Cloud Motion Prediction through Attention-Guided Frame Interpolation

Yash Agarwal (22BCE1044)
SCOPE, VIT Chennai
yash.agarwal2022b@vitstudent.ac.in

Mehul Dhingra (22BCE1008)
SCOPE, VIT Chennai
mehul.dhingra2022@vitstudent.ac.in

***Abstract*—Cloud motion prediction plays a pivotal role in weather forecasting, disaster monitoring, and satellite-based environmental analysis. Traditional frame interpolation methods often overlook the complex and deformable nature of cloud movements influenced by atmospheric variables such as wind direction and velocity. In this research, we propose a novel deep learning framework titled *Flow Enhanced Interpolation Network (FEIN)*, which leverages attention-guided autoencoders to predict intermediate satellite frames with high temporal and spatial fidelity.**

Our approach integrates multi-source satellite imagery and wind vector data from scatterometer-based sources (e.g., MOSDAC/ISRO), treating wind as a dynamic flow field that guides the temporal interpolation process. The model architecture includes a dual-branch encoder to extract features from both image sequences and wind maps, a custom attention mechanism to capture spatial-temporal correlations, and a decoder that reconstructs intermediate cloud frames. Additionally, perceptual loss using a pre-trained VGG19 network ensures high visual realism in the generated outputs.

The system achieves smooth and accurate frame transitions even under complex atmospheric dynamics, such as during cyclones or rapidly evolving cloud systems. Experimental results show significant improvements over conventional methods, both quantitatively in loss metrics and qualitatively in visual coherence. This research demonstrates the feasibility and impact of physics-informed, AI-driven solutions for real-time satellite video generation and atmospheric modeling.

***Index Terms*—Cloud Motion Prediction, Frame Interpolation, Attention Mechanism, Autoencoder, Satellite Imagery, Wind Vector Mapping, Spatiotemporal Modeling, Deep Learning, Perceptual Loss, Remote Sensing, GAN, Meteorological Visualization**

I. Introduction

Cloud motion prediction from satellite imagery plays a critical role in modern meteorological forecasting, climate modeling, and environmental monitoring. Enhancing the temporal resolution of satellite data by generating intermediate frames between captures enables a more continuous and accurate understanding of atmospheric dynamics. However, predicting the evolution of cloud structures is inherently complex due to their non-rigid, rapidly changing nature, which is heavily influenced by environmental factors such as wind speed and direction.

This research presents a deep learning-based framework—Flow Enhanced Interpolation Network (FEIN)—designed to generate high-fidelity intermediate satellite frames through attention-guided frame interpolation. Unlike conventional interpolation methods that often assume linear or uniform motion, FEIN incorporates external atmospheric variables to model non-linear cloud transformations with greater accuracy.

The proposed architecture adopts an autoencoder-based design, wherein two temporally spaced satellite images and corresponding wind vector maps are independently encoded to extract spatial and contextual features. A custom attention mechanism highlights regions of significant temporal variation between the frames, enabling the model to focus on relevant features that drive cloud movement. The decoder then synthesizes an intermediate frame that realistically represents the transition between the input frames, informed by both visual and meteorological cues. Our system leverages satellite imagery and wind vector data from the MOSDAC INSAT-3DR platform. The dataset comprises manually curated and aligned cloud images along with scatterometer-derived wind maps, allowing the model to learn the interaction between wind patterns and cloud displacement. Despite the challenges of deformable object interpolation and temporal sparsity in satellite captures, FEIN demonstrates strong generalization and visual coherence across diverse atmospheric conditions.

By integrating domain-specific environmental data with ad-

Identify applicable funding agency here. If none, delete this.

