Ref. Ares(2026)8211958 - 27/08/2026

Better Contrails Mitigation - BeCoM

European Union's Horizon Europe Research and innovation program Under Grant Agreement No 101056885

D3.3

AI algorithms have been trained, validated, and are ready for use

Delivery Date:	31/05/2026
Date of submission:	

LEAD BENEFICIARY FOR THIS DELIVERABLE			
Name:	Thales		
Contact Person:	Teodora Petrisor		
E-mail:	teodora.petrisor@thalesgroup.com		
AUTHORS			
Participants:	Thales, TU-Delft		
Work Package:	WP3		
Version:	1	Number of Pages:	

REVISION HISTORY			
VERSION	DATE	AUTHOR / REVIEWER	NOTES
V0	07-12-2025	Teodora Petrisor, Davide di Giusto, Vincent Meijer	Published as an extended abstract to the NeurReps workshop of the NEURIPS 2025 conference
V1	13-08-2026	Teodora Petrisor	Final edited version based on the extended abstract

REVIEWED AND SIGNED OFF BY			
ROLE	DATE	NAME	SIGNATURE
ENVISA participant, internal reviewer	17-08-2026	Denis Bilyarski	Signed-off by email
DWD WP2 Lead, internal reviewer	25-08-2025	Alexander Cress	Signed-off by email

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

Abstract

Contrail presence verification from observations is key to validate model predictions and is one of the main goals of WP3 in the BeCoM project. There are two main requirements for this operation: reliable annotated observations (M3.1) and robust detection algorithms (D3.2 and M3.2) limiting the additional uncertainties in the overall contrail mitigation pipeline. There are several sources of contrail observation, as achieved in M3.1, in particular ground imagers (for finer contrail localisation over a limited area) or geostationary satellites (for a more global overview of contrail coverage), each with their advantages and limitations. We have focused on geostationary observations as these are more aligned with the BeCoM objectives of evaluating the global contrail impact on climate rather than a specific or regional contrail to flight attribution. Two key aspects are highlighted in this work: the need for annotated data must be kept as low as possible, since quality annotated datasets are very costly to produce, and the contrail detection model needs to display consistent accuracy to possible variations in the inputs, regardless of the amount of training data used. To implement these considerations, we have devised a hybrid, robust-by-design AI algorithm for contrail segmentation (as presented in D3.2) that is stable to translations or rotations in the input images. This deliverable presents the obtained results, which were published as a poster and extended abstract in the NeurIPS workshop on Symmetry and Geometry in Neural Representations (NeurReps), 2025[1].

We investigated different axes of robustness such as the reliance on training data with respect to its size, representativeness or quality, as well as mathematical guarantees of covered input perturbations and their evaluation criteria. Working on these issues should ensure that the inherent uncertainties related to statistical learning methods are contained within acceptable bounds.

We considered the following frugality axes:

- **data frugality**: by studying the impact of data augmentation on training performance and proposing alternative robust-by-design methods,
- **model frugality**: by studying model size reduction in terms of number of trainable parameters and the gains of robust-by-design methods here as well,
- **transfer learning, fine-tuning**: by studying model reuse in particular in satellite imagery for contrail verification

The main findings of our algorithms show that stability of the detection performance can be achieved even with less than 5% of annotated data used for training, while using models with approximately 36× less trainable parameters than classical AI models. Moreover, we show here that in a low training data scenario the fine-tuning operation with pretrained models may not bring very significant gains or need larger AI models.

[1] ECoNets: Rotation Equivariant Contrail Detection Neural Networks in Satellite Imagery | OpenReview

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

List of abbreviations

AI	Artificial Intelligence
ML	Machine Learning
GDL	Geometric Deep Learning
CNN	Convolutional Neural Network
G-CNN	Group-Convolutional Neural Networks
GOES-16	Geostationary Operational Environment Satellite (NOAA[2] and NASA)
GOES-16-ABI	GOES-16 Advanced Baseline Imager
MSG	Meteosat Second Generation geostationary satellite (EUMETSAT[3])
MSG-SEVIRI	MSG-Spinning Enhanced Visible and Infra-Red Imager
U-Net	A type of convolutional neural network for image segmentation
ReLU	Rectified Linear Unit (a type of activation function in artificial neural networks)
ECoNets	Equivariant Contrail detection Neural Networks
WP	Work Package

[2] National Oceanic and Atmospheric Administration, USA
[3] European Organisation for the Exploitation of Meteorological Satellites

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

List of figures

List of tables

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

Table of contents

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

1 WP3 Context – Contrail verification methods

Aviation accounts for a few percent of the anthropogenic climate forcing, with more than half of its impact coming from non-CO_2 emissions, primarily condensation trails (contrails) ([11]Lee et al., 2021). Contrails are artificial cirrus clouds forming from the mixing of warm and moist aircraft exhaust with cold and dry ambient air ([19]Schumann, 1996). In specific ambient conditions, contrails can persist for several hours, causing a net warming effect on the atmosphere ([10]Kärcher, 2018). Aviation stakeholders rely on numerical weather forecasting to apply flight route optimizations and minimize contrails impact. When contrails have actually formed and persisted, they need to be detected either for contrail-to-flight attribution or to verify and improve prediction models. Given their spatial and temporal coverage, geostationary satellite imagers are well suited for persistent contrails detection. This can be viewed as a binary semantic segmentation task, akin to medical imagery due to its challenging class-unbalanced nature: young contrails are very thin linear shapes, scarcely appearing, while most of the pixels in the image belong to the background (non-contrail) class. Automated detection with satisfactory performance relies on large annotated datasets. Several efforts in contrail detection automation were done on the US geostationary GOES-16 satellite, e.g. [15]Meijer et al. (2022) who deployed U-Net architectures ([17]Ronneberger et al., 2015) for contrail segmentation, allowing to move from thresholding approaches ([13]Mannstein et al., 1999) to higher performing algorithms. The largest labelled dataset to date is OpenContrails ([16]Ng et al., 2024). This dataset consists of over 20000, 256 × 256-pixel images[4], taken from full-disk images over America. A similar training dataset over Europe would be critical for contrails detection in this very dense air-traffic region, and could rely on the Meteosat Second Generation (MSG) geostationary satellite ([1]Aminou, 2002). This lack of annotated MSG images could be compensated by training models which learn equivariant representations from fewer samples. Group Convolutional Neural Networks were introduced to exploit discrete rotation symmetries in classification ([5]Cohen and Welling, 2016), and later extended to equivariant steerable filters to implement continuous affine groups ([6]Cohen and Welling, 2017; [21]Weiler and Cesa, 2019), finding applications in image segmentation ([23]Winkens et al., 2018; [4]Chidester et al., 2019; [2]Bernander et al., 2022; [24]Zhang et al., 2025). More recently, ([8]Ghyselinck et al., 2025) benchmarked equivariant U-Nets on several datasets and found that equivariance is beneficial for arbitrarily-shaped objects with different orientations, similar to contrails in satellite images.

2 Advanced AI algorithms for contrail presence verification through imagery

Contrail predictions can be verified based on previously collected data from different sources of imagery, henceforth referred to as observations. Whole-sky imagers, for instance, can capture in high definition very "young" contrails in high definition (e.g. with temporal resolution of 30 seconds to 10 minutes) both in the visible range and in infrared depending on the different available technologies. The geographical coverage of a single camera is however very limited, usually less than 100 km and the dependence on meteorological phenomena such as snow, ice, rain, or sun flares is quite high, requiring important calibration and other types of processing. On the other hand, using images from geostationary satellites have rapidly gained traction in the AI community both due to their very wide geographic coverage and thanks to the Open Contrails labelled dataset, containing around 20,000 annotated GOES-16 images, covering the Americas. The key idea in algorithm development is to limit the amount of labelled data by using pretrained neural networks and transfer learning, as well as novel AI algorithms, such as Geometric Deep Learning[5]. These enforce knowledge about the symmetries

[4] In these images around 1% of overall pixels are contrails with an average of only about 0.5% pixels per image.

[5] Also known as Geometry Informed Neural Networks (GINNs)

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

present in an image directly in the model structure, allowing for a drastic reduction of trainable parameters, while being more data-frugal.

2.1 Image Segmentation: Contour localisation of all contrail pixels in an image

Contrail verification in images with the purpose of global contrail impact evaluation amounts to an image processing task called semantic segmentation, which is a form of pixel-based classification task – each pixel having attached a class label. In our case, images are annotated with a contrail/non-contrail label per pixel. At the time the work in WP was performed, in the contrail verification literature, the large majority of observation sources for contrails were coming from remote sensors such as polar-orbiting or geostationary satellites. Both technologies exhibit trade-offs between temporal and spatial resolutions, while providing a much more global view than ground imagers. The inconvenience is that, due to the spatial resolution of these sensors, even wide contrails occupy only very tiny surfaces in the image, with unbalanced dimensions with respect to background pixels (e.g. as little as 2×40 pixels in a geostationary satellite 256×256 image). Contrails are usually visible in the same infrared spectral bands as cirrus clouds and their concentration in ice particles allows them to be highlighted in brightness temperature differences (BTD) images and colour schemes combining all the relevant BTD channels[6] can equally be employed as a pre-filtering allowing to enhance contrail observations and ease the labelling process. The most effective methods for contrail detection at the moment are obtained with AI via artificial neural networks.

The U-Net architecture

The classical neural architecture used for this particular machine learning task is called the U-Net neural network[7] and is a model particularly suitable for medical image segmentation capturing irregular shapes in an image. It contains an "encoder" branch which extracts the relevant features (contracting path) from the input followed by a "decoder" branch aiming at precisely locating the pixels of each class.

An illustration of this architecture is given in Figure 2-1.

