Dear Dr. Billy Andrews,

We thank you for the thorough review and constructive comments. Based on these we have substantially improved the contextualisation, referencing, and clarity of presentation. We have especially focused on the Introduction and Discussion to better situate the work within the broader fracture network characterisation literature, improved figure legibility and captions, added summary/location maps and scale information, and clarified the research gap and implications throughout. Please see our detailed responses below.

Please note that to facilitate the evaluation of our revision, the page and line numbers of the reviewer's comments refer to the originally submitted manuscript while page and line numbers of our responses refer to our revised manuscript.

Best regards,
Ayoub Fatihi, Jefter Caldeira, Tom Beucler, Samuel T. Thiele, and Anindita Samsu

Reply to RC1: 'Comment on egusphere-2026-1097', **Billy Andrews**, 22 May 2026

> Dear authors,
>
> Please find attached my review of your manuscript "Towards robust fracture mapping: benchmarking automatic fracture mapping in 2D outcrop imagery". The study utilises a training dataset comprising of three published datasets to assess a range of traditional and deep-learning models to achieve fracture network interpretations. This is an area of growth in the community, and a systematic comparison of methods would be of use to those using, or assessing output from, these models. Whilst i found the results were well handled and provided a robust statistical approach to comparing datasets, the manuscript was difficult to follow and the research cap and implications poorly laid out. In particular i found the level of references to be very limited, with a wealth of literature in this, and the wider fracture network characterisation fields not called upon. This made it difficult to see where the work fit in, and how the results apply to fracture analysis more generally. As such, at present, i have to suggest that the manuscript is rejected due to the number of changes required to make it acceptable for publication. Although i am sure this is disappointing, I hope my review is helpful and i look forward to seeing a revised version of this work in the future.
>
> Please see below the attached annotated manuscript with my specific comments.
>
> Kind regards,
>
> Dr. Billy Andrews

Abstract

"Abstract." (p. 1) I suggest the Abstract is rewritten to better place the work into context, and outline the key results and their implications.

Agree: The abstract was rewritten, taking into account the comments below.

"Extracting consistent and accurate fracture traces from large volumes of high-resolution imagery remains a persistent challenge in structural analysis" (p. 1) It would be best to state why lineament analysis may be needed/useful.

Agree: We have updated the opening of the abstract to: "*Analysis of fracture geometries, orientation, distribution, and connectivity is commonly conducted to characterise the mechanical and hydraulic properties of fractured rock mass and to interpret their deformation history.*" (see L1).

We also revised the opening of the Introduction accordingly: "Mapping natural fractures in outcrops is fundamental to studies of brittle tectonics, assessments of rock stability, and predictions of subsurface fluid flow pathways for geo-energy applications(Aydin, 2000; Berkowitz, 2002; Smeraglia et al., 2021) . Recent advances in remote sensing and the uptake of uncrewed aerial vehicle (UAV) photogrammetry by the structural geology community have transformed how fracture data are collected (Bemis et al., 2014). UAV photogrammetry enables the generation of high-resolution, georeferenced orthophotographs that capture rock surfaces in remarkable detail, including outcrops that are difficult or unsafe to reach (Villarreal et al., 2022). The resulting orthophotographs enable precise tracing of thousands of fractures on a single outcrop, which can then be analysed for orientation, length, distribution, and connectivity, complementing direct field observations and measurements (Palamakumbura et al., 2020)." **(see L18)**.

"We present a harmonised benchmarking dataset," (p. 1) This is very specific and does not set your study within the wider scientific problem it is addressing. You need to walk readers through to ensure they understand where your contribution fits.

Agree: The abstract and introduction have been restructured to walk the reader through the broader problem before presenting the specific contribution. The abstract now opens with the motivation for fracture analysis, then the challenge of data extraction, and only then introduces FraXet.

"high-resolution RGB orthophotos" (p. 1) This is highly subjective.. give ranges

Agree: Addressed by replacing the subjective wording with:*"... combined high-resolution RGB orthophotographs and digital elevation models (DEMs) (ground sampling distance ranging from approximately 0.5 mm to 32 mm)."* **(see L5)**.

“FraXet curates images from three publicly available datasets, totalling 8953 256×256 RGB+DEM patches spanning diverse lithologies and imaging conditions” (p. 1) But why were these datasets used? it is not clear what the purpose of the research is from the abstract as presently written.

