CROSS-EXPERT LEARNING FOR ALL-IN-ONE IMAGE RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing mixture-of-experts (MoE) methods typically assign the learner and supervisor roles to the same experts. In this paper, we introduce **Cross-Expert Learning (CEL)**, in which each learner receives utility-weighted joint supervision from all experts. To construct this supervision, **Cross-Expert Utility Estimation (CUE)** evaluates all experts on shared input features to obtain candidate restoration responses. Each candidate response is scored by its local alignment with the restoration descent direction, which defines its utility. The utilities are input-dependent, varying with the degradation characteristics and image content of the current input. CUE estimates these utilities and uses them to combine the expert responses into supervision for the learner. Throughout this learning process, CEL requires neither degradation labels nor an external teacher. The additional expert responses are used only during training and do not enter the native prediction path. Experiments on standard and composite all-in-one restoration benchmarks show consistent performance improvements of CEL across different expert-based backbones, without additional model parameters or inference FLOPs.

1 INTRODUCTION

All-in-one (AiO) image restoration aims to handle multiple unknown degradations with a single model (Jiang et al., 2025). Different degradations require different restoration behaviors, even though much of the underlying image structure and restoration knowledge can be shared (Zhang et al., 2017; Zamir et al., 2020; 2021; Liang et al., 2021; Chen et al., 2021a; Cho et al., 2021; Zamir et al., 2022; Wang et al., 2022; Tu et al., 2022; Chen et al., 2022; Song et al., 2022; Wu et al., 2025b). Existing methods accommodate this diversity through degradation representations, prompts, adapters, and dynamic modulation (Li et al., 2022; Potlapalli et al., 2023; Conde et al., 2024; Cui et al., 2025; Tian et al., 2025; Wu et al., 2025c; Zhang et al., 2025; Ren et al., 2025). Mixture-of-experts (MoE) models provide an explicit way to model such diversity by maintaining multiple expert transformations and invoking them according to the input (Shazeer et al., 2017; Riquelme et al., 2021; Fedus et al., 2022; Zhou et al., 2022; Zamfir et al., 2025; Wang et al., 2025; 2026; Peng et al., 2026). Existing MoE studies have focused on how experts are constructed, selected, and combined, leaving a distinct question: how should learning be organized across experts?

In existing MoE methods, the same expert assignment typically determines both which experts learn from a sample and which expert behaviors are available to supervise that learning. We refer to an expert as a learner when its parameters are updated by the current sample, and as a supervisor when its response contributes to the learning signal for that sample. Here, supervisor refers to a role in the learning process rather than a separate pretrained teacher. As illustrated on the left of Figure 1(a), these two roles are coupled in standard MoE learning: the same expert assignment determines both where sample-specific knowledge is updated and which expert behaviors are available to guide that update. However, the two roles serve different purposes, suggesting that this coupling is a design choice rather than a requirement.

To examine whether this coupling is justified in practice, we analyze the MoE-based restoration model (Zamfir et al., 2025) on a held-out AIO-5 analysis split drawn from the training data and disjoint from the test sets. We sample 500 image–layer cases across the five restoration tasks, where each case corresponds to one image evaluated at one monitored MoE layer. For each case, we collect the responses of the selected learner and the non-selected experts under the same local feature state.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

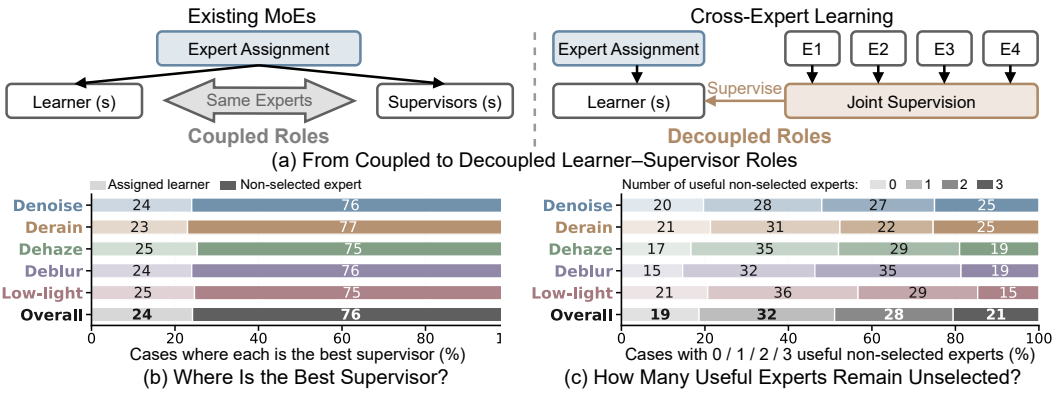


Figure 1: **Motivation for Cross-Expert Learning.** (a) CEL decouples the selected learner from the source of supervision while preserving the native prediction path. (b) Frequency with which the best supervisor is the selected learner or a non-selected expert. (c) Number of useful non-selected experts per sample-layer case.

We normalize these responses and apply them individually as equal-scale local interventions at the monitored layer, while keeping the remaining network computation unchanged. We then compare the resulting restoration loss with that of the unperturbed forward pass and regard a response as favorable if it reduces the loss. As shown in Figure 1(b), the most favorable response comes from a non-selected expert in 76% of cases. Figure 1(c) further shows that at least one non-selected expert provides a favorable response in 81% of cases, with useful responses often distributed across multiple peers. These observations show that useful supervision can arise from experts outside the learner set. This raises a natural question: *can a learner exploit useful responses from other experts without involving them in the current prediction or updating them on the same sample?*

We introduce **Cross-Expert Learning (CEL)**, which decouples the learner from the source of supervision without changing the backbone's native prediction path. As illustrated on the right of Figure 1(a), the backbone's original expert selection still determines which experts act as learners for the current sample, while responses from the full expert bank can serve as supervision during training. Thus, an expert may supervise another expert without entering the current prediction or becoming a learner for the same sample. This decoupling broadens the supervision available to each learner while preserving the backbone's original inference behavior.

For AiO restoration, we develop **Cross-Expert Utility Estimation (CUE)** to construct cross-expert supervision within CEL. CUE evaluates all experts on the same input features to obtain candidate restoration responses. Because the utility of each response varies with the degradation characteristics and image content of the current input, CUE evaluates these responses against the restoration objective and assigns them input-dependent utilities. The responses are then combined according to their estimated utilities to provide joint supervision for the learner. CUE requires neither degradation labels nor an external teacher, and all additional expert evaluations are confined to training, leaving the native prediction path, parameter count, and inference FLOPs unchanged.

Our experiments evaluate both the effectiveness and the learning mechanism of CEL. Across AIO-3, AIO-5, and CDD11, CEL consistently improves performance across different expert-based backbones, while controlled analyses validate cross-expert supervision, learner-supervisor decoupling, and CUE utility estimation.

