Deep Semantic Contrails Segmentation of GOES-16 Satellite Images: A Hyperparameter Exploration

Gabriel Jarry[*], Philippe Very[*], Amine Heffar[*]
[*] Eurocontrol, Innovation Hub
Bretigny-Sur-orge France
Email: {gabriel.jarry, philippe.very, amine.heffar}@eurocontrol.int

Valentin Torjman-Levavasseur[†][*]
[†] ENSAE - ParisTech, Palaiseau, France,
Email: valentin.tordjman–levavasseur@ensae.fr

***Abstract*—This paper presents a comprehensive study on the optimization of hyperparameters for deep semantic segmentation models aimed at detecting contrails in GOES-16 satellite imagery. The environmental impact of aviation contrails has received considerable attention due to their potential contribution to climate change. Accurate contrail detection is essential for developing strategies to mitigate these impacts. Using the OpenContrails dataset [1] and advanced computer vision techniques, we performed a greedy hyperparameter search over different neural architectures, loss functions, and preprocessing methods. Our results indicate that using CoatNet as the backbone, coupled with the Unet++ architecture and dice loss as the optimization criterion, yields superior performance in contrail segmentation. In addition, incorporating data augmentation and resizing images to 512 pixels significantly improves model accuracy and generalization. The optimized model configurations demonstrate a promising approach for improving contrail segmentation, contributing to more accurate climate impact assessments and the development of sustainable aviation practices.**

***Keywords*—Contrail Semantic Segmentation, Aviation Sustainability, GOES-16 Satellite Imagery, Neural Network, Machine Learning,**

I. Introduction

Aviation is an integral part of modern society, facilitating global connectivity and economic growth. However, the environmental impact of aviation, particularly the formation of condensation trails, has received increasing attention in recent years. Contrails, the visible line-shaped clouds produced by aircraft engine exhaust at high altitudes, probably play a significant role in climate change [2], [3]. Important uncertainties remain, so understanding and managing the complex relationship between aviation contrails and their environmental impacts is a critical challenge in contemporary scientific research and policy development.

The relation between aviation contrails and climate change has become a central focus of research. With advances in satellite imagery and machine learning, researchers have gained profound insights into the detection and characterisation of contrails and the wider environmental impact of these high-altitude phenomena.

In the quest for more accurate and comprehensive contrail detection methods, the critical role of open-source labelled datasets is increasingly recognised [1], [4]. The availability of such datasets has the potential to catalyse significant advances in contrail detection using computer vision techniques. As we delve into the nuances of contrail formation, properties and climatic impacts, it becomes increasingly clear that the collaborative efforts of the wider scientific community in labelling datasets on a larger scale are essential. This not only fosters innovation, but also ensures that detection models are robust and adaptable, ultimately contributing to a more sustainable aviation industry.

In addition, the use of the OpenContrails dataset [1] in a Kaggle competition [5] exemplifies the transformative power of community collaboration in addressing this pressing issue. Harnessing the collective intelligence of Kaggle's vast community of data scientists and enthusiasts, the competition spurred the development of innovative machine learning models tailored to contrail segmentation.

In this paper, we present a greedy hyperparameter search focusing on semantic segmentation models for single-frame contrail satellite images. Our objective is to distill and provide a comprehensive review of hyperparameters employed in state-of-the-art models for contrail satellite images segmentation. The code use in the paper is avalaible as part of an open-source library [6].

II. State of the Art

A. Contrail Detection

The complex relationship between aircraft contrails and climate change has been a focus of research, with advances in satellite imagery and machine learning providing deep insights into the detection, characterisation and environmental impact of contrail

Advances in the detection of contrails using satellite imagery have led to a multitude of studies over the last two decades. Early research had already demonstrated the ineffectiveness of using Advanced Very-High-Resolution Radiometer (AVHRR) satellite imagery to detect contrails [7]. In [8], the use of AVHRR for contrail detection revealed regional patterns over Europe, but encountered difficulties in distinguishing certain cloud structures . The Automatic Contrail Tracking Algorithm (ACTA) was introduced in [9] and has proven its effectiveness in real-world scenarios. However, shorter contrails are sometimes not detected. [10] uses Himawari-8 satellite imagery and compares two potential contrail coverages. A visual computing system that facilitates the analysis of contrails from aircraft simulations, streamlining data comparison is detailed in [11]. Convolutional neural network-based approaches have shown to be effective in detecting aircraft contrails with high accuracy, further highlighting their climatic

SESAR Innovation Days 2024
12 - 15 November 2024. Rome
ADR AEROPORTI DI ROMA enav SpA LEONARDO EUROPEAN PARTNERSHIP Co-funded by the European Union
sesar JOINT UNDERTAKING

implications in [12]. Another convolutional neural network model, ContrailMod, was optimised for contrail detection, showing a strong correlation between contrail occurrence and potential coverage [13]. A system based on a neural network, the CiPS algorithm, was presented as an effective method for recovering cirrus properties, showing greater accuracy for fine cirrus clouds in [14]. The impact of the covid-19 pandemic on the appearance of contrails in the United States and different diurnal patterns were studied in [15] using modern deep learning techniques.

