https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

ContrailStruct30: a 30 m three-band thermal-infrared dataset for contrail segmentation

Jianing Yu[1,2,4], Lixing Zhao[2], Xiaoxuan Zhou[2], Liyuan Li[3], Xinyue Ni[2], Yuqing Xue[1,4], Kefang Wang[1,4], Ying Li[1,4], Guanzheng Du[1,4], Hanwei Lu[1,4], Chaoshuai Tan[1,4], Mingxin Dai[2,4], Zhuoyue Hu[2], Xiaoyan Li[1], Fansheng Chen[1,2,3,4]

1 Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China

2 State Key Laboratory of Infrared Physics, Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China

3 Shanghai Frontiers Science Center of Smart Optoelectronics and Perception, Institute of Optoelectronics, Fudan University, Shanghai 200433, China

4 University of Chinese Academy of Sciences, Beijing 100049, China

Correspondence: Zhuoyue Hu (huzhuoyue@mail.sitp.ac.cn), Xiaoyan Li (lixiaoyan@ucas.ac.cn), and Fansheng Chen (cfs@mail.sitp.ac.cn)

Abstract

Aircraft contrails are an important source of aviation's non-CO2 climate effects, and fine-scale satellite observations provide a means of characterizing their spatial structure. Distinguishing narrow, adjacent, and overlapping contrails from complex backgrounds requires reference data with explicit pixel-level annotations. We present ContrailStruct30, a three-band thermal-infrared contrail segmentation dataset derived from Sustainable Development Science Satellite 1 (SDGSAT-1) observations at 30 m nominal sampling. The dataset contains 19,455 patches of 256 × 256 pixels from 819 source-scene entries acquired on 305 dates between August 2022 and September 2024, including 9,023 positive patches and 10,432 deliberately selected difficult negatives. Each sample provides an enhanced RGB image, corresponding three-band source values stored as uint16, and a human-corrected binary mask. Construction began with 300 manually annotated patches identified through ADS-B-assisted searches and continued through four rounds of model-assisted annotation, manual correction, and expert review. Fixed training, validation, and test partitions separate both source scenes and acquisition dates. As an application baseline, we developed a MaxViT–U-Net segmentation model and compared it with three public models under a shared training protocol to evaluate foreground recovery, centerline agreement, and difficult-background discrimination. Across three training seeds, the baseline achieved a mean positive-patch test Dice of 0.8230 ± 0.0056 (mean ± standard deviation), with strong foreground overlap and centerline agreement in the common evaluation. By combining fine-scale thermal-infrared observations, diverse contrail morphologies, expert-reviewed labels, and corresponding image representations, ContrailStruct30 provides a reusable resource for extracting fine contrail structures and developing and evaluating high-resolution segmentation methods. The dataset presented in this study corresponds to Version 3, which is publicly available via Figshare at https://doi.org/10.6084/m9.figshare.33435496.v3 (Yu et al., 2026).

1 Introduction

Aircraft contrails are ice clouds produced by aviation whose formation and evolution depend on ambient temperature, humidity, and atmospheric dynamics. When hot, moist aircraft exhaust mixes with cold ambient air, liquid droplets can form and freeze into ice crystals under suitable conditions, producing an initial contrail. Contrails may then dissipate rapidly or persist in ice-supersaturated air, spreading and deforming under wind shear and turbulence to develop into more extensive contrail cirrus (Schumann and Heymsfield, 2017; Kärcher, 2018). These anthropogenic ice clouds alter Earth's longwave and shortwave radiation

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

budgets, with their net effect depending on cloud optical properties, background conditions, and the diurnal cycle. In the global mean, contrail cirrus is an important warming component of aviation's non-CO2 climate effects. Observations of contrail distribution and morphology therefore underpin understanding of aviation's atmospheric impacts (Burkhardt and Kärcher, 2011; Lee et al., 2021).

Satellite remote sensing enables contrail observations across diverse geographical regions and atmospheric conditions. Thermal-infrared observations are independent of solar illumination and can exploit radiative contrasts between targets and their backgrounds for both daytime and nighttime identification. Contrails, however, do not always appear as distinct, isolated lines. Narrow targets may occupy only a few pixels, diffuse margins may merge with natural cirrus, and multiple contrails may lie close together, intersect, or overlap. Natural cloud edges and linear surface features can also exhibit similar local geometries. Fine-scale contrail observations thus require more than determining whether a target is present: they also require delineating its discernible extent, continuity, and boundary with the surrounding background.

Thermal-infrared contrail detection has progressed from image processing based on spectral differences and linear geometry to deep learning applied to large satellite archives (Mannstein et al., 1999; Meijer et al., 2022; Ng et al., 2024). As the task extends to pixel-level segmentation, reference labels must likewise move beyond image-level presence to describe spatial structure explicitly. This involves tracing discernible foreground, preserving background gaps between adjacent structures, and handling intersections, occlusions, and diffuse margins. For targets occupying a small fraction of an image, omitting a short segment or incorrectly joining adjacent structures can substantially change their geometry even when few pixels are affected. Representative difficult negatives are also needed to assess whether models reject backgrounds that resemble contrails. Target morphology, background composition, and annotation detail therefore jointly determine what models can learn and which errors an evaluation can reveal (Zhu et al., 2017; Ma et al., 2019; Zlateski et al., 2018).

Existing contrail datasets provide complementary resources based on different observing systems and annotation tasks (Table 1). OpenContrails combines GOES-16 ABI image sequences, pixel annotations, and flight-density information to support detection using temporal context (Ng et al., 2024). Human-labeled Landsat-8 data extend reference observations to multispectral imagery with finer spatial sampling (McCloskey et al., 2021). The MSG/SEVIRI dataset provides three independent annotations per scene alongside satellite channels and cloud properties, enabling analyses of cloud conditions and interpretation differences (Santos Gabriel et al., 2026). The ground-based visible-camera dataset GVCCS supports contrail segmentation and tracking through image sequences and temporally consistent instance identifiers, together with model application examples (Jarry et al., 2026). These resources have advanced contrail observation and automated identification, although the detail they represent depends on their spatial sampling, imaging conditions, and annotation protocols. In kilometer-scale thermal-infrared imagery in particular, mixed pixels can obscure narrow contrails and gaps between adjacent structures. This motivates finer-scale imagery and pixel-level references that retain local contrail morphology.

SDGSAT-1 provides three spatially registered thermal-infrared bands at 30 m nominal sampling, enabling detailed observations of contrails over diverse land, ocean, and cloud backgrounds (Hu et al., 2022; Guo et al., 2025). Under suitable imaging conditions, these observations reveal narrow cores, adjacent parallel structures, local gaps, and intersections, providing material for studying compact morphologies and identifying candidate early-stage contrails. Three-band combinations and enhancement also translate

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

interband radiative differences into color and texture contrasts that aid interpretation. In the enhanced representation used here, for example, some contrails appear brownish whereas some natural cirrus backgrounds appear darker; these contrasts can be interpreted together with linear geometry and full-scene context. Such colors depend on processing and scene conditions and are therefore used alongside structural evidence. The earlier SDGCONTR dataset addressed image-level recognition and classification in SDGSAT-1 imagery (Yu et al., 2024). Fine-scale morphological analysis and segmentation development require a large-scale resource that organizes these observations with pixel-level masks.

We therefore present ContrailStruct30, a dataset for three-band thermal-infrared contrail segmentation. It contains 19,455 patches spanning diverse contrail geometries and difficult-negative backgrounds, each with an enhanced RGB image, corresponding three-band uint16 source values, and a human-corrected binary mask. We describe dataset construction, annotation, and composition; reassess labels using source-scene context; and provide fixed training, validation, and test partitions that separate both source scenes and acquisition dates. We also develop a MaxViT–U-Net segmentation baseline and compare it with three public models under a shared training protocol to evaluate foreground recovery, fine-structure segmentation, and difficult-background discrimination. A controlled experiment using the corresponding RGB and uint16 images illustrates the dataset's use in preprocessing research. Together, the fine-scale thermal-infrared observations, expert-reviewed pixel labels, and companion baseline provide a reusable basis for automated extraction and evaluation of contrail structure.

Table 1. Complementary contrail reference datasets. Spatial sampling refers to the source sensor; the detail represented by labels also depends on preprocessing and annotation protocols. Dataset sizes follow the cited publications.

Dataset	Observations	Annotated data	Annotation	Context
OpenContrails (Ng et al., 2024)	GOES-16 ABI; 2 km infrared sampling at nadir	20,544 training and 1,866 validation samples	Four annotators; pixel-level consensus	Eight-frame sequences and advected flight density
Landsat-8 (McCloskey et al., 2021)	30 m cirrus band; 100 m native thermal-infrared bands	4,289 scenes	Manually drawn contrail polygons	Multispectral imagery and flight-density information
MSG/SEVIRI (Santos Gabriel et al., 2026)	SEVIRI; 3 km nominal sampling	140 scenes	Three independent annotations per scene; majority consensus	Multitemporal imagery, satellite channels, and cloud properties
GVCCS (Jarry et al., 2026)	Visible whole-sky camera	122 video sequences	Instance polygons; technical and expert review	Image sequences and flight-trajectory overlays
SDGCONTR (Yu et al., 2024)	SDGSAT-1 TIS	Image patches for classification	Image-level class labels	Enhanced three-band imagery
ContrailStruct30 (this study)	SDGSAT-1 TIS; 30 m nominal sampling	19,455 patches	Model-assisted delineation, manual correction, and expert review	Source-scene interpretation and corresponding uint16 images

2 Observations and dataset construction

2.1 Source observations and sample selection

ContrailStruct30 was constructed from observations acquired by the Thermal Infrared Spectrometer (TIS) aboard Sustainable Development Science Satellite 1 (SDGSAT-1). The instrument provides three spatially registered thermal-infrared bands spanning 8.0–10.5, 10.3–11.3, and 11.5–12.5 μm, with 30 m nominal

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



sampling and a 300 km swath (Hu et al., 2022; Guo et al., 2025). The imagery used here was acquired between 26 August 2022 and 4 September 2024 and covers diverse land, ocean, and cloud backgrounds. Figure 1 illustrates the observational setting and representative backgrounds.

Sample selection sought to cover diverse contrail morphologies and potentially confounding backgrounds. Initially, historical Automatic Dependent Surveillance–Broadcast (ADS-B) trajectories obtained from ADSB.lol helped locate aircraft and associated candidate structures. Interpretation of these candidates in full-scene imagery produced an initial set of 300 manually annotated patches. Subsequent downloads prioritized source scenes traversed by aircraft flying above 7,000 m, and manual inspection and model screening identified regions for annotation. ADS-B guided scene searches and interpretation, whereas image evidence and annotation rules determined the delineated contrail extent.

All samples were cropped to 256 × 256 pixels, approximately 7.68 km × 7.68 km at the nominal sampling interval. Extraction combined overlapping grid crops with a 128-pixel stride and patch selection within approximately 1024-pixel neighborhoods around candidate targets. This covered distinct linear targets, adjacent or intersecting structures, natural cloud backgrounds, and regions that repeatedly triggered false detections. For locally ambiguous features, annotators could examine the original full scene and consult ADS-B trajectories and ERA5 data (Hersbach et al., 2020).