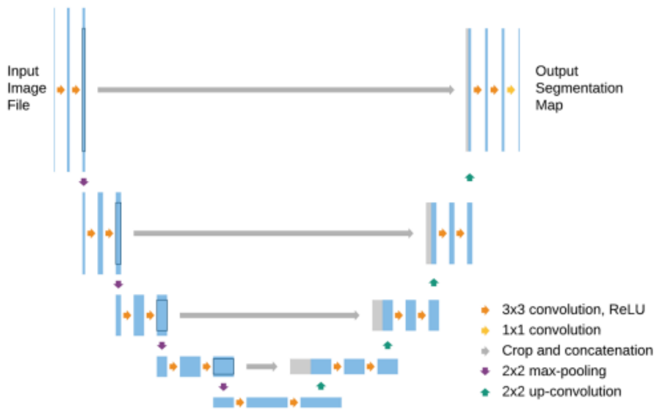

FIGURE 2. Basic U-net architecture. The arrows represent the different operations, the blue boxes represent the feature map at each layer, and the gray boxes represent the cropped feature maps from the contracting path.

Figure 2-1 Illustration of the U-net architecture, taken from [20].

[6] e.g. the channels 12.0, 10.8 and 8.7 micrometers for the European MSG/SEVIRI satellite

[7] due to its structure resembling an "U" letter

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

Equivariant architectures: ECoNets (Equivariant Contrail detection Networks)

The basic (Vanilla) U-Net architecture does not consider any physical/geometrical properties other than locality principles in the input image exploited by the regular convolutions. This leads to the neural network having to learn from scratch important information such as orientation/localisation invariances[8]. Contrail pixels form lines in the image, spanning all possible locations and orientations. This *a priori* information (mathematically expressed as symmetries in the inputs) can be exploited to develop more robust AI methods, having a stable output regardless of structural transformations (local or global) in the inputs: (rotations, reflections, translations). A unifying theory expressing symmetries in data and systems uses group theory and has been introduced in the past decade in the AI community. The main principle is constraining neural networks models to satisfy *by-design* these desirable properties of the inputs by generalizing the convolution operation to group convolutions. This leads to more trustworthy AI and also to more data frugal training models. In particular, it is a very good alternative to data augmentation techniques which are the classical way to increase training data expressivity by artificially introducing modified data during training (e.g. rotated, noisy versions, etc.). We aim at proving that stable accuracy contrail verification results can be obtained even in a very low data setting in this WP.

Given the orientation-invariant nature of contrails in satellite images, this work introduces Equivariant Contrail segmentation Networks (ECoNets), exploiting rotation and reflection symmetries based on the U-Net architecture. Our objective is to benchmark ECoNets against standard (also called Vanilla) U-Nets when varying the training dataset size.

The two main properties ensured by this modelling are invariance (the output doesn't change regardless of the transformation in the input) and equivariance (the output reflects the transform in the input exactly). In our case, a contrail appearing in the input at any given orientation/location should be mapped in the model exactly with the same orientation/location. We now present the mathematical framework employed in our algorithms into more detail.

Formally, an operator $\Phi : X \rightarrow Y$ is said to be G-*equivariant* for a given group G if there exist two *actions* π and π' on X and Y, respectively, such that: $\forall g \in G, \forall x \in X, \Phi\left(\pi_g(x)\right) = \pi'_g\left(\Phi(x)\right)$. If $\forall g \in G, \forall x \in X, \Phi(\pi_g(x)) = \Phi(x)$ the operator is *invariant*.

Rotational and translational equivariance can for instance be achieved by choosing the group $G = \mathbb{Z}^2 \times C_8$, where C_8 represents the group of rotations by angles that are multiples of $\pi/4$ radians (45°), which is a discrete and finite subgroup of the special orthogonal group, $SO(2)$. The signal operating on this chosen group is different from the initial one, therefore we stack rotated versions in a so-called lifted space before applying the group convolution. The latent representation of the inputs (the features) remain in this lifted space until projected back to the original space before the final output.

More specifically, let f be an image $f: \mathbb{Z}^2 \rightarrow \mathbb{R}^c$ that assigns a c-dimensional vector to each point $x \in \mathbb{Z}^2$. Given a filter $k: \mathbb{Z}^2 \rightarrow \mathbb{R}^c$, the cross-correlation operation $*$, usually implementing the convolution in the machine literature, is defined as $[k * f](x) = \sum_{i=1}^{c} \sum_{y \in \mathbb{Z}^2} k_i(y - x) f_i(y)$.

For simplicity we set $c = 1$. The regular (vanilla) convolution operation is equivariant with respect to the translation group with the additive operation $(\mathbb{Z}^2, +)$. Group Convolutional Neural Networks (G-CNN) introduce an additional group transformation. We focus on the discrete rotation and dihedral groups C_N and D_N of N rotations and respectively N rotations plus reflections on the plane.

[8] An object of interest should always be accurately recognized regardless of its position or orientation in the image.

This project has received funding from the European Union's Horizon Europe
Research and innovation program under grant agreement No 101056885

We work with D_N since it generalizes C_N. Mathematically, the dihedral-translation group is the semi-direct product $G = (\mathbb{Z}^2, +) \rtimes D_N$. In a G-CNN, the regular convolution is changed by a group convolution which, in the case of the roto-dihedral group, becomes:

$$[k *_G f](x, s, \theta) = \sum_{y \in \mathbb{Z}^2} \sum_{\tilde{\theta} \in \{\frac{2j\pi}{N}\}_{j=0}^{N}} \sum_{\tilde{s} \in \{-1,1\}} k(R_{\theta,s}^{-1}(y - x), \tilde{\theta} - \theta \bmod 2\pi, s\tilde{s}) f(y, \tilde{s}, \tilde{\theta})$$

Where $R_{\theta,s}^{-1}$ is a 2×2 rotation matrix with $\theta \in \{\frac{2j\pi}{N}\}_{j=0}^{N-1}$ multiplied by the reflection matrix $Diag(1, \pm s)$ and $k(\cdot, \theta, s)$ is the kernel rotated by θ and reflected by s. If this equation is applied to a scalar image $f(y)$ in the first layer of a network, the sums over $\tilde{s}, \tilde{\theta}$ disappear and f is *lifted* to a higher dimensional space indexed by the discrete angles and reflections. In practice, an equivariant network can have a similar number of trainable parameters to those of a vanilla model ([4]Chidester et al., 2019; [8]Ghyselinck et al., 2025; [7]Gerken et al., 2022) or have the same architecture with less trainable parameters ([14]Maurel et al., 2023).

In this work, we fix the architecture and plug-in the corresponding equivariant convolution layers. The group convolution means that $2N$ copies of the kernel are created (N rotations and their corresponding reflections). Thus, trainable parameters in a layer are reduced by N for C_N as in Figure 2-2. Including reflections, the D_N case, further reduces by $2N$ the trainable parameters.

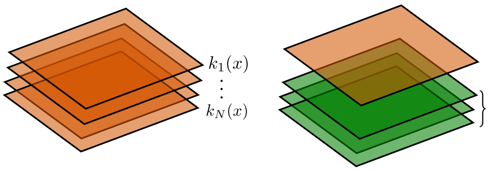

Figure 2-2 : Orange kernels are trainable while the green ones are transformed copies. Left: N trainable vanilla kernels. Right: In an equivariant C_N setting, a single trainable kernel is required, greatly reducing the number of trainable parameters

Our implementation consists of 4 encoder-decoder blocks with two 5×5 convolutional layers, followed by typical machine learning operations such as batch normalization (BN), ReLU activation function and MaxPooling. The output channel dimensions at the end of each block in the encoder path are of $64, 128, 256, 512$ and 1024, respectively. A final 1×1 convolution mixes the final 256 channels into a single segmentation map that should ideally match the annotated ground truth as accurately as possible.

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

Figure 2-1 shows into more detail our U-Net architecture.

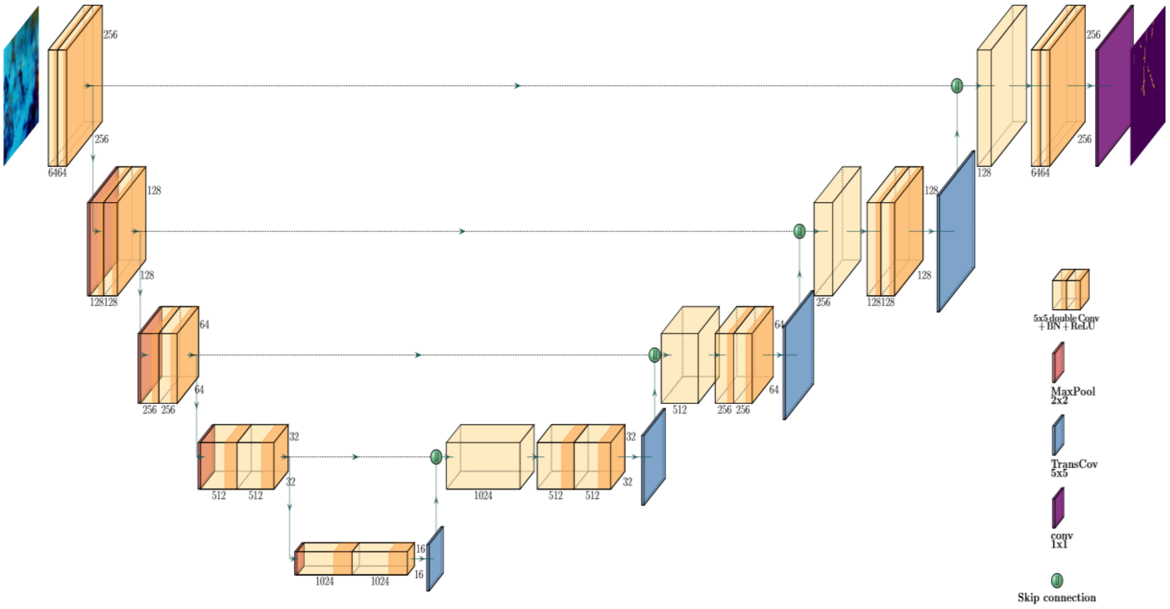

Figure 2-3 : Our U-Net implementation highlighting the main operations: filtering (conv) for feature extraction, pooling (for dimensionality reduction and information compression), residual (skip) connections for precise localisation in the reconstructed (decoded) feature maps, final transform into a binary segmentation map (conv 1x1)

All our networks, including the vanilla U-Net, use a 5×5 kernel as they can be rotated in the C_8 and D_8 setting without creating numerical artifacts.

