Towards Robust Contrail Detection by Mitigating Label Bias via a Probabilistic Deep Learning Model: A Preliminary Study*

Yejun Lee
Artificial Intelligence Graduate School (AIGS), Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea yejun@unist.ac.kr

Eun-Kyéong Kim†
Department of Urban Development and Mobility, Luxembourg Institute of Socio-Economic Research (LISER) Esch-sur-Alzette, Luxembourg eun-kyeong.kim@liser.lu

Jaeeun Yoo
Artificial Intelligence Graduate School (AIGS), Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea jaejun.yoo@unist.ac.kr

ABSTRACT
Contrails, formed by jet flights, alter Earth's energy balance, prompting research into monitoring contrails and developing satellite-based automated contrail detection. This demand has advanced deep learning (DL)-based techniques. However, training DL algorithms to detect contrails has limitations: class imbalance, labeling difficulty, and a lack of reliable labeled datasets. We propose a probabilistic DL approach using P-UNet to alleviate label bias in contrail detection. By observing model outputs based on two labeled datasets, OpenContrails and MIT-Contrails, we found the probabilistic approach robust against potentially biased labels.

CCS CONCEPTS
• Computing methodologies → Artificial intelligence; • Mathematics of computing → Probability and statistics; Neural networks

KEYWORDS
Contrail detection, Generative models, Probabilistic segmentation.

1 PROBLEM AND MOTIVATION
Contrails—linear ice clouds formed by jet-flight activity—exert significant radiative forcing on Earth's energy balance, surpassing those of CO2 [1]. The globally increasing demand for aviation accentuates the impact of aviation-driven contrails, making it crucial to understand contrail dynamics. Achieving this requires precise monitoring of individual contrail occurrences through satellite imagery. Recent advancements in deep learning (DL)-based computer vision have made automated contrail detection possible [2], [3].

Developing DL-based automated contrail detection methods is challenging due to (1) extreme class imbalance (contrails vs. non-contrail objects) and (2) labeling difficulties, resulting in a scarcity of reliable labeled data. Firstly, abundant class features (related to non-contrail objects) can overshadow scarce ones (contrails), diminishing their impact on the objective function during model training. Secondly, labeling contrails is challenging even for experts and trained labelers due to their rapidly changing nature and the limitations of low-resolution geostationary satellite images (e.g., 2km), which introduce spatial ambiguity and label uncertainty. These factors limit labeled datasets, leading to labeler bias based on subjective judgments and reducing label reliability. Such label bias can impair model generalizability and, therefore, should be assessed before model training.

This research assesses biases in labeled contrail datasets and proposes the use of a probabilistic model, P-UNet [5], to mitigate label bias. The assessment is exemplified over the contiguous United States. Quantitative and qualitative bias analyses are conducted on two labeled datasets: OpenContrails [4] and MIT-Contrails [2], [3], using self/cross-validation with the model developed by V. R. Meijer et al. [2].

2 RELATED WORK
Contrail labeling process. The contrail labeling process involves satellite data preprocessing, image sampling, guideline creation, labeler training, labeling, and validation. Geostationary satellites (e.g., GOES-16) are commonly used due to their higher temporal resolution, enabling dynamic contrail observation. Supervised learning approaches for contrail detection are relatively recent, and contrail labeling guidelines vary across studies [2], [4], potentially leading to inconsistencies in labeling rules.


Probabilistic U-Net. The Probabilistic U-Net (P-UNet) [5] is a generative segmentation model that combines a conditional Variational AutoEncoder (VAE) with a U-Net architecture. This model learns a distribution of masks conditioned on an input image and generates multiple masks by randomly sampling noise from the learned Gaussian distribution. Our approach treats these model-generated labels as multiple hypotheses from different experts during the labeling process. We applied P-UNet to the contrail segmentation task to experiment its robustness on different labeled datasets.

