

# Supplementary Materials: “Rainmer: Learning Multi-view Representations for Comprehensive Image Deraining and Beyond”

Anonymous Authors

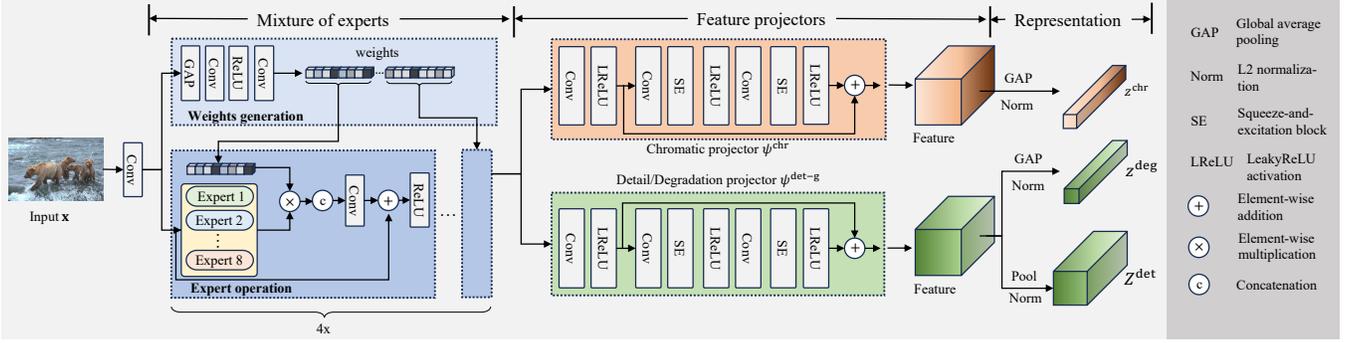


Figure 1: Details of the MoEs and feature projectors  $\psi^{\text{chr}}$ ,  $\psi^{\text{det-g}}$  in the proposed Rainmer.

## A OVERVIEW

In this supplementary material, we first present details of the MoEs and two feature projectors:  $\psi^{\text{chr}}$  and  $\psi^{\text{det-g}}$  in the proposed Rainmer (Appendix B). Subsequently, we conduct in-depth investigations of the multi-view representations for image deraining and AllinOne image restoration (Appendix C). Furthermore, we analytically quantify the relationships among different datasets, which offers insights into understanding dataset cooperation, competition, and conflicts.

## B ARCHITECTURE OF MOES AND FEATURE PROJECTORS

**MoEs.** As stated in our paper, we implement the mixture of experts (MoEs) the same as [1]. Specifically, given an input image  $x$ , we first expand the channels to 48 with a convolutional layer as shown in Fig. 1. This process results in a shallow feature  $F^{\text{sha}}$ . With the help of MoEs, the shallow feature  $F^{\text{sha}}$  is further transferred into  $F^{\text{spa}}$ :

$$F^{\text{spa}} = f_{\text{MoEs}}(F^{\text{sha}}), \quad (1)$$

where  $f_{\text{MoEs}}$  denotes the function of MoEs.

Typically, MoEs comprises four stages of mixture of expert interaction. Each stage contains eight experts [1]: an average pooling with kernel size  $3 \times 3$ , depth convolutional layers with kernel size  $1 \times 1$ ,  $3 \times 3$ ,  $5 \times 5$ ,  $7 \times 7$ , and dilation convolutional layers with kernel size  $3 \times 3$ ,  $5 \times 5$ ,  $7 \times 7$ . These eight experts are expected to extract rich information to well perceive details, degradations, and illuminations. Except for expert learning modules, MoEs utilizes a weights generation module as shown in Fig. 1 to adjust responses corresponding to each expert. The weights  $T^s, s \in \{1, 2, 3, 4\}$  for all four stages are generated from the same feature  $F^{\text{spa}}$ :

$$[T^1, T^2, T^3, T^4] = \mathcal{G} \circ \text{GAP}(F^{\text{spa}}), \quad (2)$$