Ground-based cameras calibrated with sky observations have provided valuable insights when combined with synthetic images from prediction models in [16]. Atmospheric monitoring can be done cost-effectively using all-sky cameras that use starlight absorption to map cloud structures [17]. Ground-based studies showed that the visibility of contrails in satellite imagery is largely dependent on their width, with current algorithms often miscalculating contrail and cirrus volumes [18].

In [19], a technique is proposed that uses a synthetic dataset of contrails generated using a state of the art contrail evolution model called CoCiP([20]). This technique is combined with instance segmentation models and dedicated tracking and matching algorithms to assign contrails to specific aircraft. A similar methodology combining satellite imagery and air traffic data to efficiently identify and match contrails to corresponding aircraft trajectories has been proposed by Riggi et al. [21]. The flight matching problem can be simplified by using initial altitude estimates from the contrail shadow [22].

While newer models, such as one using augmented transfer learning and dedicated SR loss, show potential, more extensive validation is required [23]. The release of a labelled dataset of Landsat-8 satellite imagery aims to advance contrail detection methods, providing potential solutions to aviation's contribution to global warming [4]. Similarly, the recent launch of the OpenContrails dataset using GOES-16 ABI imagery has set a new benchmark for contrail detection [1] and developpement of machine learning segmentation models [24], [25].

B. *Contrail avoidance*

Amidst the plethora of research and advances in contrail detection from satellite imagery and ground-based equipment, the associated climatic implications of contrails have not gone unnoticed. As the aviation industry grapples with its environmental footprint, efforts are being made not only to understand contrails, but also to develop strategies to mitigate them.

One particular study examines the potential climate impact of these contrails in Japanese airspace and highlights a dual mitigation approach that includes strategic fleet diversions combined with the implementation of new engine technologies. This dual strategy, if properly applied, could lead to a potential 91.8% reduction in the climate impact of contrails [26]. Similarly, another study in the same region explores the benefits of vertical flight diversions as a means of mitigating the adverse climate impacts of cirrus contrails. Specifically, the study suggests that such diversions could result in a remarkable 105% reduction in contrail energy forcing. In addition, these diversions appear to be associated with minimal fuel penalty and can be implemented without compromising aviation standards [27].

In a different geographical context, the North Atlantic corridor, there is ongoing research into design principles for experiments aimed at mitigating persistent heat-trapping contrails. This research provides invaluable insights into the role that such contrails play in the wider spectrum of aviation-related climate impacts. By providing detailed considerations and proposed trial methodologies, the study serves as a critical resource for stakeholders seeking to address this environmental challenge [28]. Beyond these targeted regional studies, there's a broader movement to develop models that can address the environmental footprint of aviation on a large scale [29]. One innovative approach is the introduction of a time-dependent subgraph capacity model, specifically designed to address CO2 and contrail emissions simultaneously. A real-world application of this model, as demonstrated in the French upper airspace, shows the potential to achieve significant contrail mitigation, further emphasising the need to adopt such novel models and techniques across the aviation industry [30].

C. *Segmentation*

Deep image segmentation uses sophisticated neural architectures to divide digital images into semantically meaningful regions, ensuring fine-grained recognition and detailed boundary delineation.

Semantic segmentation, a notable area in computer vision, has seen many breakthrough innovations. The Fully Convolutional Network (FCN), which uses a ”skip” architecture to merge layers of different depths, has made significant advances in semantic segmentation, notably outperforming benchmarks such as PASCAL VOC 2011 [31]. The DeepLab system has also carved out a niche with its unique blend of atrous convolution and atrous spatial pyramid pooling, leading to state-of-the-art results on platforms such as PASCAL VOC-2012 [32].

In the biomedical domain, the U-Net architecture, which strikes an optimal balance between context capture and precise localisation, has been validated to outperform existing segmentation strategies in various tasks [33]. An adaptation of this architecture, the 3D U-Net, further enhances its capabilities by addressing dense volumetric segmentations and shows clear superiority over its 2D analogues in 3D biomedical structures [34]. The V-Net, another cornerstone of volumetric medical image segmentation, presents a distinctive training approach and ”value networks” that deliver unprecedented speed and accuracy when applied to prostate MRI data [35]. UNet++, with its nested architecture, bridges the semantic gap between the encoder and decoder modules, demonstrating commendable segmentation performance in medical imaging [36].

To address the need for detailed object instance recognition, Mask R-CNN, an evolution of Faster R-CNN, introduced a mask prediction branch, setting new standards in multitask adaptability [37]. Another innovative approach is the Pyramid

Scene Parsing Network (PSPNet), which uses a pyramid pooling module to exploit global context and achieve excellence in capturing long-range dependencies [38].

Further advances in the DeepLab series, such as DeepLabv3 and DeepLabv3+, emphasised the role of atrous convolution in DCNNs and introduced efficient decoder modules, respectively. These contributions consistently outperformed benchmarks without the need for post-processing steps [39] [40] [41].