In addition, we have used the *escnn*[9] library which works in the steerable kernel paradigm: kernels are the sum of a band-limited basis. Particularly, our steerable kernels are the superposition of 11 basis filters which further reduces the trainable parameters from 25 to 11 per kernel.

Our last 1×1 convolution merges the different oriented feature maps into a single one. Instead of adding a final group pooling operation as in [5]Cohen and Welling (2016), all the orientation responses of the feature maps are taken into account. This allows a more expressive G-CNN ([23]Winkens et al., 2018).

2.2 Datasets description

After analysing potential data sources, collected and prepared them in T3.1 we have established a final setting for contrail verification in this WP based on geostationary satellite imagery, as the most advanced data source at the time this work has been performed. We have thus collected and used two datasets for developing our robust-by-design AI algorithms, which we describe here below.

OpenContrails dataset ([16]Ng et al., 2024).

The OpenContrails dataset became available during the BeCoM project and rapidly set a benchmarking standard in the scientific community. Our goal in this workpackage was to study the detection performance as a function of the available training data and with frugal and robust-by-design AI models. OpenContrails is based on the NOAA/NASA GOES-16-ABI (Advanced Baseline Imager) which has a temporal resolution of 5 to 10 minutes and a spatial resolution of 2×2 km^2 at Nadir for the infrared bands. 256×256 patches were obtained from larger images covering the

[9] escnn: Equivariant Steerable CNNs Library for Pytorch

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

North and South America regions. Data was obtained in the period April 2019 - April 2020. The dataset consists of 20,544 images for training and 1,827 for test. The training dataset has 1.2% of foreground pixels and around 55% of the images contain no contrails, while the test dataset has 0.5% of contrail pixels and 70% of contrail-free images. Two examples from the test dataset are given below in Figure 2-4 along with their ground-truth annotations.

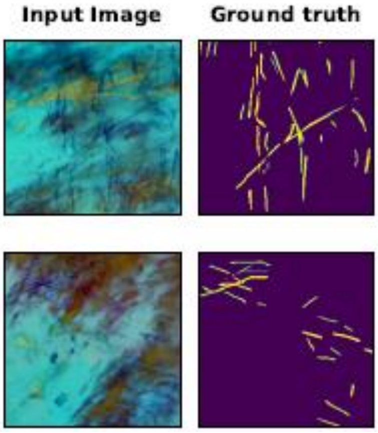

Figure 2-4 : Example of test images and their annotations in the OpenContrails dataset as binary masks with contrails represented in yellow

In Figure 2-5 we show the foreground pixels and contrail-free images proportions as functions of the dataset fraction when choosing an approximately stratified[10] dataset fractioning for the data frugality experiments done in this work.

Characteristic	Dataset fraction %				
	10%	25%	50%	75%	100%
% Contrail-free images	55.83	54.17	54.37	54.50	54.81
% Pixel contrails	1.16	1.19	1.18	1.20	1.19

Figure 2-5 : OpenContrails dataset characteristics when considering only a fraction of the data for training in an approximately stratified strategy.

This guarantees a fair comparison of models trained in a training budget setup.

The BeCoM Meteosat MSG dataset

In WP3 we have prepared a satellite imagery dataset over Europe using the data from the Meteosat Second Generation SEVIRI sensor. The MSG-SEVIRI imager has a temporal resolution of 5 − 15 minutes and a spatial resolution of 3×3 km^2, as well as slightly different infrared bands for contrail observation. This is a less favourable sensor for contrail observation due to its coarser resolution, but the newer generation recently deployed aims to palliate this limitation. We have annotated a small dataset of 256 × 256-pixel images taken by the MSG satellite between Jan 2023 and Jan 2024 containing 293 images for training and 78 for test with a balanced average contrail pixel ratio per image of around 0.3% in both sets.

We have explored two labelling strategies to build this initial dataset, following similar guidelines to OpenContrails in terms of contrail-criteria annotation (minimum size, temporal sequence, etc.)

[10] Stratified data splitting means that the statistics of the data are preserved regardless of the data quantity taken.

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

but aimed for a balanced per month temporal coverage in 2023 and contrail-image distribution, while keeping the annotation effort low. We give in Figure 2-6 two examples of these images with their respective annotations and the overall statistics of the dataset in Figure 2-7.

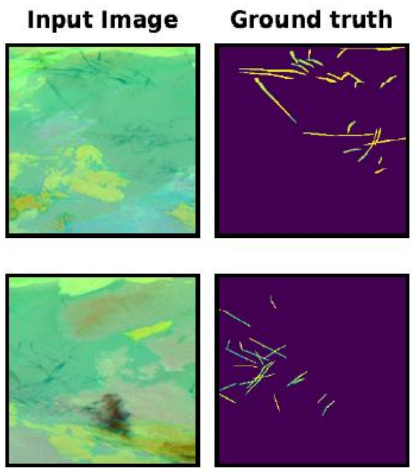

Figure 2-6 : Examples of MSG images from the test set along with the binary annotation map.

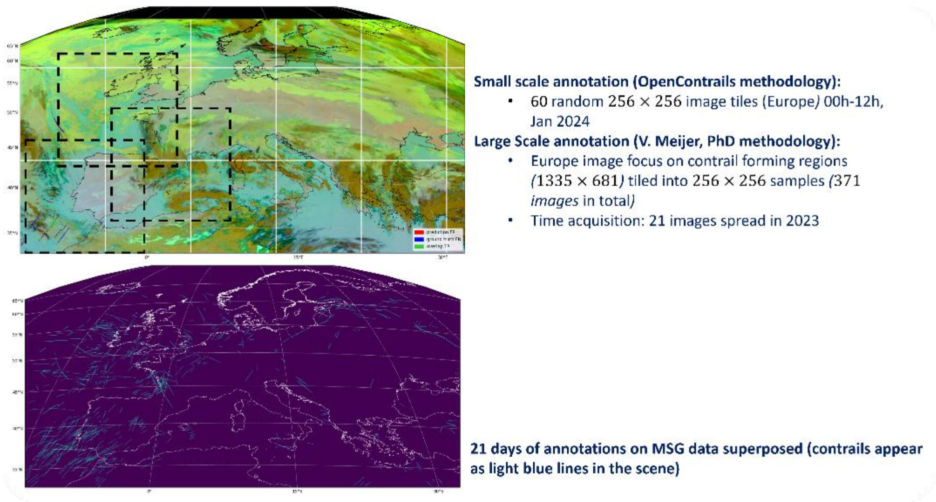

Figure 2-7 : Details on the BeCoM Meteosat MSG dataset annotation strategies (upper part) and a cumulated view of all the annotated contrails in the large images (lower part).

The purpose of this dataset was both to enhance contrail detection capabilities over Europe and to test AI algorithms within a limited training data setting.

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

3 Validation experiments

3.1 Methods

We trained vanilla U-Nets and equivariant U-Nets (ECoNets) to discrete rotations of multiples of $\frac{\pi}{2}$, respectively $\frac{\pi}{4}$ and reflections, similar to ([8]Ghyselinck et al., 2025)[11]. Models consist of 4 encoder-decoder blocks, with two 5 × 5 convolutional layers, batch normalization and ReLU per block. The output channel dimension starts at 64 and is doubled (halved) after each encoder (decoder) block. A final 1 × 1 convolution generates the final segmentation map. Equivariant layers use the regular group representation and are implemented through the escnn library ([3]Cesa et al., 2022). Similarly to [14]Maurel et al. (2023), our models share the same architecture and channels, but have a greatly reduced number of trainable parameters ([4]Chidester et al., 2019; [7]Gerken et al., 2022; [8] Ghyselinck et al., 2025)[12]. All models are trained on one A100 80G Nvidia GPU using the AdamW optimizer with convergence based on an early stopping criterion. A combo loss function is used for training ([9]Jadon, 2020), and the Global Dice score for evaluation. The initial learning rates, yielding the best training performance for each model, were set to 10^{-4} and 10^{-3} for the vanilla U-Net and the ECoNets, respectively, using a "reduce on plateau" learning rate scheduler. To assess the potential of equivariance over data-augmented vanilla U-Nets, we explore three data augmentation scenarios: no augmentation (denoted "None"), continuous rotations (denoted "Rot(0,360): Aug. 1") and continuous rotations and reflections (denoted "Flip+Rot(0,360): Aug. 2"), and we vary the percentage of original data used for training.

We have defined two benchmarks:

1. Benchmark *One*

Equivariance against data augmentation as in [7]Gerken et al. (2022) where Vanilla-Rot(0,360), (-Flip+Rot(0,360)) augmented models are compared to C_N-ECoNets (D_N-) without data augmentation.

2. Benchmark *Two*

Vanilla networks and ECoNets with the same scenario of data augmentation are compared as in [8]Ghyselinck et al. (2025)[13]. With respect to the training time, we compare the models in terms of convergence epoch, as well as wall-clock time at equivalent performance.

