Stereoscopic Dual-Stream Pseudo-Siamese Network for Cloud Top Height Retrieval from Heterogeneous Dual-Satellite Data

1st Zhaoran Fan, 2nd Ming Wu, 3rd Chuang Zhang, 4th Mingrui Xu
Beijing University of Posts and Telecommunications
Beijing, China
Emails: fzr6666662022@163.com, Wuming@bupt.edu.cn,
Zhangchuang@bupt.edu.cn, xumingrui@bupt.edu.cn

***Abstract*—Accurate retrieval of Cloud Top Height (CTH) is critical for meteorological analysis and climate monitoring. However, traditional retrieval methods relying on single geostationary satellites often suffer from limited viewing angles, lack of stereoscopic information, and sensor-specific spectral limitations. Operational observations indicate that FY-4A exhibits superior performance in retrieving high-altitude clouds, particularly cirrus, whereas FY-4B demonstrates higher accuracy in mid-to-low cloud regimes. To leverage these complementary strengths and address single-view limitations, this paper proposes a novel Stereoscopic Dual-Stream Pseudo-Siamese Network that fuses heterogeneous data from the FY-4A and FY-4B satellites. In terms of network architecture, we design a Self-Generated Prior Guidance (SPG) module to extract global contextual priors, enabling the alignment of heterogeneous features at the early encoding stage. Furthermore, a Cross-Attention Bridge (CAB) is incorporated to explicitly model the geometric relationships between the two satellite views, utilizing attention mechanisms to rectify parallax-induced discrepancies and enhance feature interaction. Extensive experiments, conducted using a dataset of synchronous FY-4A/B and CALIPSO observations collected from June 2022 to April 2023 over the Yellow and Bohai Seas, demonstrate that our dual-satellite fusion approach significantly outperforms single-satellite baselines. The proposed model achieves a Mean Absolute Error (MAE) of 1.41 km and a Pearson correlation coefficient of 0.802, proving the effectiveness of deep learning-based stereoscopic fusion for high-precision atmospheric parameter retrieval.**

***Index Terms*—Cloud Top Height Retrieval; Dual-Satellite Fusion; Pseudo-Siamese Network; Cross-Attention Mechanism**

I. Introduction

Clouds play a pivotal role in the Earth's energy budget and hydrological cycle. Cloud Top Height (CTH) is a fundamental parameter for numerical weather prediction (NWP), aviation safety, and climate monitoring. With the rapid development of the geostationary meteorological satellites, particularly FY-4A and FY-4B, the capability for high-frequency, high-resolution continuous observation has been significantly enhanced. However, retrieving accurate CTH from passive remote sensing data remains a challenging task due to the complex nonlinear relationship between satellite radiances and atmospheric states.

This work was supported by NSFC under Grant U24B20177.

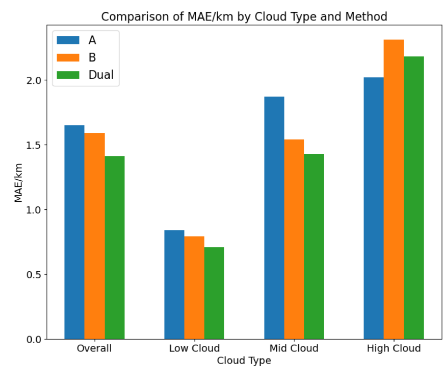

Fig. 1. Comparison of Mean Absolute Error in kilometers across different cloud types for three retrieval methods.

Traditional retrieval methods [1]–[5], such as the infrared split-window algorithm [2] and the CO_2 slicing method [4], rely heavily on physical assumptions and accurate atmospheric profiles. These methods often struggle with complex cloud conditions, such as multi-layer clouds or optically thin cirrus clouds, where model assumptions may fail. Geometric stereoscopic methods [6]–[10] using dual satellites utilize parallax [8]–[10] to calculate height geometrically, but they are sensitive to satellite synchronization and require strict matching of cloud features.

In recent years, deep learning has demonstrated superior performance in meteorological inversions. Data-driven approaches [11]–[16] can learn complex patterns without explicit physical modeling. However, existing deep learning-based CTH retrieval methods [14], [17] face two major limitations. First, most studies focus on single-satellite input [5], which inherently suffers from blind spots and a lack of stereoscopic depth information. Second, when fusing dual-satellite data, standard methods [8] often employ simple channel concatenation . This ignores the heterogeneity of the sensors (different spectral response functions) and the geometric parallax caused by different viewing angles. Direct concatenation can lead to feature confusion and failure to effectively exploit the stereoscopic cues hidden in the disparity.