Our contributions are summarized as follows:

- We identify a common restriction in existing MoE learning: the learner and supervisor roles are typically assigned to the same experts. Our analysis shows that experts outside the learner set can provide useful supervision, motivating these two roles to be separated.
- We introduce **Cross-Expert Learning (CEL)**, which allows each learner to receive joint supervision from the full expert population without changing the native prediction path. We further develop **Cross-Expert Utility Estimation (CUE)** to estimate the input-dependent utility of expert responses and construct cross-expert supervision accordingly.

- We validate CEL on standard and composite AiO restoration benchmarks and through controlled analyses of cross-expert supervision, learner-supervisor decoupling, and CUE utility estimation, without additional model parameters or inference FLOPs.

2 Related Work

2.1 All-in-One Image Restoration

All-in-one image restoration aims to handle diverse unknown degradations within a single model. Early methods learn compact degradation representations to condition a shared restoration network (Li et al., 2020; 2022; Valanarasu et al., 2022; Park et al., 2023), while later approaches broaden unified restoration through learned prompts, language instructions, diffusion-based formulations, and input-dependent modulation (Potlapalli et al., 2023; Conde et al., 2024; Luo et al., 2024; Özdenizci & Legenstein, 2023; Jiang et al., 2024; Cui et al., 2025; Wu et al., 2025c). Recent methods further strengthen this adaptation through dynamic parameterization, structured degradation modeling, and semantic or quality-aware guidance (Tian et al., 2025; Tang et al., 2025; Wu et al., 2025a; Zhang et al., 2025; Cui et al., 2026; Ren et al., 2026; He et al., 2026; Liu et al., 2026; Zhang et al., 2026; Ren et al., 2025). These methods mainly focus on adapting the restoration prediction to the degradation and content of the current input. How diverse restoration responses can instead be exploited as supervision during learning remains less explored.

2.2 Mixture-of-Experts for Image Restoration

Mixture-of-experts (MoE) methods explicitly represent restoration diversity with multiple expert transformations. One line of work promotes expert specialization through sparse, input-dependent routing (Shazeer et al., 2017; Lepikhin et al., 2020; Lewis et al., 2021; Roller et al., 2021; Riquelme et al., 2021; Fedus et al., 2022; Zhou et al., 2022; Zoph et al., 2022; Zamfir et al., 2025; Yu et al., 2024); subsequent methods improve expert selection using degradation-aware features or semantic priors (Wang et al., 2025; Peng et al., 2026); more recent designs introduce shared or collaborative experts to exchange information across specialized restoration behaviors (Wang et al., 2026). These developments mainly improve how experts are specialized, selected, and combined. However, the learner and supervisor roles are still typically determined by the same expert assignment: the experts selected to learn from a sample are also those whose responses are available to supervise that learning. Consequently, useful responses from experts outside the learner set cannot directly contribute to the supervision of the current learner.

2.3 Supervision Beyond the Learner

Learning signals need not come solely from the optimized model or expert. Knowledge distillation transfers supervision from a predefined or pretrained teacher to a separate learner (Hinton et al., 2015; Zagoruyko & Komodakis, 2016; Tian et al., 2019; Chen et al., 2021b; Zhao et al., 2022), while mutual and collaborative learning allow peer models to supervise one another during joint optimization (Furlanello et al., 2018; Zhu et al., 2018; Zhang et al., 2018; Chen et al., 2020; Wu & Gong, 2021). Recent MoE studies exploit information beyond selected experts through mutual expert distillation (Xie et al., 2024), distillation from non-activated experts (Kim et al., 2025), supplementary modules derived from unselected experts (Zhao et al., 2024), and analysis of alternative execution routes (Yoon et al., 2026). These studies show that other models or experts can provide useful learning signals. However, existing approaches variously rely on a separate teacher-student relation, jointly optimize the supervising experts, inject additional expert information into prediction, or alter the execution route. How to use peer expert responses as supervision while keeping the native prediction path and learner assignment unchanged remains less explored.

3 Cross-Expert Learning

Cross-Expert Learning (CEL) separates where expert-specific learning is applied from where the supervision for that learning is obtained. Consider an expert bank at layer l,

$$\mathcal{E}^l = \{E_1^l, \ldots, E_N^l\}. \tag{1}$$

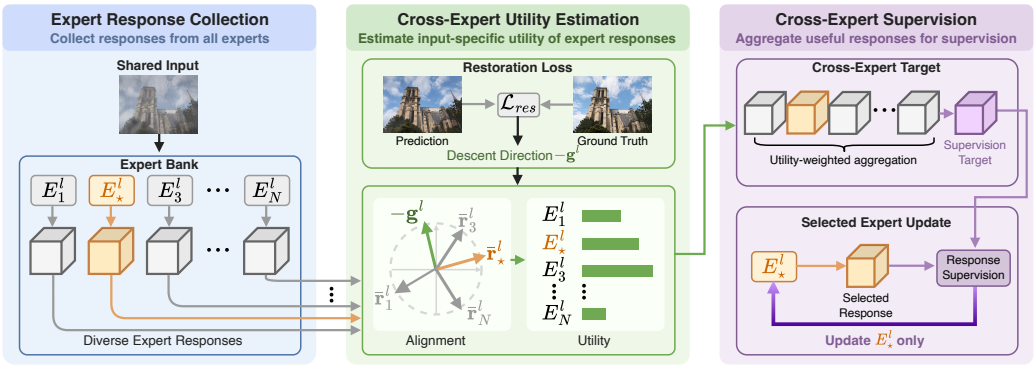


Figure 2: **Training-time realization of CEL.** Expert responses are collected under a shared local input and evaluated by CUE according to their input-specific utility. The utility-weighted responses form a detached cross-expert target that supervises only the selected learner. CEL-specific training operations are removed at inference.

For the current input, the backbone's native expert-selection mechanism determines a learner set

$$\mathcal{A}^l \subseteq \{1, \ldots, N\}. \tag{2}$$

CEL leaves this learner assignment and the native prediction path unchanged. Instead, it allows the supervision of each learner to be constructed from responses across the full expert bank. Thus, the backbone continues to determine *which experts learn*, while CEL broadens *which expert responses can supervise that learning*.

As illustrated in Figure 2, CEL realizes this principle in three stages. **Expert Response Collection** obtains comparable candidate responses from all experts under a shared local input. **Cross-Expert Utility Estimation (CUE)** evaluates the input-specific utility of these responses with respect to the restoration objective. **Cross-Expert Supervision** then combines the responses according to their utilities and transfers the resulting supervision only to the selected learner. For clarity, we describe the Top-1 setting used in our experiments, with a single learner $E^l_\star \in \mathcal{A}^l$; the construction extends to multiple selected learners by applying it independently to each learner.