Several other models, such as ICNet and RefineNet, have addressed the balance between speed and accuracy in real-time segmentation, with ICNet in particular demonstrating its ability to fuse multi-resolution branches [42], [43]. SegNet, tailored for scene understanding, has adopted a unique approach with max-pooling indices in its decoder, demonstrating its effectiveness in various scene segmentation tasks [44]. An intriguing blend of image segmentation and computer graphics rendering led to the creation of PointRend, which efficiently combines speed with high quality boundary predictions [45]. Finally, the boundary loss function introduced is tailor-made for medical images, addressing unbalanced segmentation and increasing accuracy [46].

III. Dataset Description & Analysis & Preprocessing

The dataset used in this study [1], [5] is from the GOES-16 Advanced Baseline Imager (ABI), with brightness temperatures derived from Level 1B radiances. The satellite provides coverage of North and South America and acquires a full disk image at 10-minute intervals. The spatial resolution is 2x2 km at nadir, which can be a challenge for detecting initial contrail formations. However, the limitations may be advantageous for studying contrails with more extensive heating effects.

A. *Sample Generation & labelling*

Samples are randomly taken from the visible range of GOES-16 between April 2019 and April 2020, subject to certain latitudinal and longitudinal constraints. To overcome the rarity of condensation trails, the dataset was enriched with positive examples. Aircraft tracks from terrestrial ADS-B data and wind data from ECMWF ERA5 were used to guide the sample selection. Additional layers including relative humidity over ice and the Mannstein et al. [8] contrail detection algorithm adapted for high recall were added.

An ”ash” false-color scheme was used to facilitate 24-hour labelling, highlighting ice clouds as darker shades to aid contrail identification. Each scene received annotations from at least four human labelers, with consensus reached by majority vote. An example of the ”ash” false colour and its associated human labelled mask is shown in figure 1.

B. *Description*

The dataset consists of 20,544 training samples and 1,866 validation samples. Contrail annotations are present in approximately 1.2% of the training pixels. An additional test set, proportional in size to the validation dataset, is available on the Kaggle competition website [5]. While the test data is not directly accessible, it can be used for model evaluation through a public score (representing 15% of the test dataset) and a private score (representing the remaining 85%).

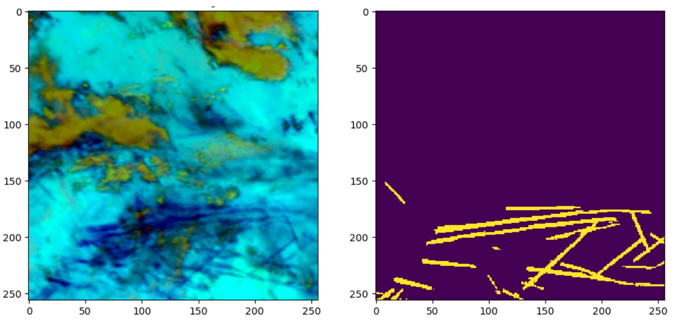

Figure 1. On the left is a false-color satellite image of ”ash”, and on the right is the corresponding human annotation.

The data are stored in .npy format. Each sample consists of a temporal sequence of images spanning the nine satellite spectral bands (08 to 16). Each image in the sequence has a resolution of 256x256 pixels. The human annotated mask is either provided or to be determined for the fourth image in the sequence. Although the dataset contains sequences, this paper specifically addresses the single-frame problem, focusing only on the fourth frame of each sequence.

C. *Analysis*

Several elements of the dataset were highlighted during the Kaggle competition [5]. The labelling process, which preferentially selected and re-centred images with a significant proportion of contrails based on Google Street View, resulted in a noticeable concentration of contrail-positive pixels in the centre of the mask, as shown in Figure 2.

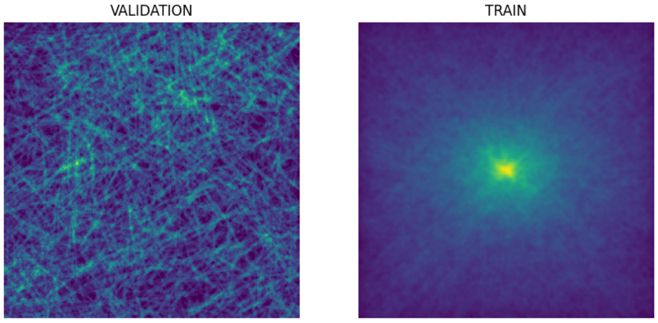

Figure 2. The figure shows the average mask for the validation set on the left and for the training set on the right. The training set has a centred property.

A significant observation was the half-pixel shift in the mask, possibly caused by the transformation from polygon coordinates to pixel masks. Specifically, the top-left coordinates might have been used instead of the centre coordinates when converting human-labelled polygons to pixel masks. This discrepancy was identified by several competitors and proved to be consequential. In particular, it affected the effectiveness of several techniques, especially those such as augmentation and test-time augmentation, which often rely on rotations. The correction for this shift is discussed in the following subsection.