Furthermore, the quality of training labels is a critical bottleneck. Existing Level-2 (L2) CTH products from individual satellites exhibit systematic biases depending on cloud type [13]. As illustrated in Fig. 1, FY-4A algorithms perform better in high-altitude clouds, while FY-4B excels in other regimes. Training a model on raw biased labels limits the upper bound of accuracy.

To address these challenges, we propose a Stereoscopic Dual-Stream Pseudo-Siamese Network for high-precision CTH retrieval. Unlike blind concatenation, our model processes FY-4A and FY-4B data through separate encoder streams to respect their heterogeneous nature. We introduce a Physics-Guided Label Fusion Strategy, leveraging cloud type classification to merge the superior cirrus retrieval of FY-4A with the robust general cloud retrieval of FY-4B, constructing a high-quality ground truth.

The main contributions of this paper are as follows:

- **Stereoscopic Dual-Stream Pseudo-Siamese Network:** We propose a dual-stream network to independently process heterogeneous FY-4A/4B inputs. This design achieves feature decoupling, extracting distinct semantic representations to effectively avoid the feature confusion caused by early concatenation.
- **Self-Generated Prior Guidance (SPG):** SPG module extracts global context from the dual-satellite input to guide the individual encoders, ensuring feature alignment across heterogeneous sensors at the early stage.
- **Cross-Attention Bridge (CAB):** A cross-attention mechanism is introduced to explicitly model the geometric correlation between the two views. It utilizes parallax to achieve soft alignment of features and enhances stereoscopic feature interaction.

II. Related Work

A. Traditional Physical Retrieval Algorithms

Traditional Cloud Top Height (CTH) retrieval [1]–[5] heavily relies on radiative transfer theory. The representative Split-Window Algorithm [2] utilizes the brightness temperature difference between the 10.8 μm and 12 μm infrared window channels to retrieve cloud parameters via radiative transfer equations. However, this method typically assumes clouds act as blackbodies, limiting its accuracy for semi-transparent or low clouds. To address this, the CO_2 Slicing method [4] was introduced, which exploits the varying sensitivity of CO_2 absorption channels in the 13-14 μm band to atmospheric height to retrieve cloud top pressure. Although these physical methods are well-established, they depend heavily on prior knowledge such as atmospheric temperature and humidity profiles. Their accuracy significantly degrades under multi-layer cloud conditions or thin cirrus clouds where model assumptions fail. Furthermore, the high computational complexity of iterative physical algorithms restricts their application in real-time operational scenarios.

B. Stereoscopic Retrieval based on Geometric Parallax

Beyond radiometric methods, geometric retrieval utilizing viewing angle differences from multiple satellites [8]–[12] constitutes another important branch. Similar to binocular vision, these methods calculate cloud height by measuring the parallax of the same cloud target observed from different satellite perspectives. For instance, Wang *et al.* applied stereoscopic observation using dual geostationary satellites [9] to estimate the height of volcanic ash clouds. While geometric methods possess the advantage of being independent of infrared calibration and atmospheric profiles, they impose extremely high requirements on temporal synchronization. Any time lag between observations can induce false parallax due to cloud motion, leading to substantial errors. Moreover, traditional geometric approaches based on rigid mathematical matching struggle to handle the non-rigid deformation and texture ambiguity of clouds.

C. Deep Learning for Cloud Parameter Retrieval

In recent years, deep learning has demonstrated superior potential over traditional physical algorithms in meteorological remote sensing due to its powerful non-linear mapping capabilities. Numerous studies [14]–[24] have achieved end-to-end cloud parameter retrieval based on Convolutional Neural Networks. For example, Zhang *et al.* utilized transfer learning [14] to predict cloud optical thickness and top height from Himawari-8 thermal infrared data, while Cheng *et al.* explored using active remote sensing data [16] as ground truth to supervise passive sensor networks. However, existing deep learning approaches still face two primary limitations. First, most existing networks rely on single-satellite input, failing to leverage the depth information inherent in stereoscopic vision. What's more, although a few studies [10] attempt to fuse dual-satellite data, they predominantly employ "Channel Concatenation". This approach ignores the spectral heterogeneity (different spectral response functions between sensors) and geometric parallax, preventing the network from effectively learning valid stereoscopic features.