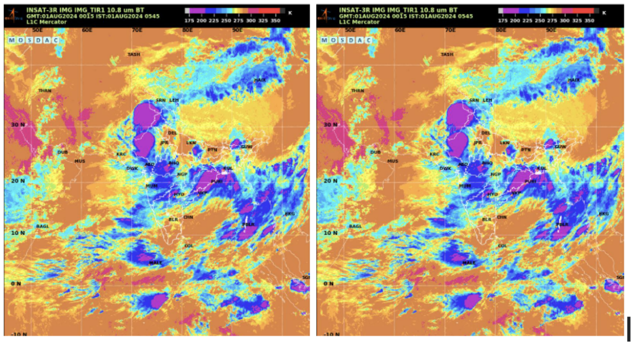

Fig. 1. Infrared BT images from the Asia Sector

vanced deep learning techniques, this work advances the state of the art in satellite-based frame interpolation and provides a valuable tool for real-time visualization and prediction of cloud dynamics.

II. Literature Review

Cloud cover estimation and forecasting are crucial to contemporary meteorology, underpinning a wide range of applications including numerical weather prediction, climate modeling, satellite-based communications, and solar irradiance estimation. The advent of high-resolution geostationary satellite imagery, such as that from the GOES, Himawari, and Meteosat missions, has enabled researchers to analyze cloud systems with improved spatial and temporal granularity.

One of the foundational contributions in this domain was presented by Escrig et al. , who utilized brightness temperature thresholds and spatiotemporal continuity constraints to classify clouds and estimate their motion using imagery from the Meteosat Second Generation satellite. Their approach enabled short-term forecasting of dynamic atmospheric phenomena through a rule-based identification and tracking framework.

With the rise of deep learning, cloud detection and segmentation have undergone a paradigm shift. Jeppesen et al. introduced RS-Net, a convolutional neural network specifically designed for remote sensing applications, which achieved notable success in distinguishing both thin and thick cloud layers under diverse weather conditions. In a parallel effort, Francis et al. developed CloudFCN, a fully convolutional network optimized for dense, pixel-wise cloud classification, outperforming conventional threshold-based methodologies.

Advection-based models, though traditional, continue to influence modern techniques. Bellerby proposed a cloud-top advection scheme leveraging visible and infrared satellite channels to estimate wind vectors and interpolate cloud positions. While physically grounded, this method established the conceptual foundation for subsequent data-driven models.

In recent years, frame interpolation techniques have gained momentum in satellite cloud motion forecasting. Vandal and Nemani presented a method that integrates optical flow with learned interpolation to synthesize intermediate satellite frames. Their work emphasized structural coherence and temporal smoothness, addressing the constraints of low temporal resolution in satellite observations.

Generative approaches, particularly those based on Generative Adversarial Networks (GANs) and transformers, have shown remarkable promise in this context. Jin et al. proposed a deformable attention-based architecture for arbitrary-time frame interpolation in natural videos. Despite being originally developed for conventional video data, the method holds considerable potential for satellite cloud dynamics where spatiotemporal variability is high.

Transformer-based models have also been effectively applied to satellite meteorology tasks. Qin et al. employed DETR (DEtection TRansformer) to identify cold fronts by analyzing large-scale cloud structures in satellite imagery. This highlighted the utility of attention mechanisms in modeling long-range dependencies in evolving atmospheric systems.

The challenge of occlusions due to thick cloud cover has been explored through inpainting and data recovery techniques. Czerkawski et al. developed a deep internal learning strategy for reconstructing cloud-occluded regions using self-similarities within the image, while Wang et al. utilized matrix completion to restore ground-level observations obscured by clouds, benefiting Earth observation applications.

The field has also benefited from benchmark datasets and hybrid deep learning architectures. Cloudcast serves as a standard dataset for evaluating spatiotemporal cloud forecasting models, offering annotated sequences for supervised learning. CDNet, tailored for Landsat-8 imagery, combines multi-spectral fusion with edge-preserving filters to enhance segmentation accuracy.

Despite these advancements, there remains a notable gap in developing an integrated framework that combines generative adversarial models, optical flow-based motion estimation, and attention mechanisms. Most contemporary solutions address cloud segmentation, motion tracking, or frame interpolation in isolation. An end-to-end architecture capable of jointly modeling these components could significantly enhance the fidelity and continuity of satellite-based cloud forecasting. This paper addresses this gap by proposing a unified generative framework for spatiotemporal cloud prediction using GANs and attention-based interpolation, aiming to advance real-time and accurate weather forecasting capabilities.