3.1 Expert Response Collection

Cross-expert supervision requires candidate responses from different experts to be directly comparable. We therefore monitor an interface at which every expert contribution can be expressed as an additive correction to a shared reference representation $\mathbf{z}^l$. Let $\mathbf{x}^l$ denote the shared local input to the expert bank. The response of expert E^l_k is written as

$$\mathbf{z}^l_k = \mathbf{z}^l + \rho^l_k(\mathbf{x}^l), \tag{3}$$

where ρ^l_k denotes the expert-specific response at the monitored interface. For residual expert blocks, ρ^l_k is the native residual response. When a backbone instead exposes a transformed expert feature, we define ρ^l_k as the expert-specific difference from the shared reference representation. In both cases, expert responses are represented in the same local space and can be interpreted as perturbation directions applied to $\mathbf{z}^l$.

We evaluate every expert using the same detached local input and collect

$$\bar{\mathbf{r}}^l_k = \mathrm{sg}\left[\rho^l_k\left(\mathrm{sg}(\mathbf{x}^l)\right)\right], \qquad k = 1, \ldots, N, \tag{4}$$

forming the response set

$$\mathcal{R}^l = \{\bar{\mathbf{r}}^l_1, \ldots, \bar{\mathbf{r}}^l_N\}. \tag{5}$$

Using a common local input removes variation caused by different input conditions, so the collected responses differ only in the expert transformation being evaluated. The stop-gradient makes this collection pass observational: experts expose candidate responses without being updated through the supervision they provide. This produces a comparable set of candidate supervision sources, but does not assume that all responses are equally useful for the current input.

3.2 CROSS-EXPERT UTILITY ESTIMATION

The usefulness of an expert response depends on whether it moves the current representation in a direction favored by the restoration objective. Let $\mathcal{L}_{\mathrm{res}}$ denote the restoration objective used to evaluate the current prediction. At the monitored interface, we compute the detached local gradient

$$\mathbf{g}^l = \mathrm{sg}(\nabla_{\mathbf{z}^l} \mathcal{L}_{\mathrm{res}}), \tag{6}$$

so that $-\mathbf{g}^l$ gives the local descent direction of the restoration objective.

To compare response directions independently of their magnitudes, we first normalize each collected response:

$$\widehat{\mathbf{r}}_k^l = \frac{\bar{\mathbf{r}}_k^l}{\|\bar{\mathbf{r}}_k^l\|_2 + \epsilon}. \tag{7}$$

Consider applying a perturbation of magnitude α along $\widehat{\mathbf{r}}_k^l$ at the monitored interface. A first-order expansion around $\mathbf{z}^l$ gives

$$\mathcal{L}_{\mathrm{res}}\big(\mathbf{z}^l + \alpha \widehat{\mathbf{r}}_k^l\big) = \mathcal{L}_{\mathrm{res}}\big(\mathbf{z}^l\big) + \alpha \left\langle \mathbf{g}^l, \widehat{\mathbf{r}}_k^l \right\rangle + \mathcal{O}(\alpha^2). \tag{8}$$

A response aligned with $-\mathbf{g}^l$ therefore produces a favorable first-order change in the restoration objective. Based on this observation, CUE defines the utility of expert E_k^l as

$$u_k^l = -\frac{\left\langle \mathbf{g}^l, \bar{\mathbf{r}}_k^l \right\rangle}{\|\mathbf{g}^l\|_2 \|\bar{\mathbf{r}}_k^l\|_2 + \epsilon}. \tag{9}$$

A larger u_k^l indicates stronger agreement with restoration descent. Cosine normalization prevents an expert from receiving higher utility simply because it produces a larger response. The utility measures local compatibility with the current restoration objective rather than predicting finite-step PSNR or loss improvement. Importantly, the restoration gradient only evaluates expert-produced responses and is not converted into the supervision target.

CUE serves a different role from the backbone's native expert selection. Native selection determines which experts act as learners for the current input, whereas CUE determines how strongly the observed expert responses should contribute to their supervision. We convert the utilities into normalized weights

$$q_k^l = \frac{\exp(u_k^l/\tau)}{\sum_{j=1}^{N} \exp(u_j^l/\tau)}, \tag{10}$$

where τ controls the concentration of the resulting utility distribution.

3.3 CROSS-EXPERT SUPERVISION

The utility weights determine how strongly each expert response contributes to the supervision of the selected learner. We construct the cross-expert target as

$$\mathbf{t}^l = \sum_{k=1}^{N} q_k^l \bar{\mathbf{r}}_k^l. \tag{11}$$

Because $q_k^l \geq 0$ and $\sum_k q_k^l = 1$, the target is a convex combination of expert-produced responses. CUE therefore determines how existing expert responses are combined rather than synthesizing a target directly from the restoration gradient.

The selected learner is then evaluated on the same detached local input while retaining gradients with respect to its parameters:

$$\mathbf{r}_{\mathrm{sel}}^l = \rho_\star^l\big(\mathrm{sg}(\mathbf{x}^l)\big). \tag{12}$$

Cross-expert supervision is transferred to the learner through

$$\mathcal{L}_{\mathrm{CEL}}^l = \big\|\mathbf{r}_{\mathrm{sel}}^l - \mathrm{sg}\big(\mathbf{t}^l\big)\big\|_1. \tag{13}$$

The response-collection pass observes the full expert population, whereas the gradient-carrying learner pass updates only the selected expert. Accordingly, for every peer expert E_k^l with $k \neq \star$,

$$\nabla_{\theta_k} \mathcal{L}_{\mathrm{CEL}}^l = 0, \qquad k \neq \star. \tag{14}$$

The detached local input further prevents the CEL-induced gradient from propagating into components preceding the monitored interface.

This construction separates the source of supervision from the destination of learning. Responses from any expert may contribute to the cross-expert target, but only the learner selected by the backbone receives the resulting CEL-induced update. An expert can therefore act as a supervisor for the current sample without entering the native prediction path or becoming a learner for that sample.

3.4 Training and Inference

Let $\mathcal{M}$ denote the set of expert layers to which CEL is applied in the current optimization step. Their learning objectives are aggregated as

$$\mathcal{L}_{\mathrm{CEL}} = \frac{1}{|\mathcal{M}|} \sum_{l \in \mathcal{M}} \mathcal{L}_{\mathrm{CEL}}^{l}, \tag{15}$$

and the model is optimized with

$$\mathcal{L} = \mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{CEL}} \mathcal{L}_{\mathrm{CEL}}, \tag{16}$$

where $\mathcal{L}_{\mathrm{base}}$ is the backbone's original training objective and λ_{CEL} controls the contribution of cross-expert learning.

Expert Response Collection, CUE, and Cross-Expert Supervision are used only during training. At inference, all CEL-specific operations are removed and the backbone retains its native expert-selection mechanism and prediction path. CEL therefore introduces no additional model parameters or inference FLOPs.