To address these gaps, this study proposes a Dual-Stream Pseudo-Siamese Network. By incorporating explicit Cross-Attention mechanisms and Self-Generated Prior Guidance, our model achieves soft alignment of heterogeneous data and interactive stereoscopic feature learning in the feature space, thereby overcoming the strict synchronization constraints of traditional geometric methods and the feature confusion issues of simple concatenation strategies.

III. Methodology

A. Dual-Stream Pseudo-Siamese Network

Traditional fusion methods typically employ either early fusion (direct concatenation at the input stage) or late fusion (averaging predictions at the output stage). However, despite both FY-4A and FY-4B being geostationary satellites, they exhibit inherent heterogeneity: their sensors possess distinct spectral response functions, and they observe the target from

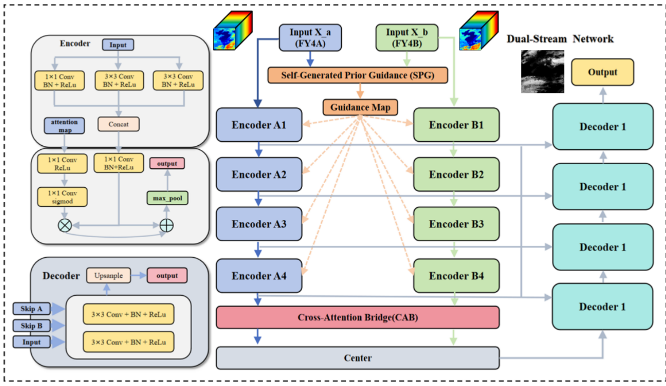

Fig. 2. The overall architecture of the proposed dual-stream fusion network. The model takes FY-4A and FY-4B satellite data as parallel inputs.

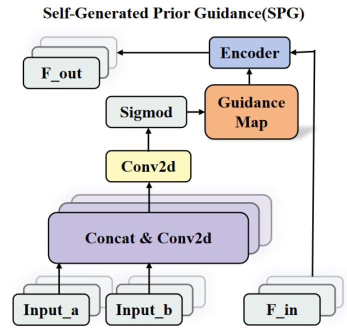

Fig. 3. Detailed structure of the Self-Generated Prior Guidance (SPG) module.

different viewing angles. Direct concatenation forces the network to simultaneously model spectral features and geometric disparities, increasing the optimization burden and leading to potential feature confusion.

To address this, we adopt a Dual-Stream Pseudo-Siamese Network. The term Pseudo-Siamese denotes that the two encoder branches share an identical architecture but maintain independent weights. This design effectively achieves feature decoupling. The network first independently maps the raw data from each satellite into their respective high-dimensional semantic manifolds. Fusion is performed at the deep stage only after each stream has extracted a robust representation of the image content. Compared to blind concatenation, this approach aligns better with physical intuition and effectively mitigates interference caused by heterogeneous data sources. As illustrated in Fig. 2, the network adopts a U-Net-like encoder-decoder structure but features two independent encoder streams to handle the spectral heterogeneity of the two satellites. The architecture begins with a Geometric-Spectral Input Representation that explicitly encodes satellite viewing geometry, followed by a Self-Generated Prior Guidance (SPG) module that extracts global contextual information to guide the feature encoding process. To capture cloud features at varying scales, we employ three parallel branches with different receptive fields. The outputs of these branches are modulated by the SPG module and then concatenated, enabling the network to robustly represent both large-scale cloud systems and fine-grained cloud top fluctuations. Cross-Attention Bridge (CAB) is introduced to explicitly model stereoscopic correlations between the two views and align their features by leveraging geometric correspondence.

B. Geometric-Spectral Input Representation

Unlike traditional methods that only use spectral radiances, we explicitly inject geometric priors into the network to facilitate stereoscopic depth learning. For each satellite input X (where $X \in \{A, B\}$), we concatenate the normalized spectral bands with three static geometric channels: pixel-wise longitude, latitude, and satellite altitude. Let $I_{\text{spectral}} \in \mathbb{R}^{C\times H\times W}$ be the multispectral data (14 channels for FY-4A, 15 for FY-4B). The augmented input X_{input} is defined as:

$$X_{\text{input}} = \text{Concat}(I_{\text{spectral}}, M_{\text{lon}}, M_{\text{lat}}, M_{\text{alt}}) \tag{1}$$

where M_{lon} , M_{lat} , and M_{alt} represent the viewing geometry maps broadcasted to the image resolution. This results in a 17-channel input for stream A and an 18-channel input for stream B, allowing the network to implicitly learn the geometric relationship between the observer (satellite) and the target (cloud).