Agree: Now reformulated to: *“FraXet curates images from three publicly available datasets, totalling 8,953 256×256 RGB+DEM patches spanning diverse lithologies and imaging conditions (ground sampling distance, illumination, etc.) to systematically assess the effectiveness of conventional image-processing approaches for fracture extraction (Canny, Sobel, Gabor, Sato, and phase congruency) and two deep-learning (DL) models (U-Net and SegFormer).”* **(see L6)**.

“image-processing filters (Canny, Sobel, Gabor, Sato, phase congruency)” (p. 1) You need to outline that image processing features have been used to create algorithm to extract fractures

Agree: Addressed by changing: *"... to systematically assess the effectiveness of conventional image-processing approaches for fracture extraction (Canny, Sobel, Gabor, Sato, and phase congruency) ..."* **(see L8)**.

“Training on the combined dataset (M_all) improves cross-site generalisation relative to models 10 trained on the individual sub-datasets.” (p. 1) How representative are the input datasets to wider geological settings.

Agree: Now reads: *“Training on the combined dataset (M_all) improves cross-site generalisation compared to models trained on individual sub-datasets used in this study.”* (see L12). And also added in the Methods (Section 2.1.1): *“We acknowledge that three datasets covering two granite types and one clastic sedimentary setting, while providing meaningful lithological contrast for a first benchmark, do not represent the full diversity of geological settings encountered in practice;”* **(see L145)**.

Introduction

The introduction was thoroughly revised to address the reviewer's comments and to better situate the study in the broader literature.

“(Twiss and Moores, 2006).” (p. 2) There are much better references that could be brought in here to highlight both the effect on rock stability and fluid flow.

Agree: We have replaced the general textbook reference with more targeted citations. The opening sentence now reads: "*Mapping natural fractures in outcrops is a key step in the study of brittle tectonics, assessing rock stability, and predicting preferred fluid flow pathways in the subsurface for geo-energy applications (Aydin, 2000; Berkowitz, 2002; Smeraglia et al., 2021).*" **(see L18)**.

Here, Aydin (2000) is cited for fractures and hydrocarbon entrapment/migration, Berkowitz (2002) for a comprehensive review of flow and transport in fractured geological media, and Smeraglia et al. (2021) for 3D DFN modelling of fluid corridors. Bemis et al. (2014) is also cited

as a supporting reference for UAV photogrammetry applications in structural geology. Additional supporting references are cited throughout the Introduction.

“Smeraglia et al., 2021).” (p. 2) Given the scope of the manuscript, more references need to be included to place this work into the wider context of lineament analysis.

Agree: We have substantially expanded the reference base. The introduction now cites, among others: Potts (2002), Bonnet (2001), Sanderson (2015) for fracture characterisation; Bemis et al. (2014), Villarreal (2022), Palamakumbura (2020) for UAV photogrammetry; Bond (2007), Scheiber (2015), Peacock (2019), Andrews (2019) for interpreter bias; Peacock and Sanderson (2026) for the influence of map resolution and quality on fault network topology; and many more throughout the classical and DL method paragraphs.

“Recent advances in remote sensing have transformed how geological fractures in outcrops are documented. Drone-based imaging now allows the creation of high-resolution, georeferenced orthophotos that capture rock surfaces in remarkable detail. These digital data enable precise mapping, measurement, and analysis of structural features like orientation and spacing, complementing traditional field observations and providing access to outcrops that are difficult or unsafe to reach. However, while data acqui25 sition has become faster and easier, the interpretation of these images remains time-consuming and subjective.” (p. 2) This is not supported by relevant references

Agree: This passage has been rewritten and is now fully referenced. The revised text reads:

"Recent advances in remote sensing and the uptake of uncrewed aerial vehicle (UAV) photogrammetry by the structural geology community have transformed how fracture data are collected (Bemis et al., 2014). UAV photogrammetry enables the generation of high-resolution, georeferenced orthophotographs that capture rock surfaces in remarkable detail, including outcrops that are difficult or unsafe to reach (Villarreal et al., 2022). The resulting orthophotographs enable precise tracing of thousands of fractures on a single outcrop, which can then be analysed for orientation, length, distribution, and connectivity, complementing direct field observations and measurements (Palamakumbura et al., 2020)." **(see L20)**.