4 Experiments

4.1 Experimental Setup

Benchmarks and evaluation. We follow standard all-in-one image restoration protocols (Zamfir et al., 2025). AIO-3 includes Gaussian denoising at $\sigma \in \{15, 25, 50\}$, deraining, and dehazing, evaluated on BSD68 (Martin et al., 2001), Rain100L (Yang et al., 2017), and SOTS (Li et al., 2018), respectively. AIO-5 further includes deblurring and low-light enhancement on GoPro (Nah et al., 2017) and LOL-v1 (Wei et al., 2018). We additionally evaluate on CDD11 (Guo et al., 2024), which contains single, double, and triple compositions of low-light, haze, rain, and snow. PSNR and SSIM (Wang et al., 2004) are computed on full RGB images.

Implementation and training. We evaluate CEL with MoCE-IR (Zamfir et al., 2025) on AIO-3, AIO-5, and CDD11, and with M2Restore (Wang et al., 2025) on CDD11. For each backbone, CEL preserves the native architecture, expert-selection mechanism, and inference path, modifying only how supervision is constructed during training. Unless otherwise specified, we set $\lambda_{\mathrm{CEL}} = 0.05$ and $\tau = 0.25$. At each optimization step, CEL is applied to one monitored expert layer per decoder stage, with the monitored layers rotated deterministically over training. For MoCE-IR, we retain the native Top-1 expert selection and optimization protocol. We use 128×128 patches, a global batch size of 32, and AdamW with an initial learning rate of 2×10^{-4}, followed by linear warm-up and cosine decay. The native training objective is $\mathcal{L}_1 + 0.1\mathcal{L}_{\mathrm{FFT}} + 0.01\mathcal{L}_{\mathrm{bal}}$. For MoCE-IR, the baseline and CEL variant use identical data sampling, optimization settings, and training budgets, with CEL as the only change to the learning procedure. For M2Restore, we use the CDD11 baseline re-trained by (Njoku & Shukla, 2026) and train the CEL variant under the M2Restore training protocol.

4.2 Overall Restoration Performance

We first evaluate CEL on AIO-3, AIO-5, and CDD11 to examine its restoration performance under both standard and composite degradation settings. For Tables 1–3, the best and second-best results are highlighted in bold and underlined, respectively.

CEL consistently improves the corresponding expert-based backbones across all three benchmarks while remaining competitive with recent all-in-one restoration methods. On AIO-3, CEL improves

Table 1: **Quantitative comparison on AIO-3.** PSNR (dB) / SSIM are reported for dehazing, deraining, and denoising at three noise levels. Best and second-best PSNR/SSIM values are highlighted in **bold** and underline, respectively. † denotes diffusion-based methods and § denotes task-specific prompt guidance for expert selection.

Method	Dehaze SOTS	Derain Rain100L	N_{15}	N_{25} BSD68	N_{50}	Avg.
PromptIR$_{\text{NeurIPS'23}}$	30.58/.974	36.37/.972	33.98/.933	31.31/.888	28.06/.799	32.06/.913
WeatherDiff$^{\dagger}_{\text{TPAMI'23}}$	25.54/.971	31.53/.961	33.26/.926	30.51/.871	26.75/.745	29.32/.895
DiffUIR$^{\dagger}_{\text{CVPR'24}}$	30.18/.973	36.78/.973	33.94/.932	31.26/.887	28.04/.797	32.04/.912
InstructIR$_{\text{ECCV'24}}$	30.22/.959	37.98/.978	34.15/.933	31.52/.890	28.30/.804	32.43/.913
UP-Restorer$^{\dagger}_{\text{AAAI'25}}$	30.68/.977	36.74/.978	33.99/.933	31.33/.888	28.07/.799	32.16/.915
Perceive-IR$_{\text{TIP'25}}$	30.87/.975	38.29/.980	34.13/.934	**31.53**/.890	28.31/.804	32.63/.917
AdaIR$_{\text{ICLR'25}}$	31.06/.980	38.64/.983	34.12/.935	31.45/.892	28.19/.802	32.69/.918
VLU-Net$_{\text{CVPR'25}}$	30.71/.980	38.93/.984	34.13/.935	31.48/.892	28.23/.804	32.70/.919
DFPIR$_{\text{CVPR'25}}$	31.87/.980	38.65/.982	34.14/.935	31.47/**.893**	28.25/**.806**	**32.88/.919**
MEASNet$^{\S}_{\text{TCSVT'25}}$	31.61/.981	**39.00/.985**	34.12/.935	31.46/.892	28.19/.803	32.85/.919
ClearAIR$_{\text{AAAI'26}}$	31.08/.981	38.61/.984	**34.18/.935**	31.50/.891	**28.31**/.804	32.74/.919
MoCE-IR$_{\text{CVPR'25}}$	31.34/.979	38.57/.984	34.11/.932	31.45/.888	28.18/.800	32.73/.917
CEL (ours) Δ**Avg. +0.12**	**31.87/.981**	38.67/.984	34.10/.932	31.44/.888	28.18/.799	32.85/.917

Table 2: **Quantitative comparison on AIO-5.** PSNR (dB) / SSIM are reported over five restoration tasks. Best and second-best PSNR/SSIM values are highlighted in **bold** and underline, respectively. † denotes diffusion-based methods and § denotes task-specific prompt guidance for expert selection. The average of DA-CLIP is computed from its five reported task-wise results.

Method	Dehaze SOTS	Derain Rain100L	Denoise BSD68 ($\sigma = 25$)	Deblur GoPro	Low-light LOL	Avg.
PromptIR$_{\text{NeurIPS'23}}$	26.54/.949	36.37/.970	31.47/.886	28.71/.881	22.68/.832	29.15/.904
DiffUIR$^{\dagger}_{\text{CVPR'24}}$	29.47/.965	35.98/.968	31.02/.885	27.50/.845	22.32/.826	29.25/.898
DA-CLIP$^{\dagger}_{\text{ICLR'24}}$	29.46/.963	36.28/.968	30.02/.821	24.86/.585	22.29/.476	28.58/.763
InstructIR$_{\text{ECCV'24}}$	27.10/.956	36.84/.973	31.40/.887	29.40/.886	23.00/.836	29.55/.907
Perceive-IR$_{\text{TIP'25}}$	28.19/.964	37.25/.977	31.44/.887	29.46/.886	22.88/.833	29.84/.909
AdaIR$_{\text{ICLR'25}}$	30.53/.978	38.02/.981	31.35/.889	28.12/.858	23.00/.845	30.20/.910
VLU-Net$_{\text{CVPR'25}}$	30.84/**.980**	**38.54**/.982	31.43/**.891**	27.46/.840	22.29/.833	30.11/.905
DFPIR$_{\text{CVPR'25}}$	**31.64**/.979	37.62/.978	31.29/.889	28.82/.873	**23.82**/.843	30.64/.913
MEASNet$^{\S}_{\text{TCSVT'25}}$	31.05/.980	38.32/.982	31.40/.888	29.41/.890	23.00/.845	30.64/.917
ClearAIR$_{\text{AAAI'26}}$	30.12/.978	38.20/.982	**31.53**/.888	29.67/.887	22.83/.846	30.47/.916
MoCE-IR$_{\text{CVPR'25}}$	30.48/.974	38.04/.982	31.34/.887	30.05/.899	23.00/.852	30.58/.919
CEL (ours) Δ**Avg. +0.38**	30.83/.979	38.49/**.984**	31.45/.888	**30.79/.911**	23.28/**.857**	**30.96/.924**