C. Self-Generated Prior Guidance (SPG)

In the satellite cloud image, clouds occupy only a fraction of the image. To handle the complex cloud textures, we introduce a Self-Generated Prior Guidance mechanism. As is shown in Fig. 3, a lightweight guidance generator first processes the concatenated raw inputs of both satellites to produce a global guidance map $G \in \mathbb{R}^{1\times H\times W}$. This map contains coarse-grained structural information of clouds and is injected into each encoder layer. The attention map produced by Guidance acts like a mask applied to the feature map: it suppresses background values (making them darker) while amplifying cloud-feature values (making them brighter). It guides the dilated convolutions to focus on the same regions of interest, thereby reducing the semantic gap between local texture features and global structural features.

Inside the encoder, the Pointwise Guided Attention Module modulates the features using G. Let F_{in} be the input feature map. The refined feature F_{out} is computed via a residual attention mechanism:

$$F_{\text{out}} = F_{\text{in}} \odot (1 + \alpha \cdot \sigma (\text{Conv}(G))) \tag{2}$$

where $\odot$ denotes element-wise multiplication, σ is the Sigmoid activation, and α is a learnable parameter initialized to 0. This design allows the network to progressively learn to emphasize informative regions (e.g., cloud edges) guided by the global prior.

D. Cross-Attention Bridge (CAB)

The core innovation for stereoscopic retrieval is the Cross-Attention Bridge, located at the deepest encoder stage where high-level semantic features are aligned. Functionally, the

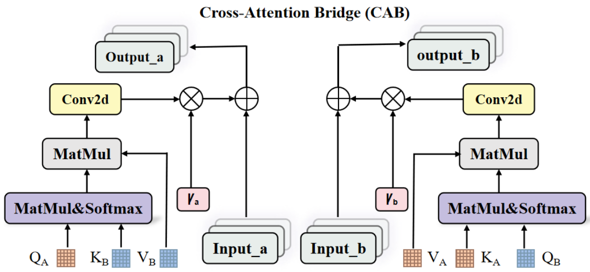

Fig. 4. Schematic illustration of the Cross-Attention Bridge (CAB) module.

CAB module acts as an implicit stereo matching and feature rectification process. Benefiting from the positional information injected at the input stage, the network utilizes geometric parallax to establish correspondence. As is shown in Fig. 4, even when cloud features in the FY-4B view are displaced due to the viewing angle, the FY-4A stream can locate the corresponding anatomical structures via similarity matching, thereby achieving soft alignment in the feature space through weighted aggregation. Furthermore, this mechanism enables Feature Hallucination: when features in the primary stream A are weak or ambiguous, it can dynamically borrow clear, high-confidence features from the auxiliary stream B. This deep, dense interaction effectively breaks the independence barrier of the dual-stream architecture, ensuring robust stereoscopic fusion.

Let F_A and F_B be the flattened feature maps from stream A and B. We compute the Query (Q), Key (K), and Value (V) projections. For stream A attending to B:

$$Q_A = F_A W_{Q_A}, \quad K_B = F_B W_{K_B}, \quad V_B = F_B W_{V_B} \tag{3}$$

The attention weights are calculated as:

$$\text{Attention}(Q_A, K_B) = \text{Softmax}\left(Q_A K_B^\top\right) \tag{4}$$

The aligned feature F'_A is obtained by aggregating information from stream B based on similarity, augmented by a learnable scale factor γ:

$$F'_A = F_A + \gamma_A \cdot (\text{Attention}(Q_A, K_B) V_B)\, W_{\text{out}} \tag{5}$$

A symmetric process is applied to update F_B using information from A. This mechanism allows the network to dynamically borrow features from the other view to correct parallax-induced mismatches and enhance feature consistency.

E. Loss Function

The task is formulated as a regression problem. We use the Mean Squared Error (MSE) loss to minimize the difference between the predicted CTH (Y_{pred}) and the fused ground truth label (Y_{gt}):

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left\| Y_{\text{pred}}^{(i)} - Y_{\text{gt}}^{(i)} \right\|^2 \tag{6}$$

where N is the number of valid pixels in the batch. The model is optimized using the AdamW optimizer with a Cosine Annealing learning rate scheduler.