III. Methodology

The proposed methodology aims to generate realistic intermediate satellite frames by leveraging a deep learning framework that learns spatiotemporal patterns influenced by wind flow. This is achieved through a custom-designed multi-input convolutional autoencoder integrated with an attention mechanism, trained using a perceptual loss function. The system accepts a pair of temporally separated satellite images along with corresponding wind vector information to predict a visually and physically consistent intermediate frame. The incorporation of wind data allows the model to go beyond simple visual interpolation and consider real-world atmospheric dynamics.

A. Dataset and Data Processing

A custom dataset was manually curated to support the development of a wind-aware satellite frame interpolation model. Over 10,000 satellite images were collected from publicly available meteorological sources, consisting of two primary image types:

- **Windy images:** Contain graphical representations of wind direction and magnitude across a geospatial grid.
- **Asia Sector Infrared 1 Brightness Temperature (BT) images:** Provide insights into atmospheric cloud coverage and surface temperature via infrared radiation.

Each image pair—composed of one infrared and one wind image—was resized to a uniform dimension of 256×256 pixels and converted to grayscale to reduce dimensionality. Pixel values were normalized to the $[0, 1]$ range to ensure stable convergence during training.

A key innovation in the preprocessing pipeline involved converting visual wind arrows into quantitative 2D wind vectors. Each arrow, encoding direction and magnitude, was decomposed into horizontal (x-axis) and vertical (y-axis) components using trigonometric analysis. This transformation yielded a dense vector field stored as a NumPy array, aligned spatially with its corresponding infrared image using affine transformation techniques.

To train the model on spatiotemporal transitions, the dataset was structured into triplets:

- **Frame 1:** Infrared satellite image at time t.
- **Frame 2:** Infrared satellite image at time $t + 30$ minutes.
- **Wind map:** Dense vector field representing environmental motion during the interval.

These inputs were compiled into structured NumPy arrays (`x_train1`, `x_train2`, `x_train_wind`), ensuring synchronized and physically consistent training data. This dataset provides both appearance-based and motion-based information essential for realistic satellite image interpolation.

B. Model Architecture

Each input stream (two images and the wind vector map) passes through an independent encoder branch composed of convolutional layers with ReLU activation and dropout for regularization. These encoders extract compact latent representations, capturing relevant spatial and temporal features.

The encoded outputs of the satellite image pair are passed through a custom self-attention mechanism that dynamically highlights regions of significant change. This is crucial for capturing non-linear cloud formation dynamics and identifying atmospheric motion cues.

The attention-enhanced feature is then concatenated with the encoded wind representation, resulting in a fused representation that combines visual transitions with physical wind cues. This composite feature map is then decoded using transposed convolutional layers to reconstruct the intermediate satellite frame.

C. Loss Function and Training Strategy

The model is optimized using a perceptual loss rather than traditional pixel-wise losses. This loss is computed by comparing the feature activations of the generated and ground truth images using a pre-trained VGG19 network, encouraging the output to be perceptually coherent and rich in structural and textural details.

Training is performed with the Adam optimizer for 50 epochs, using a batch size of 64 and a validation split of 20%. Both training and validation losses exhibit consistent convergence, indicating the model's generalization capability across varied weather conditions and sequences.

D. System Scalability and Deployment

The final system is modular and lightweight, making it suitable for deployment in both cloud-based platforms and real-time meteorological applications. Its ability to generate high-quality, physically consistent intermediate frames offers great

