Restoring Without Forgetting: Continual Learning Across Image Degradations

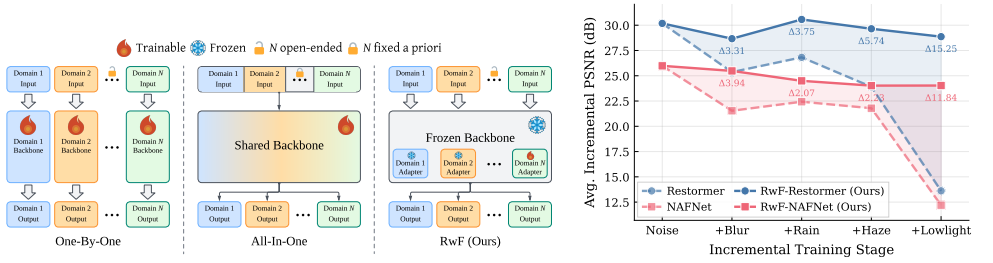
Alif Ashrafee
aa5264@rit.edu
Rochester Institute of Technology
New York, USA
Bartosz Krawczyk
bartosz.krawczyk@rit.edu

Abstract

Recent progress in image restoration has converged on all-in-one architectures that jointly handle multiple degradations within a single network. These methods are effective on static benchmarks but target a closed-world setting that assumes simultaneous access to every target degradation at training time. In practice, degradations are encountered sequentially as field-deployed systems progressively face new environmental conditions, and historical training data is often unavailable due to privacy or storage constraints. Accommodating a new degradation then requires either retraining on the union of all prior data, which is often costly or infeasible, or fine-tuning, which causes catastrophic forgetting. We formulate multi-degradation image restoration as a continual domain-incremental learning problem, in which degradations arrive incrementally and prior data is unavailable. Our proposed Restoring without Forgetting (RwF) framework learns a lightweight adapter for each new degradation, eliminating forgetting by construction at a fraction of the cost of dedicated per-domain networks. To isolate degradation learning from dataset variation, we construct a benchmark spanning five degradation domains under shared image content. At test time, an unsupervised routing mechanism identifies the appropriate restoration path for unknown inputs without requiring domain labels. Across the five-domain sequence, RwF improves final average PSNR over naive sequential fine-tuning by 15.25 dB and 11.83 dB on the Restormer and NAFNet backbones, respectively. The framework transfers to eleven canonical real-degradation benchmarks (3,465 images) at 89.5% routing accuracy with only a $+0.94$ dB oracle PSNR gap, establishing, to our knowledge, the first systematic baseline for continual multi-degradation image restoration. Code, benchmark, and weights are available at: https://github.com/AlifAshrafee/Restoring-Without-Forgetting.

1 Introduction

Image restoration has advanced rapidly in recent years, propelled by attention-based and transformer architectures that exploit long-range pixel dependencies for high-fidelity reconstruction [29, 47, 54]. Networks such as Restormer [59], NAFNet [7], SFHformer [22], and FFTformer [24] have established increasingly strong baselines across denoising, deblurring, dehazing, deraining, and low-light enhancement. Yet the dominant practice remains task-specialized: each network is trained on a single degradation type, with dedicated data,



(a) (b)

Figure 1: **(a)** Three paradigms for multi-degradation image restoration: One-By-One (one network per degradation), All-In-One (a single jointly trained network), and RwF (Ours), which freezes a pretrained backbone and attaches lightweight degradation-specific adapters routed without supervision at inference. **(b)** Average incremental PSNR across five sequentially learned degradations (Noise, +Blur, +Rain, +Haze, +Lowlight). Sequential fine-tuning forgets as degradations accumulate, while RwF holds performance across all domains.

hyperparameter tuning, and model weights for every problem [3, 13, 30, 37, 46]. This single-degradation pipeline is poorly suited to real deployment. Field-deployed sensors progressively encounter new environmental conditions over their operational lifetime, often without access to data from previously seen degradations due to storage, privacy, or licensing constraints. Naively fine-tuning an existing network on a new degradation overwrites the representations responsible for prior modes, leading to catastrophic forgetting [12, 34]. In contrast, training a fresh network from scratch for every encountered degradation is expensive, data-hungry, and increasingly unscalable as the catalog of degradations grows. Equally absent from this paradigm is a mechanism to decide which restoration network to apply to a given input at test time. Existing pipelines implicitly assume an oracle that pairs each test image with the correct network, an assumption that breaks down for intelligent systems operating in the open world.

To reduce this parameter redundancy, recent work consolidates multiple degradations into "all-in-one" networks trained jointly on the union of target tasks [15, 26, 39, 40, 52]. Effective as they are on closed-set benchmarks, these models presuppose simultaneous access to every target degradation in a single joint training run, with the degradation set fixed at design time. Extending one to a newly observed degradation then requires reassembling the full union of past and present data and retraining, which is costly at scale and infeasible when prior data cannot be retained. Weather-specialized restoration [48] narrows the problem to atmospheric phenomena and offers no clear extension to the sensor-side artifacts (noise, low-light, blur) that deployed perception stacks routinely encounter.

The challenge of accumulating capability over a sequential data stream while preserving previously learned competence is the central problem of continual learning (CL) [38, 52]. Domain-incremental learning (DIL) [49] in particular targets the setting where new domains arrive sequentially and domain identity is unavailable at test time. Yet CL research remains predominantly evaluated in discriminative, class-incremental settings on classification benchmarks, with limited engagement with dense, pixel-level generative tasks. Image restoration, acutely sensitive to distributional shifts across degradation types and demanding faithful per-pixel reconstruction, is conspicuously underexplored under the CL framework.

The central challenge is therefore not only to restore images from diverse degradations,

but to acquire these capabilities sequentially, without retaining past data and without knowing the degradation label at test time. Motivated by this, we propose RwF (Restoring without Forgetting), a parameter-efficient framework for continual multi-degradation restoration. RwF freezes a pretrained restoration backbone that provides a strong natural-image prior and incrementally attaches a lightweight low-rank adapter for each new degradation, isolating degradation-specific parameters so that previously learned degradations are preserved exactly. To resolve degradation identity at inference, a prototype-matching mechanism routes each input to the appropriate adapter without domain labels. As shown in Fig. 1(b), naive sequential fine-tuning (FT) of strong backbones suffers severe forgetting as degradations accumulate, whereas RwF preserves performance across the full sequence, improving final average PSNR by up to 15.25 dB. Our main contributions are as follows:

- **A continual learning formulation of multi-degradation image restoration.** We frame multi-degradation restoration as a domain-incremental learning problem in which qualitatively distinct degradations arrive sequentially and degradation identity is unknown at test time, a setting that has remained largely unexplored within the broader continual learning literature.
- **A parameter-efficient framework with unsupervised inference-time routing.** We introduce a backbone-frozen architecture that eliminates forgetting by construction through degradation-specific low-rank adapters at a fraction of the cost of dedicated per-domain networks, and routes test inputs to the appropriate restoration path without domain labels via a lightweight prototype-matching mechanism.
- **A content-controlled benchmark and sim-to-real evaluation.** We construct a five-domain benchmark spanning blur, noise, haze, rain, and low-light over a shared image corpus, isolating degradation learning from dataset variation, and complement it with evaluations on canonical real-degradation benchmarks to demonstrate sim-to-real transferability of the learned adapters.