IV. Methodology & Mdeling

A. *Hyper Parameter Search Methodology*

In the course of developing our model using the GOES-16 ABI dataset, we used a pseudo-greedy parameter search to fine-tune and optimise the model's hyperparameters. The primary goal of this process was to identify a combination of hyperparameters that would maximise the performance of the model in terms of global dice score without extensively exploring the entire hyperparameter space. Global Dice score ensures robust evaluation of segmentation accuracy, especially when dealing with sparse positive pixel classes, by balancing precision and recall across the entire image. We started with an initial set of hyperparameters, chosen either on the basis of prior knowledge, literature review or simple intuition. This set served as the starting point for our iterative optimisation and is described in Table I

In our study, we primarily followed the principles of greedy search, optimising one hyperparameter at a time while holding others fixed. This method allowed us to efficiently navigate the vast hyperparameter space by adjusting each parameter individually across its range and identifying its optimal value. Once an optimal value was found, it was fixed and the process was repeated for the next parameter in the sequence.

However, there were specific instances, particularly in relation to the incompatibility between the backbone architecture and the overall model architecture, where we deviated from the strict rules of the greedy approach.

The main advantage of our greedy approach is its efficiency. Instead of the combinatorial explosion of possibilities in grid search, our method significantly reduces the search space. However, it's worth noting that the greedy nature of the method can lead to local optima. This is a trade-off we have accepted for the sake of computational efficiency and time constraints.

B. *Loss function candidates*

The Dice score is a function used to quantify the similarity between two sets of data. In the context of image segmentation, these sets are the predicted segmentation and the ground truth segmentation. Mathematically, it is defined as:

$$\text{Dice score} = \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{1}$$

$$\text{Dice Loss} = 1 - \text{Dice score} \tag{2}$$

Where:

- X is the predicted set of pixels (segmentation).
- Y is the ground truth set of pixels (segmentation).
- $|X \cap Y|$ is the cardinality of the intersection of the predicted and ground truth sets.
- $|X|$ and $|Y|$ are the cardinality of the predicted and ground truth sets respectively.

This loss function is particularly useful in cases where there is class imbalance, as it provides a more robust error signal even when the classes are not equally represented. In the following, we distinguish the global dice score, which is computed at the end of an epoch using the entire data set, and the batch dice score, which is computed during training by averaging each batch score.

The Binary Cross-Entropy (BCE) loss is a widely used loss function in binary classification tasks. It quantifies the difference between two probability distributions - the actual label and the predicted probability. It is defined as

$$\text{BCE loss} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{3}$$

Where:

- N is the number of samples.
- y_i is the actual label of the i^{th} sample, which is 1 if true and 0 if false.
- p_i is the predicted probability that the i^{th} sample is in class 1.

This loss function is particularly suitable for scenarios where the model outputs a probability value, and effectively measures the deviation of the predicted probability from the actual label.

The Focal Tversky Loss function is an extension of the Tversky index used in image segmentation, and is particularly effective in cases of class imbalance. It is defined as

$$\text{Focal Tversky Loss} = \sum_{i=1}^{N} (1 - T_i)^{\gamma} \tag{4}$$

Where:

- T_i is the Tversky index for the i^{th} sample, defined as:

$$T_i = \frac{|X_i \cap Y_i|}{|X_i \cap Y_i| + \alpha |X_i \backslash Y_i| + \beta |Y_i \backslash X_i|} \tag{5}$$

- X_i and Y_i represent the predicted and ground truth sets for sample i^{th} respectively.
- α and β are hyperparameters to control the balance of false positives and false negatives.
- γ is the focusing parameter to adjust the rate at which easy examples are down-weighted.
- N is the total number of samples.

Focal Tversky Loss addresses class imbalance by modulating the loss with respect to the difficulty of each sample, focusing more on harder-to-classify examples. It is particularly useful in medical image segmentation, where unbalanced data is common.

C. *Machine Learning particular processes*

In deep learning, particularly for semantic image segmentation, the Exponential Moving Average (EMA) is a technique used to smooth parameter updates during training. It gives higher weights to more recent parameter updates, while the

TABLE I
TABLE SHOWING THE INITIAL HYPERPARAMETERS OF THE PSEUDO-GREEDY SEARCH.

CHECKPOINT	MASK	ARCHITECTURE	BACKBONE	RESIZE	EPOCHS	BATCH SIZE
Loss	Probability	Unet	Resnest26d	384	40	48

ACCUMULATION	LOSS	OPTIMIZER	SCHEDULER	WEIGHT DECAY	LEARNING RATE	WARM-UP
1	Dice	AdamW	Cosine	1e-4	5e-4	1

influence of older updates decreases exponentially. This approach helps to stabilise the training process, reduce noise, and ensure that the model adapts effectively to recent data trends without ignoring the broader historical context, thereby improving overall learning efficiency.

Batch accumulation is a technique used in deep learning to overcome hardware limitations during training. When the ideal batch size for a model exceeds the capacity of the available hardware, batch accumulation allows smaller batches to be processed in successive forward steps. The gradients from these smaller batches are accumulated, and a single backpropagation step is performed after processing multiple batches. This approach allows models to be trained with larger effective batch sizes without the need for high-end hardware, ensuring training effectiveness on more modest systems.