IV. Experiments

A. Experimental Setup

1) Dataset and Label Refinement: The dataset consists of synchronous observations from FY-4A and FY-4B satellites, spatially covering the Yellow Sea and Bohai Sea region. Temporally, the data spans from June 2022 to April 2023, ensuring a diverse representation of seasonal cloud types, from summer convective systems to winter stratocumulus. To construct high-quality ground truth, we employed a physics-based label fusion strategy. Specifically, since FY-4A sensors exhibit higher sensitivity in cirrus regions while FY-4B performs better in mid-low cloud regions, we dynamically merged their L2 CTH products based on cloud type classification to correct systematic biases. To evaluate the quality of the refined labels, we compared their performance against CALIPSO observations using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). As shown in Table I, the dual-satellite fused label achieves a lower MAE of 2.378 km and RMSE of 3.162 km, outperforming both FY-4A (MAE: 2.429 km, RMSE: 3.242 km) and FY-4B (MAE: 2.783 km, RMSE: 3.831 km). This improvement indicates that the physics-guided fusion strategy effectively mitigates systematic biases inherent in single-satellite retrievals—particularly reducing overestimation in high-cloud regions where FY-4B struggles and underestimation in mid-low clouds where FY-4A is less sensitive. The enhanced consistency with CALIPSO validates the effectiveness of our label refinement approach.

TABLE I
COMPARISON OF MAE AND RMSE BETWEEN LABEL AND CALIPSO FOR DUAL-SATELLITE, FY-4B, AND FY-4A.

label vs Cal	A	B	Dual
MAE(km)	2.429	2.783	**2.378**
RMSE(km)	3.242	3.831	**3.162**

2) Implementation Details: The model was implemented in PyTorch and trained on NVIDIA GPUs. We used the AdamW optimizer with a weight decay of 0.01. The learning rate was managed by a Cosine Annealing scheduler, initialized at 1×10^{-4} and decaying to 1×10^{-6} over 100 epochs. The input images were cropped to 1024×1024 patches. We allocated 80% of the data for training and 20% for validation.

B. Evaluation Metrics

To comprehensively evaluate the retrieval performance, we adopted three standard metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Pearson Correlation Coefficient (R). All evaluations use the refined labels as well as the high-precision CALIPSO data along the satellite ground tracks as the ground-truth reference.

C. Main Results: Single vs. Dual-Satellite Retrieval

We compared our proposed Stereoscopic Dual-Stream Network against single-satellite baselines (using only FY-4A or FY-4B).

TABLE II
COMPARISON OF PREDICTION ACCURACY AGAINST CALIPSO (PRED VS CAL) FOR SINGLE- AND DUAL-SATELLITE MODELS.

Pred vs Cal	A	B	Dual
MAE(km)	2.226	2.168	**2.109**
RMSE(km)	2.668	2.614	**2.531**

TABLE III
PERFORMANCE COMPARISON(PRED VS LABEL) ACROSS CLOUD HEIGHT REGIMES FOR SINGLE- AND DUAL-SATELLITE MODELS.

Pred vs Label	MAE (km)			RMSE (km)			R		
Satellite	A	B	Dual	A	B	Dual	A	B	Dual
Overall	1.65	1.59	**1.41**	2.26	2.20	**2.11**	0.763	0.791	**0.802**
Low Cloud (0–2 km)	0.84	0.79	**0.71**	1.21	0.98	**0.93**	0.430	0.477	**0.481**
Mid Cloud (2–6 km)	1.87	1.54	**1.43**	2.31	2.01	**1.77**	0.334	0.344	**0.348**
High Cloud (>6 km)	2.02	2.31	**2.18**	2.43	2.74	**2.54**	0.635	0.589	**0.601**

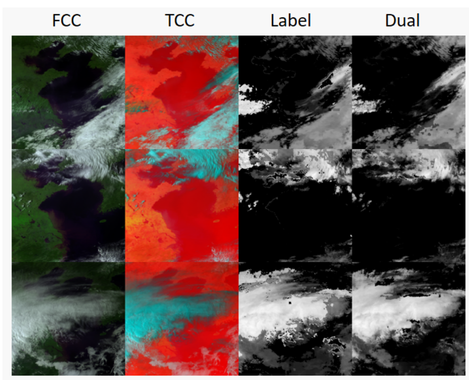

Fig. 5. Qualitative comparison of dual-satellite fusion CTH retrieval against the reference label. From left to right: false-color composite (FCC), true-color composite (TCC), label and dual-satellite fusion prediction.