2 Related Work

2.1 Image Restoration

Modern image restoration is dominated by transformer-based architectures that capture long-range dependencies for high-fidelity reconstruction. SwinIR [29] adapted shifted-window self-attention to restoration, Restormer [59] reorganized attention along the channel dimension for efficient high-resolution processing, NAFNet [7] demonstrated that careful simplification of nonlinear activations yields a competitive convolutional baseline, FFTformer [24] exploited frequency-domain representations, and SFHformer [22] fused spatial and frequency cues within a unified architecture. Despite their strong per-task performance, these networks remain single-degradation specialists, requiring a distinct trained instance for every problem.

To consolidate this redundancy, the all-in-one paradigm trains a single network jointly on the union of target degradations. AirNet [26] and TransWeather [48] pair a shared backbone with degradation-aware modules, PromptIR [39] learns visual prompts that adapt the network to the input degradation, IDR [62] decomposes degradations into shared ingredients, AdaIR [10] modulates frequency-domain features adaptively, and RAM [40] brings masked image modeling to blind multi-degradation restoration. These methods assume joint

access to every target degradation at training time, with the full degradation set specified in advance. Subsequent work scales this paradigm without departing from the assumption. FoundIR [27] trains on a million-image corpus through an incremental schedule that serves optimization stability rather than continual acquisition. DegAE [31] synthesizes degradations to pretrain transferable low-level features for downstream fine-tuning. DCPT [19] pretrains a degradation classifier over a predetermined set to initialize a joint restoration model. A nascent line of work has begun to examine continual learning in restoration more directly. CauSiam [9] performs continual test-time adaptation for defocus deblurring but not offline domain-incremental training. MINI [15] adds new restoration capabilities through a meta-convolution module, though its capacity is bounded by an embedding pool that must be fixed at design time and cannot be expanded as degradations accumulate. None of these works report a forgetting protocol across heterogeneous degradation types under unsupervised inference. We position RwF in this gap, as a domain-incremental framework that retains no past data, assigns each degradation its own restoration path under strict parameter isolation, and selects among paths by label-free routing.

2.2 Domain-Incremental Learning

Domain-incremental learning addresses sequential adaptation across new data domains without access to domain identity at inference [49, 52]. The dominant strategies fall into three families: regularization of important parameters [23, 60], replay of stored exemplars [4, 5, 21], and parameter isolation [28, 32]. Replay carries storage and privacy costs that are impractical at high resolution [51, 57], motivating a recent shift toward exemplar-free parameter-isolation approaches over frozen pretrained backbones [35, 45, 65, 66]. AdaptFormer [8] established adapters as a parameter-efficient mechanism for transformer fine-tuning, and methods such as SOYO [53] and DUCT [67] have refined exemplar-free domain selection and embedding-space calibration. Beyond classification, DIL has been extended to dense semantic segmentation [43], but the literature remains overwhelmingly discriminative. Pixel-level generative tasks such as image restoration are largely absent, which is the gap our work addresses.

3 Methodology

3.1 Problem Formulation

We cast multi-degradation image restoration as a domain-incremental learning (DIL) problem [49]. Throughout, *degradation* refers to a physical corruption operator (motion blur, sensor noise, atmospheric haze, and so on), and *domain* refers to the distribution of images it produces. Thus, each degradation type constitutes a distinct domain, and the domains are revealed to the model one at a time. Let $x \in \mathbb{R}^{H\times W\times 3}$ denote a clean image drawn from a content distribution $p(x)$, and let a degradation be described by a (possibly stochastic) operator $\mathcal{T} : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{H\times W\times 3}$ that maps a clean image to a degraded observation $y = \mathcal{T}(x)$. A restoration model seeks an approximate inverse that recovers x from y. We consider a sequence of T degradation domains $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T$, each induced by a distinct operator $\mathcal{T}_t$ and instantiated as a set of paired samples $\mathcal{D}_t = \{(x_i, y_i^{(t)})\}_{i=1}^{N_t}$ with $y_i^{(t)} = \mathcal{T}_t(x_i)$. The operators span qualitatively different physical processes. In this work we instantiate five canonical

degradations through the following forward models:

$$\text{Noise:} \quad \mathcal{T}_{\mathrm{n}}(x) = x + \mathbf{n}, \quad \mathbf{n} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}); \tag{1}$$

$$\text{Motion blur:} \quad \mathcal{T}_{\mathrm{b}}(x) = \mathbf{k}_{L,\theta,C} \otimes x + \mathbf{n}_b; \tag{2}$$

$$\text{Haze:} \quad \mathcal{T}_{\mathrm{h}}(x) = x \odot \mathbf{t} + A\,(1 - \mathbf{t}), \quad \mathbf{t} = e^{-\beta \mathbf{d}}; \tag{3}$$

$$\text{Rain:} \quad \mathcal{T}_{\mathrm{r}}(x) = (1 - v)\big(x + \alpha \textstyle\sum_{l=1}^{L} \mathbf{s}_l\big) + v A_r; \tag{4}$$

$$\text{Low-light:} \quad \mathcal{T}_{\mathrm{l}}(x) = \tfrac{1}{k}\,\mathrm{Pois}\big(k\,x^{\gamma}\big) + \mathbf{n}_r, \tag{5}$$

where $\otimes$ denotes convolution with a motion blur kernel $\mathbf{k}_{L,\theta,C}$ rasterized from a parametric trajectory of length L, direction θ, and curvature C; $\odot$ is the Hadamard product; $\mathbf{t}$ is the transmission map governed by per-pixel scene depth $\mathbf{d}$ and scattering coefficient β; A is the global atmospheric light; $\mathbf{s}_l$ are per-layer rain streak maps at increasing depth planes (with α controlling streak intensity, v the atmospheric veiling, and A_r the gray atmospheric light from droplet scattering); $\gamma > 1$ controls illumination attenuation; $\mathrm{Pois}(\cdot)$ denotes a pixelwise Poisson resampling with photon gain k; and $\mathbf{n}_b$, $\mathbf{n}_r$ are Gaussian noise terms modeling post-blur sensor readout and read noise, respectively. These models follow established degradation formulations in the restoration literature (Sec. 3.3).

