Dear Anonymous Referee #1,

We sincerely thank you for your thorough and constructive review of our manuscript. Your positive feedback on our work is greatly appreciated. We have carefully considered all your comments and suggestions, and we have revised the manuscript accordingly. Below, we provide a detailed point-by-point response to your concerns.

Major Comments

1. The manuscript should add a section to introduce the cloud categories for classification and the cloud image datasets.

Answer: We greatly appreciate your valuable suggestion. In response, we have specifically added an introduction to the WMO cloud classification standards, supplemented by the physical and morphological characteristics of different cloud genera, within the Introduction section. Meanwhile, in Section 3.1, we have explicitly distinguished between the "original image set" and the "augmented image set" (as shown in Tables 4 and 5), and we also provided the physical motivations for the data augmentation techniques (e.g., simulating varying illuminations, sensor noises, etc.) to avoid any confusion regarding the sample sizes.

The World Meteorological Organization (WMO) classifies clouds into 10 basic genera (e.g., cumulus, cumulonimbus, stratocumulus, stratus) based on their shape and structure, and has published the latest International Cloud Atlas. As cloud density gradually decreases, the boundary between clouds and the sky becomes blurred, and clouds exhibit complex textural features—posing significant challenges for ground-based cloud image classification. These ten cloud genera exhibit distinct physical and morphological attributes, which inherently dictate varying classification difficulties. For instance, Cirrus and Cirrocumulus are characterized by delicate, high-frequency fibril structures or wave-like patterns requiring fine-grained spatial feature extraction. In contrast, Stratocumulus and

Stratus possess predominantly uniform, layered structures with indistinct edges, where classification relies more on subtle cross-channel intensity variations, such as shadow depth. Furthermore, Cumulonimbus is prone to confusion with Cumulus due to their similar visual appearance at lower altitudes and the rapid vertical development of the former under unstable conditions. These distinctive structural properties impose different requirements on the network's feature extraction capability, specifically in addressing background interference, preserving global morphological contexts, and accurately capturing fine local textures.

2. The manuscript discusses the overall accuracy drop of 4.02% after INT8 quantization but does not analyze which cloud categories suffer the most significant loss from a physical perspective. Specifically, the recall of cirrocumulus (Cc) drops dramatically from 97.39% to 87.06%, yet the physical reason for this degradation is not explained. The authors should provide a detailed analysis of why cirrocumulus are more sensitive to quantization errors, and how the quantization process destroys these physical texture features. A quantitative comparison of recall changes across all cloud categories is strongly recommended.

Answer: We greatly appreciate the reviewer for raising this highly insightful physical perspective. In response, we have conducted a more detailed analysis of our experimental data and incorporated a comprehensive comparative analysis of the recall variations across all cloud categories before and after quantization into the revised manuscript.

Detailed analysis reveals that the recall of *Cirrocumulus (Cc)* drastically declines from 97.39% to 87.06% (a drop of 10.33%), and *Stratocumulus (Sc)* decreases from 98.98% to 93.02%, whereas *Stratus (St)* and *clear sky (No)* experience a marginal decline of only approximately 3%. The physical reason for this discrepancy lies in the fact that *Cc* and *Sc* exhibit high-frequency, dense textural patterns (*Cc* features tightly packed scaly

or rippled structures, while *Sc* involves subtle shadow gradients between large cloud blocks). The INT8 quantization process compresses floating-point values into merely 256 discrete integer bins, severely rounding off and smoothing out the minute grayscale gradients and fine edge details between these pixels.
Conversely, *St* and *No* are characterized by low-frequency, smooth physical features (extensive uniform veil-like structures or solid backgrounds), where the information is concentrated in the low-frequency range, making them far less susceptible to the destructive effects of quantization rounding errors. This physical analysis not only clarifies the inherent limitations of quantization in edge deployment but also provides a theoretical foundation for future optimizations on embedded platforms, such as adopting mixed-precision strategies (e.g., retaining FP16 for critical layers) or quantization-aware training (QAT) to mitigate the accuracy loss for high-frequency texture cloud classes.