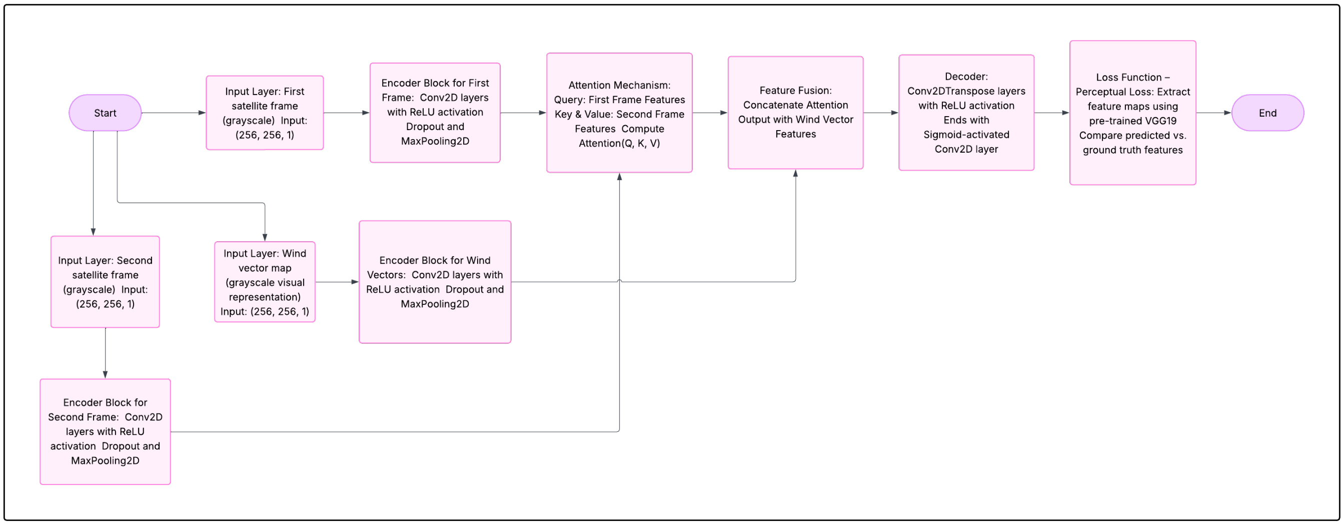
Start
Input Layer: First satellite frame (grayscale) Input: (256, 256, 1)
Encoder Block for First Frame: Conv2D layers with ReLU activation Dropout and MaxPooling2D
Attention Mechanism: Query: First Frame Features Key & Value: Second Frame Features Compute Attention(Q, K, V)
Feature Fusion: Concatenate Attention Output with Wind Vector Features
Decoder: Conv2DTranspose layers with ReLU activation Ends with Sigmoid-activated Conv2D layer
Loss Function – Perceptual Loss: Extract feature maps using pre-trained VGG19 Compare predicted vs. ground truth features
End
Input Layer: Second satellite frame (grayscale) Input: (256, 256, 1)
Input Layer: Wind vector map (grayscale visual representation) Input: (256, 256, 1)
Encoder Block for Wind Vectors: Conv2D layers with ReLU activation Dropout and MaxPooling2D
Encoder Block for Second Frame: Conv2D layers with ReLU activation Dropout and MaxPooling2D

Fig. 2. Overview of the proposed methodology.

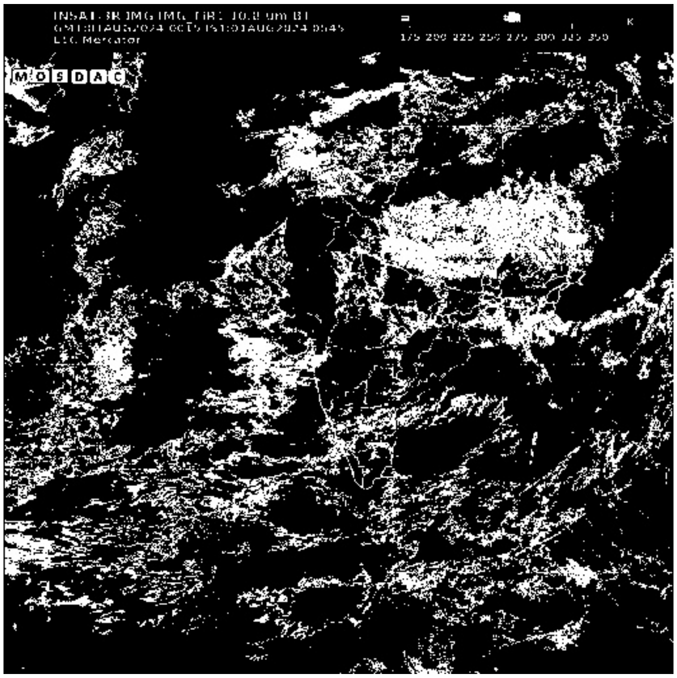
INSAT-3R IMG IMG_TIR1 10.8 um BT
GMT:01AUG2024 00:15 IST:01AUG2024 05:45
L1C Mercator
175 200 225 250 275 300 325 350
K
MOSDAC