In the DIL setting, domains are encountered strictly sequentially: at stage t only $\mathcal{D}_t$ is accessible, while all prior domains $\mathcal{D}_{<t}$ are unavailable and no exemplars are retained. Let f_θ denote a restoration network with parameters θ. At the initial stage the backbone is trained on $\mathcal{D}_1$ under an ℓ_1 reconstruction loss, as is standard for restoration [7, 59]. For any later domain, unconstrained fine-tuning from the previous parameters minimizes the current-domain loss but overwrites the representations supporting earlier domains, the hallmark of catastrophic forgetting [12, 34]. After traversing all T domains, the model parameters Θ_T should minimize the average reconstruction loss over every domain seen so far,

$$\mathcal{L}_{\mathrm{CL}} = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{N_t} \sum_{i=1}^{N_t} \big\| f_{\Theta_T}(y_i^{(t)}) - x_i \big\|_1, \tag{6}$$

subject to the constraint that domain data is never jointly available, which separates our setting from all-in-one joint training [26, 39, 62], and that no past samples are stored, which separates it from replay-based continual learning [4, 5]. Crucially, the domain index t is also withheld at test time, so the model must additionally infer which restoration behavior to apply for each input. We resolve the training-time tension in Sec. 3.4 and the test-time routing in Sec. 3.5.

3.2 Pretrained Backbone and Degradation-Isolated Learning

Modern continual learning increasingly builds on a strong pretrained backbone used as a frozen, general-purpose feature extractor, with adaptation confined to lightweight modules [35, 45, 65]. Where discriminative CL takes an ImageNet [44]-pretrained encoder as this prior, we seek its restoration analogue, and image denoising is a natural candidate: a denoiser must learn the statistics of clean natural images to separate signal from corruption, which is precisely the prior that plug-and-play and regularization-by-denoising methods reuse as a universal proximal operator across inverse problems [50, 64]. A denoising-optimized backbone therefore captures degradation-agnostic restoration primitives that transfer to other degradations. We accordingly take the publicly released Gaussian color denoising weights of

Restormer [59] off the shelf as our backbone $f_{\theta_1^*}$ and freeze them throughout, so they serve as the shared, fixed substrate over which all subsequent degradations are learned. The same scheme transfers without modification to NAFNet [7] using its released denoising weights.

The DIL formulation idealizes a domain as differing from another only in its degradation operator, but real benchmarks violate this: each is collected for a single degradation and carries its own clean-content distribution (street scenes, indoor scenes, hazy landscapes). Training one restoration path per real dataset therefore conflates the degradation operator with the dataset content, and an adapter cannot tell which it is correcting for. To make this precise, let the clean content be a latent $c \sim p(c)$ and write $y = \mathcal{T}_t(c)$, with the frozen backbone producing a routing embedding $e = g_{\theta_1^*}(y)$. Under a multi-dataset construction, domain t draws content from its own distribution $p_t(c)$, so the degraded marginal $q_t(y) = \int p(y \mid c; \mathcal{T}_t)\, p_t(c)\, dc$ varies with t through both $\mathcal{T}_t$ and $p_t(c)$. In the counterfactual where all operators are identical, $\mathcal{T}_t \equiv \mathcal{T}$, any remaining dependence of e on t is purely content-driven, and the mutual information $I(e;t)$ decomposes as

$$I(e;t) \;=\; \underbrace{I_{\text{deg}}(e;t)}_{\text{from } \{\mathcal{T}_t\}} \;+\; \underbrace{I_{\text{content}}(e;t)}_{\text{from } \{p_t(c)\}}, \qquad I_{\text{content}}(e;t) > 0 \;\text{ whenever }\; p_t(c) \not\equiv p_{t'}(c). \tag{7}$$

The router can then reach high domain-identification accuracy by recognizing content rather than degradation, and each path absorbs the dataset-specific prior $p_t(c)$ into its learned correction, entangling it with the degradation inverse $\mathcal{T}_t^{-1}$. The remedy is to draw all domains from one shared corpus, $p_t(c) = p(c)$ for all t, defining each domain solely by its operator,

$$\mathcal{D}_t = \left\{ (x_i, y_i^{(t)}) : x_i \sim p(x),\; y_i^{(t)} = \mathcal{T}_t(x_i) \right\}. \tag{8}$$

Now the counterfactual $\mathcal{T}_t \equiv \mathcal{T}$ renders e identically distributed across domains, forcing $I_{\text{content}}(e;t) = 0$ and hence

$$I(e;t) \;=\; I_{\text{deg}}(e;t). \tag{9}$$

All domain-distinguishing information, and therefore all routing accuracy, is then attributable to the degradation operators alone. Symmetrically, since the clean targets share a common manifold, each path is optimized to invert its operator $\mathcal{T}_t$ without absorbing a content prior. This identifiability is why our benchmark is synthesized over a shared corpus rather than assembled from heterogeneous real datasets.

3.3 Degradation Synthesis

To instantiate Eq. (8), we apply the forward models of Eqs. (1)–(5) to a single clean corpus (DIV2K [1]), producing one paired dataset per degradation in which every domain shares identical clean source images. Each operator is parameterized to match the degradation characteristics reported in the corresponding restoration literature, so that synthetic samples approximate their real-world counterparts in appearance and difficulty. Additive Gaussian noise follows the AWGN model used throughout learned denoising [63]. Motion blur kernels are rasterized from parametric trajectories of explicit length, direction, and curvature following [2], with mild post-blur Gaussian sensor noise to mimic long-exposure handheld capture. Haze is rendered via the Koschmieder atmospheric scattering model with per-pixel transmission derived from monocular depth estimated by MiDaS [41], matching the synthesis convention of standard dehazing benchmarks [25]. Rain is synthesized as a multi-layer composite in which each layer represents a depth plane with progressively shorter, thinner, and

more strongly out-of-focus streaks, followed by atmospheric veiling from suspended droplet scattering, combining [14] with the multi-layer rendering of [58]. Low-light is generated by gamma-based illumination attenuation followed by signal-dependent Poisson shot noise and signal-independent Gaussian read noise, following the See-in-the-Dark sensor model [6] and consistent with low-light enhancement datasets [56].

