\big), \tag{4.5.5}$$

where $d(\cdot)$ represents the distance (e.g., Mahalanobis distance) in the NSS feature space. NIQE is particularly useful for evaluating images in scenarios where ground truth references are unavailable. **Note that $N2X \mid Y$ represents ControlNet.**

4.5.3 Super-Resolution

Dataset. We use the DIV2K [76] dataset with $\times 4$ bicubic downsampling for the training and evaluation of our super-resolution experiments. The initial MMSE estimates Z are generated by state-of-the-art restoration network MambaIRv2 [61], with both Base and Large versions of the model employed to produce diverse initial estimates. The clean images X from the datasets are used as supervision targets during

rectified flow training.

Table 4.7: Performance on DIV2K using MMSE outputs generated by the MambaIRv2-Base model. All models were trained for 1000 epochs and evaluated with 20 sample steps.

Experiment	**PSNR (↑)**	**SSIM (↑)**	**LPIPS (↓)**	**FID (↓)**	**NIQE (↓)**
$Z2X \mid Y$	<u>27.8790</u>	<u>0.7916</u>	<u>0.1177</u>	11.2502	3.3298
$Z2X \mid NC$	**28.1117**	**0.7970**	0.1181	<u>11.1885</u>	3.3980
$N2X \mid Y$	26.1036	0.7320	0.1282	12.7094	<u>2.8927</u>
$Y2X \mid NC$	27.0994	0.7655	0.1752	19.8216	4.1058
$N2X \mid Z$	26.2296	0.7361	**0.1152**	**10.5094**	**2.8568**
$MMSE \quad Z$	29.7606	0.8371	0.2460	30.9108	5.3918
$Degraded \quad Y$	26.8264	0.7567	0.4150	42.0390	7.6337

Table 4.8: Performance on DIV2K using MMSE outputs generated by the MambaIRv2-Large model. All models were trained for 1000 epochs and evaluated with 20 sample steps.

Experiment	**PSNR (↑)**	**SSIM (↑)**	**LPIPS (↓)**	**FID (↓)**	**NIQE (↓)**
$Z2X \mid Y$	<u>27.9193</u>	<u>0.7931</u>	0.1132	11.0818	3.3240
$Z2X \mid NC$	**27.9609**	**0.7936**	**0.1097**	<u>10.3139</u>	3.2825
$N2X \mid Y$	25.8178	0.7255	0.1295	12.3817	**2.7943**
$Y2X \mid NC$	27.3156	0.7737	0.2123	20.4585	4.6745
$N2X \mid Z$	26.4182	0.7422	<u>0.1112</u>	**10.0713**	<u>2.8312</u>
$MMSE \quad Z$	29.9957	0.8409	0.2404	31.2273	5.3406
$Degraded \quad Y$	26.8264	0.7567	0.4150	42.0390	7.6337

Comparison with baselines.

In Tables 4.7 and 4.8, our proposed method $Z2X \mid NC$ consistently improves perceptual metrics—achieving lower LPIPS, FID, and NIQE—while maintaining competitive distortion fidelity, with the best PSNR and SSIM among all tested flows.

Although $N2X \mid Z$ achieves comparable perceptual scores, it suffers from significantly worse distortion metrics, highlighting a suboptimal balance.

The comparison in Figure 4.14 demonstrates that our method produces more detailed facial features, such as human skin texture and spots, whereas PULSE tends to generate smoother and less detailed faces. Qualitative comparisons in Figures 16 and 19 show that our method, $Z2X \mid NC$, produces cleaner outputs with well-preserved details that closely resemble the ground truth X. Visually, the outputs from $Z2X \mid Y$ are comparable to those of $Z2X \mid NC$, suggesting that the additional conditional input offers minimal benefit despite increasing model complexity and computational cost. In contrast, $N2X \mid Y$ and $N2X \mid Z$ yield lower visual quality, with noticeable artifacts—particularly in the last row of Figure 16—underscoring the drawbacks of learning to generate directly from noise instead of from the MMSE output. Finally, $Y2X \mid Z$ consistently performs the worst, with more severe distortions and a marked decline in perceptual quality relative to the MMSE baseline.

Ablation Study

An analysis of Tables 15 and 16 reveals that increasing the number of sampling steps generally improves perceptual quality, as indicated by lower LPIPS, FID, and NIQE scores. However, additional steps lead to diminished reconstruction fidelity, reflected in decreased PSNR and SSIM values. This trade-off suggests that a moderate number of sampling steps offers the most effective balance between perceptual quality and distortion. Empirically, the optimal range appears to lie between 10 and 20 steps, achieving favorable perceptual outcomes while maintaining acceptable distortion levels.

Figure 4.14: Qualitative comparisons on the DIV2K dataset. Our Last Mile (MMSE output generated by the MambaIRv2-Large model) compares with PLUSE.

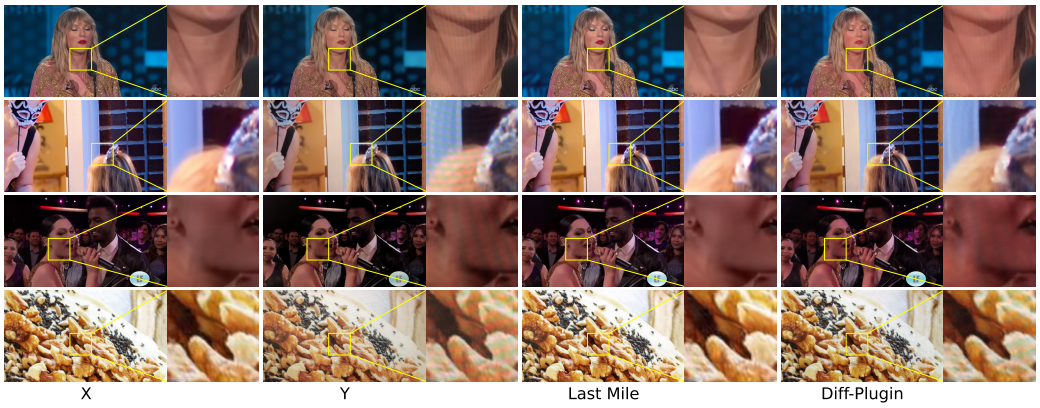


Figure 4.15: Qualitative comparisons on the FHDMi dataset. Our Last Mile (MMSE output generated by the ESDNet-Large model) compares with Diff-Plugin.

Table 4.9: Performance on FHDMi using MMSE outputs generated by the ESDNet-Base model. All models were trained for 90 epochs and evaluated with 20 sample steps.

Experiment	**PSNR (↑)**	**SSIM (↑)**	**LPIPS (↓)**	**FID (↓)**	**NIQE (↓)**
$Z2X \mid Y$	<u>23.5163</u>	<u>0.8148</u>	<u>0.1293</u>	<u>19.3015</u>	5.8283
$Z2X \mid NC$	**23.7475**	**0.8253**	**0.1218**	**15.2406**	6.0788
$N2X \mid Y$	16.9218	0.6519	0.3390	31.9723	**3.8249**
$Y2X \mid NC$	21.9214	0.7970	0.1612	26.6137	5.5821
$N2X \mid Z$	19.1732	0.7060	0.2814	33.1891	<u>4.2557</u>
$Diff - Plugin$	19.5555	0.7502	0.1945	35.9302	5.6319
$MMSE \quad Z$	24.0627	0.8340	0.1357	20.5309	6.8500
$Degraded \quad Y$	17.7393	0.7251	0.2749	44.6488	5.0810

Table 4.10: Performance on FHDMi using MMSE outputs generated by the ESDNet-Large model. All models were trained for 90 epochs and evaluated with 20 sample steps.

Experiment	**PSNR (↑)**	**SSIM (↑)**	**LPIPS (↓)**	**FID (↓)**	**NIQE (↓)**
$Z2X \mid Y$	<u>23.5834</u>	<u>0.8143</u>	<u>0.1226</u>	<u>19.0679</u>	6.0392
$Z2X \mid NC$	**24.1126**	**0.8283**	**0.1107**	**15.4219**	5.9471
$N2X \mid Y$	16.8827	0.6031	0.4018	53.1289	**3.3144**
$Y2X \mid NC$	21.8160	0.7837	0.1704	27.4793	5.3710
$N2X \mid Z$	19.9325	0.7042	0.2573	32.6380	<u>4.4910</u>
$Diff - Plugin$	19.5555	0.7502	0.1945	35.9302	5.6319
$MMSE \quad Z$	24.4255	0.8425	0.1298	20.3144	6.8538
$Degraded \quad Y$	17.7393	0.7251	0.2749	44.6488	5.0810

Compare Last Mile With Other Flows

Across both `DIV2K_bicubic_Base` and `DIV2K_bicubic_Large` experiments at step 20, our proposed *Last Mile* configuration ($Z2X \mid NC$) consistently achieves the best performance across key metrics, including PSNR, SSIM, LPIPS, and FID. It also ranks second in NIQE, demonstrating a strong balance between distortion fidelity and perceptual quality. These results underscore the effectiveness of our final rectified flow setup in delivering high-quality image restoration across different model capacities. Furthermore, comparing the two tables reveals that the further the MMSE estimate deviates from the clean image distribution, the more significant the improvement achieved by the *Last Mile* refinement—highlighting its role in bridging the final perceptual gap.

Sample Steps. Table 11 and Table 12 show that increasing flow steps generally improves perceptual quality (lower LPIPS and FID), but excessive steps (e.g., 50) can degrade FID. Step 20 and 10 offer the best perception-distortion balance, with strong LPIPS and FID scores while retaining moderate distortion fidelity. These results suggest that moderate sampling steps are optimal for rectified flow refinement. Notably, performance degradation at large step counts may arise from accumulated numerical errors and imperfect velocity field estimation over extended integration time, leading to overshooting or drift from the target data manifold.

Qualitative Comparisons Across Flows And Steps. The figures 21 and Figure 20 demonstrate the progression of image reconstructions across different flow configurations for the super resolution task on the DIV2K bicubic dataset. The Last Mile (our) $Z2X \mid NC$ achieve the best balance between distortion metrics (PSNR,

SSIM) and perceptual quality (LPIPS, FID), while $Y2X \mid NC$ prioritizes perceptual realism. Early steps produce coarse results, intermediate steps (e.g., Step 20) show significant quality improvements, and later steps may introduce artifacts.

4.5.4 Image Demoiréing

Dataset. We evaluate our method on the FHDMi dataset [63], which includes real-world images corrupted by complex moiré patterns. To generate the initial MMSE estimates, we use the state-of-the-art demoiréing network ESDNet [218]. Both the Base and Large variants of ESDNet are employed to create diverse and coarse predictions from the degraded images. The corresponding clean images are used as ground-truth references for rectified flow training.