Fig. 3. Actual Image

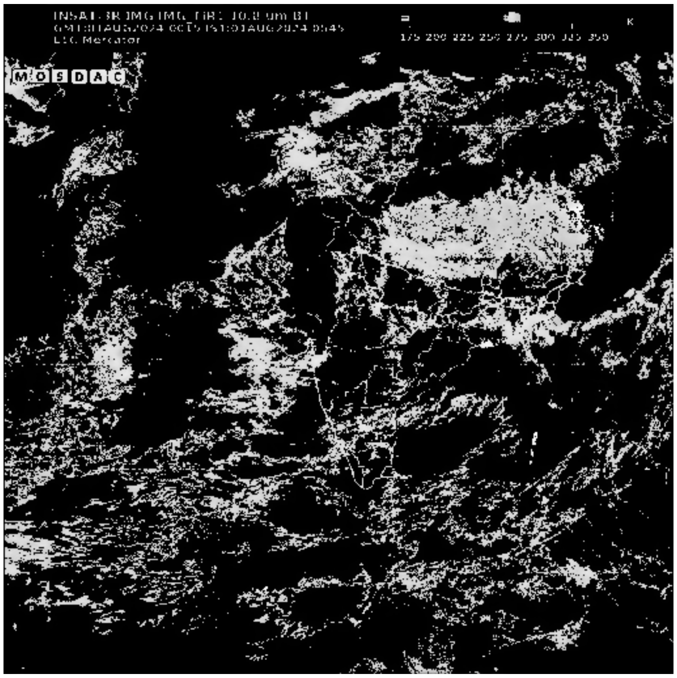
INSAT-3R IMG IMG_TIR1 10.8 um BT
GMT:01AUG2024 00:15 IST:01AUG2024 05:45
L1C Mercator
175 200 225 250 275 300 325 350
K
MOSDAC

Fig. 4. Generated Image

potential for enhancing satellite-based weather visualization and forecasting.

IV. Discussion

The proposed methodology demonstrates the effectiveness of incorporating both visual and environmental cues—specifically wind vector data—into a deep learning framework for generating intermediate satellite frames. By leveraging a custom-designed multi-branch convolutional autoencoder architecture enhanced with a self-attention mechanism and trained using a perceptual loss function, the system achieves visually coherent and physically informed interpolations.

One of the key strengths of the model lies in its **multi-modal design**, which allows it to simultaneously process satellite imagery and wind vector information. This integration is critical for overcoming the limitations of purely appearance-based interpolation models, which often fail to account for the underlying atmospheric dynamics driving visual changes. The incorporation of wind data significantly improves the model's spatiotemporal reasoning, enabling it to simulate cloud movement patterns that align more closely with real-world meteorological behaviors.

The use of a **self-attention mechanism** further enhances both the model's interpretability and performance. By enabling the network to selectively focus on regions of change between the two input frames, the model dynamically adjusts its spatial emphasis. This allows the system to effectively learn which areas are likely to exhibit motion or transformation. Such functionality is particularly beneficial for satellite imagery, where changes may be either highly localized (e.g., cumulus cloud formation) or broadly distributed (e.g., haze diffusion). Traditional convolutional filters struggle with such non-linear transitions, whereas the attention mechanism provides a more flexible and responsive modeling approach. Another notable contribution of this work is the adoption of a **perceptual loss** function, which shifts the optimization objective from low-level pixel similarity to high-level feature similarity. This significantly improves the visual quality of the synthesized frames, mitigating the common problem of overly smooth or blurry outputs associated with standard autoencoder architectures. By computing loss based on feature activations from a pre-trained VGG19 network, the generated frames maintain both structural integrity and fine-grained textural details—qualities that are essential for downstream applications such as weather forecasting or environmental monitoring.