where  $\mathcal{G}$  represents a GAP-Conv-ReLU-Conv weights generation module and GAP is the global average pooling operation. Notably,

each weight  $T^s \in \mathbb{R}^8$  characterizes the response to eight experts in the  $s$ -th stage. Denote  $F^{s-1}$  as the output feature from  $(s-1)$ -th stage, then the  $s$ -th stage outputs:

$$F^s = f_{\text{ReLU}} \left( f_{\text{Conv}} \left( T^s \odot \text{cat} \left( f_{\text{exp}} \left( F^{s-1} \right) \right) + F^{s-1} \right) \right), \quad s = 2, 3, 4 \quad (3)$$

$$F^1 = F^{\text{spa}}, \quad (4)$$

where  $f_{\text{exp}}$  indicates the function of eight experts,  $\text{cat}$  is the concatenation operation,  $f_{\text{Conv}}$  denotes a  $1 \times 1$  convolutional layer,  $f_{\text{ReLU}}$  is the ReLU activation, and  $\odot$  means element-wise multiplication with broadcasting.

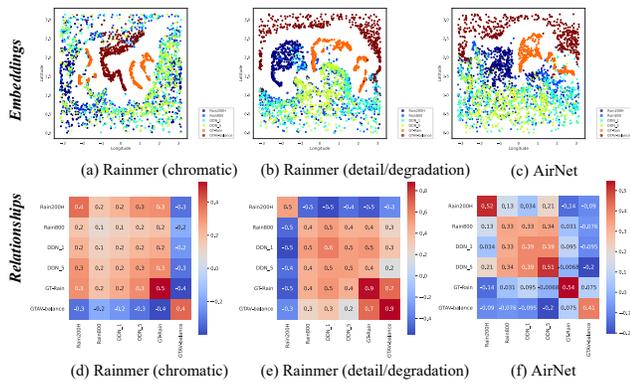
**Feature Projectors.** The feature extracted by the MoEs contains rich channel and spatial information that facilitates the perception of image details, degradations, and illuminations. Therefore, we further employ feature projectors to project the feature obtained from MoEs into specific representation spaces. As shown in Fig. 1, we employ a chromatic projector and a detail/degradation projector to obtain corresponding representations. Each projector contains a residual block with two squeeze-and-excitation layers [3] to facilitate channel interaction.

## C MULTI-VIEW REPRESENTATION ANALYSIS

In this section, we provide an in-depth analysis to investigate the effect of multi-view representations in image deraining, as well as AllinOne image restoration.

### C.1 Image Deraining

To train models on amalgamated Rain13K (*synthetic*), GT-Rain (*real-world*), and GTAV-balance (*nighttime*) datasets, we proposed to extract multi-view representations. Specifically, the detail/degradation representation is responsible for capturing image details and degradations (rain streaks, noise, blur, etc.), while the chromatic representation is introduced to perceive illuminations and color distortions caused by rain veiling effects. Hence, images with specific