3.4 Restoration Paths via Low-Rank Adaptation

Given the frozen backbone $f_{\theta_1^*}$, we learn each new degradation by attaching compact, trainable modules while leaving the backbone untouched. The atomic module is a low-rank adapter that applies a bottleneck residual correction to a block's features,

$$\mathcal{A}_\phi(\mathbf{z}) = \mathbf{z} + s\,\mathbf{W}_{\text{up}}\,\sigma\big(\mathbf{W}_{\text{down}}\,\text{LN}(\mathbf{z})\big), \tag{10}$$

where $\mathbf{z} \in \mathbb{R}^d$ is a feature vector, $\mathbf{W}_{\text{down}} \in \mathbb{R}^{r\times d}$ and $\mathbf{W}_{\text{up}} \in \mathbb{R}^{d\times r}$ form a rank-r bottleneck with $r \ll d$, σ is a ReLU, LN is layer normalization, and s is a fixed scale. Both projections are linear and act pointwise across the channel dimension. The trainable parameters are $\phi = \{\mathbf{W}_{\text{down}}, \mathbf{W}_{\text{up}}\}$, totaling $2dr + d$ per module. Following LoRA-style initialization [16, 18], $\mathbf{W}_{\text{down}}$ is Kaiming-initialized and $\mathbf{W}_{\text{up}}$ is set to zero, so each adapter is the identity at insertion and leaves the pretrained features unperturbed until training begins.

A single adapter only corrects one block, whereas inverting a degradation requires coordinated corrections along the entire encode–decode trajectory. We therefore instantiate an adapter at every block of the backbone and define a restoration path as the complete, network-spanning set of these per-block adapters for a given degradation,

$$\Phi_t = \{\phi_{t,1}, \phi_{t,2}, \ldots, \phi_{t,L}\}, \tag{11}$$

where L is the number of blocks. For a Transformer block, the path-augmented forward pass wraps the post-FFN residual,

$$\mathbf{z}'_l = \mathbf{z}_l + \text{Attn}(\text{LN}(\mathbf{z}_l)), \qquad \mathbf{z}_{l+1} = \mathcal{A}_{\phi_{t,l}}\big(\mathbf{z}'_l + \text{FFN}(\text{LN}(\mathbf{z}'_l))\big), \tag{12}$$

with Attn and FFN frozen. For convolutional blocks, the same injection follows the channel-mixing stage. Because a path is defined purely at the block level through the channel dimension d_l, the construction is backbone-agnostic, and we validate it on both a Transformer (Restormer [59]) and a fully convolutional network (NAFNet [7]) under identical path definitions. The full restoration network for domain t is $f_{\theta_1^*,\Phi_t}$: a shared frozen backbone routed through one degradation-specific path.

Sequential learning enforces strict parameter isolation. At domain t, the backbone and all committed paths $\Phi_{<t}$ are frozen and only a freshly initialized Φ_t is optimized,

$$\Phi_t^* = \arg\min_{\Phi_t} \frac{1}{N_t} \sum_{i=1}^{N_t} \big\| f_{\theta_1^*,\Phi_t}(y_i^{(t)}) - x_i \big\|_1. \tag{13}$$

Upon convergence, Φ_t^* is committed to a frozen bank and the procedure repeats. The final model state $\Theta_T = \{\theta_1^*\} \cup \{\Phi_1^*, \ldots, \Phi_T^*\}$ thus occupies mutually disjoint parameter subspaces for distinct domains, so the cross-domain gradient vanishes and forgetting is exactly zero by construction,

$$\frac{\partial \Phi_t^*}{\partial \Phi_j^*} = \mathbf{0} \quad (j \neq t) \implies \mathcal{F}_{j,t} = 0 \quad (j < t), \tag{14}$$

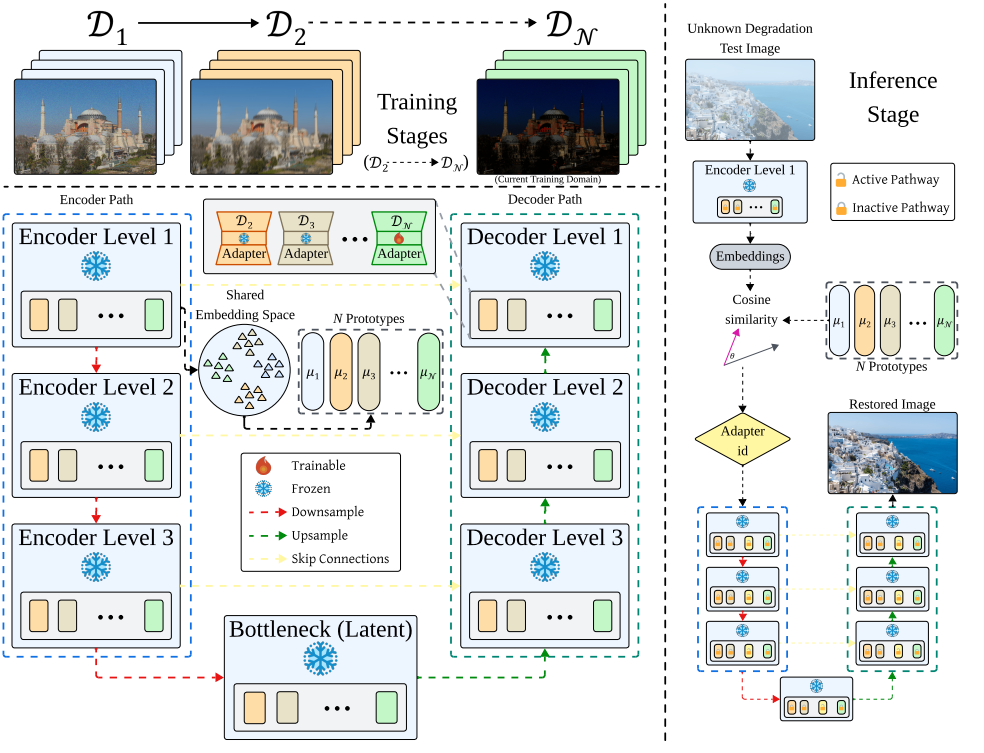


Figure 2: **RwF Framework:** a denoising-pretrained encoder–decoder backbone is frozen, and each new degradation trains a network-spanning path of low-rank adapters under strict parameter isolation, leaving prior paths and the backbone unchanged. A domain prototype is formed by pooling first-encoder-stage features. At inference, an unknown input is routed by cosine similarity to the matching path, which restores the image.

achieved without regularization penalties or replay, at a per-domain overhead of only $|\Phi_t| \approx 0.03\,|\theta_1^*|$ that scales linearly in the number of domains rather than duplicating the backbone. Because each path Φ_t and its prototype are estimated independently against the frozen backbone, with no gradient coupling across domains, the model state and the prototype bank are identical under any permutation of the domain order.