3.2 Training setup

Loss function. The sum of Dice loss and weighted Binary Cross Entropy (BCE) is known as Combo Loss [9]Jadon (2020): $L_{Combo} = \alpha L_{BCE} + (1-\alpha)L_{Dice}$,

where the BCE loss is $L_{BCE} = -\sum_{\{i=1\}}^{N} \beta y_i \log \hat{y}_i + (1-\beta)(1-y_i)\log(1-\hat{y}_i)$

[11] Ghyselinck et al. (2025) models are equivariant to the discrete rotation groups C_4 and C_8 and the dihedral group D_4. We add the D_8 group and denote discrete rotation groups by C_N and dihedral groups by D_N.

[12] Vanilla U-Net has 95.85M of trainable parameters while the C_4- C_8-, D_4- and D_8-ECoNets have 10.54M, 27M, 5.27M and 2.63M trainable parameters, respectively.

[13] The Flip+Rot(0,360) comparison includes D_N-Rot(0, 360) models since their reflection symmetry counts for the flip data augmentation. This is why the Rot(0,360) comparison does not include D_N models.

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

and the Dice loss is $L_{Dice} = 1 - \frac{\sum_{i=1}^{N} y_i \hat{y}_i + \epsilon}{\sum_{i=1}^{N} (y_i + \hat{y}_i) + \epsilon}$

where y_i is the binary annotation at the i-th pixel of the batch, $\hat{y}_i$ the CNN prediction for the same pixel, α is a weight to balance both Dice and BCE loss contributions and β regulates the weight of positive and negative pixels in the BCE loss. The small $\epsilon = 10^{-8}$ is added to the Dice loss to avoid numerical issues. The sum runs over the N pixels of the batch. We set $\alpha = \frac{1}{2}$ and $\beta = \frac{1}{2}$ and for OpenContrails. For the MSG experiments we used $\alpha = 0$ and $\beta = \frac{1}{2}$.

Hyperparameters. To determine the best hyperparameters, an extensive grid search was performed with the vanilla U-Net trained on the whole OpenContrails dataset with Flip + Rot(0, 360) data augmentation. This particular architecture and data augmentation was chosen to find the best possible Dice score for the vanilla U-Net. The learning rate, batch size, weight decay and early stopping minimum improvement in the validation loss, denoted δ_{min}, were varied across several values shown in Figure 3-1.

Hypeparameter	Grid values
Initial learning rate	$10^{-3}, 10^{-4}, 10^{-5}$
Batch size	$8, 16, 32$
Weight decay	$10^{-3}, 10^{-4}, 10^{-5}$
Early Stopping δ_{min} (%)	$5, 2, 1, 0.5, 0.1, 0.01$

Figure 3-1 : Model hyperparameters exploration options used to determine the best Vanilla model for the benchmark

This resulted in 27 experiments. Each one was performed during 100 training epochs and we simulated stopping the training for the different δ_{min}. Therefore, for each δ_{min} we obtain a stopping epoch and calculate a Global Dice Score. The maximum number of epochs was set to 100 during the hyperparameter search. Some other elements were kept fixed to reduce the search space. All experiments used the AdamW optimizer. The Early Stopping patience parameter was set to 10 epochs. Concerning the learning rate evolution, during the first epoch we used a warm-up phase where the learning rate is increased linearly from 10^{-5} to the given initial value of Figure 3-1 and then it is kept constant with the same initial value of Figure 3-1 for another epoch. The rest of the epochs are trained with a “reduce on plateau” scheduler strategy as follows: the scheduler takes place with a patience of 3 epochs, a reduction factor of 0.5 and a minimum improvement of 1% on the validation loss. If the loss does not improve by 1% after 3 epochs the learning rate is reduced by the corresponding factor. After this grid search, for the vanilla U-Net, the learning rate was set to 10^{-4}, the batch size to 16 and the weight decay to 10^{-3}. The chosen early stopping δ_{min} is of 1%. The obtained binarization threshold for the segmentation mask was 0.1 since this value maximizes the validation Global Dice score in most hyperparameter configurations. However, we noticed that the found hyperparameters for the vanilla network were suboptimal for all equivariant models. Contrary to other works ([8]Ghyselinck et al., 2025; [7]Gerken et al., 2022) where the hyperparameters are set equally for equivariant and non-equivariant models, we trained equivariant models with a different initial learning rate on the OpenContrails dataset, while leaving other hyperparameters the same for the equivariant and non-equivariant cases.

Figure 3-2 clearly shows that the vanilla U-Net has better results with an initial learning rate of 10^{-4} regardless of the data augmentation used. Similarly, equivariant models perform better with an initial learning rate of 10^{-3} regardless of the group symmetry.

Therefore we compare all models in their best performing configuration.

This project has received funding from the European Union’s Horizon Europe
Research and innovation program under grant agreement No 101056885

Model	Data Augmentation	Initial learning rate	Global Dice Score %	Stopping Epoch
Vanilla	Aug. 2	10^{-4}	62.48	43
		10^{-3}	61.46	45
	Aug. 1	10^{-4}	61.45	49
		10^{-3}	58.20	41
C_4	None	10^{-4}	58.12	18
		10^{-3}	61.70	14
C_8	None	10^{-4}	57.55	16
		10^{-3}	62.53	13
D_4	None	10^{-4}	57.75	16
		10^{-3}	63.25	14
D_8	None	10^{-4}	56.55	19
		10^{-3}	63.77	14

Figure 3-2 : U-Net models trained with different initial learning rates

To avoid bias in the final comparison, we train equivariant models with an initial learning rate of 10^{-3}, while leaving the other hyperparameters unchanged from the vanilla model. Only changes in the learning rate were studied to keep the hyperparameters values as similar as possible between the equivariant and non-equivariant models.

When pretraining on OpenContrails and fine-tunning the models on the MSG dataset, a similar hyperparameter search took place. We used the Combo loss with a value of $\alpha = 0$, leaving us only with the Dice loss. We keep the same hyperparameters fixed as those of the OpenContrails training setup while only varying the initial learning rate, batch size and weight decay as before. The values explored during the hyperparameter search are those from Figure 3-2, except for the batch size grid values which are set to 4, 8 and 16 due to the much smaller MSG dataset size. During the search, the U-Net weights trained with Flip + Rot(0, 360) on OpenContrails served as departing point for the MSG fine-tuning hyperparameter search. We found that for a vanilla U-Net fine-tuned on MSG with Flip+Rot(0,360) augmentation the best initial learning rate is 10^{-3}, the batch size is 8 and the weight decay is 10^{-3}. Similar to OpenContrails, the initial learning rate was set to 10^{-3} for the equivariant models while leaving other hyperparameters as for the U-Net Flip + Rot(0, 360). The early stopping and reduce on plateau parameters used were those used in training on OpenContrails. Models trained from scratch also kept the same optimal hyperparameters as those of the fine-tuned U-Net.

3.3 Evaluation setup

Evaluation setup. For all experiments we have chosen the *Global Dice Score* as the evaluation metric aggregating and averaging over all annotations in the test/train datasets. The Dice Score (D) is given by:

$$D = \frac{1}{P^{-1} + R^{-1}} = \frac{2|Y \cap \hat{Y}| + \epsilon}{|Y| + |\hat{Y}| + \epsilon} = \frac{2\mathrm{TP} + \epsilon}{2\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \epsilon}$$

Where P and R are the standard (pixel-wise) precision and recall metrics in an image, $|\cdot|$ denotes the pixels in a given area with Y standing for the ground truth pixels and $\hat{Y}$ standing for the predicted contrail pixel by the neural model. ϵ is again a numerical stability factor and TP, FN, FP stand for true positives (correctly predicted pixels), false negatives (missed predictions), respectively false positives (false contrail pixels) given by the model all computed by comparing Y and $\hat{Y}$.

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

The problem with the expression above is when the ground truth is empty, $|Y| = 0$. If the prediction is perfect and $|\hat{Y}| = 0$, then $D = 1$. Otherwise, $D \approx 0$ since $\epsilon = 10^{-8}$ in our implementation. Thus, individual empty ground-truth images are extreme cases.

Let $X_{dataset} = \bigcup_{i=0}^{M} X_i$ be the aggregation all images X_i in the dataset. OpenContrails and MSG images have spatial dimensions of 256×256, meaning that the assembled image $X_{dataset}$ has dimensions $256 \times 256 \times m$ where m is the dataset size. We substitute to Y and respectively $\hat{Y}$ the ensemble of ground truth masks $Y_{dataset}$ and the ensemble of predictions $\hat{Y}_{dataset}$ and thus obtain the Global Dice Score. Unless the dataset consists of only empty images, the Global Dice score is never undefined. In this work we refer sometimes to the Global Dice score simply as Dice score.

4 Results

4.2 OpenContrails dataset

Table 4-1 summarizes our ECoNets performance and training time results against the vanilla U-Nets for the defined data augmentation scenarios and varying the original training dataset size, from 10% to 100%, as illustrated in the right panel of Figure 4-1. ECoNets always give a positive Dice score gap, i.e. the % difference between equivariant and vanilla Dice scores at the end of training. This gap increases as the training dataset size is reduced which is very important if the available annotated training data for new sensors is scarce. Concerning the training time, we consider the convergence ratio, i.e. the ratio between the stopping epoch of vanilla and equivariant models, and the wall-clock time ratio, i.e. the ratio between wall-clock times when ECoNets and U-Nets reach similar Dice scores. Both time ratios show that ECoNets converge in fewer epochs and reach faster the same Dice scores than the vanilla U-Net. Faster epoch-wise convergence is also seen in the left panel of Figure 4-1.

Table 4-1: Benchmarks One and Two for the OpenContrails dataset at different training dataset fractions.