D. Training process and criteria to be explored

In our training process, we implemented a 4-fold cross-validation method to increase the robustness and generalisability of our model. The training set was divided into four different folds, each serving as a combination of training (3 folds - 75% dataset) and selection (1 fold - 25% dataset) data sets. This approach allowed for extensive training and selection across different subsets of data, ensuring that the model was exposed to a wide range of scenarios. To optimise model performance and prevent overfitting, we used a checkpoint mechanism. This technique involved saving the best model based on its performance on the selection fold. The evaluation and decision for hyper parameter search is based on an ensemble approach. The global dice score is computed on the 4-fold ensemble model (average prediction) on the independent validation dataset. This strategy provided a more robust evaluation as it took into account the collective performance of models trained on different data subsets, ensuring that the final selected model demonstrated consistent and reliable performance across different data samples.

In our greedy search approach, we examined a wide range of parameters that are critical to improving the performance and accuracy of our semantic segmentation model. This included a nuanced examination of checkpoint criteria, where we compared the effectiveness of using batch dice loss versus global dice score. A key area of investigation was the input mask format, where we evaluated the use of probability-based masks versus binary masks from vote.

Backbone architectures formed a core part of our research, with trials conducted on a range of networks including Resnet26d, EfficientnetB7, MaxVit, Resnet101, EfficientnetV2 and CoatNet. In particular, we delved into a comparative analysis between Maxvit and CoatNet for backbone refinement. The architecture of the model itself was another critical point of analysis, where we tested configurations such as DeepLabV3, Unet and Unet++.

Loss functions were carefully evaluated, ranging from Dice Loss, Cross Entropy, Focal and Focal Tversky to Lovasz, Hybrid Loss and SR loss [23]. Another area of focus was image resizing, where we experimented with dimensions of like 384, 512. Finally, we also investigated the use Data Augmentation (DA) into the training process.

V. RESULT

For all subsequent tables showing hyperparameter search steps, scores are given using the Global Dice score in percent. For both the training and validation categories, the scores are averaged across the four fold models and the standard deviation is given in parenth

The first step in the hyperparameter search is to find the best checkpoint metric between the global dice score and the batch dice loss. As shown in the table II, the batch dice loss shows better performance on the validation test set.

TABLE II
TABLE SHOWING THE PERFORMANCE OF THE MODEL WITH DIFFERENT CHECKPOINT METRICS.S

CHECKPOINT	TRAIN	VALIDATION	ENSEMBLE
Batch Dice loss	**64.18 (0.28)**	**62.80 (0.16)**	**64.87**
Global Dice score	65.05 (0.59)	62.45 (0.71)	64.21

We observe that the use of a probability mask by the average of all the labellers shows a better performance than the use of the binary mask obtained by majority voting (Table III).

TABLE III
TABLE SHOWING THE PERFORMANCE OF THE MODEL WITH DIFFERENT MASK TYPES.

MASK	TRAIN	VALIDATION	ENSEMBLE
Probability	**64.18 (0.28)**	**62.80 (0.16)**	**64.87**
Binary	66.01 (0.29)	61.57 (0.20)	63.81

Regarding the backbones (Table IV), MaxVit, a transformer backbone, presents the better performance. However, the performance on CoatNet, another transformer backbone, cannot be achieved due to convergence problems related to initial learning rate and warm-up steps. Therefore, a dedicated MaxVit vs. CoatNet analysis is performed with different learning rate and warm-up step parameters. CoatNet always outperformed MaxVit and was then selected. The best performance is displayed in Table IV.

In terms of architectures, DeepLabV3 presents incompatibilities with Transformer backone, therefore the comparison

TABLE IV
Table showing the performance of the model with different backbones.

BACKBONE	TRAIN	VALIDATION	ENSEMBLE
Resnest26d	64.18 (0.28)	62.80 (0.16)	64.87
EfficientNetB7	64.06 (0.78)	63.85 (0.22)	65.76
MaxVit	64.73 (0.70)	64.22 (0.47)	66.48
ResNet101	63.78 (0.45)	63.35 (0.26)	65.21
EfficientNetV2	64.22 (0.11)	62.71 (0.35)	64.90
CoatNet	**64.75 (0.34)**	**64.57 (0.19)**	**66.97**

is made on ResNet26Dd backone in table V. Unet++ shows the best performance, confirmed when applied with CoatNet backone.

TABLE V
Table showing the performance of the model with different architures on ResNet26d an CoatNet backbones.

BACKBONE	ARCHI	TRAIN	VALIDATION	ENSEMBLE
ResNet26d	DeepLabV3	56.69 (0.28)	53.91 (0.33)	55.92
ResNet26d	Unet	63.05 (0.36)	60.23 (0.14)	62.48
ResNet26d	Unet++	63.98 (0.28)	60.99 (0.39)	63.24
CoatNet	Unet	64.75 (0.34)	64.57 (0.19)	66.97
CoatNet	**Unet++**	**65.44 (0.70)**	**65.12 (0.22)**	**67.40**

The results indicate that the Dice loss function exhibit the best overall performance, with balanced training and validation scores and strong ensemble results, suggesting good generalization. Although other configurations achieved higher training scores, they showed signs of overfitting, evidenced by a more pronounced drop in validation performance. Consequently, the Dice-based loss emerges as the most effective choices for optimizing model performance in this context.