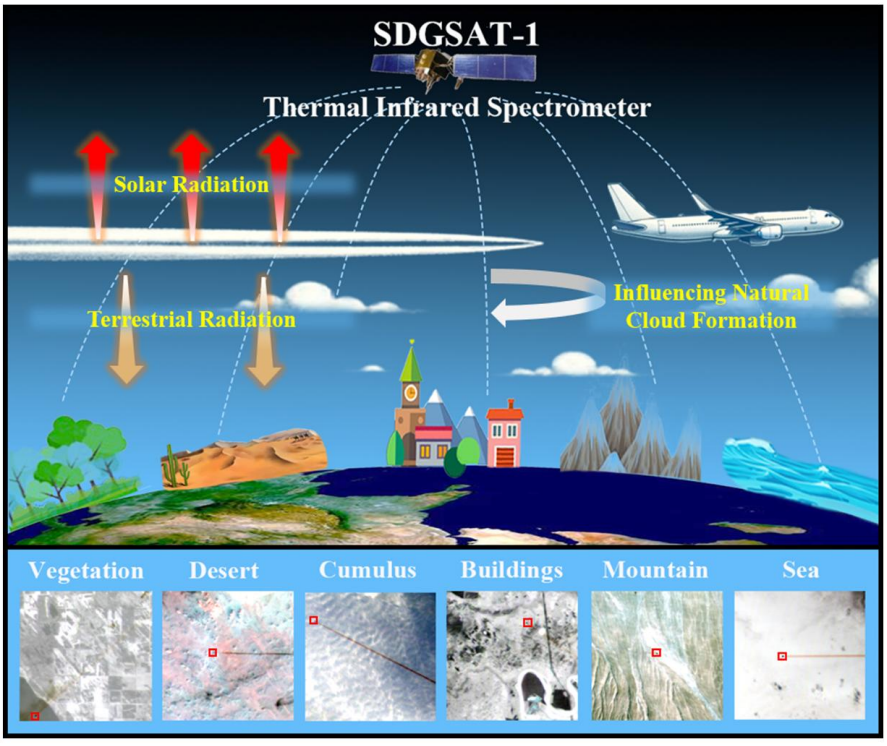


Figure 1. SDGSAT-1 TIS observational setting and representative image backgrounds. The upper schematic illustrates the satellite, aircraft-induced cloud structures, and their environment; the lower panels show vegetation, desert, cloud, built-up, mountain, and ocean backgrounds. The figure provides observational context; quantitative analyses in this study concern dataset patches, reference masks, and segmentation predictions.

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

2.2 Image enhancement and corresponding representations

Each sample provides an enhanced three-channel RGB image and corresponding three-band source values on the same pixel grid, sharing one reference mask. Enhanced images are stored as 8-bit PNG files for visual interpretation and model training, whereas source-value images are stored as uint16 TIFF files to support alternative channel transformations and normalization methods. The uint16 designation describes the storage type; the instrument uses 12-bit quantization (Hu et al., 2022).

Enhancement follows the three-band preprocessing approach of Yu et al. (2024) and is applied after patch extraction. All means and percentiles are calculated within the individual 256 × 256-pixel patch. First, zero values in each band are replaced with the mean of its nonzero pixels.

Let source band c be D_c, where $c \in \{1,2,3\}$, and let D'_c denote the band after zero replacement. Q_q denotes the qth percentile of the relevant variable within the current patch. Linear stretching between the minimum and 95th percentile of the filled band produces the initial display channel:

$$S_c = \mathrm{uint8}\left(255\mathrm{clip}\left[\frac{D'_c - Q_0(D'_c)}{Q_{95}(D'_c) - Q_0(D'_c)}, 0,1\right]\right). \quad (1)$$

Here, clip restricts values to [0, 1], and uint8 denotes conversion to an 8-bit unsigned integer. Bands 1, 2, and 3 are mapped to the red, green, and blue channels, respectively.

Spectral selection conditions calculated from the source values before zero replacement are then used to adjust saturation at selected pixels. The intermediate radiance and brightness-temperature variables used in the display transformation are

$$L_c = g_c D_c + b_c, \quad (2)$$

$$T_c = \frac{K_{2,c}}{\ln(1 + K_{1,c}/L_c)}. \quad (3)$$

Here, g_c and b_c are the gain and offset of the linear conversion, and $K_{1,c}$ and $K_{2,c}$ are brightness-temperature conversion constants (Table 2). Defining $R_{23} = D_2/(D_3 + 10^{-6})$ and $R_{21} = D_2/(D_1 + 10^{-6})$, the saturation-adjustment region M is the intersection of three conditions:

$$M = [T_2 < Q_{90}(T_2)] \cap [R_{23} > Q_{40}(R_{23})] \cap [R_{21} < Q_{50}(R_{21})]. \quad (4)$$

For pixels satisfying these conditions, saturation in the OpenCV HSV representation is multiplied by 1.2, clipped to the 8-bit range, and converted back to RGB. Computations use float32 source arrays and linearly interpolated percentiles; conversion to uint8 truncates fractional values. Zero replacement is used only in Eq. (1), whereas Eqs. (2)–(4) use the unfilled source bands. All statistics are calculated within the current patch.

This procedure combines local intensity variation and interband differences into an enhanced representation for interpretation. The same display-transformation coefficients are used across source-product processing categories in the dataset. These coefficients define the enhancement procedure; physical interpretation of source values follows the calibration relationships of the corresponding SDGSAT-1 products. Enhanced RGB images are used in the main segmentation experiments, and the corresponding uint16 images support a controlled input-representation comparison.

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Table 2. Bands and fixed coefficients used for the patch-local display transformation. These coefficients describe the enhancement implementation rather than a new product-level radiometric calibration.

Property	**Band 1**	**Band 2**	**Band 3**
Spectral range (μm)	8.0-10.5	10.3-11.3	11.5-12.5
Display channel	Red	Green	Blue
Gain, g_c	0.003947	0.003946	0.005329
Offset, b_c	0.167126	0.124622	0.222253
$K_{1,c}$	1655.627802	838.7063188	543.0579522
$K_{2,c}$ (K)	1542.760922	1342.718697	1232.021436

2.3 Model-assisted expansion, manual correction, and expert review

Dataset construction combined model-assisted annotation with manual review (Fig. 2). Building on the initial 300 patches, two senior experts jointly developed an annotation handbook covering target interpretation, boundary delineation, separation of adjacent structures, and treatment of complex cloud backgrounds. The sample pool was then expanded through four iterations. In each round, accepted samples were used to train a screening model that generated candidate masks for previously unseen scenes. Annotators corrected these masks, and reviewed samples entered the pool for the next round. This is a model-assisted annotation process centered on human interpretation and correction (Budd et al., 2021).

The screening network used for dataset construction combined a MaxViT-Base encoder with a U-Net-style decoder and a training objective comprising weighted binary cross-entropy, Tversky, and clDice losses (Ronneberger et al., 2015; Salehi et al., 2017; Shit et al., 2021; Tu et al., 2022). Full-scene inference used reflect-padded 256 × 256-pixel windows with 50% overlap. Two-dimensional Hann weights blended overlapping probability maps, and a threshold of 0.50 produced preliminary masks. This network assisted candidate discovery and delineation and was used separately from the segmentation baseline trained and evaluated on the final fixed partitions in Sect. 5.

Five annotators shared the correction work, with one annotator responsible for each patch. They consulted the enhanced patch, original full-scene image, and annotation handbook to assess whether candidate structures had coherent contrail morphology and could be distinguished from surrounding natural clouds or surface features. Depending on the interpretation, annotators retained candidate masks, adjusted boundaries, separated neighboring regions, removed false detections, or redrew targets. Cases with uncertain target identity or boundaries were referred to the experts.

After correction in each round, the two experts reviewed a random 20% of samples and all cases flagged as uncertain. Review focused on target identification, continuity of linear structures, gaps between adjacent targets, and boundary placement. Ambiguous cases were discussed jointly. Accepted labels proceeded to subsequent training and dataset assembly. Thus, the model generated candidates and initial delineations, while manual interpretation and expert review determined the final reference labels.

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

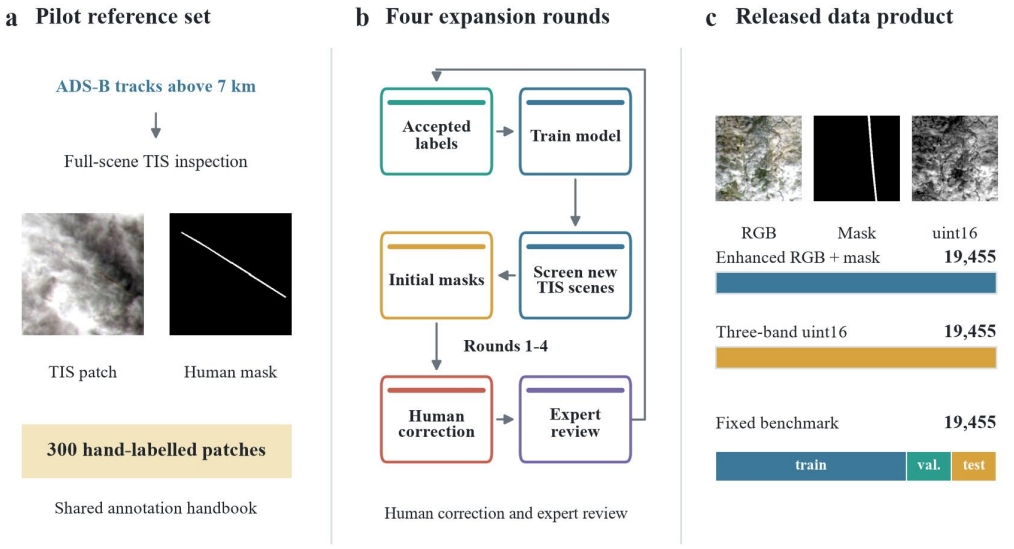


Figure 2. Construction of ContrailStruct30. (a) ADS-B-assisted source-scene interpretation supports the initial 300-patch sample set and a common annotation handbook. (b) Four expansion rounds link model training, screening of new scenes, manual correction, and expert review; accepted labels enter the next round. (c) Each final sample contains an enhanced RGB image, binary reference mask, and corresponding three-band uint16 image and belongs to a fixed training, validation, or test partition.

2.4 Annotation targets and difficult negatives

ContrailStruct30 uses binary semantic segmentation labels: regions interpreted as contrails are foreground, and all other regions are background. Target identification considers a recognizable straight or smoothly curved axis, coherent image features along that axis, and separability from the surrounding cloud field. Targets include narrow cores, parallel structures, discernible discontinuous segments, and mildly diffuse regions that retain clear axial organization. Table 3 summarizes common situations and their treatment.

Masks follow the target extent discernible in the imagery. Adjacent contrails remain separate where a background gap is clearly visible; intersecting or overlapping structures are labeled as the union of their discernible foreground. Partially occluded segments are connected only when their continuity can be established clearly; otherwise, the background gap is retained. Connected components therefore describe spatial connectivity within the annotation rather than individual aircraft or physical contrails.