U-Net Augmentation	Benchmark	ECoNets	Dice Score Gap % (Convergence ratio, Wall-clock ratio) 10%	25%	50%	75%	100%
None	Two	C_4-None	4.63 (2.2, 4.1)	5.52 (1.8, 3.6)	2.35 (1.3, 3.6)	2.96 (1.4, 3.8)	2.26 (1.7, 3.5)
		C_8-None	9.17 (2.2, 4.6)	6.18 (2.6, 4.3)	4.68 (1.8, 3.3)	4.20 (1.9, 4.1)	3.09 (1.8, 4.8)
		D_4-None	8.51 (2.5, 4.9)	5.60 (2.3, 4.3)	4.10 (1.8, 4.6)	3.83 (2.1, 3.3)	3.81 (1.7, 4.7)
		D_8-None	10.16 (1.5, 4.1)	6.30 (2.3, 3.1)	4.37 (1.7, 3.8)	4.86 (1.9, 3.3)	4.33 (1.7, 3.9)
Aug. 1	One	C_4-None	1.95 (2.9, 3.0)	4.48 (2.7, 4.4)	0.66 (1.8, 3.7)	0.5 (2.6, 1.3)	0.25 (3.5, 2.7)
		C_8-None	6.47 (2.9, 4.1)	5.14 (3.8, 6.1)	2.99 (2.5, 3.7)	1.74 (3.5, 2.2)	1.08 (3.8, 2.4)
	Two	C_4-Aug. 1	5.91 (2.2, 5.8)	4.95 (2.7, 4.4)	3.76 (1.5, 3.2)	1.62 (2.2, 2.6)	2.65 (1.8, 2.4)
		C_8-Aug. 1	7.69 (2.2, 8.6)	5.78 (2.7, 3.6)	3.94 (1.5, 3.2)	2.79 (2.2, 3.2)	3.05 (1.8, 3.0)
Aug. 2	One	D_4-None	9.78 (2.7, 5.3)	4.77 (2.7, 5.1)	3.52 (2.2, 3.8)	2.09 (3.8, 2.2)	0.82 (3.1, 2.7)
		D_8-None	11.43 (1.6, 3.4)	5.47 (2.7, 3.7)	3.79 (2.1, 3.7)	3.12 (3.6, 4.6)	1.34 (3.1, 2.4)
	Two	C_4-Aug. 2	11.34 (1.6, 4.1)	6.77 (1.4, 4.3)	5.47 (1.2, 3.7)	1.98 (1.7, 3.1)	2.37 (1.2, 2.6)
		C_8-Aug. 2	12.21 (1.8, 5.7)	7.54 (1.5, 4.3)	5.60 (1.5, 3.8)	4.61 (1.8, 3.7)	3.59 (1.6, 2.7)
		D_4-Aug. 1	11.46 (1.6, 4.1)	6.33 (1.8, 2.5)	5.60 (1.5, 2.9)	3.72 (1.8, 3.2)	2.85 (1.3, 2.9)
		D_8-Aug. 1	11.43 (1.6, 3.5)	6.55 (1.8, 3.7)	3.76 (1.7, 2.2)	4.13 (2.2, 2.6)	2.41 (2.3, 2.7)

Benchmark *One.* The Dice Score gap in favour of ECoNets without augmentation and data augmented vanilla U-Nets increases as less training data is available. This is illustrated in the right panel of Figure 4-1 where we show the evolution of the Global Dice Score for all models versus the training dataset fraction, with all data taken in a stratified manner from OpenContrails. Interestingly at 10% of training data stricter equivariance results in bigger gaps, e.g. comparing D_8- and C_4-ECoNets. This indicates that equivariance effectively replaces data augmentation in the scarce 10% regime. Convergence is faster and the time to reach the U-Nets best Dice scores is shorter for ECoNets regardless of training budget and data augmentation scenario.

Benchmark *Two.* The Dice score gaps reported in Table 4-1 are typically larger than for the first benchmark, suggesting that ECoNets benefit more from continuous data augmentation. Yet,

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

convergence ratios are now lower than in the first benchmark since data augmentation increases the training epochs. Wall-clock ratios are similar to benchmark One. Finally, precision and recall results in Table 4-2 show that generally ECoNets have gains in both metrics compared to U-Nets, precision having the stronger gaps. This suggests that equivariant models tend to have reduced false positives, i.e. a reduced number of non-contrail pixels falsely labelled as contrails, which increases the confidence in applications such as contrail-to-flight attribution, to avoid wrongly attributing contrails to flights. Area under the precision-recall curve metrics, PR-AUC, are reported as well in Table 4-2 and follow similar conclusions to the Dice Score metric. It is important to note that equivariance and steerability allow a reduction factor of over 36x between the D_8 and the *Vanilla* models. We report the Dice Score, Precision, Recall and Area under the Precision-Recall curve (PR-AUC) in percentage for each model trained on the OpenContrails for different training dataset fractions $(10\%, 25\%, 50\%, 75\% \text{ and } 100\%)$ and continuous rotation data augmentation with (Flip+Rot(0, 360)) and without reflections (Rot(0,360)) in Table 4-2.

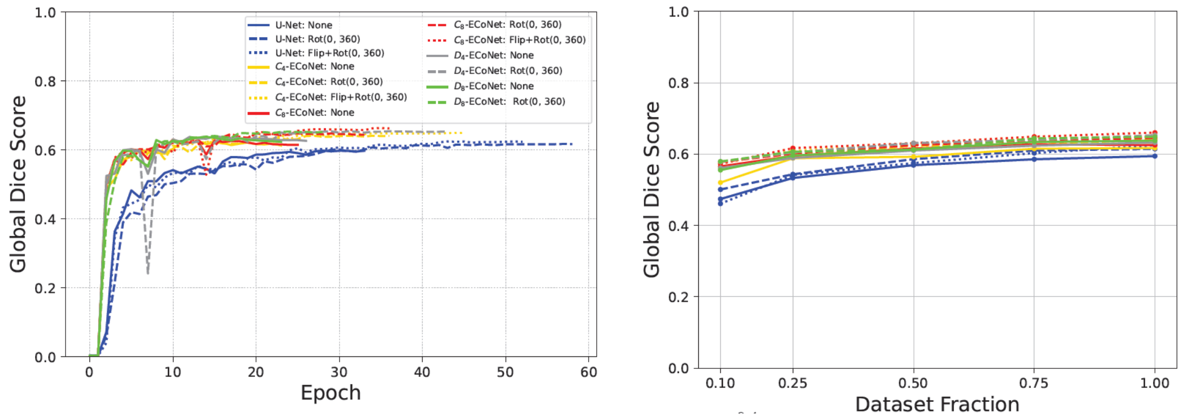

Figure 4-1: OpenContrails Test Dataset results. Left: 100% *training data used – Global Dice Score evolution for models without data augmentation (solid line), Rot(0,360) (dashed-lines) and Flip+Rot(0,360) augmentation (dotted lines).* Right: *Test Dataset Global Dice scores against the training dataset fraction, for the same models as on the left side.*

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

Table 4-2: OpenContrails test dataset segmentation scores for different levels of equivariance, data augmentation scenarios and training dataset size. Data augmentation Rot(0, 360) is denoted as Aug. 1 and Flip + Rot(0, 360) as Aug. 2

Model - Parameters (M)	Data Aug.	Dataset Size %	Dice Score %	Precision %	Recall %	PR-AUC %	Stopping Epoch	Training Time [h:mm:ss]
Vanilla - 95.86	None	10	47.39	50.93	44.31	48.22	47	1:09:02
		25	53.30	61.33	47.77	54.93	36	1:31:47
		50	56.89	59.49	54.51	58.19	31	2:18:02
		75	58.54	62.21	55.28	59.92	27	2:53:22
		100	59.44	63.19	56.11	61.01	24	3:21:50
	Aug. 1	10	50.07	51.58	48.64	50.94	61	1:26:44
		25	54.34	54.82	53.86	55.37	53	2:08:50
		50	58.58	59.01	58.15	60.01	43	3:01:23
		75	61.00	62.60	59.48	62.47	49	4:41:55
		100	61.45	61.42	61.48	62.90	49	5:17:50
	Aug. 2	10	46.10	43.23	49.37	46.53	51	1:14:32
		25	54.13	54.14	54.11	55.23	43	1:48:33
		50	57.47	56.42	58.57	58.89	38	2:44:45
		75	60.28	60.63	59.93	61.69	50	4:49:02
		100	62.48	64.20	60.84	64.04	43	5:46:33
C_4 - 10.54	None	10	52.02	68.41	41.96	56.01	21	0:56:48
		25	58.82	66.21	55.66	60.20	20	1:33:07
		50	59.24	67.10	53.04	61.20	24	3:03:57
		75	61.50	64.83	58.50	62.96	19	3:42:55
		100	61.70	67.50	56.81	63.45	14	3:51:25
	Aug. 1	10	55.98	54.05	58.05	57.32	28	1:10:23
		25	59.29	62.26	56.58	60.86	20	1:34:28
		50	62.34	64.25	60.54	63.95	29	3:29:27
		75	62.62	68.00	58.02	64.29	22	4:05:16
		100	64.10	63.74	64.47	65.47	27	6:06:35
	Aug. 2	10	57.44	56.16	58.78	58.61	32	1:14:06
		25	60.90	62.41	59.45	62.43	31	2:08:47
		50	62.94	62.78	63.11	64.58	31	3:40:14
		75	64.26	65.50	63.06	65.89	29	4:59:17
		100	64.80	64.99	64.61	66.25	36	7:40:48
C_8 - 5.27	None	10	56.54	61.39	52.40	57.81	21	0:58:21
		25	59.48	63.86	55.57	61.15	14	1:22:34
		50	61.57	66.32	57.46	63.10	17	2:37:47
		75	62.74	66.02	59.77	64.16	14	3:28:12
		100	62.53	65.24	60.04	63.98	13	4:32:03
	Aug. 1	10	57.76	60.11	55.59	59.15	28	1:11:34
		25	60.12	57.06	63.53	61.80	23	1:57:04
		50	62.52	65.03	60.19	64.21	20	3:04:30
		75	63.79	65.88	61.84	65.29	23	4:53:05
		100	64.50	67.44	61.82	65.99	24	6:34:50
	Aug. 2	10	58.31	58.61	58.01	58.90	29	1:09:46
		25	61.67	59.86	63.59	63.34	29	2:18:57
		50	63.07	60.89	65.40	64.72	25	3:36:42
		75	64.89	65.54	64.16	66.40	28	5:36:41
		100	66.02	67.64	64.68	67.57	27	7:10:12
D_4 - 5.27	None	10	55.88	61.11	51.48	57.18	19	0:57:02
		25	58.90	60.89	57.04	60.34	16	1:32:21
		50	60.99	72.51	52.63	63.75	17	2:45:52
		75	62.37	62.42	62.31	63.75	13	3:19:28
		100	63.25	61.86	64.70	64.63	14	4:46:27
	Aug. 1	10	57.56	54.85	60.55	59.35	31	1:13:42
		25	60.46	59.27	61.70	62.12	24	2:01:02
		50	63.07	61.89	64.30	64.57	25	3:40:56
		75	64.00	64.65	63.37	65.38	28	5:45:15
		100	65.28	66.90	63.74	66.71	34	8:38:06
D_8 - 2.63	None	10	54.27	61.35	48.64	56.73	31	1:26:13
		25	59.60	60.06	59.14	60.83	16	1:49:36
		50	61.26	68.98	55.59	63.13	18	3:31:46
		75	63.40	68.60	58.94	65.08	14	4:19:46
		100	63.77	63.34	63.88	65.19	14	5:36:37
	Aug. 1	10	57.53	59.40	55.77	59.48	31	1:26:13
		25	60.68	56.31	65.78	62.71	24	2:23:56
		50	61.23	52.68	73.10	65.12	22	4:03:16
		75	64.41	62.64	66.29	66.20	23	6:04:31
		100	64.84	68.60	61.48	66.74	19	6:58:24