TABLE VI
Table showing the performance of the model with different loss. Labelsmoothing is abreviated as LS and Positive Weight as PW

LOSS	TRAIN	VALIDATION	ENSEMBLE
Dice	**65.44 (0.70)**	**65.12 (0.22)**	**67.40**
BCE (PW=5)	67.57 (0.27)	64.72 (0.30)	67.26
BCE (PW=5. LS=0.05)	67.64 (0.18)	64.43 (0.22)	66.87
Focal Tsversky	67.16 (0.51)	64.15 (1.91)	66.76
Focal (PW=1)	67.64 (0.18)	64.43 (0.22)	65.01
Lovasz	66.93 (0.36)	63.03 (0.60)	65.74
Dice + BCE (PW=1)	67.80 (0.11)	65.06 (0.68)	67.29
Dice + BCE (PW=5)	67.47 (0.13)	64.89 (0.34)	67.30
SR [23]	66.41 (0.03)	64.94 (0.19)	66.96

Resizing to 512 pixels outperforms 384 pixels, yielding higher training, validation, and ensemble scores, indicating better generalization and model performance. Hence, 512-pixel resizing is recommended for optimal results.

TABLE VII
Table showing the performance of the model with resizing.

RESIZE	TRAIN	VALIDATION	ENSEMBLE
384	65.44 (0.70)	65.12 (0.22)	67.405
512	**68.27 (0.28)**	**65.24 (0.25)**	**67.54**

Incorporating data augmentation (DA) with (correction of the pixel shift) significantly enhances model performance, with improved training, validation, and ensemble scores compared to the model without augmentation. The validation score increased from 65.24 to 67.72, and the ensemble score from 67.54 to 68.56, indicating that data augmentation effectively enhances the model's generalization and robustness. Therefore, utilizing data augmentation is recommended for achieving optimal model performance.

TABLE VIII
Table showing the performance of the model with or without data augmentation.

RESIZE	TRAIN	VALIDATION	ENSEMBLE
NO DA	68.27 (0.28)	65.24 (0.25)	67.54
DA	**69.98 (0.33)**	**67.72 (0.19)**	**68.56**

VI. Discussion

The first point to note is that reported performance is based on models trained on subsets of data during cross-validation. Because each fold uses only a portion of the data for training, the results may not reflect the full potential of the model. If trained on the full dataset, the model could perform better, benefiting from more comprehensive data exposure and potentially achieving higher accuracy. Thus, the current results are likely conservative estimates of the model's optimal performance.

In addition, the model was trained on GOES-16 data, which limits its exposure to different atmospheric conditions. To improve generalization, it would be beneficial to test its performance on other satellites, such as Meteosat Third Generation (MTG) or HIMAWARI, which cover different regions and conditions. This could help to assess the adaptability of the model and improve its usefulness for global contrails.

VII. Conclusions

In this study, we explored and optimized hyperparameters for deep semantic segmentation models applied to contrail detection in GOES-16 satellite images. Our research aimed to enhance the accuracy and reliability of contrail segmentation, a critical task for understanding and mitigating aviation's environmental impact.

The results of our hyperparameter search revealed several key findings. First, we confirm that the choice a probability masks outperforms binary masks generated by majority voting. Among the backbone architectures, CoatNet emerged as the most effective, particularly when combined with the Unet++ architecture, which demonstrated superior performance across various metrics.

Our experiments with different loss functions highlighted the Dice loss as the most effective for our application, providing a good balance between training and validation performance, thus ensuring better generalization. Additionally, resizing images to 512 pixels and incorporating data augmentation further enhanced the model's accuracy and robustness, demonstrating the importance of preprocessing and data handling techniques.

In conclusion, our study underscores the importance of a systematic hyperparameter exploration in developing high-performing models for contrail detection. The optimized model configurations presented here offer a promising approach for improving contrail segmentation accuracy, contributing to more precise climate impact assessments and the development of strategies for mitigating aviation's environmental footprint. Future work should continue to refine these models and explore additional avenues for enhancing detection accuracy, such as integrating multi-frame analysis and further leveraging the full spectral range of satellite data.

References

[1] J. Y.-H. Ng, K. McCloskey, J. Cui, V. R. Meijer, E. Brand, A. Sarna, N. Goyal, C. Van Arsdale, and S. Geraedts, "OpenContrails: Benchmarking Contrail Detection on GOES-16 ABI," Apr. 2023, arXiv:2304.02122 [cs]. [Online]. Available: http://arxiv.org/abs/2304.02122

[2] D. S. Lee, D. W. Fahey, A. Skowron, M. R. Allen, U. Burkhardt, Q. Chen, S. J. Doherty, S. Freeman, P. M. Forster, J. Fuglestvedt *et al.*, "The contribution of global aviation to anthropogenic climate forcing for 2000 to 2018," *Atmospheric Environment*, vol. 244, p. 117834, 2021.

[3] R. Teoh, Z. Engberg, U. Schumann, C. Voigt, M. Shapiro, S. Rohs, and M. Stettler, "Global aviation contrail climate effects from 2019 to 2021," *EGUsphere*, vol. 2023, pp. 1–32, 2023.