the average PSNR by 0.12 dB and reaches 31.87 dB on dehazing, tying the best result among the compared methods. On AIO-5, the improvement becomes broader, with gains across all five restoration tasks and an average PSNR/SSIM of 30.96/.924, corresponding to a 0.38 dB improvement over the baseline. On CDD11, CEL improves the Single, Double, and Triple degradation groups and reaches the highest overall PSNR of 29.63 dB among the compared methods. Importantly, the improvement is consistently observed across different expert-based backbones, with average PSNR gains of 0.58 dB and 2.10 dB over the corresponding MoCE-IR and M2Restore baselines, respectively. These results demonstrate that the effectiveness of CEL is not confined to a particular backbone or degradation setting, but extends across standard restoration tasks and more challenging composite degradations.

Figure 3 further compares restoration quality on CDD11. CEL more faithfully recovers repeated architectural structures and preserves clearer object boundaries and local details under composite degradations. These visual results are consistent with the quantitative improvements and show that the gains of CEL translate into better structural and detail recovery.

Table 3: **Quantitative comparison on CDD11.** Single, Double, and Triple denote the average results over four single-, five double-, and two triple-degradation conditions, respectively. Task-wise gains over all eleven degradation conditions are provided in the appendix. PSNR (dB)/SSIM are reported. Best and second-best PSNR/SSIM values are highlighted in **bold** and underline, respectively. † denotes diffusion-based methods, and "–" denotes unavailable results.

Method	Single	Double	Triple	Avg.
AirNet$_{\text{CVPR'22}}$	25.60/.885	22.97/.796	22.02/.716	23.75/.814
Restormer$_{\text{CVPR'22}}$	30.29/.931	25.60/.851	23.86/.767	26.99/.865
WeatherDiff$^{\dagger}_{\text{TPAMI'23}}$	23.80/.860	21.98/.786	21.14/.707	22.49/.799
WGWSNet$_{\text{CVPR'23}}$	29.97/.923	25.76/.850	23.94/.772	26.96/.863
PromptIR$_{\text{NeurIPS'23}}$	28.88/.920	24.46/.834	23.53/.750	25.90/.850
OneRestore$_{\text{ECCV'24}}$	31.68/.938	27.35/.866	24.84/.790	28.47/.878
BioIR-T$_{\text{NeurIPS'25}}$	33.16/.944	28.31/**.877**	25.28/**.801**	29.52/.887
MIRAGE$_{\text{ICLR'26}}$	32.79/.944	28.00/.875	25.73/.800	29.33/**.887**
DPL$_{\text{ICCV'25}}$	–	–	–	29.55/.884
UniLDiff$^{\dagger}_{\text{CVPR'26}}$	–	27.92/.872	25.65/.792	29.35/.886
M2Restore$_{\text{TIP'25}}$	–	–	–	27.38/.861
CEL (ours) Δ**Avg. +2.10**	32.78/.938	**28.31**/.866	25.81/.786	29.48/.878
MoCE-IR-S$_{\text{CVPR'25}}$	32.54/.941	27.73/.869	25.40/.790	29.05/.881
CEL (ours) Δ**Avg. +0.58**	**33.22/.944**	28.26/.875	**25.87**/.799	**29.63**/.886



Figure 3: **Qualitative comparison on CDD11.** The red boxes indicate the regions enlarged below each example for detailed visual comparison.

4.3 Mechanism Analysis

We next examine whether the observed gains arise from CEL's learning mechanism rather than generic multi-expert aggregation or alternative supervision signals. Unless otherwise specified, all controlled experiments use AIO-5 with the same backbone and a matched 100K training budget.

Is learner-supervisor decoupling necessary? The central principle of CEL is that an expert can provide supervision without receiving the resulting update. We therefore compare learner-only CEL with a coupled control that applies the same detached target to both the selected learner and peer supervisors while matching the CEL-induced gradient norm; native expert selection and target construction remain unchanged. As shown in Table 4(a), learner-only CEL outperforms the coupled control, supporting the benefit of separating the learner and supervisor roles.

Does CUE identify useful expert responses? We evaluate CUE using the same held-out AIO-5 analysis split and equal-scale local intervention protocol as in Figure 1, covering 2,000 expert responses from 500 image-layer cases. For each response, we compare its estimated utility with the resulting local PSNR change. For visualization, responses are partitioned into equal-count bins according to their estimated utilities. Figure 4(a) shows a clear ordering between estimated utility

Table 4: **Ablation of CEL on AIO-5.** Controlled experiments analyze (a) learner-supervisor decoupling, (b) supervision design, and (c) supervision construction. All variants use the same backbone and matched 100K training budget; all Δ values are measured relative to the common baseline.

(a) Learner-Supervisor Decoupling

Update	PSNR ↑	Δ
Baseline	28.4887	–
$\mathcal{L}+\mathcal{S}$	28.5482	+0.0595
$\mathcal{L}$ Only	**28.7057**	+0.2170

(b) Supervision Design

Training Design	PSNR ↑	Δ
Baseline	28.4887	–
Uniform	28.4820	−0.0067
Router-Weighted	27.9616	−0.5271
Direct-Gradient	28.3950	−0.0937
CEL	**28.7057**	+0.2170

(c) Supervision Construction

Target	PSNR ↑	Δ
Baseline	28.4887	–
Highest	28.6113	+0.1226
Weighted	**28.7057**	+0.2170

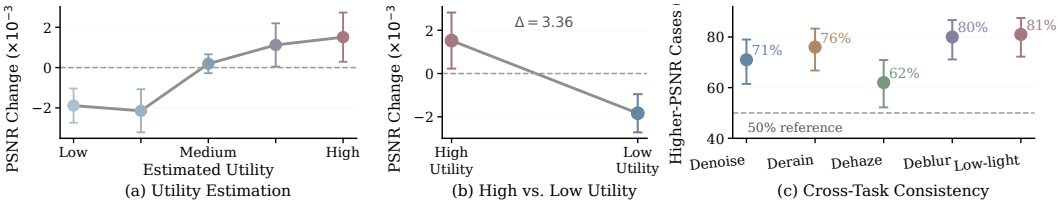


Figure 4: **Validation of CUE utility.** (a) Local PSNR change across utility groups. (b) Highest- versus lowest-utility responses. (c) Fraction of cases where the highest-utility response yields higher PSNR across AIO-5 tasks. Error bars show 95% bootstrap confidence intervals.

and local PSNR change, while Figures 4(b,c) show consistent separation between the highest- and lowest-utility responses across all AIO-5 tasks.