3. The manuscript mentions that the ECA-DP module suppresses background interference via GMP, but does not explain from a physical perspective why traditional CNNs (e.g., MobileNet) struggle with blurred cloud-sky boundaries. Furthermore, how does the self-attention mechanism in CloudMViT physically capture such gradual transitions? The authors should supplement the description of the physical characteristics of cloud edges and clarify how specific modulesrespond to these physical properties.

Answer: We greatly appreciate the reviewer's insightful focus on the physical properties of cloud imagery. In response, we have added detailed explanations in Sections 2.1 and 2.2 of the revised manuscript to elucidate the physical characteristics of the "blurred cloud-sky boundaries" and the corresponding response mechanisms of our proposed modules.

For ECA-DP: The blurred cloud-sky boundary is often a region with low-intensity contrast, making it difficult for traditional CNNs (e.g., MobileNet) to distinguish boundaries from the background due to

their limited local receptive fields and lack of global context. Physically, the Global Max Pooling (GMP) in ECA-DP captures the maximum luminance or gradient variation near the cloud boundary, effectively highlighting the sharpest structural edges regardless of the overall sky brightness. This mechanism physically suppresses the uniform and low-variance sky background, focusing the network's attention on the abrupt transitions at the cloud edges.

For Self-Attention: The gradual transition from the sky to the cloud edge is not a hard line but a continuous gradient. While GMP identifies sharp edges, the self-attention mechanism physically captures this gradual transition by computing correlations between distinct local patches and their surrounding regions. It dynamically constructs a global spatial context, where the attention weights for pixels located at the fuzzy boundary are adaptively adjusted based on their similarity to both the high-contrast cloud features and the low-contrast sky background. This enables the model to physically encode the 'softness' of the boundary and the overall morphological distribution of the clouds, rather than treating pixels as independent features.

4. The manuscript states that "DW convolution can extract subtle edge differences between cirrus (Ci) and cirrocumulus (Cc)," but does not specify how the physical texture features of these clouds are represented in the model. The authors should add a discussion explaining how specific modules (e.g., the local receptive field of depthwise separable convolution) physically match the scale characteristics of these special textures.

Answer: To further elucidate the physical basis for selecting these specific modules for fine-grained cloud texture extraction, we analyzed the synergistic matching mechanism between the network components and the unique physical textural characteristics of different cloud genera, particularly Cirrus (Ci) and Cirrocumulus (Cc).

Physical textural differences: From a meteorological

perspective, Cirrus (Ci) is characterized by sparse, directionally-oriented filamentous/fibrous structures that are typically spatially discontinuous, whereas Cirrocumulus (Cc) exhibits dense, periodic scaly or rippled textures with relatively uniform cloud-block sizes.

Physical matching of Depthwise Convolution (DW): The 3×3 depthwise convolution kernel possesses a limited local receptive field, making it ideally suited for capturing the inherent periodic local variations characteristic of Cirrocumulus's scaly patterns. This kernel scans individual scales and detects minute grayscale fluctuations between adjacent cloud blocks. Subsequently, the Pointwise Convolution (PW) aggregates these local texture primitives across the channel dimension to encode a complete morphological representation of the rippled texture.

Physical matching of Self-Attention and Patch Partitioning: The self-attention module employs a 2×2 patch partitioning strategy.
For Cirrus (Ci), this specific partition scale physically corresponds to the typical cross-sectional width of sparse fibril filaments. Each 2×2 patch can capture local, subtle fragments of the filaments. More importantly, the subsequent self-attention calculation computes long-range global dependencies among these scattered fragments. This mechanism effectively "reconstructs" the complete continuity of individual filaments across the full spatial extent of the image—a capability lacking in traditional CNNs due to their restricted receptive fields.