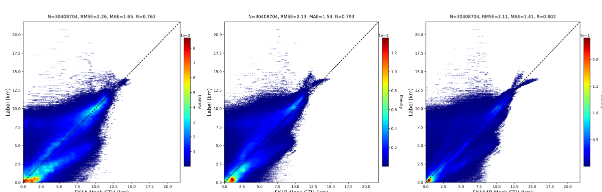

Fig. 6. Scatter-density plots of predicted versus Label CTH for 4A 4B and Dual fusion model. The diagonal dashed line indicates the 1 : 1 identity. Color bar represents sample density.

Validation against CALIPSO: Table II presents the accuracy validated against CALIPSO tracks. The proposed dual-satellite model achieves an MAE of 2.109 km and RMSE of 2.531 km. Compared to single-satellite baselines, it reduces the MAE by 0.12 km (vs. FY-4A) and 0.06 km (vs. FY-4B). This confirms that fusing heterogeneous observations effectively mitigates single-view limitations, yielding results closer to active lidar measurements.

Stratified Analysis against Refined Labels: Table III details the performance against refined labels across cloud regimes.

- **Overall Performance:** The dual model achieves the best consistency with an MAE of 1.41 km and R of 0.802, significantly outperforming both baselines.

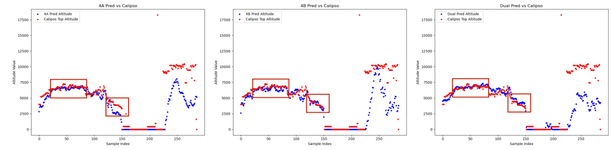

Fig. 7. Scatter plots of predicted versus Calipso CTH for 4A 4B and Dual fusion model.

- **Low/Mid Clouds:** Improvements are most robust in Low (0–2 km) and Mid (2–6 km) clouds, reducing MAE to 0.71 km and 1.43 km . This validates the efficacy of stereoscopic cues for lower tropospheric structures.
- **High Cloud Limitation:** However, for high clouds (> 6 km), although the dual-satellite model(MAE: 2.18 km, RMSE: 2.54 km) still outperforms the FY-4B baseline(MAE: 2.31 km, RMSE: 2.74 km), it is slightly inferior to FY-4A(MAE: 2.02 km, RMSE: 2.43 km). This is presumably because the merits of FY-4A in detecting extremely thin cirrus are not fully exploited, and the integration of FY-4B occasionally degrades the retrieval in these extreme conditions.

Qualitative Visualization Analysis: As shown in Fig. 5, the dual-stream network captures the spatial patterns and height distribution of cloud layers. Fig. 6 illustrates the density scatter plots for different models.The dual-satellite predictions (red core) are tightly concentrated along the 1:1 identity line. This convergence significantly exceeds that of single-satellite baselines. However, a systematic underestimation persists in the high-cloud regime. This limitation stems not from the model itself, but from inherent physical constraints of passive remote sensing in observing very high-altitude clouds. Fig. 7 shows that dual-satellite fusion yields cloud top heights much closer to CALIPSO than single-satellite, especially in complex cloud scenes. These results demonstrate the advantage of dual-satellite fusion in cloud top height retrieval.

TABLE IV
COMPARISON OF PREDICTION ACCURACY AGAINST CALIPSO (PRED VS CAL) FOR ABLATION STUDIES.

Pred vs Cal	-CAB	-SPG	Dual
MAE(km)	2.145	2.153	**2.109**
RMSE(km)	2.559	2.610	**2.531**

TABLE V
PERFORMANCE COMPARISON(PRED VS LABEL) ACROSS CLOUD HEIGHT REGIMES FOR ABLATION STUDIES.

Pred vs Label	MAE (km)			RMSE (km)			R		
Satellite	-CAB	-SPG	Dual	-CAB	-SPG	Dual	-CAB	-SPG	Dual
Overall	1.54	1.55	**1.41**	2.13	2.11	**2.11**	0.793	0.792	**0.802**
Low Cloud (0–2 km)	0.74	0.79	**0.71**	0.93	0.96	**0.93**	0.469	0.443	**0.481**
Mid Cloud (2–6 km)	1.46	1.53	**1.43**	1.76	1.91	**1.77**	0.341	0.346	**0.348**
High Cloud (>6 km)	2.45	2.29	**2.18**	2.73	2.76	**2.54**	0.553	0.525	**0.601**

D. Ablation Studies

To validate the effectiveness of the proposed modules, we conducted ablation studies by removing the Self-Generated

Prior Guidance (SPG) and the Cross-Attention Bridge (CAB) respectively.