Is CEL's supervision design necessary? We compare CUE with three alternative supervision designs. *Uniform* equally averages candidate responses, *Router-Weighted* substitutes native routing scores for CUE utilities, and *Direct-Gradient* uses the normalized negative restoration gradient as the target direction. As shown in Table 4(b), none of these alternatives reproduces the gain of CEL. This indicates that the improvement cannot be explained by generic expert aggregation, native selection preference, or direct gradient supervision.

Is utility-weighted aggregation necessary? We examine whether supervision should use only the response with the highest estimated utility or retain information from multiple useful experts. Table 4(c) shows that utility-weighted aggregation outperforms hard *Highest-Utility* supervision. This confirms that cross-expert supervision benefits from combining multiple useful responses rather than selecting a single preferred expert.

Training memory. CEL requires additional expert evaluations during training but introduces only a modest memory overhead. Under the same AIO-5 setting with four RTX 4090 GPUs, a global batch size of 32, and 128×128 patches, peak allocated memory per GPU increases from 17.10 to 18.44 GiB. All CEL-specific operations are removed after training, leaving the parameter count and inference FLOPs unchanged.

5 Limitations

CEL currently focuses on all-in-one image restoration, and the generalization of CEL to broader MoE architectures and other domains remains to be studied. CUE is a local utility criterion, and the long-term effect of cross-expert supervision on expert specialization also remains open.

6 Conclusion

We introduced Cross-Expert Learning (CEL), which decouples the learner from the source of supervision while preserving native expert execution. For all-in-one image restoration, CUE provides input-dependent cross-expert supervision and improves different expert-based backbones without additional parameters or inference FLOPs.

AI USE STATEMENT

In this work, we used large language models for language polishing. All scientific ideas, method development, experimental design, implementation, data generation, result verification, and final manuscript decisions were conducted and verified by the authors. The authors take full responsibility for the content and results of this work.

REPRODUCIBILITY STATEMENT

The response interface, utility computation, stop-gradient boundaries, and optimization objective are specified in Section 3. Dataset protocols, evaluation metrics, training settings, monitored-interface configuration, and CEL optimization details are provided in Section 4.1 and the appendix. Section 4.3 documents the matched supervision controls, the gradient-budget-matched coupled control, and direct utility interventions. The final release will include code, configurations, evaluation scripts, checkpoint-selection rules, and checkpoints required to reproduce the reported results.

REFERENCES

Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. Online knowledge distillation with diverse peers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 3430–3437, 2020.

Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *2021 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 182–192. IEEE, 2021a.

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

Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5006–5015. IEEE, 2021b.

Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4621–4630. IEEE, 2021.

Marcos V Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. In *European Conference on Computer Vision*, pp. 1–21. Springer, 2024.

Yuning Cui, Syed Waqas Zamir, Salman Khan, Alois Knoll, Mubarak Shah, and Fahad Khan. Adair: Adaptive all-in-one image restoration via frequency mining and modulation. In *International Conference on Learning Representations*, volume 2025, pp. 101306–101327, 2025.

Yuning Cui, Wenqi Ren, and Alois Knoll. Bio-inspired image restoration. *Advances in Neural Information Processing Systems*, 38:80452–80481, 2026.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International conference on machine learning*, pp. 1607–1616. PMLR, 2018.

Yu Guo, Yuan Gao, Yuxu Lu, Huilin Zhu, Ryan Wen Liu, and Shengfeng He. Onerestore: A universal restoration framework for composite degradation. In *European conference on computer vision*, pp. 255–272. Springer, 2024.

Haisen He, Xiangyu Zou, SongLin Dong, Heng Li, Yihong Gong, and Zhiheng Ma. Continuous expert assembly: Instance-conditioned low-rank residuals for all-in-one image restoration. *arXiv preprint arXiv:2605.06127*, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Junjun Jiang, Zengyuan Zuo, Gang Wu, Kui Jiang, and Xianming Liu. A survey on all-in-one image restoration: Taxonomy, evaluation and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Yitong Jiang, Zhaoyang Zhang, Tianfan Xue, and Jinwei Gu. Autodir: Automatic all-in-one image restoration with latent diffusion. In *European Conference on Computer Vision*, pp. 340–359. Springer, 2024.

Gyeongman Kim, Gyouk Chu, and Eunho Yang. Every expert matters: Towards effective knowledge distillation for mixture-of-experts language models. *arXiv preprint arXiv:2502.12947*, 2025.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. Base layers: Simplifying training of large, sparse models. In *International Conference on Machine Learning*, pp. 6265–6274. PMLR, 2021.

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.

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17431–17441. IEEE, 2022.

Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3172–3182. IEEE, 2020.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *2021 IEEE/CVF international conference on computer vision workshops (ICCVW)*, pp. 1833–1844. IEEE, 2021.

Cencen Liu, Wen Yin, Dongyang Zhang, Dongmin Li, Shan Zhao, Bing Su, Tao He, Jielei Wang, and Guoming Lu. What to remove, what to preserve: Dual-ambiguity rectification for all-in-one image restoration. *arXiv preprint arXiv:2607.28526*, 2026.

Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas Schön. Controlling vision-language models for multi-task image restoration. In *International Conference on Learning Representations*, volume 2024, pp. 16226–16246, 2024.

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, pp. 416–423. Ieee, 2001.

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*, pp. 3883–3891, 2017.

Judith Njoku and Diksha Shukla. Omnirestore: A parameter-efficient framework for universal adverse-weather image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2236–2246, 2026.

Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE transactions on pattern analysis and machine intelligence*, 45(8):10346–10357, 2023.

Dongwon Park, Byung Hyun Lee, and Se Young Chun. All-in-one image restoration for unknown degradations using adaptive discriminative filters for specific degradations. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 5815–5824. IEEE, 2023.

Shurui Peng, Xin Lin, Shi Luo, Jincen Ou, Dizhe Zhang, Lu Qi, Truong Nguyen, and Chao Ren. Sler-ir: Spherical layer-wise expert routing for all-in-one image restoration. In *European Conference on Computer Vision*, pp. 472–491. Springer, 2026.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in neural information processing systems*, 36:71275–71293, 2023.

Bin Ren, Eduard Zamfir, Zongwei Wu, Yawei Li, Yidi Li, Danda Pani Paudel, Radu Timofte, Ming-Hsuan Yang, Luc Van Gool, and Nicu Sebe. Any image restoration via efficient spatial-frequency degradation adaptation. *arXiv preprint arXiv:2504.14249*, 2025.