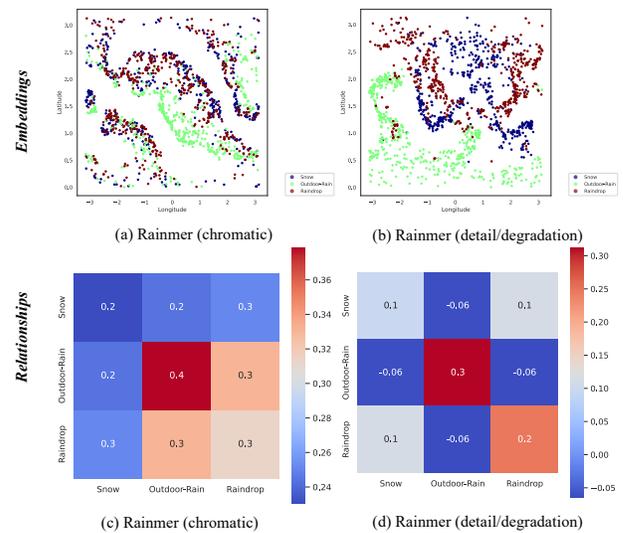


**Figure 2: Visualization of multi-view representations and dataset relationships among image deraining datasets.**

degradation factors and illuminations represent different relationships in different representation spaces, according to their similarity in detail/degradation or illumination. To investigate this hypothesis, we visualize multi-view representations among different datasets using the UMAP [6] technique. Specifically, we choose images from synthetic datasets (Rain200H [9], Rain800 [10], DDN [2]), real-world dataset (GT-Rain), and nighttime dataset (GTAV-balance) for visualization. Since rainy images in DDN contain 14 rain types, we select the first type (light rain) and fifth type (thick rain) and denote them as DDN\_1 and DDN\_5. We choose Rain200H, Rain800, and DDN because they all contribute to the synthetic dataset Rain13K. As for visualization, we randomly choose 500 rainy images from these datasets. Note that the representations lie in a high dimensional spherical surface, hence we utilize UMAP to project representations into a 2D space and then visualize the 2D longitude-latitude space (in another way, spherical coordinate system  $(r, \theta, \phi)$  with  $r = 1$ ) following [8].

Fig. 2 (a) & (b) display the embeddings of Rain200H, Rain800, DDN\_1, DDN\_5, GT-Rain, and GTAV-balance in chromatic and detail/degradation representation space, respectively. It can be seen that datasets represent different embeddings in different spaces. Specifically, in chromatic space, the GT-Rain and GTAV-balance datasets are mainly isolated from other synthetic datasets, owing to the color distortions in GT-Rain and low illuminance in GTAV-balance. However, in detail/degradation space, Rain200H and GT-Rain are nearly isolated from other datasets, due to extremely heavy rain streaks in Rain200H and complex real-world degradations in GT-Rain. In summary, the visualization results indicate that the proposed contrastive learning method successfully learned chromatic and detail/degradation representations. Additionally, we further visualize the embedding space learned by AirNet [4], a degradation-based contrastive learning method where images from different datasets are assigned with different degradation factors. The result is shown in Fig. 2 (c), which demonstrates a less compact space compared to Rainmer in Fig. 2 (b). Typically, Rain200H is mixed with other datasets in Fig. 2 (c).

In addition to visualizations of embeddings, these representations enable us to quantify the relationships among datasets, where dataset cooperation and conflict could be investigated. To this end,



**Figure 3: Visualization of multi-view representations and dataset relationships among snow, rain, and raindrop datasets.**

we further compute the relationship score between two arbitrary datasets by averaging cosine similarities of representations obtained across these datasets. The results are presented in Fig. 2 (d)-(f). As shown in Fig. 2 (d), GTAV-balance represents a negative relationship to other datasets, indicating the influence of illumination intensity. Moreover, the other datasets all share positive relationships, contributing to the cooperation of learning on daytime datasets. The result in Fig. 2 (e) provides different dataset relationships, where Rain200H shares negative relationships with other datasets while the remaining datasets represent deep cooperation. These cooperations contribute to the outstanding performances of the proposed Rainmer. In contrast, AirNet has learned worse dataset relationships where dataset conflicts between GT-Rain / GTAV-balance and synthetic datasets exist, demonstrating the poor performance of AirNet.

Both the visualizations of embeddings and quantitative dataset relationships have demonstrated the superiority of the proposed method.

## C.2 Allinone Image Restoration

We further follow Appendix C.1 to visualize learned embeddings on the AllinOne dataset. Similarly, we randomly choose 500 images from the snow, rain, and raindrop subset in AllWeather [5] for visualization. The results are presented in Fig. 3. In the chromatic space, both Fig. 3 (a) & (c) display positive relationships among different datasets. However, Fig. 3 (b) & (d) demonstrate that degradation factors of snow, rain, and raindrop are quite different, resulting in negative relationships, the same as observed in [4, 7].

## REFERENCES

- [1] Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. 2023. Learning A Sparse Transformer Network for Effective Image Deraining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5896–5905.