[4] K. McCloskey, S. Geraedts, C. Van Arsdale, and E. Brand, "A human-labeled Landsat-8 contrails dataset," Jul. 2021.

[5] J. Ng, C. Elkin, A. Sarna, W. Reade, and M. Demkin, "Google research - identify contrails to reduce global warming," 2023. [Online]. Available: https://kaggle.com/competitions/google-research-identify-contrails-reduce-global-warming

[6] G. Jarry, P. Very, R. Dalmau, and J. Sun, "Deepenv: Python library for aircraft environmental impact assessment using deep learning," 2024, https://doi.org/10.5281/zenodo.13754838, https://github.com/eurocontrol-asu/DeepEnv/tree/DeepContrails/DeepEnv/training/DeepContrails.

[7] J. Weiss, S. Christopher, and R. Welch, "Automatic contrail detection and segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 36, no. 5, pp. 1609–1619, Sep. 1998. [Online]. Available: http://ieeexplore.ieee.org/document/718864/

[8] H. Mannstein, R. Meyer, and P. Wendling, "Operational detection of contrails from NOAA-AVHRR-data," *International Journal of Remote Sensing*, vol. 20, no. 8, pp. 1641–1660, Jan. 1999. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/014311699212650

[9] M. Vazquez-Navarro, H. Mannstein, and B. Mayer, "An automatic contrail tracking algorithm," *Atmospheric Measurement Techniques*, vol. 3, no. 4, pp. 1089–1101, Aug. 2010. [Online]. Available: https://amt.copernicus.org/articles/3/1089/2010/

[10] J. Zhang, J. Shang, and G. Zhang, "Verification for Different Contrail Parameterizations Based on Integrated Satellite Observation and ECMWF Reanalysis Data," *Advances in Meteorology*, vol. 2017, pp. 1–11, 2017. [Online]. Available: https://www.hindawi.com/journals/amete/2017/8707234/

[11] N. Nipu, C. Floricel, N. Naghashzadeh, R. Paoli, and G. E. Marai, "Visual Analysis and Detection of Contrails in Aircraft Engine Simulations," *arXiv:2208.02321 [cs]*, Aug. 2022, arXiv: 2208.02321. [Online]. Available: http://arxiv.org/abs/2208.02321

[12] N. Siddiqui, "Atmospheric Contrail Detection with a Deep Learning Algorithm," p. 8.

[13] G. Zhang, J. Zhang, and J. Shang, "Contrail Recognition with Convolutional Neural Network and Contrail Parameterizations Evaluation," *SOLA*, vol. 14, no. 0, pp. 132–137, 2018.

[14] J. Strandgren, L. Bugliaro, F. Sehnke, and L. Schröder, "Cirrus cloud retrieval with MSG/SEVIRI using artificial neural networks," *Atmospheric Measurement Techniques*, vol. 10, no. 9, pp. 3547–3573, Sep. 2017. [Online]. Available: https://amt.copernicus.org/articles/10/3547/2017/

[15] V. R. Meijer, L. Kulik, S. D. Eastham, F. Allroggen, R. L. Speth, S. Karaman, and S. R. H. Barrett, "Contrail coverage over the United States before and during the COVID-19 pandemic," *Environmental Research Letters*, vol. 17, no. 3, p. 034039, Mar. 2022. [Online]. Available: https://iopscience.iop.org/article/10.1088/1748-9326/ac26f0

[16] U. Schumann, R. Hempel, H. Flentje, M. Garhammer, K. Graf, S. Kox, H. Lösslein, and B. Mayer, "Contrail study with ground-based cameras," *Atmospheric Measurement Techniques*, vol. 6, no. 12, pp. 3597–3612, Dec. 2013. [Online]. Available: https://amt.copernicus.org/articles/6/3597/2013/

[17] J. Adam, J. Buss, K. Brügge, M. Nöthe, and W. Rhode, "Cloud Detection and Prediction with All Sky Cameras," *EPJ Web of Conferences*, vol. 144, p. 01004, 2017. [Online]. Available: http://www.epj-conferences.org/10.1051/epjconf/201714401004

[18] H. Mannstein, A. Brömser, and L. Bugliaro, "Ground-based observations for the validation of contrails and cirrus detection in satellite imagery," *Atmospheric Measurement Techniques*, vol. 3, no. 3, pp. 655–669, 2010.

[19] R. Chevallier, M. Shapiro, Z. Engberg, M. Soler, and D. Delahaye, "Linear Contrails Detection, Tracking and Matching with Aircraft Using Geostationary Satellite and Air Traffic Data," *Aerospace*, vol. 10, no. 7, p. 578, Jul. 2023, number: 7 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2226-4310/10/7/578

[20] U. Schumann, "A contrail cirrus prediction model," *Geoscientific Model Development*, vol. 5, no. 3, pp. 543–580, 2012.

[21] E. Riggi, T. Dubot, C. Sarrat, J. Bedouet *et al.*, "Ai-driven identification of contrail sources: Integrating satellite observation and air traffic data," *Journal of Open Aviation Science*, vol. 1, no. 2, 2023.