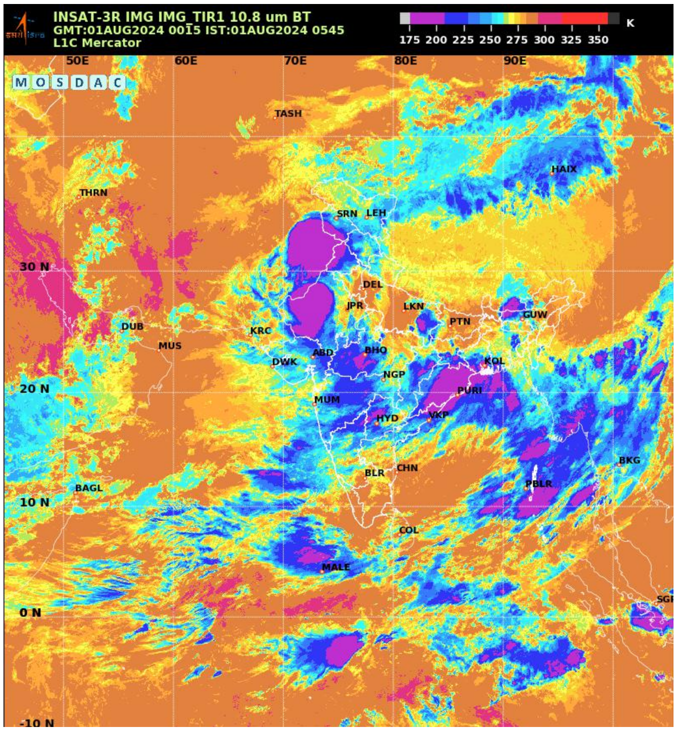

Fig. 5. Input Image 1

From a training perspective, the model exhibits **stable convergence** of both training and validation losses across multiple epochs. This suggests that the network successfully abstracts meaningful motion and environmental patterns without overfitting to specific cloud structures. The consistent generalization capability highlights the robustness of the training pipeline and validates the design choices in data preprocessing, model architecture, and loss function.

Overall, the fusion of physical wind dynamics with visual pattern recognition sets a new benchmark for satellite frame interpolation, presenting a powerful tool for enhancing temporal resolution in satellite-based observation systems.