Comparison with Baseline. Tables 4.9 and 4.10 present a comparison of various experimental configurations against the MMSE baseline Z and the degraded observation Y. Relative to the MMSE baselines produced by the Base and Large models, our method $Z2X \mid NC$ significantly improves perceptual metrics: FID is reduced from 20.5309 and 20.3144 to 15.2406 and 15.4219, and LPIPS decreases from 0.1357 and 0.1298 to 0.1218 and 0.1107, respectively. These gains come at only a slight cost in distortion metrics, with PSNR decreasing marginally from 24.0627 dB and 24.4255 dB to 23.7475 dB and 24.1126 dB, and SSIM decreasing from 0.8340 and 0.8425 to 0.8253 and 0.8283. These results demonstrate that $Z2X \mid NC$ offers the best perceptual quality with minimal compromise in reconstruction fidelity. The Diff-Plugin is less effective compared to other methods across all metrics.

Qualitative comparisons in Figures 18 show that our method, $Z2X \mid NC$, produces cleaner outputs with well-preserved details that closely resemble the ground truth X.

In contrast, methods conditioned on the degraded observation Y—such as $Z2X \mid Y$ and $Y2X \mid Z$—retain structural content but introduce moiré artifacts due to the entanglement with degraded inputs. They also fail to recover accurate colors, as illustrated in the last row of Figure 18. From Figure 4.15, we can see that the DiffPlugin failed to remove the moiré patterns.

Ablation Study of Steps. We examine the influence of the number of sampling steps on the distortion–perception tradeoff in Appendix A.2, with the corresponding results shown in Tables 13 and 14. Our proposed method, $Z2X \mid NC$, consistently achieves a strong balance between perceptual quality and distortion fidelity, even when the number of sampling steps is limited. Notably, the results indicate that using approximately 10 sampling steps yields the most favorable tradeoff, highlighting both the efficiency of our approach and its practicality for real-time or resource-constrained applications.

4.5.5 Ablation Study Across Tasks

We conduct an ablation study across both super-resolution and demoiréing tasks to evaluate the effectiveness of our approach.

Results for super-resolution are shown in Figure 23, while those for demoiréing appear in Figure 22 .

In all cases, Z2X | NC consistently outperforms existing baselines. This advantage is particularly pronounced in more challenging tasks such as image demoiréing, where the observation Y is corrupted by complex, structured artifacts (e.g., moiré patterns). In such scenarios, directly conditioning on Y can inadvertently introduce these artifacts into the output. By contrast, our method performs unconditional transport from

Z, effectively avoiding artifact propagation and producing higher-quality restorations. These findings suggest that as degradation complexity increases, the relative benefit of our approach becomes more significant. Furthermore, Z2X | NC obviates the need to encode Y during inference, thereby reducing computational overhead while preserving—or even enhancing—perceptual quality. This makes our method both practical and efficient for real-world restoration applications. For additional comparisons and implementation details, refer to Appendix A.2.

4.6 Conclusion

We proposed a principled two-stage framework for image restoration that leverages rectified flow to refine MMSE estimates through deterministic transport. By formulating the refinement process as an optimal transport problem and solving it via a supervised velocity field, our method achieves a favorable distortion–perception trade-off. Theoretical analysis justifies the use of rectified flow as an efficient and provably optimal refinement mechanism. Extensive experiments on super-resolution task validate the effectiveness and generality of our approach, demonstrating competitive or superior performance compared to existing baselines.

Chapter 5

Conclusion

This thesis explores principled and practical approaches to two long-standing challenges in machine learning systems: *heterogeneous client selection in federated learning*, and perceptual-quality enhancement in image restoration. Rather than treating these problems in isolation, our contributions reflect a broader goal to optimize model performance under distributional shifts and incomplete knowledge, by grounding the solutions in statistical and transport theoretic principles.

We advance the field in three interconnected directions:

1. **Client Selection via Gradient Similarity Modeling**

   We develop a gradient-based client selection framework for federated learning, leveraging higher-order similarity metrics such as the ℓ_4 cosine and composite moment-based distances. This enables a more nuanced understanding of gradient distributions under data heterogeneity.

2. **Efficient Image Restoration Backbones**

For restoration under severe degradations (e.g., moiré, low-resolution, or compound distortions), we design MoiréXNet, a novel backbone integrating:

- *Invertible Neural Networks (INNs)* for lossless feature transformations,
- *Learnable Frequency Enhanced Filters (LFEF)* for frequency-aware feature amplification, and
- *Test-Time Training (TTT)* linear attention modules for efficient and adaptive inference.

This architecture significantly improves restoration fidelity across both spatial and frequency domains while maintaining computational efficiency.

3. **The Last Mile: Rectified Flow for Perceptual Refinement**
   We identify a fundamental bottleneck in generative restoration models—the distortion-perception tradeoff—and propose a two-stage solution inspired by optimal transport theory. The first stage computes the MMSE estimate to minimize distortion, while the second applies a rectified flow model to transport these estimates toward the true clean-image distribution. Unlike standard diffusion or flow matching methods, rectified flow offers deterministic, non-interacting, and cost-efficient trajectories, providing a tractable and provably optimal path toward perceptual realism.

Together, these contributions establish a cohesive framework for model optimization under distributional uncertainty. By combining statistical learning, generative modeling, and optimal transport theory, this thesis provides both theoretical insights and practical tools for improving model generalization, computational efficiency, and perceptual quality in modern AI systems.

Appendix A: Last Mile

A.1 On the Suboptimality of the PULSE Algorithm

Let $Y := HX$, where X is an n-dimensional zero-mean random vector and H is an $m \times n$ matrix. Moreover, let $Z := \mathbb{E}[X|Y]$. The PULSE algorithm [138] employs a generative model with respect to p_X to identify a reconstruction $\hat{X}$ that matches X in the latent space by minimizing $\mathbb{E}[\|H\hat{X} - Y\|^2]$. As the reconstruction $\hat{X}$ obtained through posterior sampling automatically satisfies $H\hat{X} = Y$ almost surely, the PULSE algorithm can be viewed as a specific instance of posterior sampling in this particular context. In light of (4.4.3), the end-to-end distortion achieved by the PULSE algorithm is $2\mathbb{E}[\|X - Z\|^2]$ while the theoretical limit is $\mathbb{E}[\|X - Z\|^2] + W_2^2(p_Z, p_X)$. Therefore, the PULSE algorithm is optimal if and only if $\mathbb{E}[\|X - Z\|^2] = W_2^2(p_Z, p_X)$.

Let $H = U\Lambda V$ be the singular value decomposition of H, where U is an $m \times m$ unitary matrix, Λ is an $m \times n$ diagonal matrix, V is an $n \times n$ unitary matrix. Moreover, we assume that Λ is of the form $\mathrm{diag}(\lambda_1, \ldots, \lambda_k, 0, \ldots, 0)$ with $\lambda_i > 0$ for $i = 1, \ldots, k$. Let $X' := VX$ and $Z' := VZ$. As unitary transformations preserve the Euclidean distance, we have $\mathbb{E}[\|X' - Z'\|^2] = \mathbb{E}[\|X - Z\|^2]$ and $W_2^2(p_{X'}, p_{Z'}) = W_2^2(p_X, p_Z)$.

Denote X' and Z' by $(X'_1, \ldots, X'_n)^T$ and $(Z'_1, \ldots, Z'_n)^T$, respectively. It is easy to verify that $X'_i = Z'_i$ for $i = 1, \ldots, k$, and $Z'_i = \mathbb{E}[X'_i | X'_1, \ldots, X'_k]$ for $i = k+1, \ldots, n$. Clearly, if $(Z'_{k+1}, \ldots, Z'_n) = (X'_{k+1}, \ldots, X'_n)$ (i.e., $(X'_{k+1}, \ldots, X'_n)$ is a function of $(X'_1, \ldots, X'_k)$) or $(Z'_{k+1}, \ldots, Z'_n) = (0, \ldots, 0)$, then $\mathbb{E}[\|X' - Z'\|^2] = W_2^2(p_{X'}, p_{Z'})$ and consequently the PULSE algorithm is optimal. We shall show that when X is Gaussian with a positive definite covariance matrix, this condition is in fact sufficient and necessary for the optimality of the PULSE algorithm.

Assume X is Gaussian with a positive definite covariance matrix. We have $\mathbb{E}[\|X' - Z'\|^2] = W_2^2(p_{X'}, p_{Z'})$ if and only if $(Z'_{k+1}, \ldots, Z'_n) = (X'_{k+1}, \ldots, X'_n)$[1] or $(Z'_{k+1}, \ldots, Z'_n) = (0, \ldots, 0)$.

It suffices to show that $\mathbb{E}[\|X' - Z'\|^2] > W_2^2(p_{X'}, p_{Z'})$ once the condition $(Z'_{k+1}, \ldots, Z'_n) = (0, \ldots, 0)$ is violated.

Since the conditional expectation under the joint Gaussian distribution is linear, we can write

$$Z'_i = a_{1,i} X'_1 + \cdots + a_{k,i} X'_k, \quad k = k+1, \ldots + n. \tag{A.1.1}$$

If the condition $(Z'_{k+1}, \ldots, Z'_n) = (0, \ldots, 0)$ is violated, then there must exist $j \in \{1, \ldots, k\}$ and $i \in \{k+1, \ldots, n\}$ such that $a_{j,i} \neq 0$. So, without loss of generality, we assume $a_{k,n} \neq 0$.

Let $\Delta_k := X'_k - \mathbb{E}[X'_k | X'_1, \ldots, X'_{k-1}]$ and $\Delta_n := X'_n - \mathbb{E}[X'_n | X'_1, \ldots, X'_{n-1}]$. Note that $\gamma_k := \mathbb{E}[\Delta_k^2] > 0$ and $\gamma_n := \mathbb{E}[\Delta_n^2] > 0$ since the covariance matrix of X' is positive definite (which is a consequence of the assumption that the covariance matrix of X

[1] This condition cannot be satisifed under the assumption that the covariance matrix of X is positive definite. So we shall focus on the condition $(Z'_{k+1}, \ldots, Z'_n) = (0, \ldots, 0)$.

is positive definite). Let

$$\tilde{X}_i := X'_i, \quad i = 1, \ldots, k-1, \tag{A.1.2}$$

$$\tilde{X}_k := \mathbb{E}[X'_k | X'_1, \ldots, X'_{k-1}] + \tilde{\Delta}_k, \tag{A.1.3}$$

where

$$\tilde{\Delta}_k := \sqrt{\frac{\gamma_k - \epsilon^2 \gamma_n}{\gamma_k}} \Delta_k + \text{sign}(a_{k,n}) \epsilon \Delta_n. \tag{A.1.4}$$

