Detection and tracking of Medicanes through VideoMAE self-supervised vision transformer

Daniele D'Armiento[1] Stefano Sebastianelli[1] Leo Pio D'Adderio[1] Paolo Sanò[1] Daniele Casella[1] Giulia Panegrossi[1]

[1]Institute of Atmospheric Sciences and Climate, National Research Council of Italy, CNR-ISAC, Rome, Italy

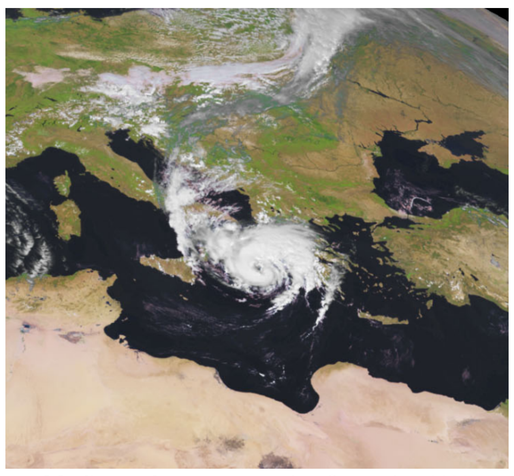
Problem & Motivation

We target the automatic detection and rotational center localization of Mediterranean Hurricanes (Medicanes) from MSG geostationary SEVIRI InfraRed image sequences. Despite the large raw data volume, Medicanes themselves are rare, this creates very limited labeled positives, affecting generalization, while manual annotation is costly and slow.

Our goal is to provide an effective algorithm for automatic detection (high POD, low FAR) and tracking (few km error) with minimal manual labels.

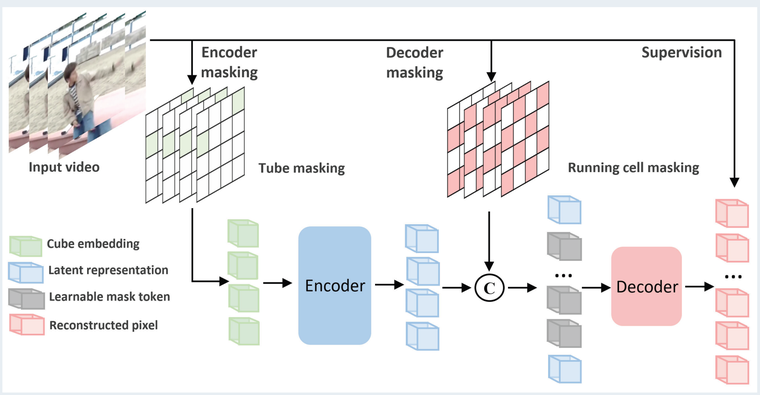
Approach & Contribution

We address label scarcity by specializing a pretrained Vision Transformer Masked AutoEncoder model for image sequences, **VideoMAE**, via self-supervised post-pretraining on unlabeled data, then fine-tune it for two tasks: (1) binary detection (cyclone vs non-cyclone), and (2) rotational center localization (regression of track coordinates). This pretrain-then-fine-tune pipeline enables label-efficient transfer from several unlabeled video to scarce labeled tasks yelding strong detection and tracking performance.



Medicane Ianos, occurred in September 2020 over the Ionian Sea. Source: EUMETSAT.

Model – VideoMAE[1] post-pretraining: specialization



VideoMAEv2 architecture encoder-decoder structure, with different masking for each one.

VideoMAE[1] is a self-supervised video foundation model based on masked autoencoding. It learns spatiotemporal representations by reconstructing masked patches from unmasked ones, enabling effective pretraining on large-scale unlabeled video datasets. The pretrained encoder can then be fine-tuned for downstream tasks with limited amount of labeled data.

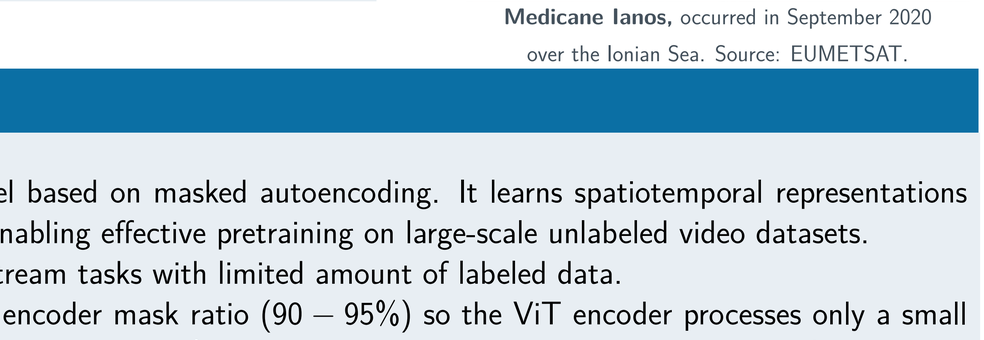
Architecture: It applies tube masking with a very high encoder mask ratio ($90 - 95\%$) so the ViT encoder processes only a small visible subset of patches; it further masks the decoder ("dual masking"), letting it reconstruct only a diverse subset of cubes, and computing the MSE loss only on patches invisible to the encoder. This shortens the decoder sequence, reducing compute and memory, enabling efficient scaling up to billion-parameter ViT under practical GPU budgets.

Our Post-Pretraining: In our setup we perform self-supervised training using pretrained ViT-g model, then reuse the trained specialized model as checkpoint for detection and center tracking regression trainings.