3.5 Prototype-Based Degradation Routing

Because the domain index is withheld at test time, the model must decide which restoration path to engage for each input. We route without any auxiliary classifier or extra trainable parameters by exploiting the geometry of the frozen backbone's feature space. A central design question is where along the backbone to read out features for prototyping. In classification CL, the deepest pre-classifier features are canonical, on the intuition that representational abstraction sharpens with depth. We find this prescription inverts for cross-degradation DIL: a restoration backbone is optimized to suppress the degradation signal en route to a clean reconstruction, so each successive encoder block, by design, attenuates precisely the cues a router needs. By the bottleneck stage, distinct degradations have largely converged toward

a shared clean-image attractor and become difficult to discriminate, whereas early-encoder features still carry visible imprints of the operator. We empirically corroborate this depth-discrimination trade-off in Sec. 4 (Tab. 3) and accordingly read out at the output of the last block of the first encoder stage, denoted $g^{(1)}_{\theta_1^*}(\cdot)$.

A second design choice is how to aggregate this spatial feature map into a fixed-length embedding. Different degradations leave signatures in different orders of spatial statistics. Noise inflates per-channel standard deviation without altering channel means, haze flattens both, rain streaks introduce structured high spatial variance over an otherwise clean background, and low-light shifts mean intensities. A first-order summary (Global Average Pooling, GAP) alone discards the variance signal, while a joint normalization of mean and standard deviation lets the larger-magnitude component dominate the cosine similarity. We therefore compute both moments and normalize them independently. Let $\mathbf{F} = g^{(1)}_{\theta_1^*}(y) \in \mathbb{R}^{C\times H\times W}$, and define

$$\mu_F = \frac{1}{HW}\sum_{h,w}\mathbf{F}_{:,h,w}, \qquad \sigma_F = \sqrt{\frac{1}{HW}\sum_{h,w}\big(\mathbf{F}_{:,h,w}-\mu_F\big)^2}. \tag{15}$$

We ℓ_2-normalize each moment independently and concatenate them into a single embedding,

$$\mathbf{e} = \Big[\,\mu_F/\|\mu_F\|_2 \;\Big\|\; \sigma_F/\|\sigma_F\|_2\,\Big] \in \mathbb{R}^{2C}. \tag{16}$$

The domain prototype is the mean embedding over a subset $\mathcal{S}_t$ of the domain's training samples, $\mu_t = \frac{1}{|\mathcal{S}_t|}\sum_{i\in\mathcal{S}_t}\mathbf{e}_i$. All embeddings are extracted through the frozen backbone with no path engaged, so prototypes inhabit a single shared space and cross-domain cosine comparisons are well-posed. At inference, an unknown input y^* is embedded via Eqs. (15)–(16) and routed to the path whose prototype it most resembles,

$$t^* = \arg\max_{t\in\{1,\dots,T\}} \frac{\langle \mathbf{e}^*, \mu_t\rangle}{\|\mathbf{e}^*\|_2\,\|\mu_t\|_2}, \tag{17}$$

after which the image is restored by a forward pass through the selected path, $\hat{x}^* = f_{\theta_1^*,\Phi_{t^*}}(y^*)$. In practice we compute each prototype from $\rho = 50\%$ of the domain's training samples, halving the prototype-extraction cost with negligible impact on routing accuracy. Fig. 2 summarizes the complete training and inference pipeline.

3.6 Continual Learning Evaluation Protocol

Standard restoration benchmarks score a model on one degradation, or under joint access to all degradations, and thus cannot reveal how performance on earlier degradations evolves as later ones are learned. We adopt the continual learning evaluation framework of [4], organized around an evaluation matrix $\mathbf{R} \in \mathbb{R}^{T\times T}$ whose entry $R_{b,j}$ is the restoration quality (PSNR or SSIM) on domain $\mathcal{D}_j$ after the model has been trained through domain $\mathcal{D}_b$, defined for $j \le b$. From this lower-triangular matrix we report three summaries,

$$A_B = \frac{1}{T}\sum_{j=1}^{T} R_{T,j}, \qquad \bar{A} = \frac{1}{T}\sum_{b=1}^{T}\frac{1}{b}\sum_{j=1}^{b} R_{b,j}, \qquad \mathcal{F} = \frac{1}{T-1}\sum_{j=1}^{T-1}\big(R_{j,j}-R_{T,j}\big), \tag{18}$$

where A_B is the average final quality across all domains after the last stage, $\bar{A}$ is the average incremental quality over all stages and thus reflects stability throughout the sequence, and $\mathcal{F}$ is the average forgetting, the mean drop on each domain relative to its peak, with $\mathcal{F} = 0$ denoting perfect retention.

4 Experiments

4.1 Experimental Setup

Backbones and adapter paths. We instantiate RwF on two restoration backbones spanning the dominant architectural families: a Transformer, Restormer [59], and a fully convolutional network, NAFNet [7]. For each, we initialize from publicly released Gaussian-denoising weights (Restormer's blind Gaussian color denoising checkpoint; NAFNet's SIDD width-32 checkpoint) and freeze the backbone throughout. These choices instantiate $\mathcal{D}_1$ directly: the pretrained backbone serves as the domain-1 restorer (Sec. 3.2), and only adapter paths are trained for $\mathcal{D}_2$–$\mathcal{D}_5$. All adapter slots use a bottleneck dimension of r=64, input LayerNorm, dropout 0.1, and a fixed scale s=1.0, applied within the FFN of each Restormer block and the channel-mixing branch of each NAFNet block. Per-domain adapter parameters total 0.98M for Restormer (3.63% of the 27.11M backbone) and 1.29M for NAFNet (4.25% of the 30.45M backbone), scaling linearly with T rather than duplicating the backbone. Aggregated over the full five-domain sequence, RwF-Restormer totals one shared 27.11M backbone plus four adapters ($\mathcal{D}_2$–$\mathcal{D}_5$), roughly 31M parameters, versus 5×27.11=135.6M for five separate per-degradation specialists (over 4× fewer), while routing adds only a single stage-1 forward pass at inference and no trainable parameters. Restormer's bias-free denoising LayerNorm [56] caused exploding gradients under the larger degradation shift, so we reintroduced the bias term. NAFNet required no analogous adjustment.

Training and evaluation protocol. Each domain stage is trained for 100k iterations with ℓ_1 pixel loss, AdamW (lr 2×10^{-4}, weight decay 10^{-4}, β=(0.9, 0.999)), and a single-cycle cosine schedule decaying to 10^{-6}, on 224×224 patches with batch size 8 and standard geometric augmentations. The starting learning rate is reduced below the from-scratch defaults of Restormer (3×10^{-4}) and NAFNet (10^{-3}) because we fine-tune over a pretrained backbone. Identical hyperparameters are used for the sequential FT baseline and the RwF variant of each backbone; the only difference is which parameters are unfrozen. We deliberately omit progressive patch-size training, larger architectural variants, and extended schedules common in single-task restoration recipes, since our aim is to establish a continual learning baseline rather than a state-of-the-art per-task result. Each run uses a single NVIDIA A100 80GB GPU. We evaluate under two settings. The *oracle* setting supplies the ground-truth domain identity at test time, and the *domain-agnostic* setting reflects the true DIL scenario, where the domain label is withheld and the prototype router (Sec. 3.5) selects an adapter from the input alone. We report PSNR [17] and SSIM [55], summarized by the metrics A_B, $\bar{A}$, and $\mathcal{F}$ (Sec. 3.6). Following denoising-evaluation convention, $\mathcal{D}_1$ is evaluated on CBSD68 [33] (the canonical test corpus for Restormer's released denoising weights), with $\mathcal{D}_1$ prototypes extracted on DIV2K-synthesized noise images to preserve a common content distribution across prototypes (Eq. (9)). $\mathcal{D}_2$–$\mathcal{D}_5$ are evaluated on held-out DIV2K-synthesized test sets.