It can be verified that the covariance matrix of $(\tilde{X}_1, \ldots, \tilde{X}_k)^T$ is the same as that of $(X'_1, \ldots, X'_k)^T$. Moreover, let $\tilde{X}_i = a_{1,i}\tilde{X}_1 + \ldots + a_{k,i}\tilde{X}_k$ for $i = k+1, \ldots, n$. By this construction, the covariance matrix of $\tilde{X} := (\tilde{X}_1, \ldots, \tilde{X}_n)^T$ is guaranteed to the same as that of Z'. Since $\mathbb{E}[\|X' - \tilde{X}\|^2] \geq W_2^2(p_{X'}, p_{\tilde{X}}) = W_2^2(p_{X'}, p_{Z'})$, the problem boils down to showing

$$\mathbb{E}[\|X' - \tilde{X}\|^2] < \mathbb{E}[\|X' - Z'\|^2]. \tag{A.1.5}$$

Clearly, we have

$$\mathbb{E}[(X'_i - \tilde{X}_i)^2] = \mathbb{E}[(X'_i - Z'_i)^2] = 0, \quad i = 1, \ldots, k-1. \tag{A.1.6}$$

Moreover,

$$\begin{aligned}
\mathbb{E}[(X'_k - \tilde{X}_k)^2] &= \mathbb{E}[(\Delta_k - \tilde{\Delta}_k)^2] \\
&= \mathbb{E}\left[\left(\left(1 - \sqrt{\frac{\gamma_k - \epsilon^2\gamma_n}{\gamma_k}}\right)\Delta_k - \text{sign}(a_{k,n})\epsilon\Delta_n\right)^2\right] \\
&= \left(1 - \sqrt{\frac{\gamma_k - \epsilon^2\gamma_n}{\gamma_k}}\right)^2 \gamma_k + \epsilon^2\gamma_n \\
&= \epsilon^2\gamma_n + o(\epsilon^2) \\
&= \mathbb{E}[(X'_k - Z'_k)^2] + \epsilon^2\gamma_n + o(\epsilon^2). \qquad \text{(A.1.7)}
\end{aligned}$$

It can be verified that

$$\begin{aligned}
\mathbb{E}[(X'_i - \tilde{X}_i)^2] &= \mathbb{E}[((X'_i - Z'_i) + a_{k,i}(\Delta_k - \tilde{\Delta}_k))^2] \\
&= \mathbb{E}[(X'_i - Z'_i)^2] + a_{k,i}^2\mathbb{E}[(\Delta_k - \tilde{\Delta}_k)^2] \\
&= \mathbb{E}[(X'_i - Z'_i)^2] + a_{k,i}^2\epsilon^2\gamma_n, \quad i = k+1, \ldots, n-1, \qquad \text{(A.1.8)}
\end{aligned}$$

where the second equality is due to the fact that $\mathbb{E}[(X'_i - Z'_i)\Delta_k] = 0$ (as $X'_i - Z'_i$ is independent of $(X'_1, \ldots, X'_k)$ while Δ_k is a linear combination of $(X'_1, \ldots, X'_k)$) and the fact that $\mathbb{E}[(X'_i - Z'_i)\Delta_n] = 0$ (as Δ_n is independent of $(X'_1, \ldots, X'_{n-1})$ while $X'_i - Z'_i$

is a linear combination of $(X'_1, \ldots, X'_{n-1})$). Finally, we have

$$\begin{aligned}
\mathbb{E}[(X'_n - \tilde{X}_n)^2] &= \mathbb{E}[((X'_n - Z'_n) + a_{k,n}(\Delta_k - \tilde{\Delta}_k))^2] \\
&= \mathbb{E}\left[\left((X'_n - Z'_n) + a_{k,n}\left(\left(1 - \sqrt{\frac{\gamma_k - \epsilon^2\gamma_n}{\gamma_i}}\right)\Delta_k - \text{sign}(a_{k,n})\epsilon\Delta_n\right)\right)^2\right] \\
&= \mathbb{E}[(X'_n - Z'_n)^2] - 2|a_{k,n}|\epsilon\mathbb{E}[(X'_n - Z'_n)\Delta_n] + a_{k,n}^2\mathbb{E}[(\Delta_k - \tilde{\Delta}_k)^2] \\
&= \mathbb{E}[(X'_n - Z'_n)^2] - 2|a_{k,n}|\epsilon\mathbb{E}[X'_n\Delta_n] + a_{k,n}^2\mathbb{E}[(\Delta_k - \tilde{\Delta}_k)^2] \\
&= \mathbb{E}[(X'_n - Z'_n)^2] - 2|a_{k,n}|\epsilon\mathbb{E}[\Delta_n^2] + a_{k,n}^2\mathbb{E}[(\Delta_k - \tilde{\Delta}_k)^2] \\
&= \mathbb{E}[(X'_n - Z'_n)^2] - 2|a_{k,n}|\epsilon\gamma_n + o(\epsilon), \qquad \text{(A.1.9)}
\end{aligned}$$

where the third equality is due to the fact that $\mathbb{E}[(X'_n - Z'_n)\Delta_k] = 0$ (as $X'_n - Z'_n$ is independent of $(X'_1, \ldots, X'_k)$ while Δ_k is a linear combination of $(X'_1, \ldots, X'_k)$), the fourth equality is due to the fact that $\mathbb{E}[Z'_n\Delta_n] = 0$ (as Δ_n is independent of $(X'_1, \ldots, X'_{n-1})$ while Z'_n is a linear combination of $(X'_1, \ldots, X'_{n-1})$), and the fifth equality is due to the fact that $X'_n = \mathbb{E}[X'_n|X'_1, \ldots, X'_{n-1}] + \Delta_n$ and $\mathbb{E}[X'_n|X'_1, \ldots, X'_{n-1}$ is independent of Δ_n. Therefore,

$$\mathbb{E}[\|X' - \tilde{X}\|^2] - \mathbb{E}[\|X' - Z'\|^2] = -2|a_{k,n}|\epsilon\gamma_n + o(\epsilon) < 0 \qquad \text{(A.1.10)}$$

when ϵ is sufficiently close to zero.

Remark 1. The assumption that X has a positive definite covariance matrix can be considerably relaxed.

A.2 Comparison for Image Super-Resolution

Comparison Figures and Tables for Last Mile.

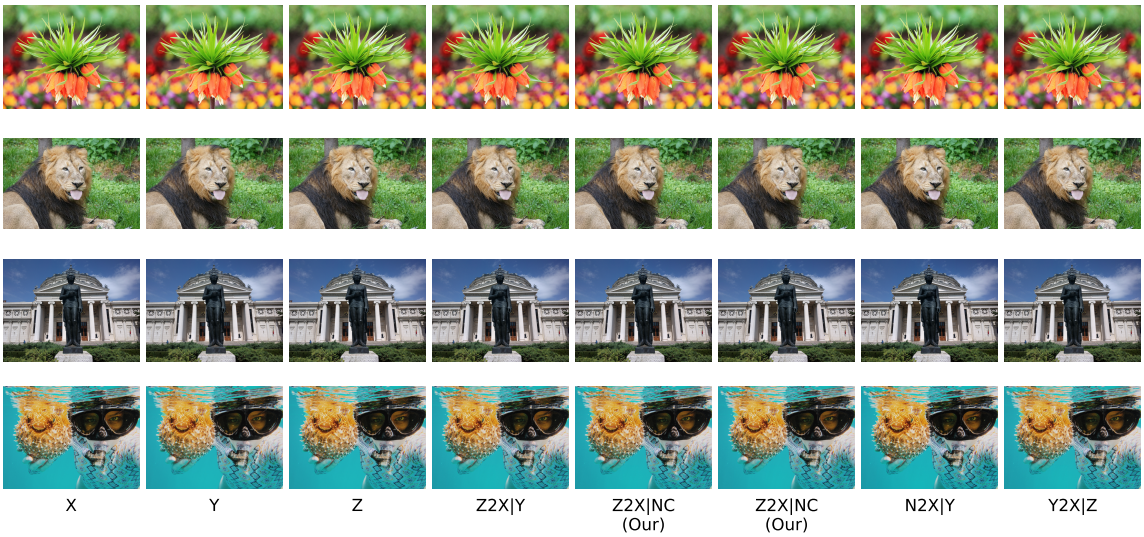


Figure 16: Full-resolution qualitative comparisons across different flow settings, where the MMSE output Z is generated by the MaMbaIRv2 Base model on the DIV2K dataset with unknown degradation.

Table 11: Performance across different flow sampling steps for `DIV2K_bicubic_Base`.

Step	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)	NIQE (↓)
1	**28.96**	**0.8271**	0.2203	0.8139	4.6750
3	<u>28.77</u>	<u>0.8209</u>	0.1960	0.7018	4.3382
5	28.58	0.8154	0.1763	**0.6903**	4.1198
10	28.19	0.8030	0.1368	<u>0.6917</u>	3.6779
20	27.78	0.7880	0.1088	0.7104	3.2444
25	27.67	0.7834	<u>0.1043</u>	0.7243	<u>3.1491</u>
50	27.38	0.7721	**0.1001**	0.7553	**3.0042**

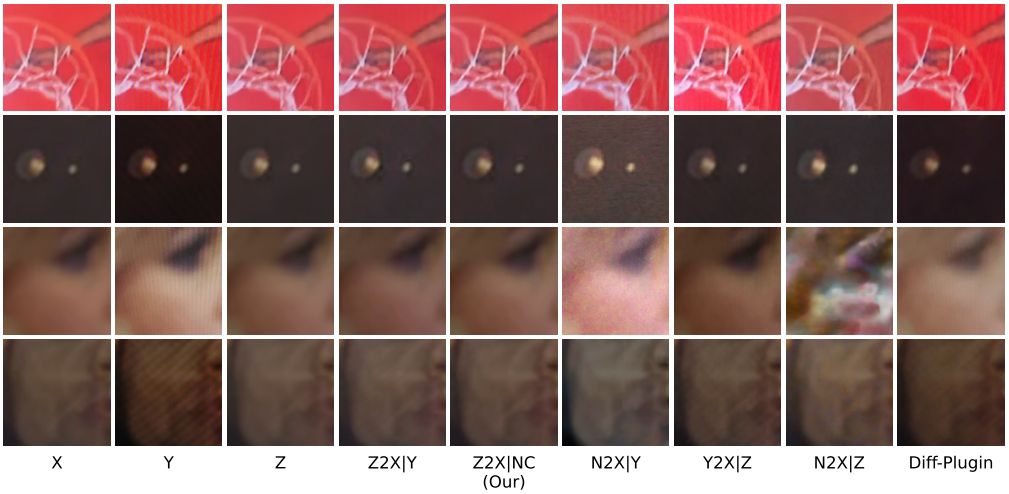


Figure 17: removeQualitative comparisons on the FHDMi dataset using MMSE output generated by the ESDNet-Large model. All models were trained for 90 epochs and evaluated with 20 sample steps.