The two experts jointly selected difficult negatives from regions prone to human misinterpretation or repeated model false detections. These include natural cirrus, elongated cloud edges, linear surface boundaries, and image artifacts. Samples were assessed using the same interpretation rules as positive patches and assigned empty masks. Their inclusion makes recovery of contrail structure and rejection of similar backgrounds joint objectives of training and evaluation.

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



Table 3. Interpretation and delineation rules for binary reference masks.

Image condition	Interpretation criterion	Mask treatment
Narrow straight or curved structure	Coherent signal along a recognizable axis	Trace discernible foreground, including narrow segments
Parallel or separated features	Distinct foreground regions and intervening background can be identified	Preserve separating gaps
Discontinuous or partly occluded structure	Visible segments and scene context jointly support continuity assessment	Retain gaps unless continuity across an occlusion is clear
Intersecting or overlapping features	Multiple axes form discernible foreground	Label their union as semantic foreground
Diffuse margins around a clear core	Axial organization and separation from the background remain discernible	Include discernible margins
Natural cirrus or broad cloud boundary	The surrounding cloud field explains the feature	Retain as background
Surface feature or image artifact	Orientation follows surface context or processing geometry	Retain as background
Ambiguous target identity or margin	More than one interpretation is plausible	Joint expert decision

3 Characteristics of ContrailStruct30

3.1 Spatial and temporal composition

ContrailStruct30 contains 19,455 patches from 819 source-scene entries acquired on 305 dates. Source-scene entries distinguish observations by product processing level and correspond to 808 source-product identifiers. Nominal source-scene positions range from 175.84° W to 169.22° E and from 46.21° S to 76.96° N (Fig. 3a). These coordinates describe the source observations and the geographical composition of the dataset.

Scene and patch distributions differ because observations contribute different numbers of candidate regions. December–February (DJF) includes 457 scene entries, March–May (MAM) 92, June–August (JJA) 57, and September–November (SON) 213. The corresponding patch counts are 5,684, 587, 969, and 12,215. Thus, DJF contributes the most scene entries, whereas SON contributes the most patches (Fig. 3b, c). This difference reflects targeted selection and the abundance of annotatable structures or backgrounds in particular observations. The two distributions describe sampling at different levels of the dataset.

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.

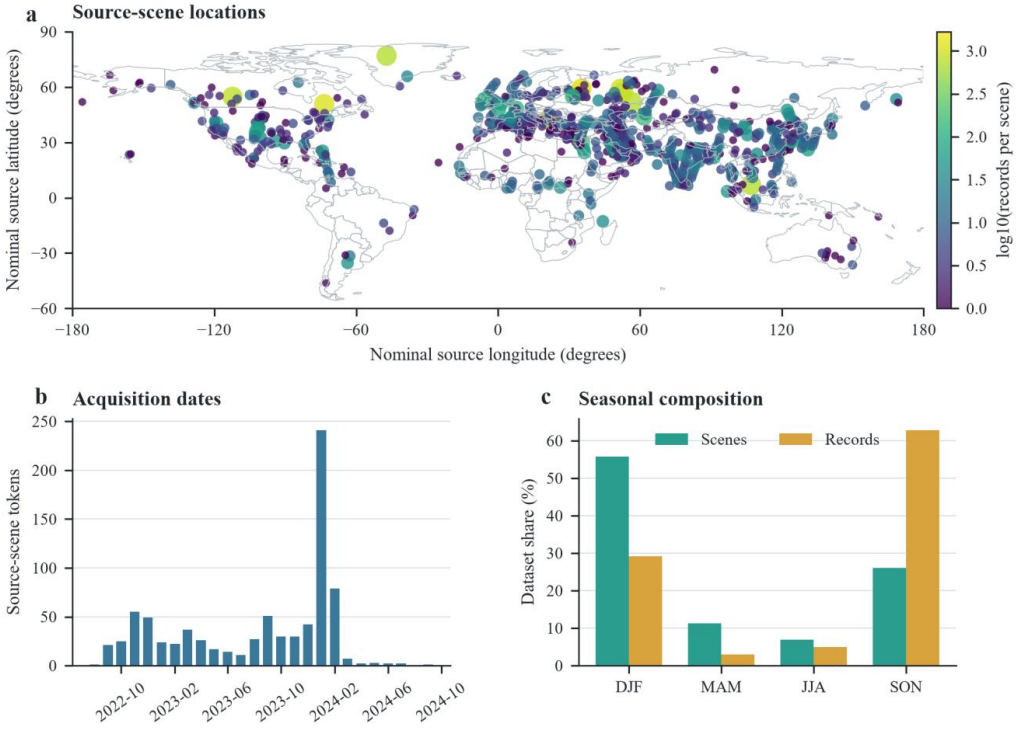


Figure 3. Spatial and temporal composition. (a) Nominal source-scene positions. Symbol area increases with the square root of the number of contributed patches, and color represents the base-10 logarithm of that count. Country outlines are from Natural Earth. (b) Monthly numbers of source-scene entries. (c) Seasonal proportions of scenes and patches. These distributions describe a purposefully assembled dataset rather than climatological contrail occurrence.

3.2 Image representations and foreground morphology

The dataset includes 9,023 positive patches and 10,432 difficult negatives. Each enhanced PNG is a three-channel uint8 image, and its corresponding TIFF contains three-band uint16 source values. The single-channel reference mask encodes foreground as 255 and background as 0. Shared pixel grids and sample identities link the two input representations to the same label. Source-scene identifiers, dates, and partition assignments describe provenance and membership, while foreground area and connected-component counts characterize reference geometry. These associations support interpretation and reuse of the images and labels (Wilkinson et al., 2016; Gebru et al., 2021).

The examples in Fig. 4 illustrate why pixel-level labels are needed. Narrow lines occupy only a small fraction of a patch but may span much of its width, and parallel features require preservation of clear background gaps. Multiple and complex foreground structures combine intersections, width variation, and spatially separated segments. Broad structures present a different delineation task, with margins interpreted against the surrounding cloud texture. These examples show geometry recorded under the handbook rules. Subsequent quantitative descriptions use foreground fraction and connectivity rather than assigning physical age from shape.

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.

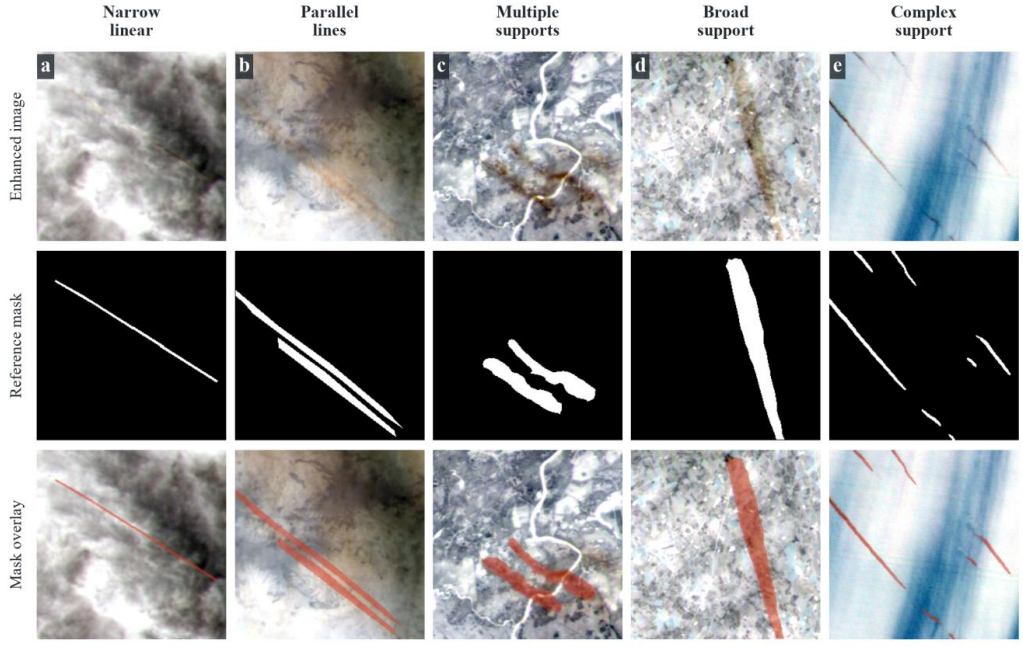


Figure 4. Representative reference geometries in ContrailStruct30. Columns show narrow linear foreground, parallel lines, multiple foreground structures, broad foreground, and complex foreground. Rows show enhanced images, binary reference masks, and overlays. Each patch contains 256 × 256 pixels, approximately 7.68 km × 7.68 km at the nominal sampling interval. Discernible gaps and intersecting foreground illustrate the semantic annotation rules.

Among positive patches, the median foreground fraction is 3.54%, with an interquartile range of 1.10%–10.85% (Fig. 5a). Many targets are therefore embedded in backgrounds much larger than themselves. This imbalance occurs within images, not merely in class frequencies: even a positive patch may contain very few contrail pixels. Both individual positive-patch metrics and pooled pixel counts are consequently needed for evaluation.

Connectivity provides a complementary description of the annotations. A single eight-connected foreground component occurs in 78.3% of positive masks, and the maximum component count is 12 (Fig. 5b). One component may comprise a broad or intersecting structure, whereas multiple components may arise from adjacent lines or discontinuous foreground. Foreground area and component count thus describe different aspects of reference geometry. Together with deliberately selected difficult negatives, they cover target delineation and target–background discrimination.

3.3 Fixed evaluation partitions

The training, validation, and test sets contain 13,213, 3,120, and 3,122 patches, respectively (Table 4; Fig. 5c). All patches from a source scene belong to one partition, and each acquisition date occurs in only one partition. Date grouping also keeps observations represented at different processing levels on the same date within the same partition. This design addresses the risk of distributing related observations across partitions when patches are split randomly (Roberts et al., 2017).

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

The test set includes 1,422 positive patches and 1,700 difficult negatives from 120 source-scene entries acquired on 51 dates. Positive test patches are stratified by foreground fraction: 318 patches at ≤1%, 488 at >1% and ≤5%, and 616 at >5% (Fig. 5d). These groups provide a direct basis for comparing sparse and larger-area targets. Both image representations share partitions and reference masks, allowing input preprocessing to change while sample content remains fixed.

Table 4. Composition of the fixed partitions. Source-scene entries and acquisition dates are assigned as groups, and every sample contains both image representations.