1) Impact of Cross-Attention Bridge (CAB): As shown in Table V, removing the CAB module results in a degradation of the overall performance: MAE increases from $1.41\,\text{km}$ to $1.54\,\text{km}$, while R drops from 0.802 to 0.793. The contribution of CAB is most critical in the High Cloud regime ($> 6\,\text{km}$). Without CAB, MAE worsens from $2.18\,\text{km}$ to $2.45\,\text{km}$, RMSE increases from $2.54\,\text{km}$ to $2.73\,\text{km}$, and R declines sharply from 0.601 to 0.553. This aligns with the physical reality that higher clouds exhibit larger parallax displacements between the FY-4A and FY-4B viewing angles. The CAB mechanism successfully captures these large displacements and transforms the geometric disparity into useful depth cues, thereby correcting the height underestimation in deep convective regions.

2) Impact of Self-Generated Prior Guidance (SPG): Removing the guidance module SPG results in an MAE of $1.55\,\text{km}$ and a decrease in R from 0.802 to 0.792. The SPG module shows consistent improvements across Low (0–$2\,\text{km}$) and Mid (2–$6\,\text{km}$) cloud regimes. Specifically, in the Low Cloud regime, the inclusion of SPG boosts R from 0.443 to 0.481 and reduces RMSE from $0.96\,\text{km}$ to $0.93\,\text{km}$. Similarly, for Mid Clouds, the RMSE decreases from $1.91\,\text{km}$ to $1.77\,\text{km}$. These gains indicate that global contextual priors help the network align heterogeneous features at the early encoding stage, enhancing robustness against texture ambiguity and improving structural consistency in lower tropospheric clouds.

Table IV further confirms the robustness of the full model against independent active remote sensing data. The full Dual-Satellite model achieves the lowest MAE of 2.109 km and RMSE of 2.531 km compared to the ablated versions (2.145 km and 2.153 km MAE for -CAB and -SPG, respectively). This demonstrates that both CAB and SPG are indispensable for achieving high-precision retrieval across all cloud regimes.

V. Conclusion

In this paper, we proposed a Stereoscopic Dual-Stream Pseudo-Siamese Network to address the challenges of Cloud Top Height retrieval using heterogeneous geostationary satellite pairs. By explicitly modeling the geometric parallax and spectral heterogeneity between FY-4A and FY-4B, our approach overcomes the limitations of traditional single-satellite inversions. We introduced a physics-guided label fusion strategy to mitigate systematic biases in operational products and designed the Cross-Attention Bridge (CAB) and Self-Generated Prior Guidance (SPG) modules to achieve robust stereoscopic feature learning and heterogeneous alignment. Validation against CALIPSO active remote sensing data confirms that our framework significantly outperforms single-satellite baselines, demonstrating the efficacy of deep learning-based stereoscopic fusion.

References

[1] C. Liu, J. Li, B. Li, and X. Ran, "A review of cloud property retrieval algorithms and products for Fengyun satellite optical imagers (invited)," *Acta Opt. Sin.*, vol. 44, no. 18, pp. 1800003-1–1800003-14, 2024.

[2] L. Duan and Y. Zhang, "Cloud top height retrieval based on satellite infrared window channel data," *Infrared*, vol. 43, no. 8, pp. 33–43, 2022.

[3] G. Ma, J. Huang, X. Gong, and Z Hua, "Development of clear sky channel identification techniques in satellite infrared hyperspectral data assimilation," *Adv. Earth Sci.*, vol. 40, no. 5, p. 487, 2025.

[4] V. Ramanathan, "The role of Earth radiation budget studies in climate and general circulation research," *J. Geophys. Res., Atmos.*, vol. 92, no. D4, pp. 4075–4095, Apr. 1987.

[5] Y. Dong, "From single-satellite observation to system-level synergy: Technical features and prospects of the intelligent collaborative observation system of Fengyun satellites," *Aerosp. Shanghai*, vol. 41, no. 3, pp. 37–46, 2024.