For Dice, Precision and Recall metrics a threshold of 0.1 was applied to binarize the predictions[14]. This value was obtained from maximizing the validation Dice score at the stopping epoch. The stopping epoch corresponds to the point where the validation loss is at its lowest and the Training Time is the wall-clock time that takes the model to converge. Note that this value includes the early stopping patience of ten epochs.

We analysed the convergence speed as ratios of the stopping epochs presented in Table 4-2. For a deeper comparison, we include models' wall-clock times at similar Dice score performances as done in [7]Gerken et al. (2022). Since equivariant epochs are longer, similar performance wall-clock

[14] The output of the neural network gives a prediction in the interval [0,1] which needs to be binarized to match the ground-truth labels per pixel. The threshold used for this operation is a hyperparameter of the model as explained in the methods section.

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

times are a fairer comparison criterion. Table 4-3 summarizes our findings in the frameworks of Benchmarks *One* and *Two*. Notice that the performance of ECoNets was chosen to approximately match that of the baseline U-Net in this case. It is not always possible to exactly match the U-Net Dice Score since the metric can change abruptly from one epoch to another. However, we chose ECoNets Dice Scores to the U-Nets Dice ones.

Table 4-3: Benchmarks One and Two in terms of wall-clock time for the OpenContrails dataset at different training dataset fractions. ECoNets wall clock times are obtained such that the performance is similar to the Vanilla baseline.

	Model	Dice % (Clock Wall Time / Epoch)				
	Data Aug.	10 %	25 %	50%	75%	100%
Bench. One	U-Net: Aug. 1	50.07 (1:09:02/47)	54.34 (2:08:50/53)	58.58 (3:01:23/43)	61.00 (2:53:22/27)	61.45 (5:17:50/49)
	C_4: None	51.02 (0:23:04/12)	55.15 (0:29:05/09)	58.58 (0:49:46/09)	61.07 (2:06:41/16)	62.24 (2:01:00/11)
	C_8: None	50.68 (0:17:46/09)	54.95 (0:21:05/06)	59.01 (0:49:19/08)	61.18 (1:19:20/09)	61.35 (1:31:05/08)
	U-Net: Aug. 2	46.10 (1:14:32/51)	54.13 (1:48:33/43)	57.47 (2:44:45/38)	60.28 (4:49:02/50)	62.48 (5:46:33/43)
	D_4: None	48.53 (0:14:55/08)	54.77 (0:21:28/06)	58.51 (0:43:19/07)	60.68 (1:10:47/08)	62.33 (2:08:33/11)
	D_8: None	47.14 (0:22:30/12)	56.57 (0:29:53/07)	57.88 (0:44:20/06)	60.70 (1:03:54/06)	62.54 (2:22:38/10)
Bench. Two	U-Net: None	47.39 (1:09:02/47)	53.30 (1:31:47/36)	56.89 (2:18:02/31)	58.54 (2:53:22/27)	59.44 (3:21:50/24)
	C_4: None	46.32 (0:17:22/09)	53.57 (0:25:52/08)	57.31 (0:38:35/07)	58.42 (0:45:58/06)	59.87 (0:57:48/06)
	C_8: None	47.09 (0:15:47/08)	54.97 (0:21:05/06)	57.22 (0:42:46/07)	59.60 (0:42:23/05)	59.18 (0:42:56/04)
	D_4: None	48.53 (0:14:55/08)	54.77 (0:21:28/06)	56.49 (0:30:22/05)	59.53 (0:52:06/06)	59.93 (0:43:23/06)
	D_8: None	49.65 (0:21:12/10)	56.57 (0:29:53/07)	56.82 (0:36:18/05)	59.52 (0:52:18/05)	59.86 (0:52:23/04)
	U-Net: Aug. 1	50.07 (1:26:44/61)	54.34 (2:08:50/53)	58.58 (3:01:23/43)	61.00 (4:41:55/49)	61.45 (5:17:50/49)
	C_4: Aug. 1	49.96 (0:25:55/13)	54.75 (0:29:31/09)	58.57 (0:56:14/10)	61.26 (1:49:39/14)	61.41 (2:11:15/13)
	C_8: Aug. 1	50.08 (0:17:39/09)	53.95 (0:36:01/10)	58.99 (0:56:17/09)	60.68 (1:29:12/10)	61.04 (1:44:30/19)
	U-Net: Aug. 2	46.10 (1:14:32/51)	54.13 (1:48:33/43)	57.47 (2:44:45/38)	60.28 (4:49:02/50)	62.48 (5:46:33/43)
	C_4: Aug. 2	46.59 (0:18:41/10)	54.12 (0:25:53/08)	57.58 (0:44:44/08)	60.22 (1:33:15/12)	62.34 (2:11:47/13)
	C_8: Aug. 2	45.89 (0:13:01/07)	54.39 (0:25:37/07)	57.41 (0:43:18/07)	60.31 (1:19:29/19)	62.62 (2:08:46/11)
	D_4: Aug. 1	46.89 (0:18:31/10)	54.60 (0:43:50/12)	57.92 (0:57:20/09)	60.24 (1:30:01/12)	62.44 (1:58:06/10)
	D_8: Aug. 1	47:83 (0:21:37/10)	56.61 (0:29:45/07)	58.36 (1:16:32/09)	60.19 (1:50:32/10)	62.53 (2:09:45/09)

For completeness, we show the Global Precision and Recall scores as functions of the training dataset size in Figure 5. As observed, except for the D_4-ECoNet with Flip+Rot(0,360) at 50% of the dataset fraction and some ECoNets at 25%, the precision of ECoNets is always higher than that of U-Nets regardless of the training budget and data augmentation. Generally, ECoNets recall is higher than U-Nets as well but precision remains the stronger ECoNets point.

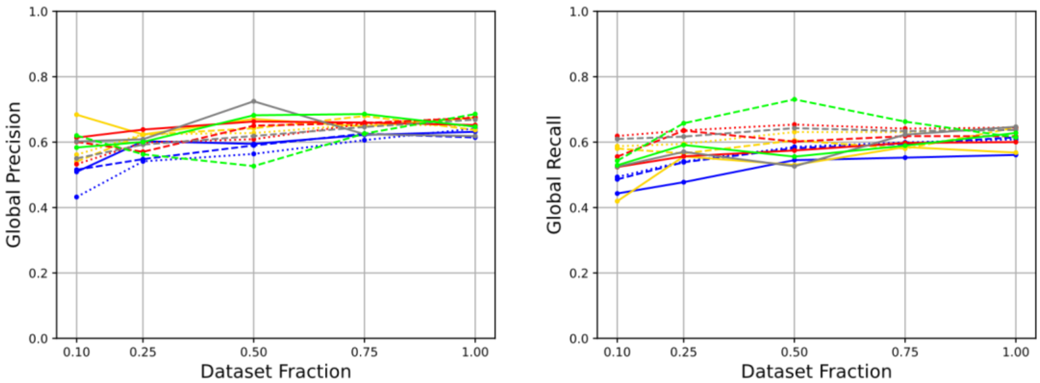

Figure 4-2: OpenContrails test dataset results. Left*: Global Precision against the training dataset fraction.* Right*: Global Recall against the training dataset fraction [models without data augmentation (solid line), Rot(0,360) (dashed-lines) and Flip+Rot(0,360) augmentation (dotted lines); Vanilla U-net in blue]*

Precision and Recall curves for different dataset fractions are illustrated in Figure 4-3. The curves are obtained at the stopping epoch for each model. These results complement the main conclusions from the Global Dice Score discussion since the PR-AUC metric is threshold independent. As expected, U-Nets have a lower PR-AUC than ECoNets. This is particularly noticeable in the low-data regime.