Dataset – AirmassRGB: Pre-processing and Post-processing

Collected data built from IR measurements by EUMETSAT SEVIRI MSG Rapid Scan Service (RSS) with 5 minutes frequency availability, and 3 km spatial resolution.

AirmassRGB composite enhances the visual identification of cyclone dynamics and cloud features, by integrating data from infrared channels, according to EUMETSAT guidelines[2].

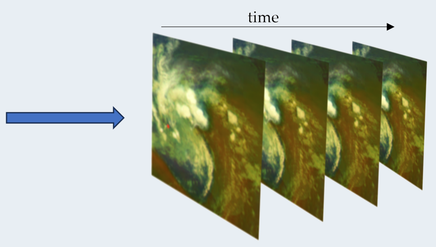


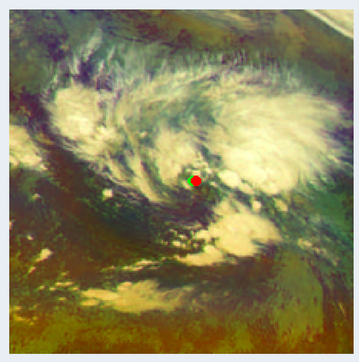
Pre-processing

Source Dataset: over **860K** frames, total time 7.5 years, spanning from 2010 to 2023, over the Mediterranean basin, resolution 1290×420 pixels.

Working Dataset: each complete image is split into partially overlapping tiles (224×224 pixels); then 16 frame tiles are stacked to build a single video tile (1h 20' time span).

Labeling: Using *Tracks_CL* database[3] to label positive tiles when they contain cyclone's center track.





Post-processing

Self-supervised specialization samples: **Train set** $\approx 80K$, **Test set** $\approx 20K$. No post processing.

Supervised classification samples: initially taken from *Tracks_CL7* database[3], encompassing heterogeneous extra-tropical cyclones, then reduced to a smaller number, high quality Medicanes with clear rotating cloud bands and cloud-free eye, after many rounds of thorough visual inspection.

Train set: 1238, **Balanced validation set:** 354, **Unbalanced validation set:** 2400

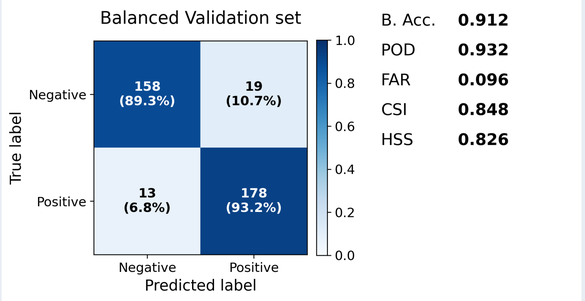
Detection (binary classification)

Head – Loss: Linear head – Binary Cross Entropy.

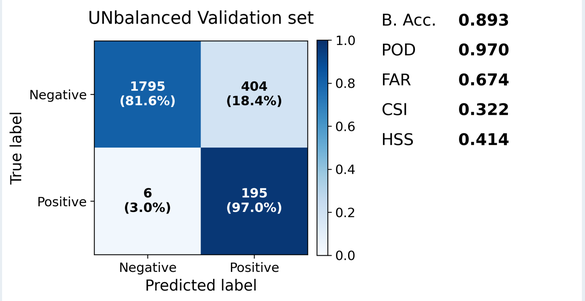
Training: Class-balanced sampling; validation sets: balanced and unbalanced.

Metrics: Balanced Accuracy – Probability of detection – False Alarms.

Best Results: 91% Balanced Accuracy



B. Acc.	**0.912**
POD	**0.932**
FAR	**0.096**
CSI	**0.848**
HSS	**0.826**



B. Acc.	**0.893**
POD	**0.970**
FAR	**0.674**
CSI	**0.322**
HSS	**0.414**

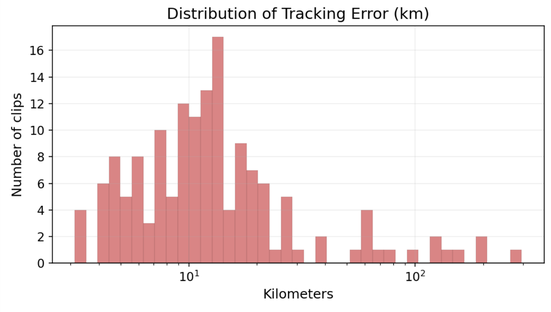
Tracking (rotational center regression)

Head – Loss: Linear Regression head – MSE.

Training: Dataset similar to that for detection, with only one validation set.

Metrics: Absolute positional error in km, $\|\hat{\mathbf{c}} - \mathbf{c}\|_2$.

Best Results: median error = **12.06 km**



Future improvements

- Detection has very few misses but high false alarm rate. Introduce a third intermediate class for tiles neighboring to the one containing the cyclone center.
- Tracking doesn't need particular improvement, except to enlarge the dataset.
- Study neural network's embedding space – reduce space dimension and cluster data for unsupervised classes discovery.

References

[1] Wang, L., et al. (2023). *VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking*. CVPR.

[2] EUMETSAT (2020). *RGB recipes for air-mass imagery*. Available at `https://eumetrain.org/sites/default/files/2020-05/RGB_recipes.pdf`.

[3] Flaounas, E., et al. (2023). *A composite approach to produce reference datasets for extratropical cyclone tracks: application to Mediterranean cyclones*. Weather and Climate Dynamics, 4(3), 639–661. https://doi.org/10.5194/wcd-4-639-2023.

Acknowledgement

The present work is supported by the European Space Agency's (ESA) Medicanes project. `https://medicanes.isac.cnr.it/` — ESA Contract No. 4000144111/23/I-KE.