Baselines. We compare RwF against four references. Sequential FT of the same two backbones provides the natural reference for catastrophic forgetting under unconstrained parameter updates. We add two regularization-based continual learning baselines, EWC [23] and LwF [28], applied to both backbones under conditions identical to sequential FT, with their regularization strengths tuned once at the $\mathcal{D}_1\to\mathcal{D}_2$ transition and held fixed thereafter. As these methods share a single model, we evaluate them against RwF in the oracle setting. We further report a Joint all-in-one model trained on the union of all five domains as a

non-continual reference. We do not retrain specialized all-in-one networks (PromptIR [39], AirNet [26], IDR [62], AdaIR [10], TransWeather [48], DCPT [19]) under our sequential protocol, treating them as complementary closed-world methods rather than direct competitors. Deployment on canonical real-degradation benchmarks is evaluated in Sec. 4.3.

4.2 Comparison with Continual Learning Baselines

Table 1: Continual learning comparison across the five-domain sequence ($\mathcal{D}_1 \to \mathcal{D}_5$: noise, blur, rain, haze, low-light) under the *oracle* setting.

		PSNR (dB)			SSIM		
Method	Backbone	$A_B\uparrow$	$\bar{A}\uparrow$	$\mathcal{F}\downarrow$	$A_B\uparrow$	$\bar{A}\uparrow$	$\mathcal{F}\downarrow$
Sequential FT	NAFNet	12.20	20.79	15.83	0.5697	0.6262	0.3287
	Restormer	13.62	24.16	22.05	0.6148	0.7102	0.3705
EWC	NAFNet	11.63	20.78	16.00	0.5376	0.6256	0.3366
	Restormer	12.00	21.26	14.43	0.4521	0.6133	0.3193
LwF	NAFNet	18.10	22.14	0.22	0.5559	0.6591	0.0305
	Restormer	17.59	23.25	3.04	0.5361	0.6922	0.1351
Joint (All-in-one)	NAFNet	25.49	–	–	0.7809	–	–
	Restormer	26.72	–	–	0.8230	–	–
RwF (Ours)	NAFNet	24.03	24.80	**0.00**	0.8003	0.7722	**0.0000**
	Restormer	**28.87**	**29.59**	**0.00**	**0.8744**	**0.8658**	**0.0000**

Tab. 1 summarizes the five-domain continual trajectory under the oracle setting. Sequential FT suffers severe forgetting on both backbones. $\mathcal{F}$ reaches 22.05 dB for Restormer and 15.83 dB for NAFNet, and the final-stage average A_B collapses to 13.62 dB and 12.20 dB respectively, far below the per-domain peaks captured in $\bar{A}$. This gap between A_B and $\bar{A}$, exceeding 10 dB on Restormer and 8 dB on NAFNet, tracks the progressive overwriting of earlier-domain representations as each new degradation is absorbed. The standard continual-learning remedies fare little better under this degree of inter-degradation shift. EWC does not improve on naive fine-tuning, its A_B of 12.00 dB on Restormer and 11.63 dB on NAFNet sitting at or below sequential FT, with forgetting reduced only on Restormer. LwF holds forgetting to $\mathcal{F}$ of 3.04 and 0.22 dB, but restricting a single shared model to every domain caps its final quality at A_B of 17.59 and 18.10 dB. Regularization of this kind, whether over gradients or outputs, preserves competence in class-incremental classification but not when successive domains demand qualitatively different restoration behaviors.

RwF removes this collapse by construction. With zero forgetting, final A_B rises to 28.87 dB for Restormer, a +15.25 dB gain over sequential FT, and to 24.03 dB for NAFNet (+11.83 dB), narrowing the gap between A_B and $\bar{A}$ to under 1 dB on both backbones, with SSIM following the same trend. These results are competitive with joint all-in-one training over the five domains, which reaches 26.72 and 25.49 dB while optimizing on all degradations at once. RwF-Restormer exceeds its joint counterpart and RwF-NAFNet trails it only slightly, so assigning each degradation an isolated path recovers the quality of joint optimization without revisiting past data, and on the stronger backbone surpasses it by sidestepping cross-degradation interference. The behavior is consistent across architectures, holding for

both a Transformer and a convolutional backbone under identical hyperparameters, with Restormer leading NAFNet by close to 5 dB under RwF.

4.3 Domain-Agnostic Inference on Real Degradations

Under the domain-agnostic setting, the router must infer adapter identity (Sec. 3.5) from the degraded input alone. To test whether this transfers beyond the synthetic training corpus, we deploy RwF-Restormer on eleven canonical real-degradation benchmarks across all five domains without any retraining or recalibration: CBSD68 [33], Kodak24 [11], and Urban100 [20] with synthetic additive Gaussian noise at σ=25 for $\mathcal{D}_1$; RealBlur-J and RealBlur-R [42] for $\mathcal{D}_2$; Rain100H, Rain100L [58], and Test100 [61] for $\mathcal{D}_3$; SOTS-Indoor and SOTS-Outdoor [25] for $\mathcal{D}_4$; and LOL-v1 [56] for $\mathcal{D}_5$, totaling 3,465 images. Tab. 2 reports per-benchmark routing accuracy and reconstruction quality.

Table 2: Domain-agnostic evaluation of RwF-Restormer on eleven real-degradation benchmarks. Routing accuracy and reconstruction quality are reported under both *domain-agnostic* (Pred.) and *oracle* settings.