Partition	Patches	Positive	Difficult negative	Scene entries	Dates
Training	13,213	6,243	6,970	580	203
Validation	3,120	1,358	1,762	119	51
Test	3,122	1,422	1,700	120	51
Total	19,455	9,023	10,432	819	305

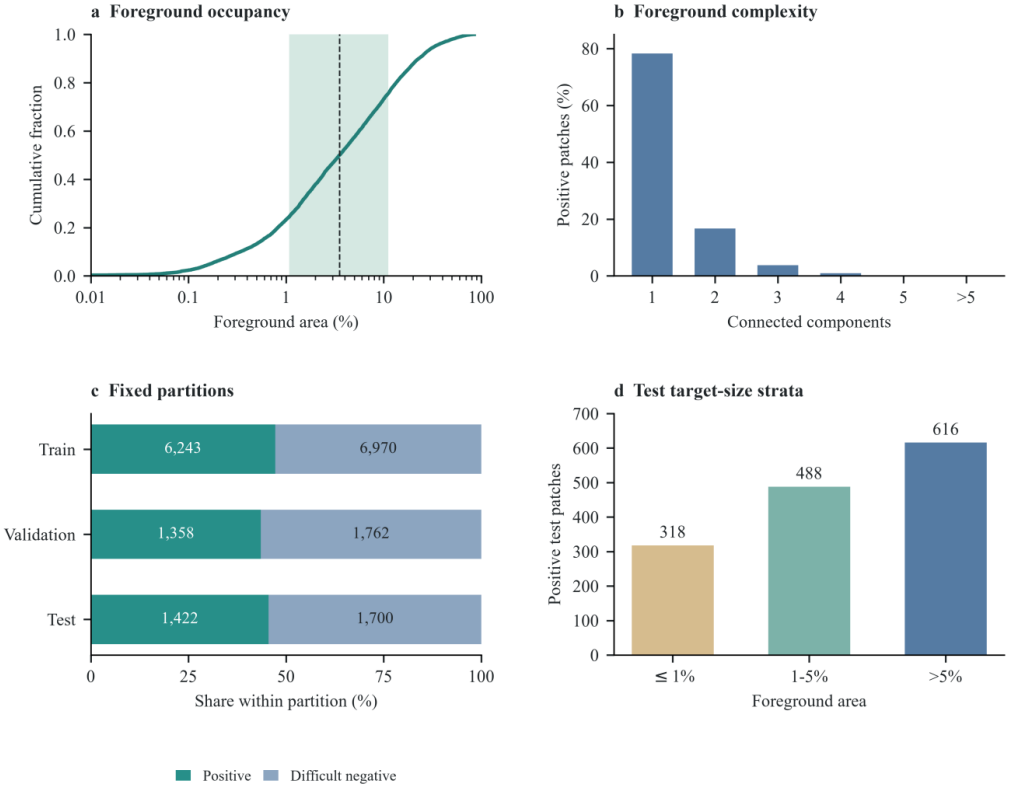


Figure 5. Foreground geometry and evaluation composition. (a) Cumulative distribution of foreground fraction for 9,023 positive patches; shading denotes the interquartile range and the dashed line the median. (b) Number of eight-connected foreground components per positive patch, with counts above five grouped together. (c) Proportions of positive patches and difficult negatives in each partition; numbers inside bars give sample counts. (d) Positive test-patch counts in the three foreground-area strata.

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



4 Quality assessment

4.1 Contextual reassessment design

Two additional experts who had not participated in the original annotation reassessed 100 patches using the common handbook and recoverable source-scene context. Sampling used seed 20260909, selecting one patch per source-scene entry and covering 94 distinct source-product identifiers. A coverage group contained 60 cases: 10 positive patches from each of the three foreground-area strata and 30 difficult negatives. A challenge group contained 40 cases, with 10 each selected for screening disagreement, multiple connected components, thin lines or low foreground fraction, and broad structures or high foreground fraction. Challenge cases were selected first; coverage samples were then drawn from the remaining source-scene entries. This mixed design examined a range of interpretation conditions, including difficult cases.

The first expert initially viewed the enhanced patch, its position in the full scene, and a local source-image neighborhood with reference masks hidden, recording the target as present, absent, or uncertain. Mask-visible inspection then covered all initially present or uncertain cases and all initially absent cases with nonempty reference masks. Cases initially judged absent with empty reference masks were recorded as concordant in the first stage. The second expert inspected every initially uncertain or presence-discordant case and a random 20% of the remaining concordant cases, rounding the sample size upward. This expert could see the masks but not the first expert's responses.

Initial agreement compared determinate presence judgments with whether reference masks contained foreground; uncertain judgments were counted separately. Subsequent acceptance proportions used only the masks actually inspected by each expert as denominators. Because the coverage and challenge groups served different sampling purposes, their results are reported separately, consistent with the principle that accuracy interpretations should match the sampling design (Olofsson et al., 2014). Segmentation experiments used the reference masks in the fixed dataset.

4.2 Interpretation agreement and reassessment examples

The first expert initially judged 59 cases as present, 31 as absent, and 10 as uncertain (Table 5; Fig. 6g). Of the 90 determinate judgments, 77 agreed with reference-mask foreground presence, corresponding to 85.6% agreement among determinate cases. This proportion describes interpretation agreement in the mixed reassessment sample, not label accuracy across the dataset. Agreement was 48 of 54 determinate cases in the coverage group and 29 of 36 in the challenge group. The 13 disagreements comprised seven cases initially judged present with empty reference masks and six initially judged absent with nonempty reference masks.

The two disagreement types reflect different interpretation challenges. The seven empty-mask cases were subsequently accepted as background under the handbook criteria. The six nonempty-mask cases were accepted as foreground after attention was directed to the labeled locations. The annotation team's retrospective explanation linked the former to reinterpretation of cloud or background features and the latter to small targets overlooked during initial inspection. Figure 6a–f shows one actual example of each type, placing the enhanced patch and reference overlay alongside the source-image neighborhood.

The first expert accepted all 75 masks requiring second-stage inspection; the remaining 25 cases had already shown agreement between an absent judgment and an empty mask in the first stage. The second expert accepted all 39 inspected masks, comprising the 23 initially discordant or uncertain cases and 16 sampled

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.

concordant cases. All 100 reference labels were ultimately retained. These results separately document initial interpretation differences and subsequent acceptance in context.

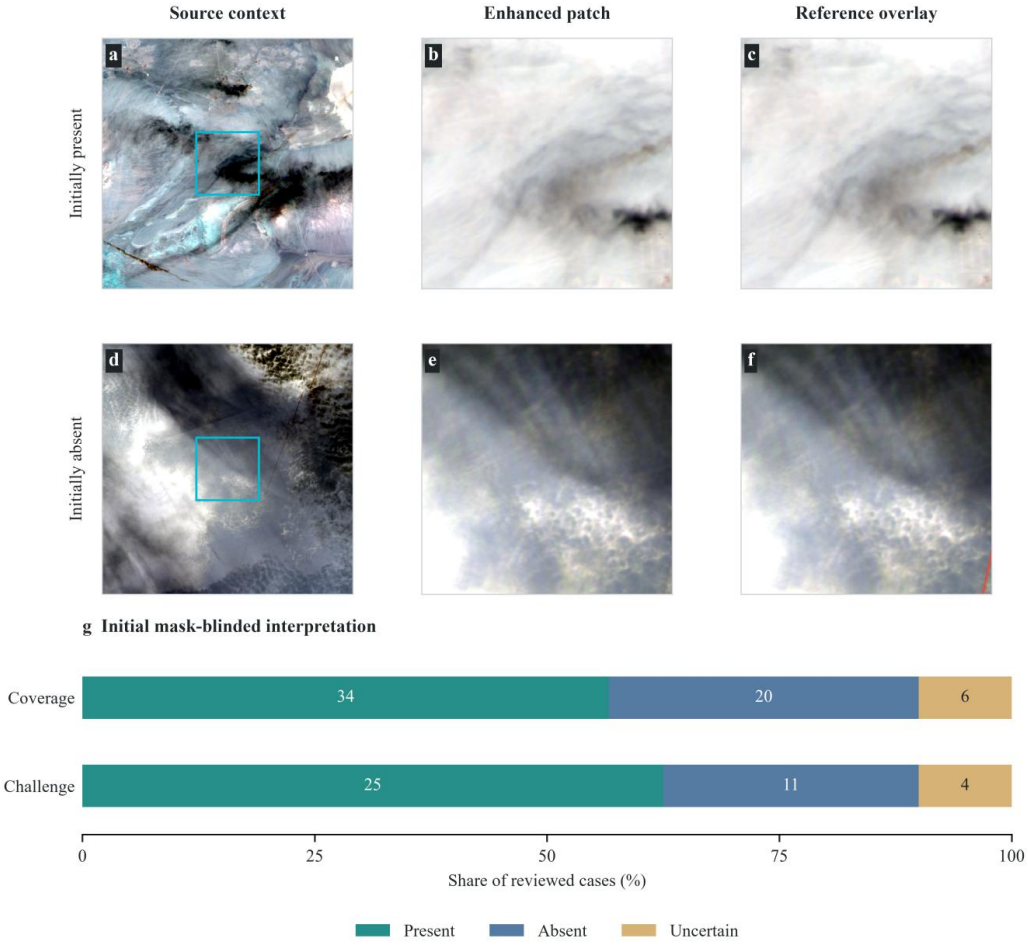


Figure 6. Source-scene context and initial expert judgments. (a–c) A case initially judged present with an empty reference mask and subsequently accepted as background. (d–f) A case initially judged absent with a nonempty reference mask and subsequently accepted as foreground. Columns show source-scene context, the enhanced patch, and the reference-mask overlay. Cyan boxes mark patch windows, and coral regions indicate reference foreground. Source-image neighborhoods are displayed using per-band 2nd–98th percentile stretching and represent another visualization of the same observation. (g) Initial mask-blinded judgment distributions in the coverage and challenge groups; numbers inside bars give case counts.

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Table 5. Contextual reassessment results by sampling group. Reference masks were hidden during initial interpretation. Subsequent acceptance results come from mask-visible inspection, with denominators restricted to cases actually inspected at that stage.

Outcome	Coverage	Challenge	Total
Reassessed patches	60	40	100
Initially present	34	25	59
Initially absent	20	11	31
Initially uncertain	6	4	10
Agreement among determinate judgments	48/54	29/36	77/90
Initially present; empty reference	5	2	7
Initially absent; nonempty reference	1	5	6
Expert 1 acceptance after mask-visible inspection	41/41	34/34	75/75
Expert 2 acceptance after mask-visible inspection	22/22	17/17	39/39
Reference labels ultimately retained	60	40	100

5 Contrail segmentation baseline and performance evaluation

To support use of ContrailStruct30 and future method comparisons, we evaluated a MaxViT–U-Net contrail segmentation baseline alongside three public segmentation models under a shared training and validation-selection protocol. Each model was evaluated over three training seeds. The experiments examine recovery of narrow, sparse, and complex structures together with false detections in difficult backgrounds. A separate input-representation experiment illustrates the use of corresponding RGB and uint16 images for preprocessing research.

5.1 MaxViT–U-Net segmentation baseline

Contrails often form elongated, locally discontinuous, or overlapping structures, requiring segmentation methods to recognize local texture while using broader spatial relationships. We use MaxViT as the feature encoder and a U-Net-style progressive upsampling decoder to construct a baseline for three-band enhanced imagery (Ronneberger et al., 2015; Tu et al., 2022). The baseline combines multiscale feature extraction and pixel-level prediction using existing public architectural components.