*Corresponding author: Eun-Kyéong Kim, Ph.D. (eun-kyeong.kim@liser.lu).
†This research was funded by the Cartography and Geographic Information Society (CaGIS), under CaGIS Rising grant program, and Liser (Project acronym: COVID-Contrail; Principal investigator: Eun-Kyéong Kim, Ph.D.). SIGSPATIAL ’23, November 13–16, 2023, Hamburg, Germany © 2023 Copyright held by the owner/author(s).
https://doi.org/10.1145/3589132.3628364
3 ANALYSIS

We conducted experiments that demonstrated the superior performance of leveraging multiple segmentation hypotheses generated by the probabilistic model P-UNet [5] in contrail detection when compared to the deterministic model by V. R. Meijer et al. [2]. This approach is novel in contrail detection research, as it represents the first instance of using generative segmentation models to assess model generalization across different datasets.

Experimental Setting. Two models were trained using two datasets, MIT-Contrails [2], [3], and OpenContrails [4]. Both datasets were based on GOES-16 satellite data and transformed into images using the Ash RGB color recipe. V. R. Meijer et al. and P-UNet were trained according to their original work [2], [5], each using its specific training settings. For each model, the model weights with the highest Intersection over Union (IoU) were selected for inference. The same sets of training and test images were used for both models. Each trained model was employed in self/cross-validation to assess its performance.

4 RESULTS

Figure 1 visualizes the outcomes of self/cross-validation. While inference with the trained V. R. Meijer et al. yields a single output (a-b), P-UNet generates multiple inference outcomes by varying the number of output samples from the trained model (c-e). The result reveals three noteworthy patterns. First, V. R. Meijer et al., when trained with the 'OpenContrails' data (b), detected more contrails than when trained with the 'MIT-Contrails' data (a). Second, P-UNet, when trained with the 'OpenContrails' data (c-e), detected linear-shaped and more contrails than V. R. Meijer et al. (b). Third, when comparing the performance of (a) and (b), a notable decrease in recall was observed with the 'MIT-Contrails' data, while precision remained relatively consistent. Table 1 illustrates further inference outcomes with P-UNet on the 'OpenContrails' and 'MIT-Contrails' datasets. The results show that an increased number of output samples drawn from P-UNet decreased precision but improved recall. The primary reason for decreased precision was the increased number of false-positive pixels as well as thicker or longer contrail labels. Nevertheless, P-UNet-based inferences still retained contrail-like shapes and patterns qualitatively. For the interpretation, we considered performance measures differently from traditional DL approaches. Contrail labeling tasks harbor significant uncertainty, even domain experts might label contrail masks differently; subtle differences in the edges and shapes of contrail masks can have a substantial impact on performance measures. Therefore, we assumed that a single labeled dataset cannot fully represent the ground-truth contrails but only forms a single hypothesis of contrails. It is more critical for the contrail detection model not to miss contrails in various datasets, which is signified by the higher recall measure across multiple datasets. Hence, we conclude that the P-UNet model has the potential to be more robust and generalizable when inferring contrails from diverse, new satellite image datasets.

![Figure 1: Visual outcomes and performance measures for a single example of V. R. Meijer et al. and P-UNet.](image)

Table 1: Cross-validation results of P-UNet.

<table>
<thead>
<tr>
<th>Inference data</th>
<th>P-UNet (training dataset: OpenContrails)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
</tr>
<tr>
<td>OpenContrails</td>
<td>0.5016</td>
</tr>
<tr>
<td>12 samples</td>
<td></td>
</tr>
<tr>
<td>MIT-Contrails</td>
<td>0.2856</td>
</tr>
<tr>
<td>12 samples</td>
<td></td>
</tr>
<tr>
<td>MIT-Contrails</td>
<td>0.2602</td>
</tr>
<tr>
<td>50 samples</td>
<td>(-0.0254)</td>
</tr>
</tbody>
</table>

REFERENCES