				PSNR (dB) ↑		SSIM ↑	
Domain	Test set	N	ID Acc. (%)	Pred.	Oracle	Pred.	Oracle
$\mathcal{D}_1$ (noise)	CBSD68	68	92.6	28.76	30.18	0.8448	0.8713
	Kodak24	24	100.0	31.66	31.66	0.8754	0.8754
	Urban100	100	90.0	28.74	30.75	0.8622	0.8906
$\mathcal{D}_2$ (blur)	RealBlur-J	980	93.6	26.55	26.57	0.8349	0.8347
	RealBlur-R	980	94.9	32.78	33.88	0.8962	0.9379
$\mathcal{D}_3$ (rain)	Rain100H	100	100.0	16.92	16.92	0.4696	0.4696
	Rain100L	100	83.0	30.18	32.13	0.8942	0.9291
	Test100	98	70.4	21.20	21.64	0.6399	0.6762
$\mathcal{D}_4$ (haze)	SOTS-Indoor	500	91.2	19.94	20.58	0.8728	0.8946
	SOTS-Outdoor	500	72.4	25.50	27.93	0.9305	0.9625
$\mathcal{D}_5$ (low-light)	LOL-v1	15	46.7	12.13	16.65	0.4495	0.7356
Overall		**3,465**	**89.5**	**26.96**	**27.90**	**0.8568**	**0.8808**

Overall routing accuracy reaches 89.5% (3,101 / 3,465 images correctly routed), with a residual +0.94 dB PSNR gap to the oracle ceiling. Seven of the eleven benchmarks exceed 90% routing accuracy, and two (Kodak24 noise, Rain100H) achieve 100% with predicted and oracle reconstruction quality matching exactly. The framework therefore transfers to real degradations without adaptation, supporting the content-controlled design rationale (Sec. 3.2): isolating adapter learning from dataset-specific content priors lets the restoration paths generalize to unseen content of the same physical family. Crucially, misrouting is also graceful rather than catastrophic. As adjacent operators share restoration primitives (rain

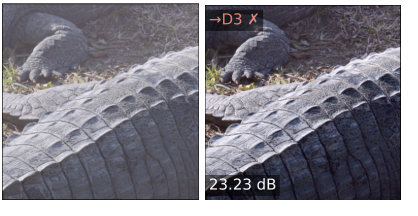


Figure 3: A $\mathcal{D}_4$ hazy input (left) misrouted to the $\mathcal{D}_3$ derain path is still largely restored (right).

synthesis includes atmospheric veiling, low-light includes sensor noise; Sec. 3.3), a wrong route still applies a related correction rather than an arbitrary one, as in Fig. 3.

Two cases, however, warrant attention. LOL-v1 is the weakest at 46.7% routing accuracy, the regime our depth-discrimination argument predicts will be hard, since low-light's signature is a global intensity statistic that early encoder features partially equalize. Oracle routing on the same data recovers 16.65 dB, so the failure is localized to routing, not the adapter. Rain100H shows the opposite imbalance, routing perfectly (100%) but reaching only 16.92 dB because the dataset's heavy-rain severity exceeds our $\mathcal{D}_3$ synthesis distribution, an adapter generalization gap rather than a routing failure.

4.4 Ablation: Feature Readout Depth and Aggregation

Two design choices govern the quality of the routing embedding: where along the backbone to read out the feature map, and how to summarize the resulting spatial tensor into a fixed-length vector. We ablate each independently on RwF-Restormer using the synthetic DIV2K-derived test set, which provides controlled comparison across configurations on identical content. The real-benchmark transfer of the selected configuration is reported in Sec. 4.3.

Table 3: Routing quality vs. feature readout depth on RwF-Restormer.

Readout location	ID Acc. (%)	Oracle gap (dB)
Encoder stage 1 (Ours)	**79.7**	**+2.01**
Encoder stage 2	70.7	+2.54
Encoder stage 3	69.2	+2.81
Latent (bottleneck)	58.8	+4.14

Table 4: Routing quality vs. embedding aggregation.

Aggregation	ID Acc. (%)
GAP only	73.9
GAP + std (joint ℓ_2)	78.4
GAP + std (sep. ℓ_2, Ours)	**79.7**

Tab. 3 reports routing accuracy and oracle PSNR gap as a function of readout depth, holding aggregation fixed at GAP + std with separate ℓ_2 normalization. Routing quality degrades monotonically with depth: encoder stage 1 achieves 79.7% accuracy (+2.01 dB gap), while the bottleneck collapses to 58.8% and +4.14 dB. This reverses the classification-CL convention that deeper pre-classifier features yield more discriminative prototypes. A restoration backbone, by design, suppresses the very signal that distinguishes the domains, so late features grow informative about the clean image but less so about what was removed. Aggregating across scales, as adopted for degradation classification in DCPT [19], does not overcome this. Concatenating stages 1–3 yields 77.1% accuracy and adding the latent stage drops it to 72.0%, both below the single shallow readout. This proves that combining scales only reintroduces the deeper features that dilute the degradation cue.

Tab. 4 reports the effect of aggregation at the fixed encoder-stage-1 readout. GAP alone reaches 73.9% accuracy. Adding per-channel spatial standard deviation (GAP + std, jointly ℓ_2-normalized) lifts it to 78.4%, confirming that second-order spatial statistics carry meaningful degradation signal beyond the mean, and normalizing the two moments separately before concatenation lifts it further to 79.7%. We adopt the separately-normalized variant on the basis of its higher domain-identification accuracy.

4.5 Qualitative Analysis

Fig. 4 visualizes the practical consequence of the quantitative gap in Tab. 1 on a single shared scene degraded under each of the five operators. Sequential fine-tuning recovers only $\mathcal{D}_5$, the

Figure 4: Qualitative comparison across all five degradation domains for a shared scene. **Top:** degraded inputs. **Middle:** sequentially fine-tuned Restormer after the full $\mathcal{D}_1 \to \mathcal{D}_5$ schedule. **Bottom:** RwF-Restormer under domain-agnostic inference, with the router's prediction ($\to \mathcal{D}_t$ ✓). Per-image PSNR is shown in the lower-left of each restored panel.

most recently trained domain (25.05 dB). On every earlier domain, the final model outputs heavily artifacted reconstructions, collapsing to 3.78 dB on $\mathcal{D}_4$ where subsequent training stages have overwritten the model's haze-removal capability. RwF, by contrast, produces visually consistent reconstructions across all five domains (23.84 – 35.15 dB), with the prototype router correctly selecting the appropriate restoration path for every input despite the substantial appearance differences between degraded variants.

5 Conclusion

We formulated multi-degradation image restoration as a continual domain-incremental learning problem and introduced Restoring without Forgetting (RwF), a parameter-efficient framework that freezes a denoising-pretrained backbone, attaches network-spanning low-rank adapter paths under strict parameter isolation, and routes unknown inputs to the matching path via unsupervised prototype matching. On a content-controlled five-domain benchmark spanning noise, blur, rain, haze, and low-light, RwF eliminates catastrophic forgetting by construction and improves final-stage average PSNR by up to +15.25 dB over naive sequential fine-tuning, on both a Transformer and a convolutional backbone. The framework transfers without modification to eleven canonical real-degradation benchmarks (3,465 images), reaching 89.5% routing accuracy and a +0.94 dB oracle PSNR gap, establishing, to our knowledge, the first systematic baseline of its kind. Two directions remain open. First, routing accuracy degrades on degradations whose signature is primarily global rather than spatially localized, low-light enhancement in particular. A degradation-aware embedding head or a differentiable router could absorb the remaining accuracy gap on such regimes. Second, designing adapter pathway primitives that explicitly capture degradation-specific structure through frequency-domain projections, kernel-aware re-parameterizations, or richer cross-block interactions, is a promising direction for further per-domain quality gains, particularly on real benchmarks whose distribution exceeds the synthesis range of the training corpus.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017.