Through this synergistic mechanism, the model achieves precise physical scale alignment at the cloud texture level, significantly enhancing the discriminative accuracy between the visually similar Ci and Cc categories.

The above analysis has been incorporated into Section 2.3 "DW-PW Collaborative Feature Decoupling" of the revised manuscript.

Specific Comments

1.Many citation marker were missing in the literature; for

example, line 55 Wang et al.; and in Section 2 Method, during the introduction of each method I couldn't see any references

Answer: We acknowledge that the foundational modules in the Methodology section lacked proper references. To address this issue, we have added the corresponding original citations for the core mechanisms. Specifically, we have cited [Vaswani et al., 2017] for the self-attention mechanism in Section 2.2, [Wang et al., 2020] for the ECA module in Section 2.1, and [Han et al., 2020] for the Ghost module in Section 2.4. The corresponding text has been revised accordingly.

2.In Section 3.2, the manuscript states "total number of iterations was 100." However, deep learning training typically uses "epochs" rather than "iterations." The authors should clarify the intended meaning and unify the terminology to avoid confusion.

Answer: We thank you for pointing out this terminology issue. In deep learning, the term "epochs" is conventionally used to denote the number of times the model iterates over the entire training dataset. To avoid confusion, we have revised the phrase "total number of iterations was 100" in **Section 3.2** to "**the total number of training epochs was set to 100**," ensuring consistency with standard deep learning terminology.

3. In Fig. 1, σ points to the output of the sigmoid activation function. In Eq. (1), η=σ(Vkγ), where σ already denotes the sigmoid function. However, Fig. 1 also shows a multiplication operation (⊗) after σ, which is not reflected in Eq. (1). The authors should check and correct this inconsistency.

Answer: We sincerely thank you for pointing out the detailed issue of symbol inconsistency between Eq. (1) and Fig. 1. We fully agree with this comment and have revised the manuscript accordingly, with the following corrections made: The original phrase "these attention weights σ" in the main text was inaccurate, as σ

represents the sigmoid activation function rather than the attention weights. We have corrected it to "these attention weights η" to ensure that the textual description is consistent with the symbols used in Eq. (1) and Fig. 1.

The aforementioned revisions guarantee complete symbol uniformity and logical correspondence among Eq. (1), Fig. 1, and the main text, thereby avoiding any potential confusion for readers. Please refer to Section 2.1 and Eq. (1) in the revised manuscript for the specific changes.

4. The Conclusion states that the inference time is "only 0.006 s," but Table 10 lists an inference time of 30.14 s for CloudMViT (total time). Readers cannot easily derive the per-image inference time of 0.006 s. The authors should verify and correct the data, or explicitly state the calculation basis for the per-image inference time.

Answer: We sincerely thank you for identifying this inconsistency in the reported inference time. The reviewer's calculation is absolutely correct.

Upon re-examination, the 30.14 s reported in Table 10 is indeed the total inference time on the entire PC-side test set, which contains 25,118 images. Therefore, the correct per-image inference time should be calculated as: The original 0.006 s stated in the Conclusion was erroneously derived using the test set size of the subsequent embedded deployment experiment (5,028 images), i.e., 30.14/5,028≈0.006. This cross-referencing of datasets indeed led to a misleading figure and confusion.

We have made the following corrections in the revised manuscript: the GPU per-image inference time stated in the Conclusion has been revised to 0.0012 s (1.2 ms), with a clear note that this value is calculated based on the full test set of 25,118 images. Meanwhile, in Table 10, the original heading “Inference Time” has been changed to “Total Inference Time on Test Set” to ensure full consistency between the table and the main text. We greatly appreciate the

reviewer’s rigorous attention to detail, which has significantly improved the data consistency of our paper.

5. In Eq. (4) and Eq. (5), the parameters band rare only stated as "set to 1 and 2" in the text, but their physical meaning or derivation is not explained. Readers cannot understand why these specific values were chosen. The authors should clarify the physical significance or the basis for determining these parameters.