“increasingly appealing alternative.” (p. 2) this needs examples

Agree: We added further references; the sentence now reads: *“Rapid advances in computer vision and machine learning (ML) have made automated fracture detection an increasingly appealing alternative to manual and semi-automated methods (Thiele et al., 2017b; Mattéo et al., 2021; Chudasama et al., 2024).“* **(see L46)**.

“—” (p. 2) ,

Agree: Replaced with a comma.

“imaging conditions” (p. 2) image quality

Clarify: In this instance we are referring to the conditions under which the images were acquired which, although a strong control on image quality, also influences other properties (e.g., white balance, illumination patterns, resolution, etc.). We have therefore chosen to keep the term “imaging conditions”, and clarified in the manuscript what we mean by this: *“imaging conditions (e.g., white balance, illumination patterns, resolution, etc.)*” **(see L91)**.

“Table 1. Summary of methods and their characteristics for fracture mapping” (p. 3) This is a good table, but it should be integrated better into the text, which at present does not clearly outline the existing literature or how your contribution fills the research gap.

Agree: We added several paragraphs explaining and contextualising the table. The introduction now includes dedicated paragraphs on classical approaches:

"*Among the classical approaches, colorimetry-based segmentation exploits the contrast between fractures and host rock but is highly sensitive to lighting and shadowing (Ren, 2003; Vasuki, 2017). DEM and DTM-based feature extraction methods, such as segment tracing and directional detection algorithms (Koike, 1995; Raghavan, 1993, 1995), can reveal terrain-related structures but depend heavily on the resolution and quality of the underlying elevation data and are computationally demanding…*" **(see L53)**.

And on deep learning approaches:

"*Deep learning and machine learning approaches represent a distinct category that has grown rapidly in recent years. These include convolutional neural network architectures such as U-Net for pixel-wise fracture segmentation (Chudasama et al., 2024), incrementally trained networks designed to adapt to new geological settings such as the Geological Adaptive Incremental Network (Zhang et al., 2024), hybrid approaches that combine DL-based feature extraction with classical line detection (Oliveira et al., 2019)…*" **(see L70)**.

The research gap and our contribution are then stated explicitly in the final paragraph of the Introduction (quoted above in response to the "We present a harmonised benchmarking dataset" comment) **(see L109)**.

“However, a persisting issue is that automated fracture detection still struggles with false positives and negatives, particularly under natural and uncontrolled imaging conditions.” (p. 5) e.g.

Agree: We have added concrete examples to illustrate this point. The revised text now reads:

"Several issues still need to be addressed as DL-based automated fracture detection continues to develop. First, these methods still struggle with false positives and negatives, particularly under natural and uncontrolled imaging conditions (e.g., white balance, illumination patterns, resolution, etc.) (Mattéo et al., 2021; Yaqoob et al., 2024; An et al., 2023), e.g., topographic shadows mimicking fracture traces, vegetation creating false linear features, or weathering textures producing edges indistinguishable from fractures." **(see L90)**.

“Most fracture analyses also require instance-level outputs—such as fracture length, orientation, spacing, and connectivity—which depend directly on the correct topology of each trace.” (p. 5) See Peacock and Sanderson, 2026: Peacock, D.C.P. and Sanderson, D.J., 2026. Influences of map resolution, quality and interpretation on fault network topology. Earth-Science Reviews, p.105461.

Agree: We thank the reviewer for pointing us to this highly relevant paper. This point has been relocated to Section 1 (persisting issues) and now reads: “While these inconsistencies may influence analyses of fracture lengths, orientations, and network topology, they have an even greater impact on model training and evaluation accuracy. This is important because annotation scale and interpretation can significantly affect the derived properties of fracture networks (Peacock and Sanderson, 2026)” **(see L99)**.

“As a result,” (p. 5) As written, this does not follow

Agree: We have rewritten that paragraph for better logical flow: *“Third, ground-truth labels used for training are often drawn at a coarser scale than the maximum image resolution allows: a time-saving compromise that leads to misalignment between the labelled fracture and the pixels that represent the fracture in the orthophotograph (Fig. 1a,b).”* **(see L96)**.

“Furthermore, the ground-truth labels used for training are often drawn at coarser scales than pixel-level 45 precision, leading to misalignment between the labelled and actual fracture edges (Figure 1).” (p. 5) See comment on the figure, it is often useful to simplify fracture traces!