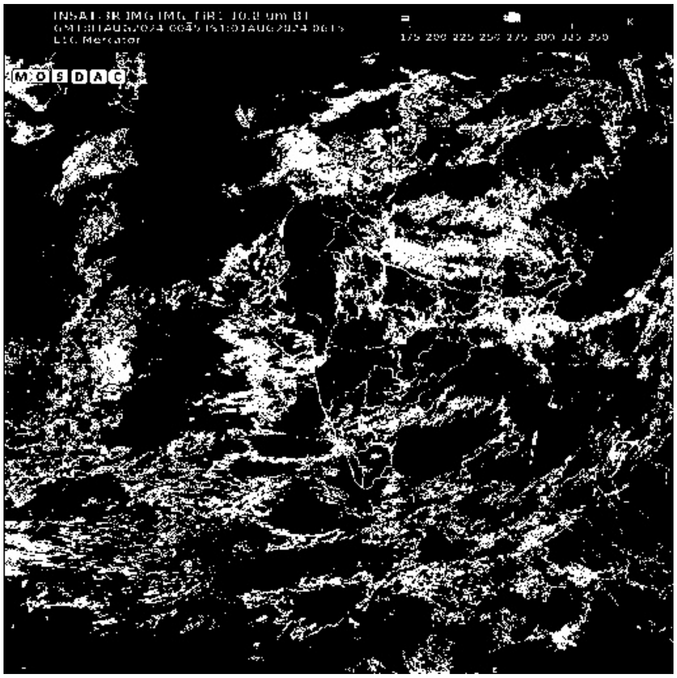

Fig. 6. Output : Masked Image

V. Results

To evaluate the effectiveness of various deep learning architectures for intermediate satellite frame generation, we implemented and compared five distinct models: a baseline Convolutional Autoencoder (CAE), the proposed multi-input attention-enhanced autoencoder with wind guidance, a GAN-based interpolation framework, a ConvLSTM-based spatiotemporal model, and a Variational Autoencoder (VAE).

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT MODELS FOR INTERMEDIATE SATELLITE FRAME GENERATION

Model	SSIM ↑	PSNR (dB) ↑	LPIPS ↓	Loss Function	Val Loss ↓	Params (M)	Training Time	Prediction Time
Baseline CAE	0.791	24.21	0.211	MSE	0.0192	1.2	0.5 hrs	∼30 sec
Proposed CAE + Attention + Wind	**0.864**	**26.74**	**0.128**	Perceptual (VGG19)	**0.0114**	4.5	1.0 hrs	∼1 min
GAN-based Frame Interpolation	0.851	25.93	0.140	Perceptual + Adv.	0.0158	12.3	2.5 hrs	∼2.5 min
ConvLSTM Autoencoder	0.839	25.48	0.149	MSE	0.0173	6.8	1.2 hrs	∼1.5 min
Variational Autoencoder (VAE)	0.803	24.75	0.175	KL + Recon.	0.0206	3.1	**17 hrs**	**3 min**

All models were trained and validated on the same custom-curated dataset comprising temporally spaced satellite image pairs and corresponding wind vector maps, following uniform preprocessing steps and input dimensions.

The **baseline CAE**, trained using mean squared error (MSE), performed adequately in reconstructing the coarse structure of intermediate frames. However, it struggled with retaining fine-grained cloud textures and smooth motion transitions, often producing slightly blurred and visually dull outputs.

In contrast, the **proposed attention-guided multi-branch convolutional autoencoder**, which incorporates explicit wind vector inputs and utilizes perceptual loss derived from VGG19 feature maps, exhibited significant improvements. This model preserved spatial sharpness and demonstrated superior capability in modeling coherent atmospheric transitions, particularly in regions of dynamic cloud movement.

The **GAN-based framework** produced visually sharper and more high-contrast outputs than the baseline, but occasionally suffered from minor spatial artifacts, particularly in low-texture or sparsely cloudy areas. These inconsistencies, although infrequent, detracted from the physical plausibility of some frames.

The **ConvLSTM model** effectively captured temporal continuity across sequential frames, showcasing a strong understanding of motion trajectories. Nevertheless, due to memory bottlenecks, it occasionally underperformed in preserving finer spatial textures, resulting in slightly smeared outputs.

The **Variational Autoencoder (VAE)** offered reconstructions that were globally coherent but often appeared washed out—consistent with the known effects of variational sampling. Despite this drawback, the VAE maintained meaningful spatial structures and generalized well across different cloud formations, albeit at the expense of longer training times.

Quantitative evaluations based on Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS) confirmed the superior performance of the proposed model. It achieved the highest SSIM and PSNR scores while maintaining the lowest LPIPS value, indicating excellent perceptual and structural fidelity.

An **ablation study** further highlighted the critical contribution of both the self-attention mechanism and wind vector inputs to overall performance. Removing either component led to noticeable degradation in output quality.

In terms of **computational efficiency**, the baseline CAE was the fastest to train, while the VAE incurred the longest training time due to its stochastic nature. The proposed model achieved a favorable balance between accuracy and efficiency, making it well-suited for practical deployment in real-time meteorological analysis and forecasting systems.