The encoder is timm's maxvit_base_tf_512.in21k_ft_in1k, initialized with weights pretrained on ImageNet-21K and fine-tuned on ImageNet-1K. The segmentation network extracts the stem and first three stages, yielding features at 1/2, 1/4, 1/8, and 1/16 of the input resolution with 64, 96, 192, and 384 channels, respectively. The decoder has four upsampling blocks; the first three fuse encoder features at the corresponding scales through skip connections. Each block applies two 3 × 3 convolutions with batch normalization and ReLU. Decoder output channels are 384, 192, 96, and 64, and a final 3 × 3 convolution produces single-channel logits at the input resolution. A sigmoid converts these logits into foreground probabilities. The network contains 83,947,541 trainable parameters and follows a dense pixel-prediction formulation (Long et al., 2015).

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

The baseline takes enhanced RGB patches of 256 × 256 pixels, scaled by division by 255, and fine-tunes all parameters. The same training procedure is applied to all four models: unweighted binary cross-entropy plus soft Dice loss, AdamW with an initial learning rate of 2×10^{-4} and weight decay of 10^{-4}, and batch size 12 (Loshchilov and Hutter, 2019). Augmentation consists of horizontal and vertical flips applied jointly to images and masks (Shorten and Khoshgoftaar, 2019). Training uses mixed precision for at most 24 epochs, with early stopping after six consecutive epochs without improvement in validation loss. ReduceLROnPlateau halves the learning rate when validation loss plateaus, with patience set to one. The checkpoint with the lowest validation loss is retained. All four models use seeds 42, 123, and 2025.

5.2 Comparison models and evaluation protocol

U-Net–ResNet34, DeepLabV3+–ResNet50, and SegFormer-B0 represent commonly used encoder–decoder, atrous-convolution, and Transformer segmentation approaches (Ronneberger et al., 2015; He et al., 2016; Chen et al., 2018; Xie et al., 2021). All three use public implementations from segmentation-models-pytorch 0.5.0 and encoders pretrained on ImageNet-1K (Deng et al., 2009). They share the partitions, RGB/255 inputs, augmentation, loss, optimizer, training budget, and checkpoint-selection rule described in Sect. 5.1 with MaxViT–U-Net. This controls the training procedure while retaining the architectures and their available pretrained encoders. MaxViT uses ImageNet-21K pretraining followed by ImageNet-1K fine-tuning; parameter counts also differ (Table 7).

All segmentation thresholds are selected on validation data using the same rule. Thresholds from 0.01 to 0.95 in steps of 0.01 are ranked by mean positive-patch validation Dice, subject to a negative-patch activation rate of at most 5%. If no threshold satisfies the constraint, a prespecified fallback maximizes validation Dice over the full scanned range; ties favor lower negative activation and then the higher threshold. All 12 runs required this fallback. Thresholds are fixed before test evaluation, without test-based seed selection or test-time augmentation. For the prespecified seed-42 visualizations, thresholds are 0.34, 0.30, 0.24, and 0.30 for MaxViT–U-Net, U-Net, DeepLabV3+, and SegFormer, respectively.

The test set contains 3,122 patches from 51 acquisition dates: 1,422 positives and 1,700 difficult negatives. The primary metric is mean positive-patch Dice, calculated within each positive patch and then averaged arithmetically. Mean positive-patch IoU is calculated independently in the same way. Micro Dice, IoU, precision, and recall instead use true-positive, false-positive, and false-negative pixel counts pooled across all test patches (Taha and Hanbury, 2015; Reinke et al., 2024; Maier-Hein et al., 2024). Foreground fractions of ≤1%, >1% and ≤5%, and >5% define strata for assessing sparse and larger-area targets. Each metric is calculated separately for each fitted model; the main comparison reports the arithmetic mean of the three seed-specific estimates, not an ensemble prediction.

Linear-structure evaluation uses clDice and reference-centerline recall. The latter is the fraction of reference skeleton pixels covered by predicted foreground. clDice is the harmonic mean of this fraction and the fraction of predicted skeleton pixels falling within reference foreground (Shit et al., 2021). Both metrics are calculated per positive patch and averaged, assigning zero to empty predictions. Difficult-negative activation is the fraction of the 1,700 empty-reference patches containing at least one predicted foreground pixel; the negative-pixel false-positive rate instead uses all pixels in these patches as its denominator. Pixel-level precision–recall curves are calculated using 4,096 probability bins.

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



Percentile 95% confidence intervals are obtained from 2,000 paired acquisition-date-clustered bootstrap resamples. Each resample draws 51 dates with replacement and retains all patches from the selected dates, including multiplicities. The same date draws are applied across models and seeds. Metrics are calculated separately for each seed within a resample and then averaged across the three seeds; paired model differences are calculated from these averages (Efron, 1979; Field and Welsh, 2007). These intervals describe test-date sampling variation conditional on the fitted runs. Across-seed sample standard deviations are reported separately. A sensitivity analysis uses 5,000 paired resamples of the 120 source-scene entries. Intervals are pointwise and are not adjusted for multiple comparisons. Common evaluation uses PyTorch 2.5.1, timm 1.0.27, and float16 CUDA inference.

5.3 Comparison with other segmentation models

Under the shared protocol, MaxViT–U-Net had the highest mean positive-patch Dice, at 0.8230, compared with 0.7978 for U-Net, 0.7798 for DeepLabV3+, and 0.7903 for SegFormer (Table 6; Fig. 7). The corresponding micro Dice values were 0.8479, 0.8112, 0.8008, and 0.8155. MaxViT–U-Net also had the highest point estimates of mean positive-patch IoU and clDice, at 0.7399 and 0.8775, respectively. Its positive-patch Dice advantages over the three comparison models were 0.0252, 0.0432, and 0.0327, with paired date-clustered 95% intervals of 0.0183–0.0406, 0.0350–0.0604, and 0.0254–0.0531. These results support stronger foreground recovery by the tested baseline under the common training procedure.

Differences were also apparent for sparse targets. For patches with foreground occupying at most 1% of the image, the three-seed mean positive-patch IoU was 0.649 for MaxViT–U-Net, compared with 0.602, 0.553, and 0.563 for U-Net, DeepLabV3+, and SegFormer (Fig. 7c). Above 5% foreground, the corresponding values were 0.800, 0.778, 0.763, and 0.776. All models performed less well on small-area targets, and MaxViT–U-Net retained the highest point estimate in each of the three tested strata.

Figure 8 compares reference masks and predictions for four fixed test patches representing a narrow target, approximately parallel components, complex foreground, and a texture-rich negative. The cases were selected from reference geometry before prediction inspection, using seed 20260914, and come from different source scenes. The same cases are shown for every model, using the prespecified training seed 42 rather than the best-performing test run.

At their validation-selected operating points, MaxViT–U-Net also had the lowest three-seed mean difficult-negative activation, at 34.57%, compared with 44.24% for U-Net, 53.90% for DeepLabV3+, and 53.45% for SegFormer. Its negative-pixel false-positive rate was 0.272%, compared with 0.614%, 0.619%, and 0.542%, respectively. Thus, the baseline combined stronger foreground recovery with lower average background activation in this comparison. Nevertheless, more than one-third of difficult-negative patches contained at least one predicted pixel on average, showing that a low false-positive pixel fraction does not eliminate the burden of candidate-image screening.

All four models recovered the principal narrow segment in Fig. 8a, with local differences in width and extent. In the parallel-structure example (Fig. 8b), they recovered the broader upper feature but only part of the neighboring thin line. The examples therefore show that thin-segment completeness remains challenging even when patch-level overlap is relatively high; aggregate rankings do not imply that one model performs best on every individual case.

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

All four models recovered the broad transverse feature in Fig. 8c, with differences at its margins and in peripheral components. In the difficult-negative example (Fig. 8d), MaxViT–U-Net predicted 43 foreground pixels, U-Net and DeepLabV3+ predicted none, and SegFormer predicted one. This example complements the dataset-level results by showing how a small local false detection can trigger negative-patch activation, even for the model with the lowest average activation.

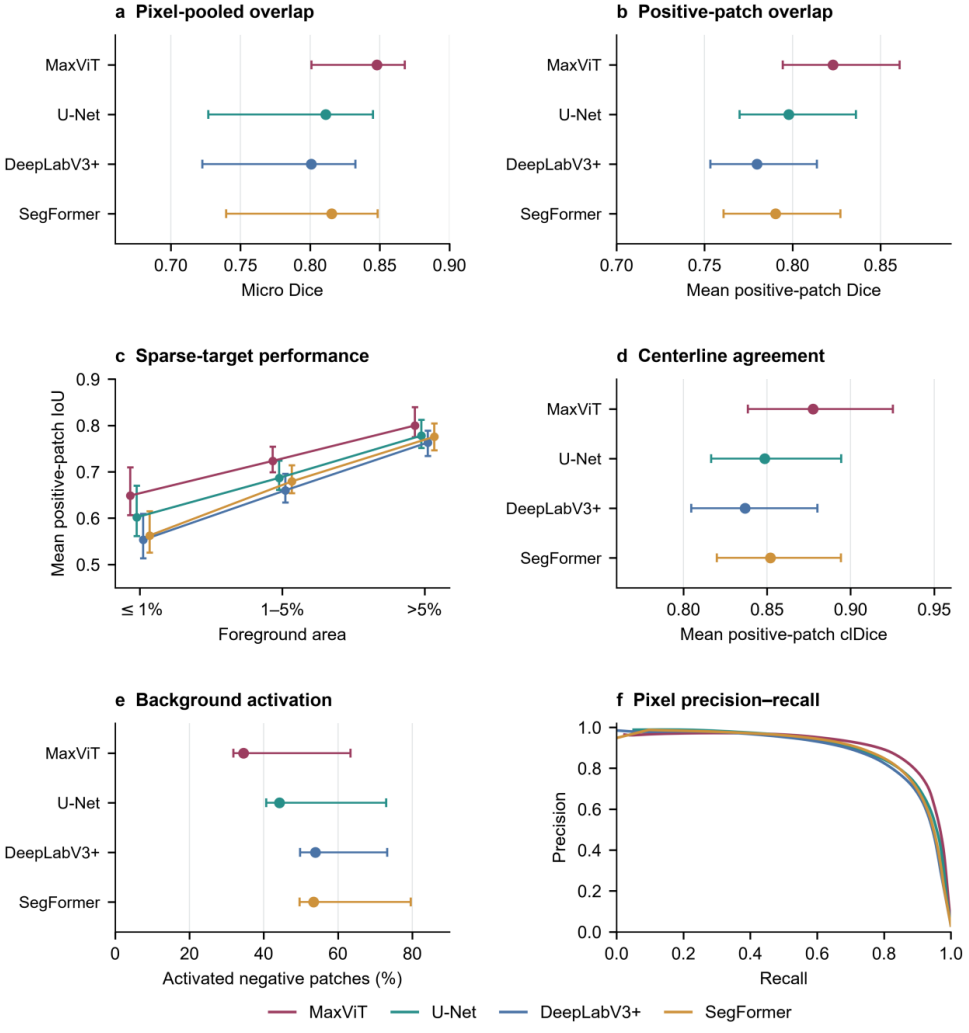


Figure 7. Segmentation performance under a shared training protocol. (a) Micro Dice across all test pixels. (b) Mean positive-patch Dice. (c) Mean positive-patch IoU by foreground fraction: ≤1%, >1% and ≤5%, and >5%. (d) Mean positive-patch clDice. (e) Difficult-negative activation, the fraction of empty-reference patches containing at least one predicted foreground pixel. In (a–e), points are means of three seed-specific estimates (42, 123, and 2025); error bars are percentile 95% intervals from 2,000 paired acquisition-date-clustered bootstrap resamples, calculating metrics per seed before averaging. (f) Pixel-level precision–recall curves for the prespecified seed 42, calculated using 4,096 probability bins. The test set contains 3,122 patches, including 1,422 positives and 1,700 difficult negatives from 51 dates. All threshold-dependent metrics use each run's fixed validation-selected threshold. Model names are abbreviated in the panels.