[6] Z. Tan, S. Ma, and C. Liu , "Advances in cloud base height retrieval algorithms using spaceborne multispectral radiometer imagers," *Acta Opt. Sin.*, vol. 44, no. 24, pp. 2400001-1–2400001-10, 2024.

[7] H. Ma, Y. Li, and X. Wu , "Daytime land fog detection using Himawari-8/AHI satellite data," *Remote Sens. Technol. Appl.*, vol. 39, no. 4, pp. 784–792, 2024.

[8] Q. Li, X. Sun, and X. Wang, "Reliability evaluation of the joint observation of cloud top height by FY-4A and Himawari-8," *Remote Sens.*, vol. 13, no. 19, p. 3851, 2021.

[9] P. Wang, H. Tan, S. Bi , and X.Yi, "PF-LRM: Pose-free large reconstruction model for joint pose and shape prediction," *arXiv preprint arXiv:2311.12024*, 2023.

[10] X. Huang, Y. Chen, and P. Li, "Dual-satellite stereoscopic retrieval of cloud top height using FY-4A and FY-4B," *J. Meteorol. Res.*, vol. 38, no. 6, pp. 1141–1149, 2024.

[11] M. Zhang, N. Xu, and Z. Zheng , "Algorithm and accuracy validation of FY-3D MERSI-II cloud-top products," *J. Trop. Meteorol.*, vol. 38, no. 6, pp. 823–834, 2022.

[12] T. Zhang, S. Hu, and F. Tao , "Joint retrieval method for cloud top height using ground-based cloud radar and FY-4A satellite data," *Meteorol. Mon.*, vol. 50, no. 7, pp. 859–867, 2024.

[13] Z. Yu, Z. Tan, and S. Ma , "Nowcast for cloud top height from Himawari-8 data based on deep learning algorithms," *Meteorol. Appl.*, vol. 30, no. 3, p. e2130, 2023.

[14] J. Li, F. Zhang, and W. Li, "Transfer-learning-based approach to retrieve the cloud properties using diverse remote sensing datasets," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–10, 2023.

[15] T. Li, N. Chen, F. Tao ,S Hu, and J Xue, "Cloud top height retrieval from FY-4A data: A residual module and genetic algorithm approach," *Atmosphere*, vol. 15, no. 6, p. 643, 2024.

[16] Y. Cheng, H. He, and Q. Xue , "Remote sensing retrieval of cloud top height using neural networks and data from Cloud-Aerosol Lidar with Orthogonal Polarization," *Sensors*, vol. 24, no. 2, p. 541, 2024.

[17] G. Wang, H. Wang, and Y. Zhuang , "Tropical overshooting cloud-top height retrieval from Himawari-8 imagery based on random forest model," *Atmosphere*, vol. 12, no. 2, p. 173, 2021.

[18] Z. Wang, X. Kong, and Z. Cui , "VecNet: A spectral and multi-scale spatial fusion deep network for pixel-level cloud type classification in Himawari-8 imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2021, pp. 4083–4086.

[19] J. Shao, H. Letu, and X. Ri , "Estimation of surface downward longwave radiation and cloud base height based on infrared multichannel data of Himawari-8," *Atmosphere*, vol. 14, no. 3, p. 493, 2023.

[20] X. Shi, J. Zhang, and G. Liu , "Self-attention based cloud top height retrieval for intelligent meteorological service recommendation," *Inf. Sci.*, vol. 713, p. 122192, 2025.

[21] X. Wang, H. Iwabuchi, and T. Yamashita, "Cloud identification and property retrieval from Himawari-8 infrared measurements via a deep neural network," *Remote Sens. Environ.*, vol. 275, Art. no. 113026, Jun. 2022.

[22] M. Min, J. Li, F. Wang, Z. Liu, and W. P. Menzel, "Retrieval of cloud top properties from advanced geostationary satellite imager measurements based on machine learning algorithms," *Remote Sens. Environ.*, vol. 239, Art. no. 111616, Mar. 2020.

[23] Z. Tan , "Estimating cloud base height from Himawari-8 based on a random forest algorithm," *Int. J. Remote Sens.*, vol. 42, no. 7, pp. 2485–2501, Apr. 2021.

[24] K. Chen, M. Xu, W. Li, F. Mo, Y. Huang, M. Wu, and C. Zhang, "SirP: Semi-recurrent paradigm-based prediction network for satellite image sequences," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 1–15, 2025.