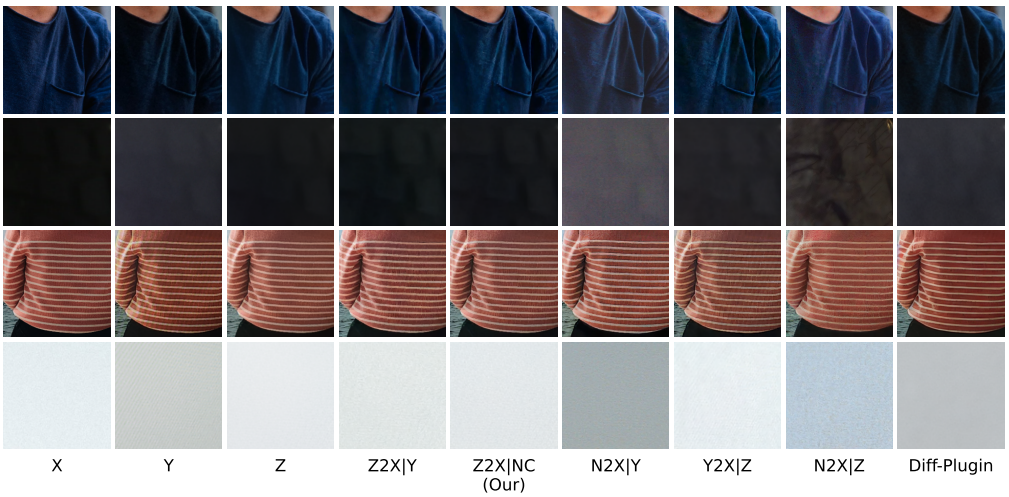


Figure 18: Qualitative comparisons on the FHDMi dataset using MMSE output generated by the ESDNet-Base model. All models were trained for 90 epochs and evaluated with 20 sample steps.

Figure 19: Full-resolution qualitative comparisons across different flow settings, where the MMSE output Z is generated by the MaMbaIRv2 Large model on the DIV2K dataset with unknown degradation.

A.2.1 Comparison Figures and Tables for Image Demoiréing

Tables 13 and 14 report the performance of various flow sampling steps on the FHDMi dataset, specifically for the image demoiriéing task.

A.2.2 Comparison Figures and Tables for Image Super-Resolution

Tables 15 and 16 report the performance of various flow sampling steps on the DIV2K dataset, specifically for the image super-resolution task.

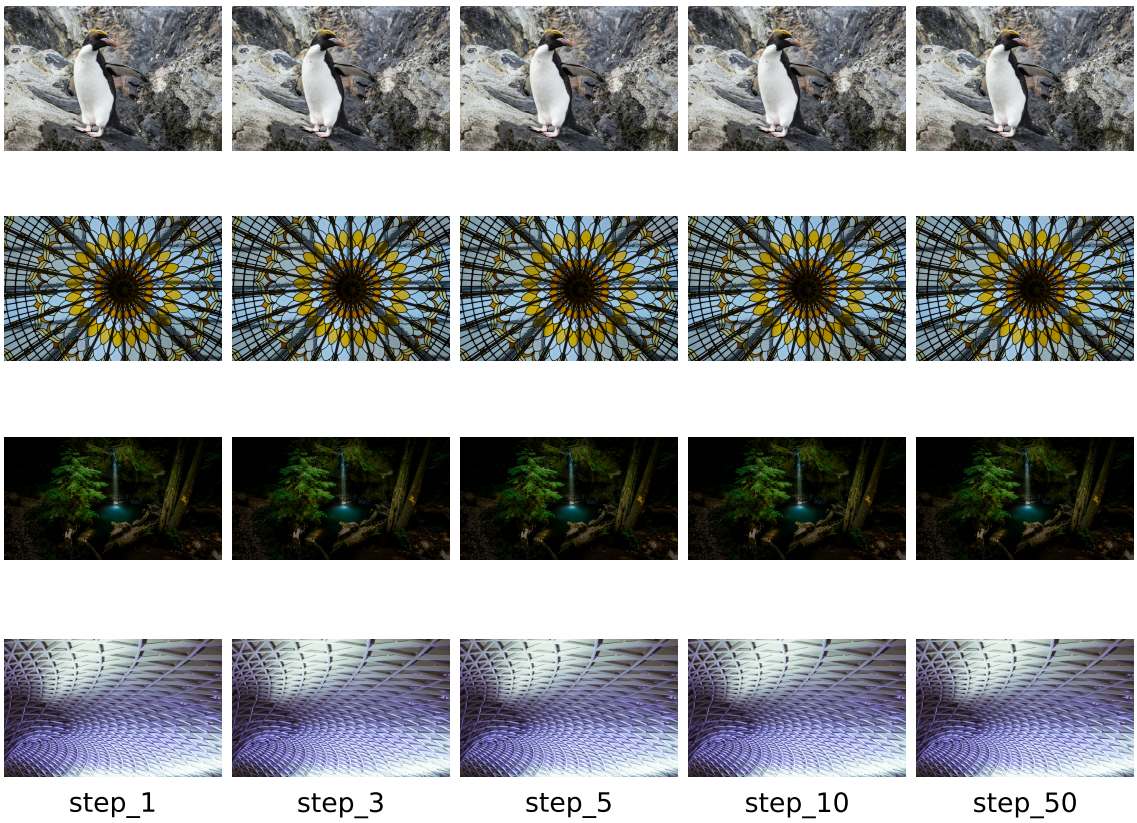


Figure 20: Full-resolution qualitative comparisons across different flow steps, where the MMSE output Z is generated by the MaMbaIRv2 Large model on the DIV2K dataset with bicubic degradation.

Summary

Comparing Figure 23 , we observe that `Z2X|Y` aligns more closely with `Z2X|NC` when the MMSE predictions Z are generated by the larger model. This suggests that conditioning on Y—which may introduce artifacts due to its degraded nature—becomes less critical as the quality of the MMSE estimate Z improves. In other words, a stronger prior (from a larger model) reduces the dependence on the corrupted input, thereby mitigating the propagation of degradation artifacts.

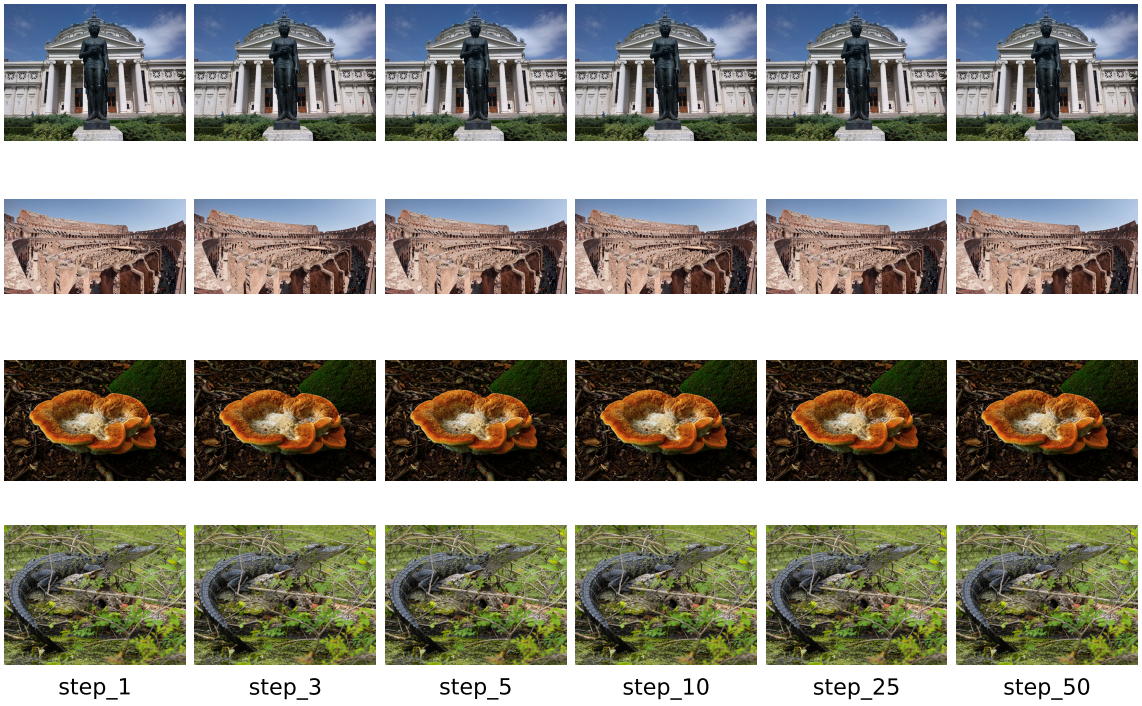


Figure 21: Full-resolution qualitative comparisons across different flow steps, where the MMSE output Z is generated by the MaMbaIRv2 Base model on the DIV2K dataset with bicubic degradation.

However, comparing Figure 22 , in more challenging restoration tasks such as image demoiring—where the degradation patterns (e.g., moiré artifacts) are highly complex and structured—our unconditional refinement method `Z2X|NC` demonstrates a more significant advantage over `Z2X|Y`. This is because conditioning on Y in such cases risks embedding structured artifacts into the final output. Furthermore, by eliminating the need to encode Y during inference, our approach reduces computational cost, offering both practical and performance benefits.

Table 12: Performance across different flow sampling steps for `DIV2K_bicubic_Large`.

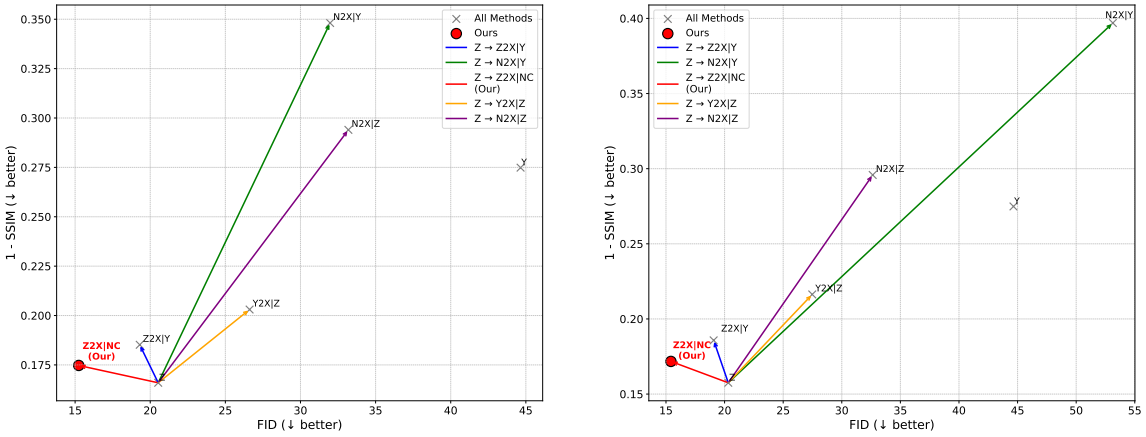
Step	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)	NIQE (↓)
1	**28.94**	**0.8286**	0.2151	0.8381	4.6263
3	<u>28.89</u>	<u>0.8242</u>	0.1921	**0.7141**	4.3393
5	28.74	0.8194	0.1740	<u>0.7344</u>	4.1421
10	28.43	0.8090	0.1389	0.7805	3.7355
20	28.10	0.7972	<u>0.1115</u>	0.8241	<u>3.3232</u>
50	27.81	0.7863	**0.0995**	0.8615	**3.0665**

Table 13: Performance of various sample steps on FHDMi using MMSE outputs generated by the ESDNet-Base model. Model was trained for 90 epochs using the $Z2X \mid NC$ configuration.