“Figure 1. Example of annotation misalignment. (a) Outcrop orthophoto (Nordbäck and Ovaskainen, 2025) overlaid with fracture traces in white from the ground-truth labels (Ovaskainen and Nordbäck, 2022). (b) Zoomed-in view showing how manually drawn fracture lines deviate from the actual visible fracture edges.” (p. 5)The simplification of fracture traces is releative to the scale the fractures are mapped at, much like a coastline is fractal in length, the roughness of a fracture trace is also such that you get more roughness as you zoom in. It is perfectly applicable to simplify traces based on the scale of observation - in this example, it is very close to the trace and would have negligible impact on any analysis.

Clarify: We agree that trace simplification is appropriate and standard practice depending on the mapping scale. Our point is not that simplification is wrong, but that when these annotations are used as pixel-level ground truth for training deep learning models, the spatial offset between the simplified trace and the actual fracture pixels introduces label noise that degrades both model training and metric evaluation. And showed in the manuscript by: “*While these inconsistencies … they have an even greater impact on model training and evaluation accuracy.*” **(see L99)**.

(p. 5) Figure has no scale, which is integral when discussing scale of observation.

Agree: The figure was initially intended only to illustrate annotation misalignment, but a scale bar is included for completeness **(see Figure 1)**.

"(Sharma et al., 2022)." (p. 5) Please provide examples of where both have been used

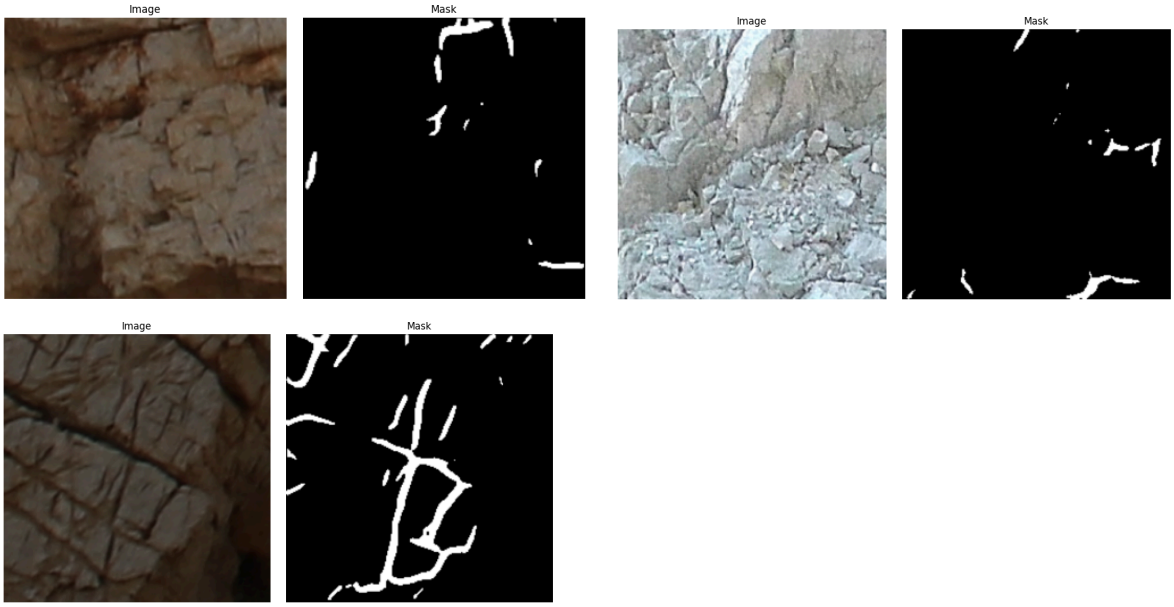
Agree: This paragraph was moved to the Conclusions. To our knowledge, existing studies apply semantic (pixel) segmentation first (Chudasama et al., 2024; Mattéo et al., 2021; Prabhakaran et al., 2019; among others), and none use instance segmentation. This is broadly justified—fractures are difficult to treat as discrete instances—but worth noting.

"GeoCrack" (p. 5) elaborate further what this is

Agree: An explanatory sentence was added: *"While initiatives like GeoCrack (Yaqoob et al., 2024), a publicly available benchmark dataset for automated geological fracture detection in RGB images of rock outcrops"* **(see L105)**.

"incomplete annotations." (p. 5) Unclear what you mean by this

Agree: After exploring the GeoCrack dataset we found missing annotations in some patches (see figures below). The text now reads: *"... and contain incomplete annotations in some image patches where only a subset of visible fractures are labelled."* **(see L107)**.