Overall, the attention-enhanced convolutional autoencoder with wind integration emerged as the most promising candidate. Its ability to blend visual fidelity with physically grounded environmental cues pushes the boundaries of satellite-based frame interpolation, offering a robust and scalable solution for high-resolution atmospheric monitoring.

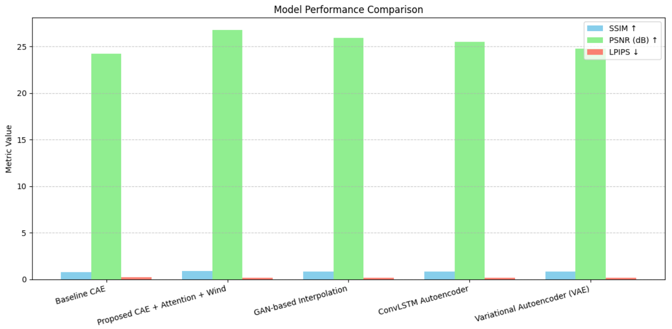

Fig. 7. Comparison between different models

REFERENCES

[1] T. J. Bellerby, ”A semi-automated technique for cloud-top advection tracking in geostationary satellite imagery,” *Int. J. Remote Sens.*, vol. 27, no. 6, pp. 1177–1192, 2006.

[2] M. Czerkawski, C. Pelletier, S. Lefèvre, and J. Inglada, ”Inpainting cloud-occluded satellite images with deep internal learning,” *ISPRS J. Photogramm. Remote Sens.*, vol. 189, pp. 81–93, 2022.

[3] M. Escrig, J. Vázquez, D. Salguero, and J. García, ”Cloud classification and motion estimation in Meteosat Second Generation satellite images,” *Atmos. Meas. Tech.*, vol. 6, no. 12, pp. 3201–3212, 2013.

[4] M. Francis, F. Khalid, and S. Rehman, ”CloudFCN: A deep learning-based approach for cloud detection,” *Remote Sens.*, vol. 11, no. 22, p. 2632, 2019.

[5] J. H. Jeppesen, M. Dyrmann, and R. N. Jørgensen, ”RS-Net: A convolutional neural network for cloud detection in remote sensing images,” *Remote Sens.*, vol. 11, no. 4, p. 467, 2019.

[6] X. Jin, Y. Xu, C. Xu, R. Ranftl, and Z. Liu, ”Towards unifying video frame interpolation and prediction,” *arXiv preprint arXiv:2401.08979*, 2025.

[7] E. Nielsen, T. Vandal, and R. Nemani, ”CloudCast: A benchmark dataset for cloud classification and forecasting using satellite imagery,” in *Proc. CVPRW*, 2021.

[8] H. Qin, Y. Kong, X. Guo, and H. Wang, ”Cold front detection using detection transformer and cloud satellite images,” *Remote Sens.*, vol. 16, no. 3, p. 420, 2024.

[9] T. Vandal and R. Nemani, ”Deep learning interpolation of satellite images for real-time monitoring,” in *Machine Learning for Earth and Space Sciences Workshop, NeurIPS*, 2021.

[10] Y. Wang, Z. Zhang, and Y. Lin, ”Recovering ground observations from cloud-contaminated satellite data using matrix completion,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2598–2610, 2016.

[11] Y. Yang, B. Zheng, Z. Zhao, and Y. Tian, ”CDnet: Cloud detection for Landsat-8 imagery using a convolutional neural network,” *Remote Sens.*, vol. 11, no. 5, p. 514, 2019.

[12] X. Shi et al., ”Convolutional LSTM network: A machine learning approach for precipitation nowcasting,” in *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[13] A. Dosovitskiy et al., ”FlowNet: Learning optical flow with convolutional networks,” in *Proc. IEEE ICCV*, pp. 2758–2766, 2015.

[14] C. Ledig et al., ”Photo-realistic single image super-resolution using a generative adversarial network,” in *Proc. IEEE CVPR*, pp. 4681–4690, 2017.

[15] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, ”Image-to-image translation with conditional adversarial networks,” in *Proc. IEEE CVPR*, pp. 1125–1134, 2017.