Bin Ren, Yawei Li, Xu Zheng, Yuqian Fu, Danda Pani Paudel, Hong Liu, Ming-Hsuan Yang, Luc Van Gool, and Nicu Sebe. Efficient degradation-agnostic image restoration via channel-wise functional decomposition and manifold regularization. In *International Conference on Learning Representations*, volume 2026, pp. 126404–126430, 2026.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in neural information processing systems*, 34:8583–8595, 2021.

Stephen Roller, Sainbayar Sukhbaatar, Jason Weston, et al. Hash layers for large sparse models. *advances in neural information processing systems*, 34:17555–17566, 2021.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *arXiv preprint arXiv:2204.03883*, 2022.

Xiaole Tang, Xiang Gu, Xiaoyi He, Xin Hu, and Jian Sun. Degradation-aware residual-conditioned optimal transport for unified image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6764–6779, 2025.

Xiangpeng Tian, Xiangyu Liao, Xiao Liu, Meng Li, and Chao Ren. Degradation-aware feature perturbation for all-in-one image restoration. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28165–28175. IEEE, 2025.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5759–5770. IEEE, 2022.

Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2343–2353. IEEE, 2022.

Shiwei Wang, Yongzhen Wang, Bingwen Hu, Liyan Zhang, Xiao-Ping Zhang, and Mingqiang Wei. M2ir: proactive all-in-one image restoration via mamba-style modulation and mixture-of-experts. *arXiv preprint arXiv:2603.14816*, 2026.

Yongzhen Wang, Yongjun Li, Zhuoran Zheng, Xiao-Ping Zhang, and Mingqiang Wei. M2restore: Mixture-of-experts-based mamba-cnn fusion framework for all-in-one image restoration. *IEEE Transactions on Image Processing*, 34:8086–8100, 2025.

Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17662–17672. IEEE, 2022.

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.

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

Gang Wu, Junjun Jiang, Kui Jiang, Xianming Liu, and Liqiang Nie. Beyond degradation redundancy: Contrastive prompt learning for all-in-one image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Gang Wu, Junjun Jiang, Kui Jiang, Xianming Liu, and Liqiang Nie. Dswinir: Rethinking window-based attention for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

Gang Wu, Junjun Jiang, Kui Jiang, Xianming Liu, and Liqiang Nie. Learning dynamic prompts for all-in-one image restoration. *IEEE Transactions on Image Processing*, 34:3997–4010, 2025c.

Guile Wu and Shaogang Gong. Peer collaborative learning for online knowledge distillation. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 35, pp. 10302–10310, 2021.

Zhitian Xie, Yinger Zhang, Chenyi Zhuang, Qitao Shi, Zhining Liu, Jinjie Gu, and Guannan Zhang. Mode: A mixture-of-experts model with mutual distillation among the experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16067–16075, 2024.

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 *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1685–1694. IEEE, 2017.

Youngsik Yoon, Siwei Wang, Wei Chen, and Jungseul Ok. When are experts misrouted? counterfactual routing analysis in mixture-of-experts language models. *arXiv preprint arXiv:2605.07260*, 2026.

Xiaoyan Yu, Shen Zhou, Huafeng Li, and Liehuang Zhu. Multi-expert adaptive selection: Task-balancing for all-in-one image restoration. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(5):4619–4634, 2024.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yuedong Tan, Danda Pani Paudel, Yulun Zhang, and Radu Timofte. Complexity experts are task-discriminative learners for any image restoration. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12753–12763. IEEE, 2025.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *European conference on computer vision*, pp. 492–511. Springer, 2020.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 14816–14826. IEEE, 2021.

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*, pp. 5728–5739, 2022.

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.

Xu Zhang, Jiaqi Ma, Guoli Wang, Qian Zhang, Huan Zhang, and Lefei Zhang. Perceive-ir: Learning to perceive degradation better for all-in-one image restoration. *IEEE Transactions on Image Processing*, 35:2018–2033, 2025.

Xu Zhang, Huan Zhang, Guoli Wang, Qian Zhang, and Lefei Zhang. Clearair: A human-visual-perception-inspired all-in-one image restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12861–12869, 2026.

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4320–4328. IEEE, 2018.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11943–11952. IEEE, 2022.

Hao Zhao, Zihan Qiu, Huijia Wu, Zili Wang, Zhaofeng He, and Jie Fu. Hypermoe: Towards better mixture of experts via transferring among experts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10605–10618, 2024.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

Xiatian Zhu, Shaogang Gong, et al. Knowledge distillation by on-the-fly native ensemble. *Advances in neural information processing systems*, 31, 2018.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022.

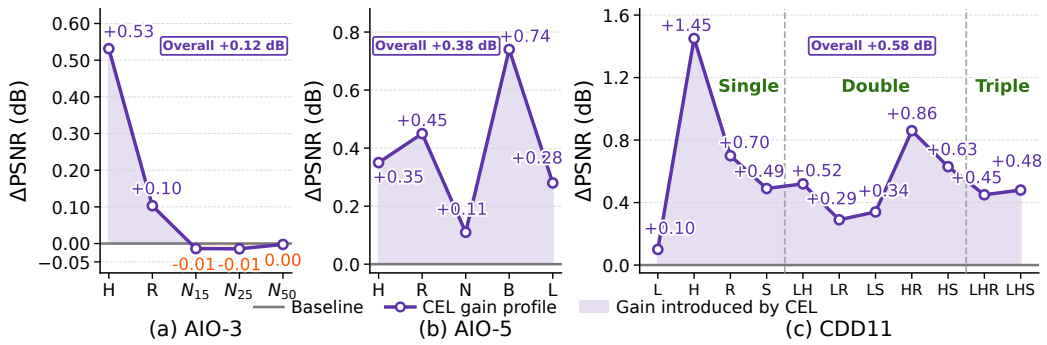


Figure A1: **Task-wise performance gains of CEL.** ΔPSNR is measured relative to the corresponding MoCE-IR baseline on AIO-3, AIO-5, and CDD11.

A Task-wise Performance Gains

To examine whether the improvement of CEL is concentrated on a small subset of tasks or consistently observed across degradation conditions, we report task-wise PSNR gains over the corresponding MoCE-IR baselines on AIO-3, AIO-5, and CDD11.

As shown in Figure A1, the gains of CEL are distributed across diverse restoration conditions rather than being driven by a single task. On AIO-3, the clearest improvements occur in dehazing ($+0.53$ dB) and deraining ($+0.10$ dB), while denoising remains nearly unchanged. On AIO-5, CEL improves all five restoration tasks, yielding an average PSNR gain of $+0.38$ dB. On CDD11, CEL improves all eleven single and composite degradation conditions, including gains of $+1.45$ dB on haze and $+0.86$ dB on haze–rain, and increases the overall PSNR by $+0.58$ dB. These results show that the benefit of CEL is not dominated by an isolated degradation type and remains positive across both standard and composite restoration settings.

