Dear Anonymous Referee #2,

We sincerely thank you for the positive assessment of our work. We greatly appreciate your recognition of the engineering value and thorough experimental design of this study. We also highly value the insightful comments regarding clarity, mathematical rigor, and standard terminology. In the revised manuscript, we have meticulously addressed all these points, as detailed in our point-by-point responses below.

Major Issues ：

1、Inconsistent dataset sample numbers require clarification and justification of augmentation strategies.

The main text clearly states that the HBMCD dataset contains 25,118 images and the GCD dataset contains 12,000 images. However, in Tables 4 and 5, the "Total Samples" columns sum to 125,590 and 60,000, respectively. It is evident that the tables report post-augmentation totals, whereas the text reports the original captured sample sizes. The authors should explicitly distinguish between the "original image set" and the "augmented image set" in both the text and the table captions. In addition, a brief justification should be provided for the four augmentation techniques (rotation, contrast enhancement, Gaussian noise, gamma correction).

Answer: We thank you for pointing out the inconsistency in dataset sample numbers. In the revised manuscript, we have explicitly distinguished between the "original image set" and the "augmented image set." In **Section 3.1**, we clarify that the original HBMCD and GCD datasets contain 25,118 and 12,000 images, respectively, and we further state that after data augmentation, the total numbers of training samples reached 125,590 and 60,000, respectively. To eliminate any confusion, we have also added the (Augmented) label to the headers of **Tables 4 and 5**. Additionally, we provided a physical justification for the four augmentation techniques: random rotation, contrast enhancement, Gaussian noise, and gamma correction are employed to simulate varying atmospheric conditions, changing solar illuminations, sensor noises, and camera exposure differences encountered in real-world deployments, thereby effectively mitigating overfitting and enhancing model robustness against complex environments.

We employed four data augmentation techniques, including image rotation, contrast enhancement, Gaussian noise addition, and gamma correction. These augmentation methods are physically motivated to simulate varying atmospheric conditions, changing solar illuminations, sensor noises, and camera exposure differences encountered in real-world deployments, thereby effectively mitigating overfitting and enhancing model robustness against complex environments. Consequently, the total numbers of samples for the HBMCD and GCD datasets after augmentation reached 125,590 and 60,000, respectively.

2、The mathematical formulation of Global Max Pooling (GMP) is not rigorous.

The physical meaning of global max pooling is to take the maximum value over the entire spatial feature map in order to capture the most salient feature responses. However, Eq. (2) adopts a 'sum-then-max' formulation, which is effectively equivalent to taking the maximum of a scalar (the result of global average pooling) again. This completely defeats the original physical intention of GMP in highlighting edge information from locally strong activation regions. This equation should be carefully checked and corrected.

Answer: We sincerely thank you for pointing out the lack of rigor in the mathematical formulation of GMP in Eq. (2). We carefully re-examined the equation, and you are absolutely correct that the previous 'sum-then-max' formulation entirely contradicts the physical intention of GMP, which is to take the maximum value over the spatial dimensions to extract salient features. We have fully accepted this suggestion and have corrected Eq. (2) to the standard spatial global max pooling formula in the revised manuscript: $\gamma_{\mathrm{gmp}} = \max_{i,j}(X_{ij})$， $X \in \mathbb{R}^{w \times h \times c}$, where i, j denote the spatial coordinates of the feature map, and c represents the channel index.

Specifically, we have made the following revisions in the updated manuscript:

（1）Removed the summation symbol ∑ in Eq. (2) and directly applied the maximum operation over all spatial positions.

（2）Added a physical explanation below the revised formula, clarifying that the operation retains the strongest activations and prevents edge information from being smoothed out by averaging.

（3）Included the following clarification: "Here, $\max_{i,j}$ denotes the maximum value across all spatial positions (i,j) in the feature map. This formulation physically aligns with the purpose of Global Max Pooling (GMP), which captures the most strongly activated features to highlight edge information, preventing critical details from being smoothed out by averaging."

3、The assumption of a '1/2 proportion of original features' in the CloudGhost module lacks justification.

In Section 2.4, when deriving the compression ratio of the

CloudGhost module, the authors directly assume in Eqs. (13) and (14) that 'the proportion of original features in the output cloud features is 1/2.' However, no rationale is provided for this specific choice. It is suggested that the authors clarify the basis for determining this ratio.

Answer: We thank you for pointing out that the assumption of a 1/2 proportion in the CloudGhost module lacked justification. In the revised manuscript, we have added a detailed explanation in **Section 2.4**. This ratio originates from the core design philosophy of the Ghost module (Han et al., CVPR 2020): generating half of the output feature maps via cheap operations and the other half via pointwise convolution achieves an optimal trade-off between parameter efficiency and representational capacity. Previous studies have also validated that this empirical split significantly reduces FLOPs while effectively preserving the feature diversity essential for fine-grained texture recognition. We have explicitly cited the original GhostNet literature to support this design choice.

4、The introduction lacks a systematic summary of the core innovations.

The introduction effectively reviews the evolution of cloud classification methods and identifies three key challenges, but it ends abruptly without explicitly synthesizing how the proposed work addresses these challenges. I strongly recommend adding a dedicated paragraph at the end of the introduction that concisely summarises the main contributions. Such a summary would significantly improve the readability and narrative coherence of the paper.