Step	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)	NIQE (↓)
1	23.5396	0.8152	0.1567	23.7437	6.4758
3	23.7619	0.8260	0.1340	16.5014	6.1852
5	<u>23.7869</u>	**0.8272**	0.1295	15.8363	6.1666
7	**23.7942**	<u>0.8272</u>	0.1267	15.5214	6.1469
10	23.7839	0.8268	0.1245	15.3882	6.1244
15	23.7646	0.8260	0.1228	15.2707	6.1047
20	23.7475	0.8253	0.1218	<u>15.2406</u>	6.0788
25	23.7444	0.8248	<u>0.1211</u>	**15.2405**	<u>6.0611</u>
50	23.7245	0.8233	**0.1200**	15.3003	**5.9963**

Table 14: Performance of various sample steps on FHDMi using MMSE outputs generated by the ESDNet-Large model. Model was trained for 90 epochs using the $Z2X \mid NC$ configuration.

Step	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)	NIQE (↓)
1	23.6003	0.8167	0.1576	26.7510	6.5276
3	24.0965	0.8310	0.1278	17.3649	6.2396
5	24.1500	**0.8320**	0.1211	16.2115	6.1727
7	**24.1652**	0.8318	0.1171	15.7284	6.1153
10	24.1597	0.8310	0.1141	15.4514	6.0542
15	24.1270	0.8291	0.1116	**15.4080**	5.9767
20	24.1126	0.8283	0.1107	15.4219	5.9471
25	24.0829	0.8271	0.1104	15.5864	5.9121
50	24.0365	0.8250	**0.1102**	15.8627	**5.8565**



(a) PD curve for ESDNet-Base (90 epochs, 20 steps)

(b) PD curve for ESDNet-Large (90 epochs, 20 steps)

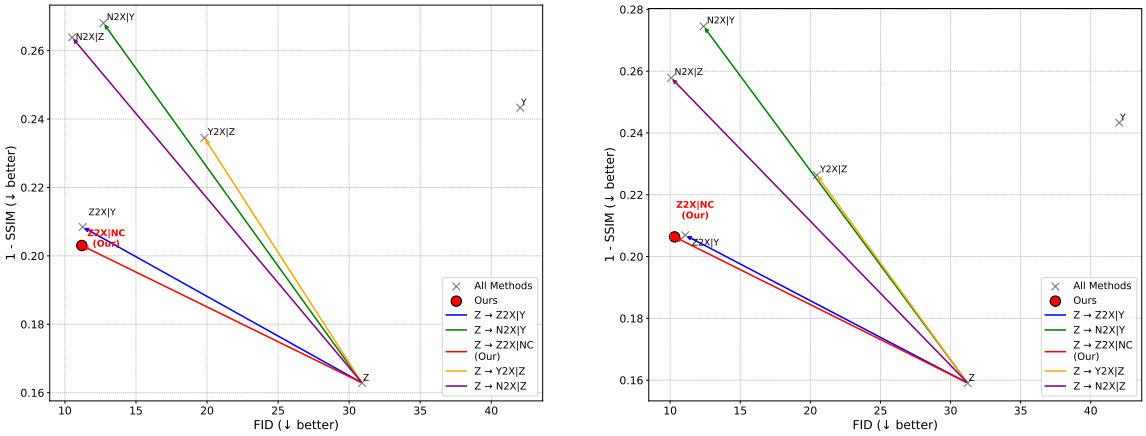
Figure 22: Perception–distortion comparison on FHDMi: left is the Base model (trained 90 epochs, 20 sampling steps), right is the Large model (same training and sampling), showing FID vs. 1–SSIM.

Table 15: Performance of various sample steps on DIV2K using MMSE outputs generated by the MambaIRv2-Base model. Model was trained for 1000 epochs using the $Z2X \mid NC$ configuration.

Step	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)	NIQE (↓)
1	**28.9401**	**0.8265**	0.2174	29.3701	4.6276
3	<u>28.8167</u>	<u>0.8212</u>	0.1952	23.7827	4.3461
5	28.6918	0.8169	0.1786	20.0583	4.1655
7	28.5668	0.8129	0.1634	17.2565	3.9999
10	28.4166	0.8078	0.1456	14.5267	3.7943
15	28.2407	0.8016	0.1279	12.2615	3.5545
20	28.1117	0.7970	0.1181	11.1885	3.3980
25	28.0302	0.7939	<u>0.1129</u>	<u>10.6583</u>	<u>3.3052</u>
50	27.8255	0.7862	**0.1040**	**9.8395**	**3.1165**

Table 16: Performance of various sample steps on DIV2K using MMSE outputs generated by the MambaIRv2-Large model. Model was trained for 1000 epochs using the $Z2X \mid NC$ configuration.

Step	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)	NIQE (↓)
1	**28.9523**	**0.8283**	0.2123	29.4720	4.5719
3	<u>28.8309</u>	<u>0.8227</u>	0.1887	23.6561	4.2568
5	28.6688	0.8176	0.1702	19.5663	4.0554
7	28.5081	0.8125	0.1531	16.4071	3.8737
10	28.3224	0.8064	0.1351	13.6005	3.6659
15	28.1054	0.7988	0.1180	11.3023	3.4269
20	27.9609	0.7936	0.1097	10.3139	3.2825
25	27.8596	0.7900	<u>0.1054</u>	<u>9.8509</u>	<u>3.1941</u>
50	27.6295	0.7812	**0.0992**	**9.2846**	**3.0349**



(a) PD curve for MambaIRv2-Base

(b) PD curve for MambaIRv2-Large

Figure 23: Perception–distortion comparison: (a) is the Base model (trained 1000 epochs, 20 sampling steps) showing FID vs. 1–SSIM; (b) is the Large model (same training and sampling).

Bibliography

[1] Seyoung Ahn, Soohyeong Kim, Yongseok Kwon, Joohan Park, Jiseung Youn, and Sunghyun Cho. Communication-efficient diffusion strategy for performance improvement of federated learning with non-iid data. *arXiv preprint arXiv:2207.07493*, 2022.

[2] Ekin Akyürek, Afra Feyza Akyürek, and Jacob Andreas. Learning to recombine and resample data for compositional generalization. *arXiv preprint arXiv:2010.03706*, 2020.

[3] Rana Albelaihi, Akhil Alasandagutti, Liangkun Yu, Jingjing Yao, and Xiang Sun. Deep-reinforcement-learning-assisted client selection in nonorthogonal-multiple-access-based federated learning. *IEEE Internet of Things Journal*, 10 (17):15515–15525, 2023.

[4] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

[5] Fabian Altekrüger, Alexander Denker, Paul Hagemann, Johannes Hertrich, Peter Maass, and Gabriele Steidl. Patchnr: learning from very few images by patch normalizing flow regularization. *Inverse Problems*, 39(6):064006, 2023.

[6] Lynton Ardizzone, Carsten Lüth, Jakob Kruse, Carsten Rother, and Ullrich Köthe. Guided image generation with conditional invertible neural networks. *arXiv preprint arXiv:1907.02392*, 2019.

[7] Siavash Arjomand Bigdeli, Matthias Zwicker, Paolo Favaro, and Meiguang Jin. Deep mean-shift priors for image restoration. *Advances in neural information processing systems*, 30, 2017.

[8] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017.

[9] Muhammad Asim, Max Daniels, Oscar Leong, Ali Ahmed, and Paul Hand. Invertible generative models for inverse problems: mitigating representation error and dataset bias. In *International conference on machine learning*, pages 399–409. PMLR, 2020.

[10] Ravikumar Balakrishnan, Tian Li, Tianyi Zhou, Nageen Himayat, Virginia Smith, and Jeff Bilmes. Diverse client selection for federated learning via submodular maximization. In *International Conference on Learning Representations*, 2022.

[11] Zahra Batool, Kaiwen Zhang, and Matthew Toews. Block-racs: Towards reputation-aware client selection and monetization mechanism for federated learning. *ACM SIGAPP Applied Computing Review*, 23(3):49–65, 2023.

[12] Claudio Battiloro, Paolo Di Lorenzo, Mattia Merluzzi, and Sergio Barbarossa.

Lyapunov-based optimization of edge resources for energy-efficient adaptive federated learning. *IEEE Transactions on Green Communications and Networking*, 7(1):265–280, 2022.

[13] Heli Ben-Hamu, Omri Puny, Itai Gat, Brian Karrer, Uriel Singer, and Yaron Lipman. D-flow: Differentiating through flows for controlled generation. *arXiv preprint arXiv:2402.14017*, 2024.

[14] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018.

[15] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018.

[16] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *International conference on machine learning*, pages 537–546. PMLR, 2017.

[17] Geoffrey Burton. Review of *Topics in Optimal Transportation* (graduate studies in mathematics, vol. 58) by cédric villani. *Bulletin of the London Mathematical Society*, 36(2):285–286, 2004.

[18] Chentao Cao, Zhuo-Xu Cui, Yue Wang, Shaonan Liu, Taijin Chen, Hairong Zheng, Dong Liang, and Yanjie Zhu. High-frequency space diffusion model for accelerated mri. *IEEE Transactions on Medical Imaging*, 43(5):1853–1865, 2024.

[19] Antonin Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical imaging and vision*, 20:89–97, 2004.

[20] Antonin Chambolle, Ronald A De Vore, Nam-Yong Lee, and Bradley J Lucier. Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Transactions on image processing*, 7(3):319–335, 1998.

[21] Stanley H Chan, Xiran Wang, and Omar A Elgendy. Plug-and-play admm for image restoration: Fixed-point convergence and applications. *IEEE Transactions on Computational Imaging*, 3(1):84–98, 2016.

[22] Tony Chan, Selim Esedoglu, Frederick Park, A Yip, et al. Recent developments in total variation image restoration. *Mathematical Models of Computer Vision*, 17(2):17–31, 2005.

[23] Jiadi Chen, Chunjiang Duanmu, and Huanhuan Long. Large kernel frequency-enhanced network for efficient single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6317–6326, 2024.

[24] Ke Chen, Liangyan Li, Huan Liu, Yunzhe Li, Congling Tang, and Jun Chen. Swinfsr: Stereo image super-resolution using swinir and frequency domain knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1764–1774, June 2023.

[25] Ke Chen, Liangyan Li, Huan Liu, Yunzhe Li, Congling Tang, and Jun Chen. Swinfsr: Stereo image super-resolution using swinir and frequency domain

knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1764–1774, 2023.

[26] MN Chen, KY Ho, YN Hung, CC Su, CH Kuan, HC Tai, NC Cheng, and CC Lin. Pre-treatment quality of life as a predictor of distant metastasis-free survival and overall survival in patients with head and neck cancer who underwent free flap reconstruction. *European Journal of Oncology Nursing*, 41: 1–6, 2019.