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.

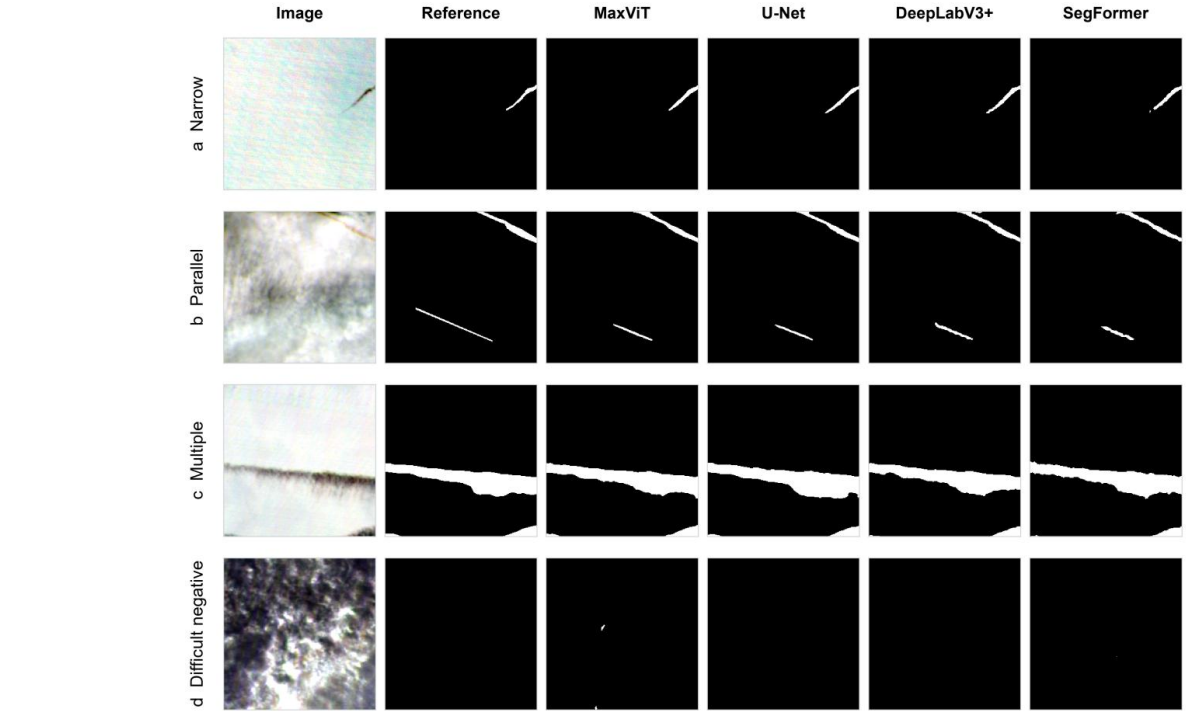


Figure 8. Example predictions under the shared training protocol. Rows show a narrow contrail, parallel structures, multicomponent foreground, and a difficult negative. Columns show enhanced RGB images, reference masks, and predictions from MaxViT–U-Net, U-Net–ResNet34, DeepLabV3+–ResNet50, and SegFormer-B0. White denotes foreground and black denotes background. Each patch contains 256 × 256 pixels. The four cases were selected before prediction inspection and are unchanged across models. All models use the prespecified seed 42 and their fixed validation-selected thresholds: 0.34, 0.30, 0.24, and 0.30, respectively.

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



Table 6. Four-model comparison under the shared training protocol. Values are means of separately calculated results for seeds 42, 123, and 2025; parentheses give percentile 95% intervals from 2,000 paired acquisition-date-clustered bootstrap resamples. Each resample calculates metrics per seed before averaging. Micro metrics use all 3,122 test patches, positive-patch metrics use 1,422 positives, and negative metrics use 1,700 difficult negatives. Thresholds are selected separately on validation data under the common rule. Intervals describe test-date sampling variation rather than across-seed standard deviations or ensemble performance.

Metric	MaxViT–U-Net	U-Net–ResNet34	DeepLabV3+–ResNet50	SegFormer-B0
Micro Dice	0.848 (0.801-0.868)	0.811 (0.727-0.845)	0.801 (0.723-0.833)	0.816 (0.740-0.848)
Micro IoU	0.736 (0.668-0.767)	0.682 (0.571-0.732)	0.668 (0.566-0.713)	0.689 (0.587-0.737)
Mean positive-patch Dice	0.823 (0.794-0.861)	0.798 (0.770-0.836)	0.780 (0.753-0.814)	0.790 (0.761-0.827)
Mean positive-patch IoU	0.740 (0.711-0.778)	0.707 (0.676-0.746)	0.681 (0.649-0.715)	0.695 (0.660-0.731)
Pixel precision	0.825 (0.731-0.861)	0.757 (0.618-0.804)	0.746 (0.626-0.786)	0.773 (0.645-0.818)
Pixel recall	0.873 (0.835-0.936)	0.874 (0.842-0.916)	0.865 (0.823-0.906)	0.863 (0.838-0.899)
Mean positive-patch clDice	0.878 (0.839-0.925)	0.849 (0.817-0.894)	0.837 (0.805-0.880)	0.852 (0.820-0.894)
Mean positive-patch centerline recall	0.908 (0.868-0.958)	0.891 (0.864-0.930)	0.883 (0.860-0.917)	0.881 (0.853-0.917)
Negative-patch activation (%)	34.57 (31.87-63.41)	44.24 (40.71-72.95)	53.90 (49.82-73.29)	53.45 (49.69-79.56)
Negative-pixel false-positive rate (%)	0.272 (0.176-1.091)	0.614 (0.429-2.057)	0.619 (0.471-1.552)	0.542 (0.379-1.699)

5.4 Stability across seeds and sampling units

Mean positive-patch Dice was 0.8230 ± 0.0056 for MaxViT–U-Net, 0.7978 ± 0.0039 for U-Net, 0.7798 ± 0.0026 for DeepLabV3+, and 0.7903 ± 0.0047 for SegFormer across the three training seeds (mean ± sample standard deviation; Table 7). MaxViT–U-Net had the highest positive-patch Dice in each same-seed comparison. Its mean clDice and centerline recall were 0.8775 and 0.9083, respectively. Paired resampling of source-scene entries retained the direction of the positive-patch Dice differences observed with date-level resampling.

Background activation varied more than foreground overlap across MaxViT–U-Net runs. Its negative-patch activation ranged from 26.59% to 48.41%, yielding 34.57% ± 12.04 percentage points, compared with 44.24% ± 2.61, 53.90% ± 4.05, and 53.45% ± 2.65 for the other three models. All runs used the same threshold-selection fallback, so this variability combines independent training with the resulting validation-selected operating points. The parameter counts range from 3.71 million for SegFormer-B0 to 83.95 million for MaxViT–U-Net. The matched protocol therefore supports comparison under common training rules, while computational scale and pretraining remain relevant to interpretation.

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



Table 7. Model size and across-seed variability under the shared protocol. Results are mean ± sample standard deviation across seeds 42, 123, and 2025, with each run evaluated at its fixed validation-selected threshold. Dice and clDice are averaged over positive patches. Parameter counts are in millions (M), and all parameters are fine-tuned. Standard deviations of percentage metrics are in percentage points. These summaries describe variation among fitted runs, not bootstrap confidence intervals.

Model	Params. (M)	Mean positive-patch Dice	Mean positive-patch clDice	Negative-pixel false-positive rate (%)	Negative-patch activation (%)
MaxViT–U-Net	83.95	0.8230 ± 0.0056	0.8775 ± 0.0045	0.272 ± 0.092	34.57 ± 12.04
U-Net–ResNet34	24.44	0.7978 ± 0.0039	0.8485 ± 0.0034	0.614 ± 0.065	44.24 ± 2.61
DeepLabV3+–ResNet50	26.68	0.7798 ± 0.0026	0.8368 ± 0.0015	0.619 ± 0.130	53.90 ± 4.05
SegFormer-B0	3.71	0.7903 ± 0.0047	0.8520 ± 0.0053	0.542 ± 0.124	53.45 ± 2.65

5.5 Enhanced RGB versus uint16 inputs

A separate U-Net–ResNet34 experiment compares the corresponding enhanced RGB and uint16 representations while keeping sample content and labels fixed. Both inputs use seeds 3407, 3408, and 3409 and share partitions, initialization sources, geometric and intensity augmentation, and validation-selection rules. This representation experiment uses positive-class-weighted binary cross-entropy plus soft Dice, AdamW at an initial learning rate of 10^{-4}, batch size 32, a maximum of 30 epochs, and early-stopping patience of eight epochs. Its results are compared within this experiment rather than pooled with the matched four-model experiment. Enhanced RGB values are divided by 255. For uint16 images, fill values are replaced with the corresponding training-channel median; values are then clipped at the training-set 1st and 99th percentiles and scaled to [0, 1]. Fixed lower bounds for the three channels are 657, 812, and 582, and upper bounds are 2112, 2437, and 1657. Both representations subsequently undergo ImageNet channel standardization.

Enhanced RGB achieved a mean micro Dice of 0.776 ± 0.008, compared with 0.693 ± 0.089 for the tested uint16 normalization (Table 8). Mean positive-patch IoU was 0.634 ± 0.016 and 0.571 ± 0.012, respectively. Negative-patch activation averaged 38.69% and 37.24%, with appreciable variation across seeds. Under this protocol, enhanced inputs therefore yielded higher and more stable overlap metrics, but background activation did not improve in the same way. The comparison concerns two complete input-processing procedures rather than an isolated effect of storage bit depth.

Table 8. Input-representation comparison using U-Net–ResNet34. Results are mean ± sample standard deviation for seeds 3407, 3408, and 3409. Both representations share samples, reference masks, partitions, initialization sources, and training budgets, with thresholds selected separately on validation data. Dice is micro-averaged over all test pixels, whereas IoU is averaged over positive patches. The standard deviation of activation is expressed in percentage points.

Metric	Enhanced RGB	Normalized uint16
Micro Dice	0.776 ± 0.008	0.693 ± 0.089
Mean positive-patch IoU	0.634 ± 0.016	0.571 ± 0.012
Difficult-negative activation (%)	38.69 ± 5.98	37.24 ± 16.71

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



6 Discussion

6.1 Contribution to fine-scale contrail observations

The principal contribution of ContrailStruct30 is to turn three-band thermal-infrared observations at 30 m nominal sampling into a contrail image resource with pixel-level reference labels. Narrow lines, adjacent parallel structures, intersections, and local gaps allow research to move beyond target presence toward the spatial form discernible in the imagery. Compared with image-level class labels, segmentation masks retain target location, extent, and relationships between adjacent structures, supporting analyses of visible continuity, local width variation, and spatial organization. These morphological descriptions concern discernible contrail regions, whose detail depends jointly on spatial sampling, imaging conditions, and target–background contrast.