Methods

"This framework enables systematic comparison of detection performance across diverse datasets, geological settings, and imaging conditions. By harmonising data processing, parameter tuning, and evaluation metrics, it promotes transparency, reproducibility, and comparabil65 ity in automated fracture mapping research, while also providing a useful test dataset and benchmark for future developments." (p. 6) Is 3 datasets really sufficient to create a lithologically diverse training set?

Agree: We acknowledge this limitation explicitly and have added a dedicated paragraph in the Methods (Section 2.1.1) that reads: *"We acknowledge that three datasets covering two granite types and one clastic sedimentary setting, while providing meaningful lithological contrast for a first benchmark, do not represent the full diversity of geological settings encountered in practice; notably absent are carbonate, metamorphic, and volcanic lithologies. FraXet is intended as a living benchmark to be expanded as further annotated public datasets become available; including these other lithological settings is a priority for future work."* **(see L145)**.

"FraXet (Fatihi et al., 2025a)" (p. 6) As this is the dataset you have used within this work - rephrase to "we integrate three annotated outcrop datasets to create FraXet, a training dataset used to X, Y, Z" or similar.
(p. 6) Make this sentence clearer please - splitting may help.

Agree: Rephrased to: *"We designed a benchmarking workflow (Fig.3) applied to three publicly available, annotated outcrop datasets … that make up the first version of FraXet. FraXet is a curated benchmark dataset used for training and evaluating automatic fracture trace extraction models, comprising high-resolution orthophotographs and DEMs across diverse lithologies and outcrop geometries."* **(see L121)**.

"We standardise and pre-process all patches, then apply a suite of traditional image-processing filters (Canny, Sobel, Gabor, Sato, phase congruency) and deep learning models (U-Net, SegFormer) to generate per-pixel fracture probability maps. All methods are evaluated using a set of image-quality, segmentation and proposed similarity metrics on held-out test splits, with deep models tuned on validation subsets before final testing." (p. 6) This needs to be introduced better within the introduction so non-specialists can recognize the value of the contribution.

Agree: We have included a summary of the workflow at the end of the Introduction, highlighting the value of the contribution: *"The primary goal of this study is to establish a standardised evaluation framework for pixel-wise fracture segmentation from 2D outcrop imagery, incorporating both traditional computer vision and deep learning methods. Here we present FraXet, comprising three curated, heterogeneous benchmark datasets on which a suite of traditional image-processing filters (Canny, Sobel, Gabor, Sato, and phase congruency) and deep learning models (U-Net, SegFormer) were applied to generate per-pixel fracture probability maps…"* **(see L109)**.

(p. 6) Datasets are very small in this flowchart, consider increasing the size so the reader can see the similarities and differences

Agree: New figures showing the datasets in detail were added to the appendix **(see Figs A1-A3)**.

(p. 6) This figure has promise, but text size is quite small at present.

Agree: The workflow figure has been reworked for legibility **(see Fig. 3)**.

“Ovaskainen22 captures fracture networks in the Mesoproterozoic Wiborg Rapakivi Granite Batholith of southeastern Finland, mostly isotropic coarse-grained wiborgite facies granite with K-feldspar megacrystals up to 5–10 cm (Härmä, 2020). Matteo21 (Mattéo et al., 2021) presents data from the Granite Dells, Arizona, an anorogenic granite of similar age (~1.40 Ga) but 80 with medium-grained, and generally equigranular textures showing weak porphyritic tendencies (Krieger, 1965). Samsu19 (Samsu et al., 2019) map fractures in a very different geological setting: a fluvial–lacustrine syn-rift sediment succession in the Lower Cretaceous Strzelecki Group (Southeastern Australia), comprising mud- and sand-dominated siliciclastic strata with conglomeratic or organic-rich interbeds (Constantine, 2001). Together, these three datasets cover varied geological contexts (isotropic crystalline to anisotropic sedimentary lithologies) 85 with fractures mapped at different scales, providing a diverse testbed for evaluating model generalisation.” (p. 7) What scale were fractures mapped at? a comparison table i think would help here. Also what extent was mapped and importantly what was the studies purpose (for example, a connectivity paper will be different to one that is only interested in geometry)

Agree: We added new columns (**mapping scale** and **study purpose**) to the table describing the data **(see Table 3)**.