[27] Ting Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023.

[28] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017.

[29] Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1256–1272, 2016.

[30] Zhilu Chen and Xinming Huang. End-to-end learning for lane keeping of self-driving cars. In *2017 IEEE intelligent vehicles symposium (IV)*, pages 1856–1860. IEEE, 2017.

[31] Zhong-Jing Chen, Eduin E Hernandez, Yu-Chih Huang, and Stefano Rini. Communication-efficient federated dnn training: Convert, compress, correct. *IEEE Internet of Things Journal*, 2024.

[32] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 7–10, 2016.

[33] Xi Cheng, Zhenyong Fu, and Jian Yang. Multi-scale dynamic feature encoding network for image demoiréing. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3486–3493, 2019. doi: 10.1109/ICCVW.2019.00432.

[34] Xi Cheng, Zhenyong Fu, and Jian Yang. Multi-scale dynamic feature encoding network for image demoiréing. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3486–3493. IEEE, 2019.

[35] Xinlong Cheng, Tiantian Cao, Guoan Cheng, Bangxuan Huang, Xinghan Tian, Ye Wang, Xiaoyu He, Weixin Li, Tianfan Xue, and Xuan Dong. Consistent diffusion: Denoising diffusion model with data-consistent training for image restoration. *arXiv preprint arXiv:2412.12550*, 2024.

[36] Yijia Cheng, Xin Liu, and Jingyu Yang. Recaptured raw screen image and video demoireing via channel and spatial modulations. *Advances in Neural Information Processing Systems*, 36:40414–40425, 2023.

[37] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33:4479–4488, 2020.

[38] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and

Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

[39] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12413–12422, 2022.

[40] Hyungjin Chung, Jeongsol Kim, and Jong Chul Ye. Direct diffusion bridge using data consistency for inverse problems. *Advances in Neural Information Processing Systems*, 36:7158–7169, 2023.

[41] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198, 2016.

[42] Peng Dai, Xin Yu, Lan Ma, Baoheng Zhang, Jia Li, Wenbo Li, Jiajun Shen, and Xiaojuan Qi. Video demoireing with relation-based temporal consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17622–17631, 2022.

[43] Tao Dai, Jianping Wang, Hang Guo, Jinmin Li, Jinbao Wang, and Zexuan Zhu. Freqformer: Frequency-aware transformer for lightweight image super-resolution. *IJCAI. ijcai. org*, 2024.

[44] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[45] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[46] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp, 2017. URL `https://arxiv.org/abs/1605.08803`.

[47] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199, 2014.

[48] David L Donoho. De-noising by soft-thresholding. *IEEE transactions on information theory*, 41(3):613–627, 2002.

[49] David L Donoho and Iain M Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association*, 90(432): 1200–1224, 1995.

[50] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2020.

[51] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[52] Berthy T Feng, Jamie Smith, Michael Rubinstein, Huiwen Chang, Katherine L

Bouman, and William T Freeman. Score-based diffusion models as principled priors for inverse imaging. In *ICCV*, pages 10520–10531, 2023.

[53] Jessica Fridrich. Digital image forensics. *IEEE Signal Processing Magazine*, 26 (2):26–37, 2009.

[54] Tianyu Gao, Yanqing Guo, Xin Zheng, Qianyu Wang, and Xiangyang Luo. Moiré pattern removal with multi-scale feature enhancing network. In *Proceedings of the IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 240–245, 2019.

[55] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

[56] Jack Goetz, Kshitiz Malik, Duc Bui, Seungwhan Moon, Honglei Liu, and Anuj Kumar. Active federated learning, 2019. URL `https://arxiv.org/abs/1909.12641`.

[57] Aidan N Gomez, Mengye Ren, Raquel Urtasun, and Roger B Grosse. The reversible residual network: Backpropagation without storing activations. *Advances in neural information processing systems*, 30, 2017.

[58] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL `https://arxiv.org/abs/1406.2661`.

[59] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[60] Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast monocular depth estimation with flow matching. *arXiv preprint arXiv:2403.13788*, 2024.

[61] Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, and Yawei Li. Mambairv2: Attentive state space restoration. *arXiv preprint arXiv:2411.15269*, 2024.

[62] Bin He, Ce Wang, Boxin Shi, and Ling-Yu Duan. Mop moire patterns using mopnet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2424–2432, 2019.

[63] Bin He, Ce Wang, Boxin Shi, and Ling-Yu Duan. Fhde2net: Full high definition demoireing network. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII*, page 713–729, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58541-9. doi: 10.1007/978-3-030-58542-6_43. URL `https://doi.org/10.1007/978-3-030-58542-6_43`.

[64] Bin He, Ce Wang, Boxin Shi, and Ling-Yu Duan. Fhde 2 net: Full high definition demoireing network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 713–729. Springer, 2020.

[65] Bin He, Ce Wang, Boxin Shi, and Ling-Yu Duan. Fhde 2 net: Full high definition demoireing network. In *Computer Vision–ECCV 2020: 16th European*

Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16, pages 713–729. Springer, 2020.

[66] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *TPAMI*, 33(12):2341–2353, 2010.

[67] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. corr abs/1512.03385 (2015), 2015.

[68] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[69] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2017.

[70] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[71] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL `https://arxiv.org/abs/2006.11239`.

[72] Chen Hou, Guoqiang Wei, and Zhibo Chen. High-fidelity diffusion-based image editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2184–2192, 2024.

[73] Lingyu Huang, Liang Feng, Handing Wang, Yaqing Hou, Kai Liu, and Chao

Chen. A preliminary study of improving evolutionary multi-objective optimization via knowledge transfer from single-objective problems. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1552–1559, 2020. doi: 10.1109/SMC42975.2020.9283151.

[74] Samuel Hurault, Arthur Leclaire, and Nicolas Papadakis. Gradient step denoiser for convergent plug-and-play, 2022. URL `https://arxiv.org/abs/2110.03220`.

[75] Samuel Hurault, Arthur Leclaire, and Nicolas Papadakis. Proximal denoiser for convergent plug-and-play optimization with nonconvex regularization. In *International Conference on Machine Learning*, pages 9483–9505. PMLR, 2022.

[76] Andrey Ignatov, Radu Timofte, et al. Pirm challenge on perceptual image enhancement on smartphones: report. In *European Conference on Computer Vision (ECCV) Workshops*, January 2019.

[77] 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.

[78] Jörn-Henrik Jacobsen, Arnold Smeulders, and Edouard Oyallon. i-revnet: Deep invertible networks. *arXiv preprint arXiv:1802.07088*, 2018.

[79] Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Towards understanding biased client selection in federated learning. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of*

Machine Learning Research, pages 10351–10375. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/jee-cho22a.html.

[80] Daniel M Jimenez G, David Solans, Mikko Heikkila, Andrea Vitaletti, Nicolas Kourtellis, Aris Anagnostopoulos, and Ioannis Chatzigiannakis. Non-iid data in federated learning: A systematic review with taxonomy, metrics, methods, frameworks and future directions. *arXiv e-prints*, pages arXiv–2411, 2024.

[81] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[82] 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.

[83] Bahjat Kawar, Gregory Vaksman, and Michael Elad. Snips: Solving noisy inverse problems stochastically. *Advances in Neural Information Processing Systems*, 34:21757–21769, 2021.

[84] Bahjat Kawar, Gregory Vaksman, and Michael Elad. Stochastic image denoising by sampling from the posterior distribution, 2021. URL https://arxiv.org/abs/2101.09552.

[85] Sangmin Kim, Hyungjoon Nam, Jisu Kim, and Jechang Jeong. C3net: Demoiréing network attentive in channel, color and concatenation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 426–427, 2020.

[86] Yunhee Kim, Gilbae Park, Seong-Woo Cho, Jae-hyun Jung, Byoungho Lee, Yoonsun Choi, and Moon-Gyu Lee. Integral imaging with reduced color moiré pattern by using a slanted lens array. In *Stereoscopic Displays and Applications XIX*, volume 6803, pages 541–548. SPIE, 2008.

[87] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions, 2018.

[88] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL `https://arxiv.org/abs/1312.6114`.

[89] Jakub Konečnỳ. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[90] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). 2009. URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

[91] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[92] Fan Lai, Xiangfeng Zhu, Harsha V Madhyastha, and Mosharaf Chowdhury. Oort: Efficient federated learning via guided participant selection. In *15th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 21)*, pages 19–35, 2021.

[93] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan

Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017.

[94] Jian Li, Tongbao Chen, and Shaohua Teng. A comprehensive survey on client selection strategies in federated learning. *Computer Networks*, page 110663, 2024.

[95] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data, 2020. URL `https://arxiv.org/abs/1907.02189`.

[96] Xiang Li, Soo Min Kwon, Ismail R Alkhouri, Saiprasad Ravishankar, and Qing Qu. Decoupled data consistency with diffusion purification for image restoration. *arXiv preprint arXiv:2403.06054*, 2024.

[97] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021.

[98] Kang Liao, Zongsheng Yue, Zhouxia Wang, and Chen Change Loy. Denoising as adaptation: Noise-space domain adaptation for image restoration. *arXiv preprint arXiv:2406.18516*, 2024.

[99] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5404–5411, 2024.

[100] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt

Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[101] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=PqvMRDCJT9t`.

[102] Bolin Liu, Xiao Shu, and Xiaolin Wu. Demoir\'eing of camera-captured screen images using deep convolutional neural network. *arXiv preprint arXiv:1804.03809*, 2018.

[103] Huan Liu, Zijun Wu, Liangyan Li, Sadaf Salehkalaibar, Jun Chen, and Keyan Wang. Towards multi-domain single image dehazing via test-time training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5831–5840, 2022.

[104] Jiaming Liu, Salman Asif, Brendt Wohlberg, and Ulugbek Kamilov. Recovery analysis for plug-and-play priors using the restricted eigenvalue condition. *Advances in Neural Information Processing Systems*, 34:5921–5933, 2021.

[105] Lin Liu, Jianzhuang Liu, Shanxin Yuan, Gregory Slabaugh, Aleš Leonardis, Wengang Zhou, and Qi Tian. Wavelet-based dual-branch network for image demoiréing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 86–102. Springer, 2020.

[106] Lin Liu, Junfeng An, Shanxin Yuan, Wengang Zhou, Houqiang Li, Yanfeng

Wang, and Qi Tian. Video demoiréing with deep temporal color embedding and video-image invertible consistency. *IEEE Transactions on Multimedia*, 2024.

[107] Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.

[108] Shuai Liu, Chenghua Li, Nan Nan, Ziyao Zong, and Ruixia Song. Mmdm: Multi-frame and multi-scale for image demoiréing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1751–1759, 2020. doi: 10.1109/CVPRW50498.2020.00225.

[109] Shuai Liu, Chenghua Li, Nan Nan, Ziyao Zong, and Ruixia Song. Mmdm: Multi-frame and multi-scale for image demoireing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[110] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[111] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32 (11):7505–7517, 2022. doi: 10.1109/TCSVT.2022.3189545.