[22] E. Roosenbrand, J. Sun, and J. Hoekstra, "Contrail altitude estimation based on shadows detected in landsat imagery," in *13th SESAR Innovation Days*, 2023.

[23] J. Sun and E. Roosenbrand, "Flight contrail segmentation via augmented transfer learning with novel sr loss function in hough space," *arXiv preprint arXiv:2307.12032*, 2023.

[24] J. P. Hoffman, T. F. Rahmes, A. J. Wimmers, and W. F. Feltz, "The application of a convolutional neural network for the detection of contrails in satellite imagery," *Remote Sensing*, vol. 15, no. 11, p. 2854, 2023.

[25] Y. Lee, E.-K. Kim, and J. Yoo, "Towards robust contrail detection by mitigating label bias via a probabilistic deep learning model: A preliminary study," in *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, 2023, pp. 1–2.

[26] R. Teoh, U. Schumann, A. Majumdar, and M. E. Stettler, "Mitigating the climate forcing of aircraft contrails by small-scale diversions and technology adoption," *Environmental Science & Technology*, vol. 54, no. 5, pp. 2941–2950, 2020.

[27] R. Teoh, U. Schumann, and M. E. J. Stettler, "Beyond Contrail Avoidance: Efficacy of Flight Altitude Changes to Minimise Contrail Climate Forcing," *Aerospace*, vol. 7, no. 9, p. 121, Aug. 2020. [Online]. Available: https://www.mdpi.com/2226-4310/7/9/121

[28] J. Molloy, R. Teoh, S. Harty, G. Koudis, U. Schumann, I. Poll, and M. E. J. Stettler, "Design Principles for a Contrail-Minimizing Trial in the North Atlantic," *Aerospace*, vol. 9, no. 7, p. 375, Jul. 2022. [Online]. Available: https://www.mdpi.com/2226-4310/9/7/375

[29] C. Demouge, M. Mongeau, N. Couellan, and D. Delahaye, "Climate-aware air traffic flow management optimization via column generation," *Transportation Research Part C: Emerging Technologies*, vol. 166, p. 104792, 2024.

[30] ——, "A time-dependent subgraph-capacity model for multiple shortest paths and application to CO2/contrail-safe aircraft trajectories," Jun. 2023. [Online]. Available: https://enac.hal.science/hal-03900872

[31] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *arXiv:1411.4038 [cs]*, Mar. 2015, arXiv: 1411.4038. [Online]. Available: http://arxiv.org/abs/1411.4038

[32] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," May 2017, arXiv:1606.00915 [cs]. [Online]. Available: http://arxiv.org/abs/1606.00915

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," May 2015, arXiv:1505.04597 [cs]. [Online]. Available: http://arxiv.org/abs/1505.04597

[34] O. Cicek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," Jun. 2016, arXiv:1606.06650 [cs]. [Online]. Available: http://arxiv.org/abs/1606.06650

[35] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation,” Jun. 2016, arXiv:1606.04797 [cs]. [Online]. Available: http://arxiv.org/abs/1606.04797

[36] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: A Nested U-Net Architecture for Medical Image Segmentation,” Jul. 2018, arXiv:1807.10165 [cs, eess, stat]. [Online]. Available: http://arxiv.org/abs/1807.10165

[37] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” Jan. 2018, arXiv:1703.06870 [cs]. [Online]. Available: http://arxiv.org/abs/1703.06870

[38] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid Scene Parsing Network,” Apr. 2017, arXiv:1612.01105 [cs]. [Online]. Available: http://arxiv.org/abs/1612.01105

[39] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs,” Jun. 2016, arXiv:1412.7062 [cs]. [Online]. Available: http://arxiv.org/abs/1412.7062

[40] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking Atrous Convolution for Semantic Image Segmentation,” Dec. 2017, arXiv:1706.05587 [cs]. [Online]. Available: http://arxiv.org/abs/1706.05587

[41] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation,” Aug. 2018, arXiv:1802.02611 [cs]. [Online]. Available: http://arxiv.org/abs/1802.02611

[42] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, “ICNet for Real-Time Semantic Segmentation on High-Resolution Images,” Aug. 2018, arXiv:1704.08545 [cs]. [Online]. Available: http://arxiv.org/abs/1704.08545

[43] G. Lin, A. Milan, C. Shen, and I. Reid, “RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, Jul. 2017, pp. 5168–5177. [Online]. Available: http://ieeexplore.ieee.org/document/8100032/

[44] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,” Oct. 2016, arXiv:1511.00561 [cs]. [Online]. Available: http://arxiv.org/abs/1511.00561

[45] A. Kirillov, Y. Wu, K. He, and R. Girshick, “PointRend: Image Segmentation as Rendering,” Feb. 2020, arXiv:1912.08193 [cs]. [Online]. Available: http://arxiv.org/abs/1912.08193

[46] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, “Boundary loss for highly unbalanced segmentation,” Oct. 2020, arXiv:1812.07032 [cs, eess]. [Online]. Available: http://arxiv.org/abs/1812.07032