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

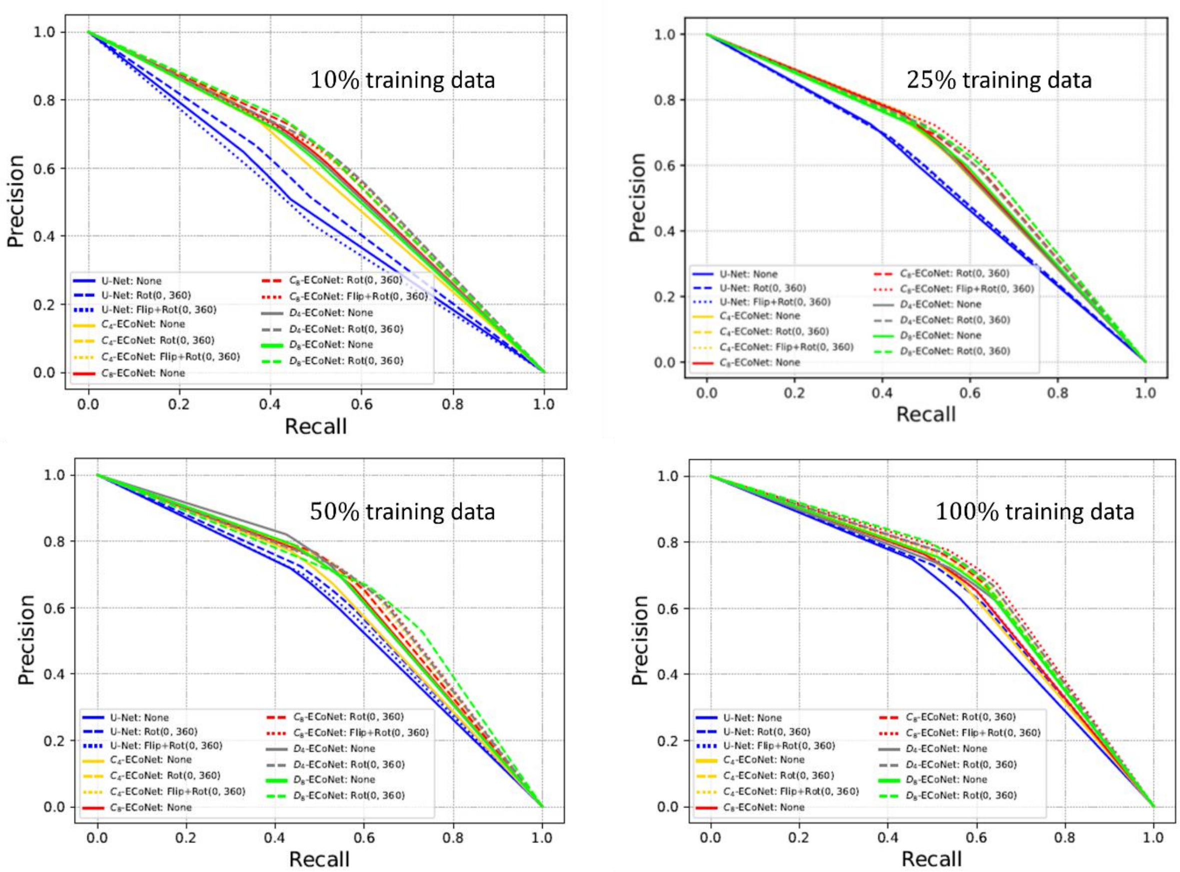

Figure 4-3 : Open Contrails Precision-Recall curves for ECoNets and U-Nets trained with 10%, 25%, 50% and 100% *of the dataset size. Thresholds of* $\{0.1i\}_{i=0}^{10}$ *are applied.*

Finally, for illustration, we show in the inferences (contrail predictions) of some models on a OpenContrail test image when trained with only 10% of the data.

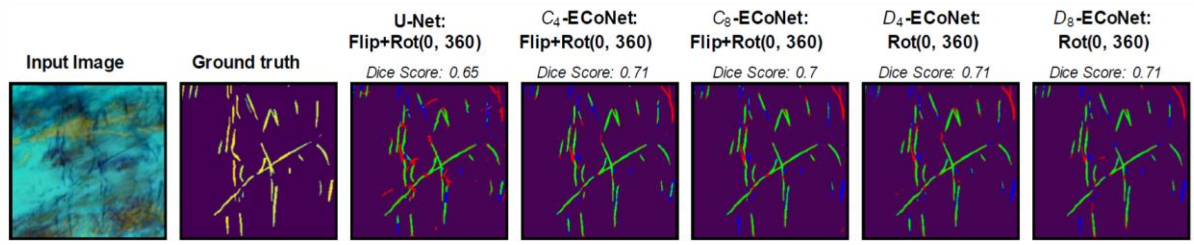

Figure 4-4 : Illustrations of model predictions for Vanila and ECoNets on one test image from the OpenContrails dataset. Note that the equivariant models (in particular the D_8 *one) tend to miss and overestimate less contrails than Vanilla [green = contrail hits, red = contrail overestimation, blue = missed contrails]*

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

4.3 MSG dataset results

Going further in data frugality and assessing domain adaptation capabilities, models trained on OpenContrails[15] are further retrained on our very reduced size dataset[16], MSG, of contrail-segmented images over Europe. We have performed two experiments: fine-tuning as explained above and training from scratch only using the MSG dataset with no pretraining in order to assess the equivariant models' potential at even less than 2% images of a regular training dataset. We thus aim to emulate an operational usage where a new sensor is introduced in the contrail verification pipeline with very few annotated data[17].

Regardless of the benchmark, fine-tuned ECoNets generally present an improved performance over Vanilla models as shown in Table 4-4, albeit lower than on the American dataset due to the training dataset size.

Table 4-4 : MSG results for the different neural models when using early stopping

Model	Data Aug.	Dice Score %	Training Time [mm:ss]	Stopping Epoch
Vanilla	None	45.82	08:03	16
	Aug. 1	49.16	13:20	15
	Aug. 2	47.44	08:17	13
C_4	None	50.72	02:36	11
	Aug. 1	51.72	04:30	24
	Aug. 2	52.72	06:12	10
C_8	None	51.17	07:08	17
	Aug. 1	51.69	04:59	10
	Aug. 2	53.58	09:55	11
D_4	None	52.23	04:46	4
	Aug. 1	52.31	14:40	23
D_8	None	52.08	07:51	16
	Aug. 1	53.81	07:46	26

Figure 4-5 illustrates the Global Dice Score evolution for several ECoNets and Vanilla U-Nets. Epoch-wise convergence and shorter wall-clock times for ECoNets are less significant possibly due to the fine-tuning strategy which retrained the entire network. In future work only the last layers could be re-trained to speed-up convergence. Interestingly, ECoNets trained from scratch (i.e. only on the 293 MSG training images), with data augmentation, reach similar Dice scores to the Vanila fine-tuned models. This is remarkable considering the very-low data regime and may indicate that in usecases with very scarce annotated data and where no similar larger datasets are available, equivariance with thoughtful data augmentation can substitute fine-tuning.

[15] OpenContrails stopping epoch weights are used as starting point
[16] Our European dataset, MSG, amounts to less than 1.5% of the OpenContrails dataset size
[17] Note that the labelling effort for only a couple of hundreds of images, considering day, geographical region selection, temporal sequences and labelling strategies amount to a workload of several months (training and data download included).

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

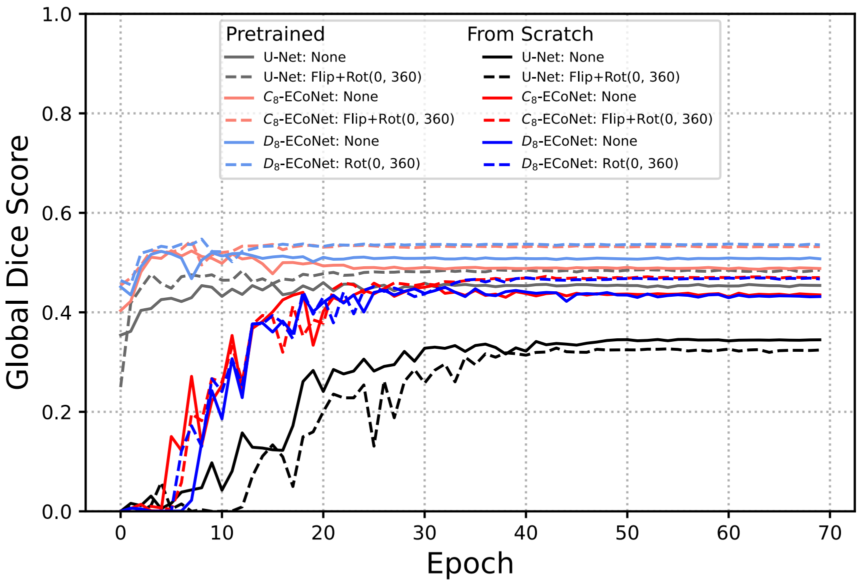

Figure 4-5 : Global Dice Score evolution for models trained on the MSG dataset from scratch (darker colours) and fine-tuned (lighter colours) on MSG at a fixed number of epochs.

To train the models from scratch we used the corresponding hyperparameters of the fine-tuning. There are two interesting points. Firstly, if we consider only models trained from scratch, the Dice gap between ECoNets and U-Nets is very important. This is expected by extrapolating the results of the OpenContrails Dice score gap that increases as the training budget diminishes below the 10% and as low as 1.5%. Secondly, U-Net-None pretrained on the OpenContrails dataset reaches a Dice Score of 45.60% at the epoch 70 when fine-tuned on the MSG european dataset. As explained above, ECoNets trained from scratch on this extreme low-size dataset, MSG, with data augmentation reach similar scores. For example, D_8-ECoNet+Rot(0, 360) has a score of 46.79% at epoch 70. Thus ECoNets with data augmentation can reach performances similar to pretrained U-Nets. In the MSG fine-tuning, only ECoNets with data augmentation show a notable gain.