This observational capability is particularly relevant to small contrails. The median foreground fraction among positive patches is only 3.54%, indicating that many targets occur as sparse linear structures within extensive backgrounds. In the stratified evaluation, all models performed less well on the smallest foreground group than on larger-area targets. Image examples likewise show that a broad structure may be recovered while an adjacent thin line remains partly omitted. Detection and continuous delineation of small contrails thus remain important challenges. By retaining these structures in the reference masks, ContrailStruct30 makes them explicit targets for training and focused evaluation rather than allowing their contribution to be obscured by pooled pixel-overlap metrics.

Narrow, compact morphologies also provide candidate observations for studies of newly formed and early-stage contrails. Post-formation spreading, deformation, and interaction with surrounding clouds depend on atmospheric conditions and contrail development (Schumann and Heymsfield, 2017; Kärcher, 2018). Discernible narrow lines, parallel structures, and local broadening can provide spatial references for case studies that incorporate flight trajectories and meteorological data. Formation stage must still be established from aircraft position, observation time, and transport conditions. The dataset's value for such studies is therefore its detailed morphology, which can be linked to additional evidence, rather than an age assignment based on shape alone.

6.2 Contrail interpretation and segmentation in complex backgrounds

Fine-scale observations reveal not only more contrail detail but also richer background texture. Natural cirrus, elongated cloud edges, and linear surface features may locally resemble contrails. Segmentation must therefore address two coupled tasks: recovering targets completely and rejecting similar backgrounds. By combining expert-selected difficult negatives with diverse positive examples, ContrailStruct30 makes background discrimination part of the learning task rather than an issue addressed only after model deployment.

The matched comparison illustrates the need to interpret foreground recovery and background response jointly. MaxViT–U-Net combined the highest mean foreground overlap with the lowest mean negative-patch activation among the tested implementations, but its activation still averaged 34.57% and varied appreciably across seeds. A small false-positive pixel fraction can therefore coexist with activation of many negative patches. Such scattered detections increase the inspection burden when masks are used to screen candidate images. The dataset exposes these complementary aspects of performance, providing more informative test conditions than a single overlap ranking (Maier-Hein et al., 2018).

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



Expert reassessment identified two further difficulties in local interpretation: background texture can be mistaken for a target, while small or low-contrast contrails can be overlooked. Source-scene context helps establish whether a local line belongs to a broader natural cloud structure and whether a candidate has coherent spatial morphology. The reassessment supports the plausibility of the inspected masks in their image context. Acceptance after mask disclosure and agreement between independently redrawn boundaries nevertheless represent different forms of evidence; independent references and boundary-sensitive evaluations address differences in pixel delineation (Warfield et al., 2004; Cheng et al., 2021; Gourmelon et al., 2022). For diffuse margins, occlusions, and gaps between adjacent structures, the masks should be understood as expert interpretations under common rules. Targeted scene searches and difficult-background selection make the dataset primarily suited to these segmentation tasks rather than estimating contrail occurrence in natural scenes.

Model screening and preliminary masks may also influence candidate composition and initial boundaries. Manual correction and expert review can reduce this influence, but participation of the same architectural family in dataset construction remains relevant to interpreting model comparisons. The shared protocol controls input processing, training settings, and validation-selection rules; it does not equalize pretraining data, parameter counts, or computational cost. The results consequently characterize complete pretrained implementations under common training conditions rather than isolate an encoder-architecture effect or each model's fully tuned performance ceiling.

6.3 Data reuse through corresponding image representations

Providing enhanced RGB images alongside corresponding three-band uint16 source values allows the same annotations to support different image-processing and modeling approaches. Enhanced images facilitate visual inspection and can be used directly by common visual models, while source values retain scope for alternative contrast transformations, channel combinations, and normalization. Both representations share sample content, pixel grids, and reference masks, enabling input-processing comparisons on fixed observations without reconstructing labels.

The U-Net comparison shows that input representation materially affects segmentation performance. Under the tested protocol, enhanced RGB produced higher pixel overlap with less cross-seed variation, whereas difficult-negative activation did not show the same improvement. Transformations that favor foreground recovery therefore also need to be evaluated for background discrimination. The difference may reflect local enhancement, training-set normalization, and the input statistics expected by pretrained models. It characterizes two complete processing procedures, not the relative merits of storage bit depths alone.

Corresponding representations thus provide more than two file formats: they make preprocessing a component that can be compared and studied reproducibly. Together with fixed partitions, expert-reviewed labels, and reference segmentation results, ContrailStruct30 supports research from input design and fine-structure recovery to difficult-background discrimination, providing reusable data for fine-scale thermal-infrared contrail analysis.

7 Data and code availability

Version 3 of ContrailStruct30 is publicly available via Figshare at https://doi.org/10.6084/m9.figshare.33435496.v3 (Yu et al., 2026). The dataset contains 19,455 samples,

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



each comprising an enhanced three-channel RGB image, a corresponding three-band uint16 source-value image, and a human-corrected binary reference mask, together with associated metadata and fixed training, validation, and test partitions.

All dataset components distributed in this Figshare release, including the enhanced RGB images, uint16 TIFF source-value images, binary reference masks, partition files, and metadata, are made available under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence. Redistribution of the SDGSAT-1-derived uint16 source-value images has been authorized through the SDGSAT-1 Open Science Program.

The companion segmentation benchmark code, including the model implementations, training and evaluation procedures, and associated reproducibility materials, is publicly available at https://github.com/jianingyu777/contrail-structural-benchmark.

The original full-scene SDGSAT-1 observations can be obtained from the SDGSAT-1 data portal (https://data.sdgsat.ac.cn/) under the portal's applicable terms. Historical ADS-B trajectories used for candidate-scene searches were obtained from ADSB.lol (https://adsb.lol/).

8 Conclusions

ContrailStruct30 provides 19,455 pixel-labeled, three-band thermal-infrared patches at 30 m nominal sampling, with each reference mask linked to enhanced RGB and uint16 images. Model-assisted expansion, manual correction, and expert review combine fine and complex foreground geometries with deliberately selected difficult negatives in one resource. Fixed evaluation partitions separate both source scenes and acquisition dates.

Application examples demonstrate complementary evaluation of foreground overlap, sparse linear structures, and difficult-background response. A matched comparison across four segmentation implementations and three training seeds provides reference performance, with MaxViT–U-Net offering strong foreground recovery and lower average negative-patch activation under the tested protocol. A separate RGB–uint16 comparison illustrates controlled input-preprocessing research. Together with expert-reviewed masks and corresponding image representations, these evaluation conditions make ContrailStruct30 a reusable resource for developing thermal-infrared contrail segmentation methods.

Author contributions

JY: conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, and writing of the original draft. LZ: methodology, validation, and data curation. XZ, LL, and XN: data curation and validation. YX, KW, YL, GD, HL, CT, and MD: validation. ZH: methodology and supervision. XL: supervision and funding acquisition. FC: conceptualization, supervision, project administration, funding acquisition, and review and editing. All authors reviewed the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

We thank the International Research Center of Big Data for Sustainable Development Goals for providing observations through the SDGSAT-1 Open Science Program. We thank ADSB.lol for the historical ADS-B

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



trajectories used to locate candidate targets and the annotators and experts for their contributions to label construction and reassessment.

Financial support

This research was supported by the National Natural Science Foundation of China (grant no. 62575297).

References

Budd, S., Robinson, E. C., and Kainz, B.: A survey on active learning and human-in-the-loop deep learning for medical image analysis, Med. Image Anal., 71, 102062, https://doi.org/10.1016/j.media.2021.102062, 2021.

Burkhardt, U. and Kärcher, B.: Global radiative forcing from contrail cirrus, Nat. Clim. Change, 1, 54-58, https://doi.org/10.1038/nclimate1068, 2011.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Computer Vision - ECCV 2018, Lect. Notes Comput. Sc., 11211, 833-851, https://doi.org/10.1007/978-3-030-01234-2_49, 2018.

Cheng, B., Girshick, R., Dollar, P., Berg, A. C., and Kirillov, A.: Boundary IoU: Improving Object-Centric Image Segmentation Evaluation, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15329-15337, https://doi.org/10.1109/cvpr46437.2021.01508, 2021.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L.: ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248-255, https://doi.org/10.1109/cvpr.2009.5206848, 2009.

Efron, B.: Bootstrap Methods: Another Look at the Jackknife, Ann. Stat., 7, https://doi.org/10.1214/aos/1176344552, 1979.

Field, C. A. and Welsh, A. H.: Bootstrapping clustered data, J. R. Stat. Soc. B, 69, 369-390, https://doi.org/10.1111/j.1467-9868.2007.00593.x, 2007.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K.: Datasheets for datasets, Commun. ACM, 64, 86-92, https://doi.org/10.1145/3458723, 2021.

Gourmelon, N., Seehaus, T., Braun, M., Maier, A., and Christlein, V.: Calving fronts and where to find them: a benchmark dataset and methodology for automatic glacier calving front extraction from synthetic aperture radar imagery, Earth Syst. Sci. Data, 14, 4287-4313, https://doi.org/10.5194/essd-14-4287-2022, 2022.

Guo, H., Dou, C., Liang, D., Fu, B., Chen, H., Zou, Z., Huang, P., Li, X., Chen, F., Han, C., Jing, J., Hu, T., Yan, L., Hu, Y., Tang, Y., Jiang, N., Feng, X., Ding, H., Zhang, H., Qiao, E., and Zhou, B.: The SDGSAT-1 mission and its role in monitoring SDG indicators, Remote Sens. Environ., 328, 114885, https://doi.org/10.1016/j.rse.2025.114885, 2025.

He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, 770-778, https://doi.org/10.1109/CVPR.2016.90, 2016.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Q. J. Roy. Meteor. Soc., 146, 1999-2049, https://doi.org/10.1002/qj.3803, 2020.

Hu, Z., Zhu, M., Wang, Q., Su, X., and Chen, F.: SDGSAT-1 TIS prelaunch radiometric calibration and performance, Remote Sens., 14, 4543, https://doi.org/10.3390/rs14184543, 2022.

Jarry, G., Dalmau, R., Very, P., Ballerini, F., and Bocu, S.-D.: GVCCS: a dataset for contrail identification and tracking on visible whole sky camera sequences, Earth Syst. Sci. Data, 18, 1037-1059, https://doi.org/10.5194/essd-18-1037-2026, 2026.

Kärcher, B.: Formation and radiative forcing of contrail cirrus, Nat. Commun., 9, 1824, https://doi.org/10.1038/s41467-018-04068-0, 2018.