[112] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022.

[113] Yangyi Liu, Huan Liu, Liangyan Li, Zijun Wu, and Jun Chen. A data-centric solution to nonhomogeneous dehazing via vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1406–1415, June 2023.

[114] Yangyi Liu, Huan Liu, Liangyan Li, Zijun Wu, and Jun Chen. A data-centric solution to nonhomogeneous dehazing via vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1406–1415, 2023.

[115] Yangyi Liu, Sadaf Salehkalaibar, Stefano Rini, and Jun Chen. M22: Rate-distortion inspired gradient compression. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10097231.

[116] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2025.

[117] Yuhao Liu, Zhanghan Ke, Fang Liu, Nanxuan Zhao, and Rynson WH Lau. Diff-plugin: Revitalizing details for diffusion-based low-level tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4197–4208, 2024.

[118] Zhenpeng Liu, Sichen Duan, Shuo Wang, Yi Liu, and Xiaofei Li. Mflces: Multi-level federated edge learning algorithm based on client and edge server selection. *Electronics*, 12(12):2689, 2023.

[119] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *ECCV*, 2020.

[120] Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021.

[121] Sebastian Lunz, Ozan Öktem, and Carola-Bibiane Schönlieb. Adversarial regularizers in inverse problems. *Advances in neural information processing systems*, 31, 2018.

[122] Xiaotong Luo, Jiangtao Zhang, Ming Hong, Yanyun Qu, Yuan Xie, and Cuihua Li. Deep wavelet network with domain adaptation for single image demoireing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 420–421, 2020.

[123] Michael Lustig, David Donoho, and John M Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

[124] Weimin Lyu, Xinyu Dong, Rachel Wong, Songzhu Zheng, Kayley Abell-Hart, Fusheng Wang, and Chao Chen. A multimodal transformer: Fusing clinical notes with structured ehr data for interpretable in-hospital mortality prediction. In *AMIA Annual Symposium Proceedings*, volume 2022, page 719. American Medical Informatics Association, 2022.

[125] Weimin Lyu, Songzhu Zheng, Lu Pang, Haibin Ling, and Chao Chen. Attention-enhancing backdoor attacks against bert-based models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10672–10690, 2023.

[126] Weimin Lyu, Jiachen Yao, Saumya Gupta, Lu Pang, Tao Sun, Lingjie Yi, Lijie Hu, Haibin Ling, and Chao Chen. Backdooring vision-language models with out-of-distribution data. *arXiv preprint arXiv:2410.01264*, 2024.

[127] Xiaodong Ma, Jia Zhu, Zhihao Lin, Shanxuan Chen, and Yangjie Qin. A state-of-the-art survey on solving non-iid data in federated learning. *Future Generation Computer Systems*, 135:244–258, 2022.

[128] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

[129] Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. *arXiv preprint arXiv:2305.04391*, 2023.

[130] Razvan V Marinescu, Daniel Moyer, and Polina Golland. Bayesian image reconstruction using deep generative models. *arXiv preprint arXiv:2012.04567*, 2020.

[131] Ouiame Marnissi, Hajar El Hammouti, and El Houcine Bergou. Client selection in federated learning based on gradients importance. In *AIP Conference Proceedings*, volume 3034. AIP Publishing, 2024.

[132] Ségolène Martin, Anne Gagneux, Paul Hagemann, and Gabriele Steidl. Pnp-flow: Plug-and-play image restoration with flow matching. *arXiv preprint arXiv:2410.02423*, 2024.

[133] Brendan McMahan and Daniel Ramage. Federated learning: Collaborative machine learning without centralized training data. *Google Research Blog*, 3, 2017.

[134] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[135] Manan Mehta and Chenhui Shao. A greedy agglomerative framework for clustered federated learning. *IEEE Transactions on Industrial Informatics*, 19(12): 11856–11867, 2023.

[136] Tim Meinhardt, Michael Moller, Caner Hazirbas, and Daniel Cremers. Learning proximal operators: Using denoising networks for regularizing inverse imaging problems. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1781–1790, 2017.

[137] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik P. Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14297–14306. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01374. URL `https://doi.org/10.1109/CVPR52729.2023.01374`.

[138] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, pages 2437–2445, 2020.

[139] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[140] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[141] Ihab Mohammed, Shadha Tabatabai, Ala Al-Fuqaha, Faissal El Bouanani, Junaid Qadir, Basheer Qolomany, and Mohsen Guizani. Budgeted online selection of candidate iot clients to participate in federated learning. *IEEE Internet of Things Journal*, 8(7):5938–5952, 2020.

[142] Pierre Moulin and Juan Liu. Analysis of multiresolution image denoising schemes using generalized gaussian and complexity priors. *IEEE transactions on Information Theory*, 45(3):909–919, 1999.

[143] A Tuan Nguyen, Philip Torr, and Ser Nam Lim. Fedsr: A simple and effective domain generalization method for federated learning. *Advances in Neural Information Processing Systems*, 35:38831–38843, 2022.

[144] Hung T Nguyen, Vikash Sehwag, Seyyedali Hosseinalipour, Christopher G Brinton, Mung Chiang, and H Vincent Poor. Fast-convergent federated learning. *IEEE Journal on Selected Areas in Communications*, 39(1):201–218, 2020.

[145] Yuzhen Niu, Zhihua Lin, Wenxi Liu, and Wenzhong Guo. Progressive moire removal and texture complementation for image demoireing. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[146] Yuzhen Niu, Rui Xu, Zhihua Lin, and Wenxi Liu. Std-net: Spatio-temporal decomposition network for video demoiréing with sparse transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[147] Guy Ohayon, Tomer Michaeli, and Michael Elad. Posterior-mean rectified flow: Towards minimum mse photo-realistic image restoration. *arXiv preprint arXiv:2410.00418*, 2024.

[148] Yan-Tsung Peng, Chih-Hsiang Hou, You-Cheng Lee, Aiden J Yoon, Zihao Chen, Yi-Ting Lin, and Wei-Cheng Lien. Image demoiréing via multi-scale fusion networks with moiré data augmentation. *IEEE Sensors Journal*, 2024.

[149] Ashwini Pokle, Matthew J Muckley, Ricky TQ Chen, and Brian Karrer. Training-free linear image inverses via flows. *arXiv preprint arXiv:2310.04432*, 2023.

[150] Yuhui Quan, Haoran Huang, Shengfeng He, and Ruotao Xu. Deep video demoiréing via compact invertible dyadic decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12677–12686, 2023.

[151] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in neural information processing systems*, 34:980–993, 2021.

[152] 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.

[153] Stefan Roth and Michael J Black. Fields of experts: A framework for learning image priors. In *CVPR*, pages 860–867, 2005.

[154] Daniel Ruderman and William Bialek. Statistics of natural images: Scaling in the woods. *Advances in neural information processing systems*, 6, 1993.

[155] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[156] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[157] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

[158] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of

diffusion models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL `https://openreview.net/forum?id=TIdIXIpzhoI`.

[159] Ryoji Sasada, Masahiko Yamada, Shoji Hara, Hideya Takeo, and Kazuo Shimura. Stationary grid pattern removal using 2d technique for moire-free radiographic image display. In *Medical Imaging 2003: Visualization, Image-Guided Procedures, and Display*, volume 5029, pages 688–697. SPIE, 2003.

[160] Johannes Schusterbauer, Ming Gui, Pingchuan Ma, Nick Stracke, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Fmboost: Boosting latent diffusion with flow matching. In *European Conference on Computer Vision*, pages 338–355. Springer, 2024.

[161] Zhenfeng Shao and Jiajun Cai. Remote sensing image fusion with deep convolutional neural network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(5):1656–1669, 2018.

[162] Yuxin Shi, Zelei Liu, Zhuan Shi, and Han Yu. Fairness-aware client selection for federated learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 324–329, 2023. doi: 10.1109/ICME55011.2023.00063.

[163] Hasib Siddiqui, Mireille Boutin, and Charles A Bouman. Hardware-friendly descreening. *IEEE Transactions on Image Processing*, 19(3):746–757, 2009.

[164] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[165] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL `https://arxiv.org/abs/1409.1556`.

[166] Durga Sivasubramanian, Lokesh Nagalapatti, Rishabh Iyer, and Ganesh Ramakrishnan. Gradient coreset for federated learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2648–2657, 2024.

[167] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

[168] Binbin Song, Jiantao Zhou, Xiangyu Chen, and Shile Zhang. Real-scene reflection removal with raw-rgb image pairs. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[169] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[170] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023.

[171] Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. *arXiv preprint arXiv:2310.14189*, 2023.

[172] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[173] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023.

[174] Yiren Song, Cheng Liu, and Mike Zheng Shou. Omniconsistency: Learning style-agnostic consistency from paired stylization data. *arXiv preprint arXiv:2505.18445*, 2025.

[175] Suhas Sreehari, S Venkat Venkatakrishnan, Brendt Wohlberg, Gregery T Buzzard, Lawrence F Drummy, Jeffrey P Simmons, and Charles A Bouman. Plug-and-play priors for bright field electron tomography and sparse interpolation. *IEEE Transactions on Computational Imaging*, 2(4):408–423, 2016.

[176] Abeda Sultana, Md Mainul Haque, Li Chen, Fei Xu, and Xu Yuan. Eiffel: Efficient and fair scheduling in adaptive federated learning. *IEEE Transactions on Parallel and Distributed Systems*, 33(12):4282–4294, 2022.

[177] Bin Sun, Shutao Li, and Jun Sun. Scanned image descreening with image redundancy and adaptive filtering. *IEEE Transactions on Image Processing*, 23 (8):3698–3710, 2014.

[178] Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, et al. Learning to (learn at test time): Rnns with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024.

[179] Yujing Sun, Yizhou Yu, and Wenping Wang. Moiré photo restoration using multiresolution convolutional neural networks. *IEEE Transactions on Image Processing*, 27(8):4160–4172, 2018.

[180] Xavier Tan, Wei Chong Ng, Wei Yang Bryan Lim, Zehui Xiong, Dusit Niyato, and Han Yu. Reputation-aware federated learning client selection based on stochastic integer programming. *IEEE Transactions on Big Data*, 2022.

[181] Minxue Tang, Xuefei Ning, Yitu Wang, Jingwei Sun, Yu Wang, Hai Li, and Yiran Chen. Fedcor: Correlation-based active client selection strategy for heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10102–10111, 2022.

[182] Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[183] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[184] Singanallur V. Venkatakrishnan, Charles A. Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 945–948, 2013. doi: 10.1109/GlobalSIP.2013.6737048.

[185] Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE global conference on signal and information processing*, pages 945–948. IEEE, 2013.

[186] Luisa Verdoliva. Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020.

[187] An Gia Vien, Hyunkook Park, and Chul Lee. Dual-domain deep convolutional neural networks for image demoireing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 470–471, 2020.