[2] Giacomo Boracchi and Alessandro Foi. Modeling the performance of image restoration from motion blur. *IEEE Transactions on Image Processing*, 21(8):3502–3517, 2012.

[3] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE transactions on image processing*, 25(11):5187–5198, 2016.

[4] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018.

[5] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

[6] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3291–3300, 2018.

[7] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33. Springer, 2022.

[8] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

[9] Shuang Cui, Yi Li, Jiangmeng Li, Xiongxin Tang, Bing Su, Fanjiang Xu, and Hui Xiong. Continual test-time adaptation for single image defocus deblurring via causal siamese networks. *International Journal of Computer Vision*, 133(7):4134–4157, 2025.

[10] Yuning Cui, Syed Waqas Zamir, Salman Khan, Alois Knoll, Mubarak Shah, and Fahad Shahbaz Khan. Adair: Adaptive all-in-one image restoration via frequency mining and modulation. In *13th international conference on learning representations, ICLR 2025*, pages 57335–57356. International Conference on Learning Representations, ICLR, 2025.

[11] Rich Franzen. Kodak lossless true color image suite, 1999.

[12] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.

[13] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3855–3863, 2017.

[14] Kshitiz Garg and Shree K Nayar. Vision and rain. *International Journal of Computer Vision*, 75(1):3–27, 2007.

[15] Xiaoxuan Gong and Jie Ma. A minimalistic unified framework for incremental learning across image restoration tasks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[17] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[19] JiaKui Hu, Lujia Jin, Zhengjian Yao, and Yanye Lu. Universal image restoration pre-training via degradation classification. In *International Conference on Learning Representations*, volume 2025, pages 64647–64668, 2025.

[20] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015.

[21] Kishaan Jeeveswaran, Elahe Arani, and Bahram Zonooz. Gradual divergence for seamless adaptation: A novel domain incremental learning method. *arXiv preprint arXiv:2406.16231*, 2024.

[22] Xingyu Jiang, Xiuhui Zhang, Ning Gao, and Yue Deng. When fast fourier transform meets transformer for image restoration. In *European conference on computer vision*, pages 381–402. Springer, 2024.

[23] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[24] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5886–5895, 2023.

[25] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE transactions on image processing*, 28(1):492–505, 2018.

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

[27] Hao Li, Xiang Chen, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Foundir: Unleashing million-scale training data to advance foundation models for image restoration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12626–12636, 2025.

[28] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[29] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[30] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[31] Yihao Liu, Jingwen He, Jinjin Gu, Xiangtao Kong, Yu Qiao, and Chao Dong. Degae: A new pretraining paradigm for low-level vision. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23292–23303. IEEE, 2023.

[32] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018.

[33] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, volume 2, pages 416–423. Ieee, 2001.

[34] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[35] Mark D McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton Van den Hengel. Ranpac: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems*, 36:12022–12053, 2023.

[36] Sreyas Mohan, Zahra Kadkhodaie, Eero P Simoncelli, and Carlos Fernandez-Granda. Robust and interpretable blind image denoising via bias-free convolutional neural networks. *arXiv preprint arXiv:1906.05478*, 2019.

[37] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017.

[38] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113: 54–71, 2019.

[39] Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. Promptir: prompting for all-in-one blind image restoration. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[40] Chu-Jie Qin, Rui-Qi Wu, Zikun Liu, Xin Lin, Chun-Le Guo, Hyun Hee Park, and Chongyi Li. Restore anything with masks: Leveraging mask image modeling for blind all-in-one image restoration. In *European Conference on Computer Vision*, pages 364–380. Springer, 2024.

[41] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44 (3):1623–1637, 2020.

[42] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *European conference on computer vision*, pages 184–201. Springer, 2020.

[43] Xue Rui, Ziqiang Li, Yang Cao, Ziyang Li, and Weiguo Song. Dilrs: Domain-incremental learning for semantic segmentation in multi-source remote sensing data. *Remote Sensing*, 15(10):2541, 2023.

[44] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115 (3):211–252, 2015.

[45] Hai-Long Sun, Da-Wei Zhou, Hanbin Zhao, Le Gan, De-Chuan Zhan, and Han-Jia Ye. Mos: Model surgery for pre-trained model-based class-incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20699–20707, 2025.

[46] Chunwei Tian, Yong Xu, and Wangmeng Zuo. Image denoising using deep cnn with batch renormalization. *Neural Networks*, 121:461–473, 2020.

[47] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5769–5780, 2022.

[48] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2353–2363, 2022.

[49] Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

[50] 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.

[51] Guankun Wang, Long Bai, Yanan Wu, Tong Chen, and Hongliang Ren. Rethinking exemplars for continual semantic segmentation in endoscopy scenes: Entropy-based mini-batch pseudo-replay. *Computers in Biology and Medicine*, 165:107412, 2023.

[52] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383, 2024.

[53] Qiang Wang, Xiang Song, Yuhang He, Jizhou Han, Chenhao Ding, Xinyuan Gao, and Yihong Gong. Boosting domain incremental learning: Selecting the optimal parameters is all you need. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4839–4849, 2025.

[54] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022.

[55] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[56] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

[57] Cheng Xu, Weiwen Zhang, Hongrui Zhang, Xuemiao Xu, Huaidong Zhang, Jing Zou, and Jing Qin. Fr2seg: Continual segmentation across multiple sites via fourier style replay and adaptive consistency regularization. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, pages 8815–8823, 2025.

[58] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1357–1366, 2017.

[59] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[60] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. Pmlr, 2017.

[61] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Image de-raining using a conditional generative adversarial network. *IEEE transactions on circuits and systems for video technology*, 30(11):3943–3956, 2019.

[62] Jinghao Zhang, Jie Huang, Mingde Yao, Zizheng Yang, Hu Yu, Man Zhou, and Feng Zhao. Ingredient-oriented multi-degradation learning for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5825–5835, 2023.

[63] 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.

[64] 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.

[65] Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23554–23564, 2024.

[66] Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *International Journal of Computer Vision*, 133(3):1012–1032, 2025.

[67] Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, Lijun Zhang, and De-Chuan Zhan. Dual consolidation for pre-trained model-based domain-incremental learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20547–20557, 2025.