Lee, D. S., Fahey, D. W., Skowron, A., Allen, M. R., Burkhardt, U., Chen, Q., Doherty, S. J., Freeman, S., Forster, P. M., Fuglestvedt, J., Gettelman, A., De Leon, R. R., Lim, L. L., Lund, M. T., Millar, R. J., Owen, B., Penner, J. E., Pitari, G., Prather, M. J., Sausen, R., and Wilcox, L. J.: The contribution of global aviation to anthropogenic climate forcing for 2000 to 2018, Atmos. Environ., 244, 117834, https://doi.org/10.1016/j.atmosenv.2020.117834, 2021.

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



674 Long, J., Shelhamer, E., and Darrell, T.: Fully convolutional networks for semantic segmentation, in: 2015 IEEE
675 Conference on Computer Vision and Pattern Recognition (CVPR), 3431-3440,
676 https://doi.org/10.1109/cvpr.2015.7298965, 2015.

677 Loshchilov, I. and Hutter, F.: Decoupled weight decay regularization, in: International Conference on Learning
678 Representations, https://arxiv.org/abs/1711.05101, 2019.

679 Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., and Johnson, B. A.: Deep learning in remote sensing applications: A meta-
680 analysis and review, ISPRS J. Photogramm. Remote Sens., 152, 166-177,
681 https://doi.org/10.1016/j.isprsjprs.2019.04.015, 2019.

682 Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley,
683 A. P., Carass, A., Feldmann, C., Frangi, A. F., Full, P. M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek,
684 M., Landman, B. A., März, K., Maier, O., Maier-Hein, K., Menze, B. H., Müller, H., Neher, P. F., Niessen, W.,
685 Rajpoot, N., Sharp, G. C., Sirinukunwattana, K., Speidel, S., Stock, C., Stoyanov, D., Taha, A. A., van der Sommen,
686 F., Wang, C.-W., Weber, M.-A., Zheng, G., Jannin, P., and Kopp-Schneider, A.: Why rankings of biomedical image
687 analysis competitions should be interpreted with care, Nat. Commun., 9, 5217, https://doi.org/10.1038/s41467-018-
688 07619-7, 2018.

689 Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek,
690 J., Kozubek, M., Reyes, M., Riegler, M. A., Wiesenfarth, M., Kavur, A. E., Sudre, C. H., Baumgartner, M.,
691 Eisenmann, M., Heckmann-Nötzel, D., Rädsch, T., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A.,
692 Blaschko, M. B., Cardoso, M. J., Cheplygina, V., Cimini, B. A., Collins, G. S., Farahani, K., Ferrer, L., Galdran, A.,
693 van Ginneken, B., Haase, R., Hashimoto, D. A., Hoffman, M. M., Huisman, M., Jannin, P., Kahn, C. E.,
694 Kainmueller, D., Kainz, B., Karargyris, A., Karthikesalingam, A., Kofler, F., Kopp-Schneider, A., Kreshuk, A., Kurc,
695 T., Landman, B. A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A. L., Mattson, P., Meijering, E., Menze, B.,
696 Moons, K. G. M., Müller, H., Nichyporuk, B., Nickel, F., Petersen, J., Rajpoot, N., Rieke, N., Saez-Rodriguez, J.,
697 Sánchez, C. I., Shetty, S., van Smeden, M., Summers, R. M., Taha, A. A., Tiulpin, A., Tsaftaris, S. A., Van Calster,
698 B., Varoquaux, G., and Jäger, P. F.: Metrics reloaded: recommendations for image analysis validation, Nat. Methods,
699 21, 195-212, https://doi.org/10.1038/s41592-023-02151-z, 2024.

700 Mannstein, H., Meyer, R., and Wendling, P.: Operational detection of contrails from NOAA-AVHRR-data, Int. J.
701 Remote Sens., 20, 1641-1660, https://doi.org/10.1080/014311699212650, 1999.

702 McCloskey, K., Geraedts, S., Jackman, B., Meijer, V. R., Brand, E., Fork, D., Platt, J. C., Elkin, C., and Van Arsdale, C.:
703 A human-labeled Landsat-8 contrails dataset, in: Proceedings of the ICML 2021 Workshop on Tackling Climate
704 Change with Machine Learning, https://www.climatechange.ai/papers/icml2021/2, 2021.

705 Meijer, V. R., Kulik, L., Eastham, S. D., Allroggen, F., Speth, R. L., Karaman, S., and Barrett, S. R. H.: Contrail
706 coverage over the United States before and during the COVID-19 pandemic, Environ. Res. Lett., 17, 034039,
707 https://doi.org/10.1088/1748-9326/ac26f0, 2022.

708 Ng, J. Y. H., McCloskey, K., Cui, J., Meijer, V. R., Brand, E., Sarna, A., Goyal, N., Van Arsdale, C., and Geraedts, S.:
709 Contrail detection on GOES-16 ABI with the OpenContrails dataset, IEEE T. Geosci. Remote, 62, 1-14,
710 https://doi.org/10.1109/TGRS.2023.3345226, 2024.

711 Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., and Wulder, M. A.: Good practices for
712 estimating area and assessing accuracy of land change, Remote Sens. Environ., 148, 42-57,
713 https://doi.org/10.1016/j.rse.2014.02.015, 2014.

714 Reinke, A., Tizabi, M. D., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Kavur, A. E., Rädsch, T., Sudre, C.
715 H., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Buettner, F., Cardoso, M. J., Cheplygina, V., Chen, J.,
716 Christodoulou, E., Cimini, B. A., Farahani, K., Ferrer, L., Galdran, A., van Ginneken, B., Glocker, B., Godau, P.,
717 Hashimoto, D. A., Hoffman, M. M., Huisman, M., Isensee, F., Jannin, P., Kahn, C. E., Kainmueller, D., Kainz, B.,
718 Karargyris, A., Kleesiek, J., Kofler, F., Kooi, T., Kopp-Schneider, A., Kozubek, M., Kreshuk, A., Kurc, T.,
719 Landman, B. A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A. L., Meijering, E., Menze, B., Moons, K. G. M.,
720 Müller, H., Nichyporuk, B., Nickel, F., Petersen, J., Rafelski, S. M., Rajpoot, N., Reyes, M., Riegler, M. A., Rieke,
721 N., Saez-Rodriguez, J., Sánchez, C. I., Shetty, S., Summers, R. M., Taha, A. A., Tiulpin, A., Tsaftaris, S. A., Van
722 Calster, B., Varoquaux, G., Yaniv, Z. R., Jäger, P. F., and Maier-Hein, L.: Understanding metric-related pitfalls in
723 image analysis validation, Nat. Methods, 21, 182-194, https://doi.org/10.1038/s41592-023-02150-0, 2024.

724 Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J.,
725 Schroder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F.: Cross-validation strategies
726 for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography, 40, 913-929,
727 https://doi.org/10.1111/ecog.02881, 2017.

https://doi.org/10.5194/essd-2026-759
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



728 Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional networks for biomedical image segmentation, in:
729 Medical Image Computing and Computer-Assisted Intervention, 234-241, https://doi.org/10.1007/978-3-319-24574-
730 4_28, 2015.

731 Salehi, S. S. M., Erdogmus, D., and Gholipour, A.: Tversky loss function for image segmentation using 3D fully
732 convolutional deep networks, in: Machine Learning in Medical Imaging (MLMI 2017), Lecture Notes in Computer
733 Science, 10541, 379–387, Springer, Cham, ISBN 978-3-319-67389-9, https://doi.org/10.1007/978-3-319-67389-
734 9_44, 2017.

735 Santos Gabriel, V., Bugliaro, L., Montag, M., Ries, S., Wang, Z., Widmaier, K., Arico, M., Unterstrasser, S., Mayer, J.,
736 Menekay, D., Marsing, A., de la Torre Castro, E., Megill, L., Scheibe, M., and Voigt, C.: A manually labeled
737 contrail dataset from MSG/SEVIRI, Earth Syst. Sci. Data, 18, 2397-2412, https://doi.org/10.5194/essd-18-2397-
738 2026, 2026.

739 Schumann, U. and Heymsfield, A. J.: On the life cycle of individual contrails and contrail cirrus, Meteorol. Monogr., 58,
740 3.1-3.24, https://doi.org/10.1175/AMSMONOGRAPHS-D-16-0005.1, 2017.

741 Shit, S., Paetzold, J. C., Sekuboyina, A., Ezhov, I., Unger, A., Zhylka, A., Pluim, J. P. W., Bauer, U., and Menze, B. H.:
742 clDice: a novel topology-preserving loss function for tubular structure segmentation, in: 2021 IEEE/CVF
743 Conference on Computer Vision and Pattern Recognition, 16555-16564,
744 https://doi.org/10.1109/CVPR46437.2021.01629, 2021.

745 Shorten, C. and Khoshgoftaar, T. M.: A survey on Image Data Augmentation for Deep Learning, J. Big Data, 6, 60,
746 https://doi.org/10.1186/s40537-019-0197-0, 2019.

747 Taha, A. A. and Hanbury, A.: Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool,
748 BMC Med. Imaging, 15, 29, https://doi.org/10.1186/s12880-015-0068-x, 2015.

749 Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., and Li, Y.: MaxViT: multi-axis vision transformer, in:
750 Computer Vision - ECCV 2022, Lect. Notes Comput. Sc., 13684, 459-479, https://doi.org/10.1007/978-3-031-
751 20053-3_27, 2022.

752 Warfield, S. K., Zou, K. H., and Wells, W. M.: Simultaneous Truth and Performance Level Estimation (STAPLE): An
753 Algorithm for the Validation of Image Segmentation, IEEE Trans. Med. Imaging, 23, 903-921,
754 https://doi.org/10.1109/tmi.2004.828354, 2004.

755 Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W.,
756 da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O.,
757 Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S.,
758 Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer,
759 A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T.,
760 Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A.,
761 Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B.: The FAIR Guiding Principles for scientific data
762 management and stewardship, Sci. Data, 3, 160018, https://doi.org/10.1038/sdata.2016.18, 2016.

763 Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P.: SegFormer: Simple and efficient design for
764 semantic segmentation with transformers, Adv. Neur. In., 34, 12077–12090,
765 https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf, 2021.

766 Yu, J., Zhou, X., Li, L., Gao, L., Li, X., Pan, W., Ni, X., Wang, Q., and Chen, F.: High-resolution thermal infrared
767 contrails images identification and classification method based on SDGSAT-1, Int. J. Appl. Earth Obs., 131, 103980,
768 https://doi.org/10.1016/j.jag.2024.103980, 2024.

769 Yu, J., Xue, Y., Li, Y., Du, G., Tan, C., Lu, H., Wang, K., Li, X., and Chen, F.: ContrailStruct30: a 30 m three-band
770 thermal-infrared dataset for contrail segmentation, Figshare [data set],
771 https://doi.org/10.6084/m9.figshare.33435496.v3, 2026.

772 Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., and Fraundorfer, F.: Deep Learning in Remote Sensing: A
773 Comprehensive Review and List of Resources, IEEE Geosci. Remote Sens. Mag., 5, 8-36,
774 https://doi.org/10.1109/mgrs.2017.2762307, 2017.

775 Zlateski, A., Jaroensri, R., Sharma, P., and Durand, F.: On the Importance of Label Quality for Semantic Segmentation,
776 in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1479-1487,
777 https://doi.org/10.1109/cvpr.2018.00160, 2018.