[188] Ce Wang, Bin He, Shengsen Wu, Renjie Wan, Boxin Shi, and Ling-Yu Duan. Coarse-to-fine disentangling demoiréing framework for recaptured screen images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[189] Chenyu Wang, Qiong Wu, Qian Ma, and Xu Chen. A buffered semi-asynchronous mechanism with mab for efficient federated learning. In *2022 International Conference on High Performance Big Data and Intelligent Systems (HDIS)*, pages 180–184, 2022. doi: 10.1109/HDIS56859.2022.9991371.

[190] Hailing Wang, Qiaoyu Tian, Liang Li, and Xiaojie Guo. Image demoiréing with a dual-domain distilling network. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[191] Peijuan Wang, Bulent Bayram, and Elif Sertel. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Science Reviews*, 232:104110, 2022.

[192] Xinrui Wang and Yan Jin. Work process transfer reinforcement learning: Feature extraction and finetuning in ship collision avoidance. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 86212, page V002T02A069. American Society of Mechanical Engineers, 2022.

[193] Xinrui Wang and Yan Jin. Transfer reinforcement learning: Feature transferability in ship collision avoidance. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 87318, page V03BT03A071. American Society of Mechanical Engineers, 2023.

[194] Xinrui Wang and Yan Jin. Exploring causalworld: Enhancing robotic manipulation via knowledge transfer and curriculum learning. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 88360, page V03AT03A013. American Society of Mechanical Engineers, 2024.

[195] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[196] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[197] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data, 2021. URL `https://arxiv.org/abs/2107.10833`.

[198] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022.

[199] Yuejiao Wang, Zhidong Cao, Daniel Dajun Zeng, Qingpeng Zhang, and Tianyi Luo. The collective wisdom in the covid-19 research: Comparison and synthesis of epidemiological parameter estimates in preprints and peer-reviewed articles. *International Journal of Infectious Diseases*, 104:1–6, 2021.

[200] Yuwei Wang and Burak Kantarci. A novel reputation-aware client selection scheme for federated learning within mobile environments. In *2020 IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pages 1–6. IEEE, 2020.

[201] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

[202] Xinyi Wei, Hans Van Gorp, Lizeth Gonzalez-Carabarin, Daniel Freedman, Yonina C Eldar, and Ruud JG van Sloun. Deep unfolding with normalizing flow priors for inverse problems. *IEEE Transactions on Signal Processing*, 70:2962–2971, 2022.

[203] Zhouping Wei, Jian Wang, Helen Nichol, Sheldon Wiebe, and Dean Chapman. A median-gaussian filtering framework for moiré pattern noise removal from x-ray microscopy image. *Micron*, 43(2-3):170–176, 2012.

[204] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms, 2017.

[205] Zeyu Xiao, Zhihe Lu, and Xinchao Wang. P-bic: Ultra-high-definition image

moiré patterns removal via patch bilateral compensation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8365–8373, 2024.

[206] Dejia Xu, Yihao Chu, and Qingyan Sun. Moiré pattern removal via attentive fractal network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 472–473, 2020.

[207] Shuning Xu, Xina Liu, Binbin Song, Xiangyu Chen, Qiubo Chen, and Jiantao Zhou. Demmamba: Alignment-free raw video demoireing with frequency-assisted spatio-temporal mamba. *arXiv preprint arXiv:2408.10679*, 2024.

[208] Shuning Xu, Binbin Song, Xiangyu Chen, Xina Liu, and Jiantao Zhou. Image demoireing in raw and srgb domains. In *European Conference on Computer Vision*, pages 108–124. Springer, 2024.

[209] Shuning Xu, Binbin Song, Xiangyu Chen, and Jiantao Zhou. Direction-aware video demoireing with temporal-guided bilateral learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6360–6368, 2024.

[210] Shuning Xu, Binbin Song, Xiangyu Chen, Xina Liu, and Jiantao Zhou. Image demoireing in raw and srgb domains. In *European Conference on Computer Vision*, pages 108–124. Springer, 2025.

[211] Xiangyu Xu, Jinshan Pan, Yu-Jin Zhang, and Ming-Hsuan Yang. Motion blur kernel estimation via deep learning. *IEEE Transactions on Image Processing*, 27(1):194–205, 2017.

[212] Xiangyu Xu, Deqing Sun, Jinshan Pan, Yujin Zhang, Hanspeter Pfister, and

Ming-Hsuan Yang. Learning to super-resolve blurry face and text images. In *ICCV*, pages 251–260, 2017.

[213] Xingyu Xu and Yuejie Chi. Provably robust score-based diffusion posterior sampling for plug-and-play image reconstruction. *arXiv preprint arXiv:2403.17042*, 2024.

[214] Yang Xu, Zhida Jiang, Hongli Xu, Zhiyuan Wang, Chen Qian, and Chunming Qiao. Federated learning with client selection and gradient compression in heterogeneous edge systems. *IEEE Transactions on Mobile Computing*, 2023.

[215] Shengke Xue, Wenyuan Qiu, Fan Liu, and Xinyu Jin. Faster image super-resolution by improved frequency-domain neural networks. *Signal, Image and Video Processing*, 14(2):257–265, 2020.

[216] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873, 2010.

[217] Jingyu Yang, Fanglei Liu, Huanjing Yue, Xiaomei Fu, Chunping Hou, and Feng Wu. Textured image demoiréing via signal decomposition and guided filtering. *IEEE Transactions on Image Processing*, 26(7):3528–3541, 2017.

[218] Xin Yu, Peng Dai, Wenbo Li, Lan Ma, Jiajun Shen, Jia Li, and Xiaojuan Qi. Towards efficient and scale-robust ultra-high-definition image demoiréing. In *European Conference on Computer Vision*, pages 646–662. Springer, 2022.

[219] Xin Yu, Peng Dai, Wenbo Li, Lan Ma, Jiajun Shen, Jia Li, and Xiaojuan Qi.

Towards efficient and scale-robust ultra-high-definition image demoiréing. In *European Conference on Computer Vision*, pages 646–662. Springer, 2022.

[220] Shanxin Yuan, Radu Timofte, Gregory Slabaugh, and Ales Leonardis. Aim 2019 challenge on image demoireing: Dataset and study, 2019. URL `https://arxiv.org/abs/1911.02498`.

[221] Huanjing Yue, Yijia Cheng, Fanglong Liu, and Jingyu Yang. Unsupervised moiré pattern removal for recaptured screen images. *Neurocomputing*, 456:352–363, 2021.

[222] Huanjing Yue, Yijia Cheng, Yan Mao, Cong Cao, and Jingyu Yang. Recaptured screen image demoiréing in raw domain. *IEEE Transactions on Multimedia*, 2022.

[223] Huanjing Yue, Yijia Cheng, Xin Liu, and Jingyu Yang. Recaptured raw screen image and video demoir\'eing via channel and spatial modulations. *arXiv preprint arXiv:2310.20332*, 2023.

[224] Shaolei Zhai, Xin Jin, Ling Wei, Hongxuan Luo, and Min Cao. Dynamic federated learning for gmec with time-varying wireless link. *IEEE Access*, 9:10400–10412, 2021. doi: 10.1109/ACCESS.2021.3050172.

[225] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin. Swinfir: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution. *arXiv preprint arXiv:2208.11247*, 2022.

[226] George Zhang, Jingjing Qian, Jun Chen, and Ashish Khisti. Universal rate-distortion-perception representations for lossy compression. *NeurIPS*, 34:11517–11529, 2021.

[227] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

[228] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6360–6376, 2021.

[229] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *IEEE International Conference on Computer Vision*, pages 4791–4800, 2021.

[230] Lin Zhang, Li Shen, Liang Ding, Dacheng Tao, and Ling-Yu Duan. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10174–10183, 2022.

[231] Peiying Zhang, Chao Wang, Chunxiao Jiang, and Zhu Han. Deep reinforcement learning assisted federated learning algorithm for data management of iiot. *IEEE Transactions on Industrial Informatics*, 17(12):8475–8484, 2021. doi: 10.1109/TII.2021.3064351.

[232] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang.

The unreasonable effectiveness of deep features as a perceptual metric. *CVPR*, 2018.

[233] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[234] Yasi Zhang, Peiyu Yu, Yaxuan Zhu, Yingshan Chang, Feng Gao, Ying Nian Wu, and Oscar Leong. Flow priors for linear inverse problems via iterative corrupted trajectory matching. *Advances in Neural Information Processing Systems*, 37: 57389–57417, 2025.

[235] Yuxin Zhang, Mingbao Lin, Xunchao Li, Han Liu, Guozhi Wang, Fei Chao, Shuai Ren, Yafei Wen, Xiaoxin Chen, and Rongrong Ji. Real-time image demoireing on mobile devices. *arXiv preprint arXiv:2302.02184*, 2023.

[236] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5906–5916, 2023.

[237] Bolun Zheng, Shanxin Yuan, Gregory Slabaugh, and Ales Leonardis. Image demoireing with learnable bandpass filters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3636–3645, 2020.

[238] Bolun Zheng, Shanxin Yuan, Chenggang Yan, Xiang Tian, Jiyong Zhang, Yaoqi Sun, Lin Liu, Aleš Leonardis, and Gregory Slabaugh. Learning frequency domain priors for image demoireing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7705–7717, 2021.

[239] Yunshan Zhong, Yuyao Zhou, Yuxin Zhang, Fei Chao, and Rongrong Ji. Learning image demoiréing from unpaired real data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7623–7631, 2024.

[240] Zhaohui Zhou, Shijie Shi, Fasong Wang, Yanbin Zhang, and Yitong Li. Joint client selection and cpu frequency control in wireless federated learning networks with power constraints. *Entropy*, 25(8):1183, 2023.

[241] Zhenyu Zhou, Defang Chen, Can Wang, Chun Chen, and Siwei Lyu. Simple and fast distillation of diffusion models. *Advances in Neural Information Processing Systems*, 37:40831–40860, 2024.

[242] Bo Zhu, Jeremiah Z Liu, Stephen F Cauley, Bruce R Rosen, and Matthew S Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487–492, 2018.

[243] Hongbin Zhu, Yong Zhou, Hua Qian, Yuanming Shi, Xu Chen, and Yang Yang. Online client selection for asynchronous federated learning with fairness consideration. *IEEE Transactions on Wireless Communications*, 22(4):2493–2506, 2022.

[244] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[245] Song Chun Zhu and David Mumford. Prior learning and gibbs reaction-diffusion. *TPAMI*, 19(11):1236–1250, 1997.

[246] Xiaobin Zhu, Zhuangzi Li, Xiao-Yu Zhang, Changsheng Li, Yaqi Liu, and Ziyu Xue. Residual invertible spatio-temporal network for video super-resolution. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5981–5988, 2019.

[247] Yixuan Zhu, Wenliang Zhao, Ao Li, Yansong Tang, Jie Zhou, and Jiwen Lu. Flowie: Efficient image enhancement via rectified flow, 2024.

[248] Daniel Zoran and Yair Weiss. From learning models of natural image patches to whole image restoration. In *2011 international conference on computer vision*, pages 479–486. IEEE, 2011.