Answer: We greatly appreciate your constructive suggestion. We fully agree with your observation that the original Introduction ended too abruptly after presenting the key challenges. In the revised manuscript, we have added a dedicated paragraph at the end of the Introduction to provide a systematic summary of the key challenges and core innovations, as follows:

In summary, current ground-based remote sensing cloud classification algorithms mainly face two key challenges: (1) insufficient fine-grained feature extraction leading to low recall rates for certain cloud genera; (2) excessive model parameters hindering deployment on mobile meteorological observation devices. To systematically tackle these issues while fully harnessing the physical characteristics of various cloud types, the proposed CloudMViT framework is deliberately designed to align with these specific challenges. Specifically, the ECA-DP module employs a dual-pooling strategy to suppress uniform sky background while highlighting high-contrast edge features, effectively tackling the blurred cloud-sky boundaries seen in layered clouds and the fragmented, non-contiguous cloud blocks separated by clear-sky backgrounds. The cross-scale self-attention mechanism is incorporated to build long-range spatial dependencies beyond local receptive fields, thereby grasping the elongated, globally distributed fibril structures of Cirrus-like clouds. Meanwhile, the DW-PW collaborative decoupling module extracts precise local texture variations and inter-channel correlations, preventing feature loss from subtle cross-channel shadow variations in layered clouds. Finally, to meet the demands of portable meteorological observation for computational efficiency, the CloudGhost cascade compression strategy further reduces model parameters to ensure seamless deployment on edge devices. Through these targeted architectural innovations, this work aims to promote the intelligent development of ground-based remote sensing cloud observation methods and devices.

Minor Issues:

1、The manuscript uses names such as exSwish, exLayerNorm, and exSoftmax (in Table 2). These should be clearly identified as standard operators (SiLU, LayerNorm, Softmax) at their first occurrence. I suggest using the format "SiLU (exSwish)" in the operator column of Table 2 to improve accessibility for a broader readership.

Answer: Thank you for this suggestion. In the revised manuscript, we have updated the operator names in **Table 2** to the standard format: "SiLU (exSwish)", "LayerNorm (exLayerNorm)", and "Softmax (exSoftmax)". This approach retains the original operator specifics generated by the framework while ensuring easy identification of the standard operators for a broader readership.

2、Throughout the paper (Tables 10 and 11), units such as "FLOPs/M" and "G" are used ambiguously. To conform to community standards, please adopt MFLOPs (mega floating-point operations) or GFLOPs (giga floating-point operations) consistently.

Answer: We thank you for your meticulous attention to this detail. We have strictly standardized the units for FLOPs in the revised manuscript. Specifically, the units in Table 10 have been unified to MFLOPs (mega floating-point operations), while the units in Table 11 have been unified to GFLOPs (giga floating-point operations), ensuring consistency in units throughout the manuscript.

3、In Section 2.1, immediately after Eq. (5), the phrase "Here, |t| odd denotes the odd integer closest to t" contains a duplicated "odd" and is stylistically awkward.

Answer: We thank you for pointing out this awkward phrasing. In the revised manuscript, we have corrected this sentence in **Section 2.1** to a clearer and more natural expression: " Here, $|t|_{odd}$ represents the nearest odd integer to the value t, ensuring that the kernel size of the 1D convolution is an odd number."

4、Section 2.1 states: "Finally, an activation function k is used to calculate channel-wise weights for the cloud feature map, and these attention weights σ are multiplied by the original input feature map." This is misleading: k is the kernel size of the 1D convolution, not an activation function. The activation function is the sigmoid function σ. The sentence should be rephrased to accurately reflect that the sigmoid function generates the normalized attention weights.

Answer: We sincerely thank you for this precise correction. We fully accept this comment. The original sentence misleadingly referred to the 1D convolution kernel size kk as an activation function. We have revised this in **Section 2.1** to: " Finally, a sigmoid activation function σ is applied to generate channel-wise attention weights η for the cloud feature map. These weights η are then multiplied element-wise with the original input feature map to produce the final output cloud feature map.”

5、In Sections 2.4 and the abstract, the authors refer to the transformations used within the CloudGhost module to

generate redundant features as 'linear operations.' However, these transformations typically involve depthwise convolution followed by batch normalization (BN) and activation functions, and are therefore not purely linear. It is suggested that this terminology be standardized to more accurately reflect the actual operations.

Answer: We thank you for pointing out this terminology issue. We fully agree with your comment that the transformations are not purely 'linear' due to the inclusion of Batch Normalization and activation functions. Therefore, as suggested, we have replaced the term 'linear operations' with **'cheap operations'** (or **' efficient CloudGhost '**) throughout the manuscript, particularly in **Section 2.4**.

In Section 2.4, we also explicitly state that these cheap operations typically consist of a depthwise convolution, followed by Batch Normalization (BN) and an activation function. We emphasize that although these transformations incorporate non-linear components, their effective computational cost remains significantly lower than that of standard convolutions.

We sincerely hope that our revisions fully address your concerns. We believe that the manuscript has been significantly strengthened in terms of clarity, mathematical rigor, and terminology accuracy. We look forward to your positive consideration.

Sincerely,
Wei Xu