B Additional Implementation and Evaluation Details

B.1 CEL Application Details

CEL is applied at expert interfaces where the outputs of different experts can be expressed as additive responses relative to a common local representation, following the response construction defined in Section 3. For the MoCE-IR backbone, the native Top-1 expert-selection mechanism is retained without modification. At each optimization step, one expert layer in each decoder stage is monitored, and the monitored layers are rotated deterministically over training.

At a monitored interface, the same detached local input is provided to the full expert bank to collect candidate responses. These responses are used only to construct the cross-expert supervision target. The CEL-induced auxiliary gradient is applied only to the learner selected by the native backbone; peer responses and representations preceding the detached local interface do not receive gradients from $\mathcal{L}_{\mathrm{CEL}}$. The native restoration objective, routing mechanism, and prediction path remain unchanged.

Unless otherwise specified, we use $\lambda_{\mathrm{CEL}} = 0.05$ and $\tau = 0.25$ throughout the experiments. All CEL-specific response collection, utility estimation, and cross-expert supervision operations are used only during training and are removed at inference. Consequently, CEL introduces no additional model parameters or inference FLOPs.

B.2 Gradient-Flow Verification

We explicitly verify the gradient boundaries induced by CEL at every monitored interface. First, the routed learner receives a non-zero auxiliary gradient from $\mathcal{L}_{\mathrm{CEL}}$. Second, non-selected peer experts receive zero CEL-induced gradient in the default formulation. Third, modules preceding the detached local interface receive zero CEL-induced gradient. Finally, when $\mathcal{L}_{\mathrm{CEL}}$ is removed, the resulting gradient recovers that of the native backbone up to numerical tolerance.

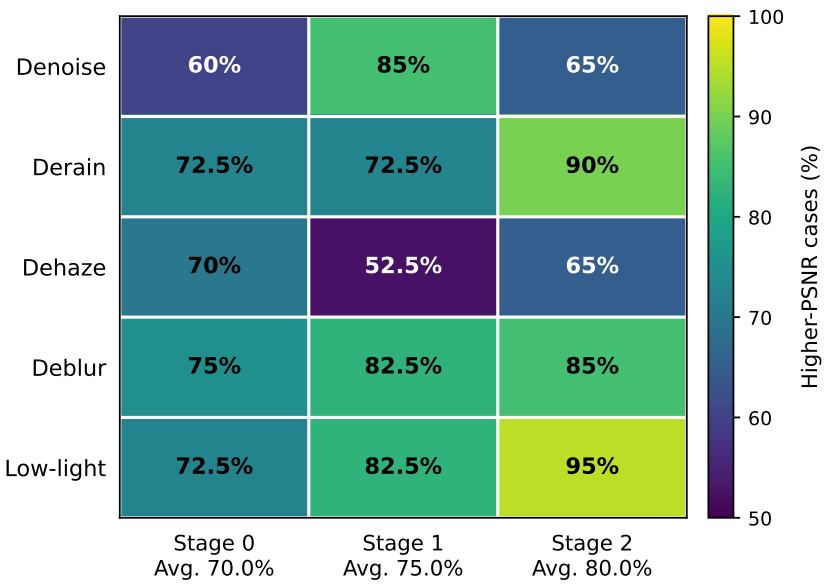


Figure A2: **CUE utility ordering across restoration tasks and decoder stages.** Each cell reports the percentage of sample–layer cases in which the highest-utility expert response produces a higher local PSNR than the lowest-utility response under the same fixed intervention. Monitored expert blocks belonging to the same decoder stage are pooled for reporting. The values below the stage labels denote the corresponding averages over all five restoration tasks.

These checks directly verify the stop-gradient construction in Eq. (14) and ensure that an expert can contribute its response as supervision without being updated by the CEL objective on the same sample. They also separate the training-only supervision path introduced by CEL from the backbone's native restoration gradient.

B.3 Coupled-Control Protocol

The coupled control in Table 4(a) isolates the effect of separating the source of supervision from the destination of learning. The cross-expert target is kept identical and detached in both the learner-only CEL variant and the coupled control. Native expert selection, candidate-response construction, and utility estimation are also kept unchanged.

The only intervention concerns where the auxiliary update is applied. In learner-only CEL, the detached cross-expert target supervises only the routed learner. In the coupled control, the same target is additionally applied to the peer experts contributing to the supervision process. Because updating multiple experts would otherwise increase the total auxiliary update magnitude, the aggregate CEL-induced gradient of the coupled variant is rescaled to match the gradient norm of the learner-only variant.

Both variants therefore use the same backbone, data, optimization settings, supervision target, and matched 100K training budget. This construction isolates learner–supervisor decoupling from differences in target quality or auxiliary gradient budget.

B.4 CUE Utility Ordering across Decoder Stages

The main-paper analysis evaluates whether the utility estimated by CUE reflects the relative restoration value of expert responses across different restoration tasks. We further examine whether this relationship depends on where the responses are observed in the decoder.

We use the same held-out AIO-5 analysis split and the same local-intervention protocol as in Figure 4. The analysis contains 500 image–layer cases and 2,000 expert responses in total. For each image–layer case, we identify the responses with the highest and lowest estimated utilities and compare their resulting local PSNR changes under the same fixed intervention. We then group the monitored expert layers according to their decoder stage and report the percentage of cases in which

the highest-utility response yields a higher PSNR than the lowest-utility response. Each cell in Figure A2 therefore measures the utility-ordering agreement for one restoration task at one decoder stage.

As shown in Figure A2, the utility ordering remains positive across all fifteen task–stage combinations. Every combination exceeds the 50% reference, indicating that the relationship between estimated utility and local restoration effect is not confined to a particular degradation type or decoder stage. The stage-level agreement rates are 70.0%, 75.0%, and 80.0% for decoder Stages 0–2, respectively.

The behavior is particularly stable for deblurring and low-light enhancement, where the highest-utility response outperforms the lowest-utility response in 75.0–85.0% and 72.5–95.0% of cases, respectively. The weakest task–stage combination is dehazing at Stage 1, with an agreement rate of 52.5%. This variation indicates that CUE does not impose a uniformly strong ordering on every individual response. Instead, consistent with its formulation in Section 3, CUE provides an input-dependent local compatibility signal whose ordering remains informative at the population level.

We additionally examine the magnitude of the high–low separation after pooling cases within each decoder stage. The mean PSNR differences between the highest- and lowest-utility responses are 4.116×10^{-3}, 3.723×10^{-3}, and 1.102×10^{-3} dB for Stages 0–2, respectively, and remain positive at every stage. These values are used only as local diagnostics rather than as estimates of end-to-end restoration gains.

Together with Figure 4, these results provide complementary evidence for the utility estimated by CUE. The main-paper analysis shows that higher estimated utility is associated with more favorable restoration behavior across degradation types, while the stage-wise analysis shows that this relationship persists across different decoder stages. This supports the use of CUE as a restoration-relevant criterion for ranking candidate expert responses across both task and network depth.