Epoch-wise convergence and wall-clock times at similar performance are reported in Table 4-4 as well. Since scores at the stopping epoch are similar for fine-tuned models, no deeper analysis is needed for wall-clock times. For both temporal metrics, epoch-wise and performance-wise wall-clock, ECoNets confirm their advantage. However, the overall performance of the fine-tuning operation with our considered strategy are limited, probably by the small MSG dataset size and the image nature. Contrails in the MSG dataset usually correspond to older contrails in real life due to the temporal resolution of the sensor and are usually smaller and thinner than those in the OpenContrails dataset at the same age due to the spatial resolution of the sensor. The addition of the MSG images to the training is maybe in this situation is perhaps hindered by the very small number of samples. In future work combinations of OpenContrails and MSG data during training or only training final layers in the architecture could give different strategies of assessing overall model scalability.

To corroborate our results on the MSG trained from scratch case, we additionally performed an experiment on OpenContrails with only 2% of the training dataset size, similar in size to MSG. The Dice gap with 2% of the OpenContrails dataset was comparable to the UNet-ECoNet difference observed for MSG. Even more, with only 2% of the OpenContrails data, our best ECoNet (C_8 - Flip+Rot(0, 360)) reaches a score of 49.41% comparable to U-Net - None trained on 10% of the OpenContrails dataset. ECoNets show a more stable performance over training budgets from 2% (around 400 images) to 100% (around 20000 images) than U-Nets.

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

Finally, for illustration purposes, we show a couple of inference results for models fine-tuned with Flip+Rot(0,360) or equivalent data augmentation for the MSG dataset in Figure 4-6.

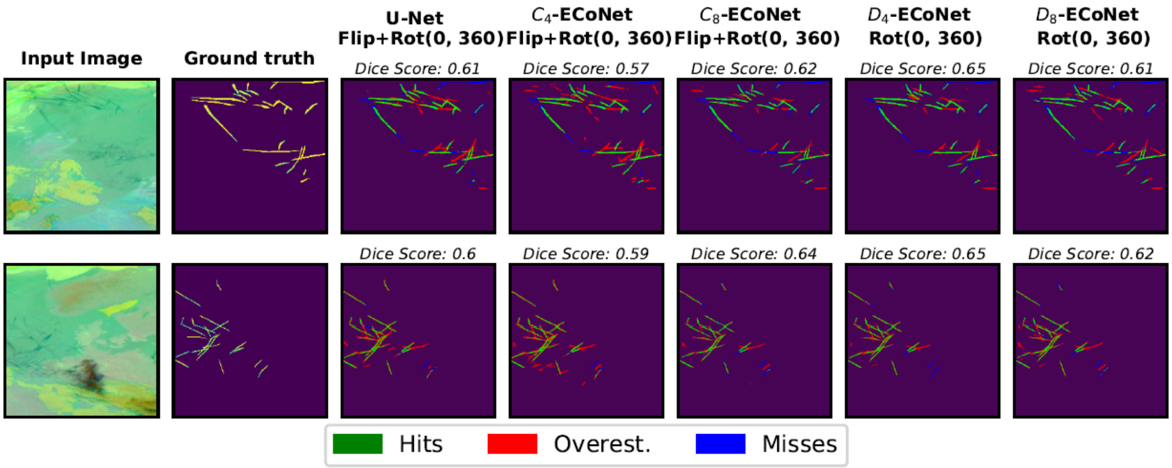

Figure 4-6: Inferences on two images from the MSG test dataset. The individual image Dice score is given for illustrating an on-the-fly contrail detection system. Green pixels correspond to correct model predictions, red pixels to false positive model predictions and blue to contrail pixels not predicted by the model.

5 Conclusion

In summary, in this deliverable we showed the potential of equivariant models in a complex use-case with scarce annotated data. Equivariant models with far fewer trainable parameters demonstrate higher performance, quicker epoch-wise convergence and shorter wall-clock training time than classic U-Nets for different training dataset sizes taken from OpenContrails and data augmentation scenarios.

When fine-tuning on the MSG dataset, moderate equivariant performance gains and shorter wall-clock times are obtained at similar Dice scores. Compared to models trained from scratch, only fine-tuned ECoNets with data augmentation show significant score gains. The current limitations to discrete symmetries can be addressed In future work, e.g. by introducing continuous rotation ($SO(2)$-) with possibly reflections ($O(2)$-) ECoNets for improved performance. These will help enforce an even greater uncertainty reduction in the contrail verification, by guaranteeing exact (up to the software discretisation precision) equivariance at all angles.

Even more importantly, ECoNets have a less significant Dice Score drop than U-Nets as the training budget is reduced, which is the stability (robustness-by-design) that we aimed for and have highlighted in this workpackage deliverable.

References

[1] Donny Aminou. MSG's SEVIRI instrument. *ESA Bull.*, 111, 08 2002.

[2] Karl Bengtsson Bernander, Joakim Lindblad, Robin Strand, and Ingela Nystr¨om. Rotationequivariant semantic instance segmentation on biomedical images. In Guang Yang, Angelica Aviles-Rivero, Michael Roberts, and Carola-Bibiane Sch¨onlieb, editors, *Medical Image Understanding and Analysis*, pages 283–297. Springer International Publishing, 2022. ISBN 978-3-031-12053-4.

[3] Gabriele Cesa, Leon Lang, and Maurice Weiler. A program to build E(N)-equivariant steerable CNNs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=WE4qe9xlnQw.

[4] Benjamin Chidester et al. Enhanced Rotation-Equivariant U-Net for Nuclear Segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

(CVPRW), pages 1097–1104, Long Beach, CA, USA, June 2019. IEEE. ISBN 9781728125060. doi: 10.1109/CVPRW.2019.00143.

[5] Taco Cohen and Max Welling. Group equivariant convolutional networks. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2990–2999, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/cohenc16.html.

[6] Taco S. Cohen and Max Welling. Steerable CNNs. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rJQKYt5ll.

[7] Jan Gerken, Oscar Carlsson, Hampus Linander, Fredrik Ohlsson, Christoffer Petersson, and Daniel Persson. Equivariance versus augmentation for spherical images. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7404–7421. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/gerken22a.html.

[8] Robin Ghyselinck, Valentin Delchevalerie, Bruno Dumas, and Benoit Frenay. On the effectiveness of rotation-equivariance in u-net: A benchmark for image segmentation. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=UcrVnXBdZI.

[9] Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7, 2020. doi: 10.1109/CIBCB48159.2020.9277638.

[10] Bernd K¨archer. Formation and radiative forcing of contrail cirrus. 9(1):1824, 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-04068-0. URL https://www.nature.com/articles/s41467-018-04068-0.

[11] D.S. Lee, D.W. Fahey, A. Skowron, M.R. Allen, U. Burkhardt, Q. Chen, S.J. Doherty,

[12] S. Freeman, P.M. Forster, J. Fuglestvedt, A. Gettelman, R.R. De Le´on, L.L. Lim, M.T. Lund, R.J. Millar, B. Owen, J.E. Penner, G. Pitari, M.J. Prather, R. Sausen, and L.J. Wilcox. The contribution of global aviation to anthropogenic climate forcing for 2000 to 2018. *Atmospheric Environment*, 244:117834, 2021. ISSN 1352-2310. doi: https://doi.org/10.1016/j.atmosenv.2020.117834. URL https://www.sciencedirect.com/science/article/pii/S1352231020305689.

[13] Hermann Mannstein, Richard Meyer, and Peter Wendling. Operational detection of contrails from noaa-avhrr-data. *International Journal of Remote Sensing*, 20(8): 1641–1660, 1999. doi: 10.1080/014311699212650. URL https://doi.org/10.1080/ 014311699212650.

[14] Benjamin Maurel, Samy Blusseau, Santiago Velasco-Forero, and Teodora Petrisor. Rototranslation equivariant YOLO for aerial images. In *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*, 2023. URL https://openreview.net/forum? id=EoyeHdfJ6l.

[15] Vincent R Meijer, Luke Kulik, Sebastian D Eastham, Florian Allroggen, Raymond L Speth, Sertac Karaman, and Steven R H Barrett. Contrail coverage over the united states before and during the covid-19 pandemic. *Environmental Research Letters*, 17(3):034039, mar 2022. doi: 10.1088/1748-9326/ac26f0. URL https://dx.doi.org/10.1088/1748-9326/ ac26f0.

[16] Joe Yue-Hei Ng, Kevin McCloskey, Jian Cui, Vincent R. Meijer, Erica Brand, Aaron Sarna, Nita Goyal, Christopher Van Arsdale, and Scott Geraedts. Contrail detection on goes16 abi with the opencontrails dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. doi: 10.1109/TGRS.2023.3345226.

[17] Olaf Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation.

[18] In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885

[19] Ulrich Schumann. On conditions for contrail formation from aircraft exhausts. *Meteorologische Zeitschrift*, 5(1):4–23, 03 1996. doi: 10.1127/metz/5/1996/4. URL http://dx.doi.org/10.1127/metz/5/1996/4.

[20] N. Siddique et al. U-Net and Its Variants for Medical Image Segmentation: A review of theory and applications*, IEEEAccess*, 2021

[21] Maurice Weiler and Gabriele Cesa. General e(2)-equivariant steerable cnns. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlch´e-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates,

[22] Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/45d6637b718d0f24a237069fe41b0db4-Paper.pdf.

[23] Jim Winkens, Jasper Linmans, Bastiaan S. Veeling, Taco S. Cohen, and Max Welling. Improved semantic segmentation for histopathology using rotation equivariant convolutional networks. 2018. URL https://openreview.net/forum?id=SyXbz1hiM.

[24] Jiazhen Zhang, Yuexi Du, Nicha C. Dvornek, and John A. Onofrey. Improved vessel segmentation with symmetric rotation-equivariant u-net. abs/2501.14592, 2025. URL https://doi.org/10.48550/arXiv.2501.14592.

This project has received funding from the European Union's Horizon Europe Research and innovation program under grant agreement No 101056885