Answer: We sincerely thank you for pointing out the lack of explanation regarding the parameters in Eqs. (4) and (5). We fully agree that the derivation and basis for setting b and r were insufficient in the original manuscript.

In the revised manuscript, we have added a clarifying paragraph in Section 2.1 as follows:

（1）Definition of parameters: We explicitly stated that b and r are hyperparameters controlling the nonlinear mapping between the channel dimension c and the 1D convolution kernel size k. They are not physical constants specific to cloud physics.

(2) Basis of selection: The mapping function and the specific values (r=2 and b=1) originate from the empirical optimization in the ECA-Net paper (Wang et al., CVPR 2020). According to the original study, these values were determined through grid search. The configuration r=2 and b=1 maps common channel dimensions (e.g., 16 to 1024) to appropriate odd kernel sizes (e.g., 3, 5, 7), effectively capturing local cross-channel interactions without introducing extra parameter redundancy.

(3) Applicability: Since the ECA mechanism and this mapping strategy have been widely validated for general computer vision tasks, we directly adopted this baseline configuration in our ECA-DP module and verified its effectiveness for the specific task of cloud image classification.

We have updated the text in the revised manuscript to provide the necessary context for readers, and we greatly appreciate this valuable suggestion.

6. Fig. 6 labels step1, step2, and step3. Section 2.5 describes them as "Step 1: Extraction of local detailed features,""Step 2: Extraction of global features," and "Step 3: Enhancement of local detailed features." However, step1 and step3 have identical structures in the figure, and step2 is positioned in the middle. The text states that step3 "repeats step1," but the spatial arrangement in the figure conflicts with the semantic meaning of "repetition." The authors should revise either the figure or the text to ensure consistency.

Answer: We sincerely thank you for pointing out the semantic inconsistency between Fig. 6 and the text regarding the term "repeated". We fully agree that the word "repeated" might mislead readers into thinking Step 3 is a cyclic or feedback mechanism, which contradicts the forward cascaded layout shown in the figure.

In the revised manuscript, we have updated the description in Section 2.5 as follows:

Avoided the term "repeat": We replaced "The same steps as Step 1 are repeated" with a more precise statement.

Clarified the structural relationship: We now explicitly state that "Step 3 adopts the identical architectural design as Step 1". We also clarify that although the internal modules (PW-Conv, BatchNorm, SiLU, DW-Conv, BatchNorm, SiLU, and CloudGhost) are identical, Step 3 is positioned at a deeper stage of the network. Its purpose is to further refine the local detailed features before the final classification stage.

This revision accurately reflects the forward, cascaded structure shown in Fig. 6 while eliminating the semantic ambiguity. No

modifications to the figure itself are required. We have updated the text accordingly in the revised manuscript.

7. The "Inference Time/s" column in Table 10 shows large values (e.g., 30.14 s for CloudMViT), but the manuscript does not specify that this is the total time for the entire test set . Readers may mistakenly interpret it as the per-image inference time, leading to misunderstanding of the model's speed. The authors should clearly label the column as "Total Inference Time on test set" in the table header or footnote, and also provide the per-image inference time in the main text.

Answer: We thank you for pointing out the potential ambiguity in Table 10. We completely agree that the value 30.14 s listed in the table represents the total inference time on the entire test set (containing 25,118 images), not the per-image inference time. Without a clear label, this could indeed mislead readers.

To resolve this issue, we have made the following corrections in the revised manuscript:

Modified the table header: The column Inference Time/s has been renamed to Total Inference Time on Test Set (s).

Updated the main text: We have revised the statements in the Conclusion to explicitly mention both the per-image inference time and the total inference time,explicitly mention the per-image inference time alongside the total time, ensuring consistency between the table and the description.

We believe these revisions will effectively eliminate any ambiguity and help readers accurately interpret both the total and per-image inference speeds. We sincerely appreciate the reviewer's meticulous attention to detail.