(p. 7) I find this confusing as written, it is not clear where this dataset is used.

Agree: We have clarified the role of the Lapiés di Bou and Bingie Bingie sites. The revised text now reads: *“Neither the Lapiés di Bou nor the Bingie Bingie data were included in the training, validation, or test splits; they were used exclusively for qualitative visual comparison and runtime evaluation, respectively.”* **(see L153)**.

“To ensure reliable model assessment and minimise spatial autocorrelation—a known issue in geospatial machine learning where spatial overlap inflates accuracy metrics by reducing the statistical independence of training and testing data (Wang et al., 2023)—we have defined spatially disjoint data splits by dividing each of the three datasets into training, validation, and 95 test regions (with no spatial overlap). Specifically, here the Ovaskainen22 dataset consists of Orrengrund (training), Kasaberget (validation), and Kampuslandet (testing) islands off the coast of the Loviisa region; the Samsu19 dataset comprises Eagles Nest (training) and Harmers Haven north (split between validation and testing); and the Matteo21 dataset is split into sites A and B (training), held-out regions of A and B (validation), and site C (testing). Only tiles containing at least one labelled fracture pixel were retained, to reduce challenges with class imbalance and to focus learning on relevant features.” (p. 7) it is hard to visualize the applicability of the areas chosen as there is no summary maps of the input datasets.

Agree: New figures were added to the appendix **(see Figs A1-A3)**.

Table 3 (p. 8) Some of this is useful, but there are key things missing -
1) what resolution were the datasets mapped at (for your masks) - this will directly effect the output parameters and should be considered.
2) where are the patches relative to each other - fracture networks are rarely homgenous,

without seeing a visual of where they are relative to the overall structure it is very difficult to assess the applicability of the input dataset.

Agree:
1) We added new columns describing the data (mapping scale, study purpose) **(see Table 3)**.

2) Patches were extracted on a regular grid with no overlap, retaining only patches that contain at least one labelled fracture pixel. An example is shown now in a new figure **(see Fig. 2)**.

"Wide-fracture expansion. Preliminary experiments revealed that the models underperformed on wide and visually prominent fractures. Many such fractures were annotated as 1-pixel-wide lines, underrepresenting their true width. To address this, we developed a region-growing algorithm that widens the single-pixel annotations to fill thick fracture zones. This algorithm 135 (Figure 3c) fills dark linear regions (in the green channel) by applying a maximum filter followed by morphological dilation and then merging the resulting (dark and connected) regions with the original binary mask. Only expanded areas connected to existing labelled pixels are retained to prevent false positives. This operation improved annotation fidelity for wide fractures. Multi-scale label smoothing. Manual fracture tracing also often leads to small misalignments between annotations and" (p. 9) This needs further explanation - the workflow is good but rather hidden in your methods.

Agree: We have expanded the description of both preprocessing steps in the revised Methods (Section 2.2). The wide-fracture expansion now specifies all parameters: *"This algorithm (Fig. 4c) fills dark linear regions (in the green channel) resulting from a thresholding operation (value of 30 on a 0 to 255 scale) by applying a maximum filter with a kernel size of 2 pixels followed by morphological dilation and then merging the resulting (dark and connected) regions with the original binary mask. Only expanded areas connected to existing labelled pixels are retained to prevent false positives."* **(see L201)**.

And the multi-scale label smoothing now specifies the kernel sizes and weights: *"First, labels are expanded with a distance of 2 pixels to fill small gaps and buffer the reference traces. This is followed by three successive morphological dilations applied to the expanded mask using disk-shaped structuring elements with radii of 2, 5, and 7 pixels. These bands are then combined into a graded representation using decreasing weights of 1/3, 1/5, and 1/9 respectively. This produces a smooth boundary around fractures (Fig. 4d)."* **(see L207)**.

§ (p. 10) ?
"§2.4.1" (p. 11) ?

Agree: Section-symbol (§) hyperlinks have been replaced with written section/subsection references throughout the manuscript.

(p. 12) Figure 5 and 6 are both very hard to follow and it is unclear and the text is very small making it difficult to read. Many of the terms in Figure 6 are not defined. Please vastly expand on the figure captions.

Agree: The figures have been redrawn **(Fig. 6 & 7)** and the captions have been substantially expanded. The SegFormer caption now defines all terms:

"Overview of the SegFormer architecture combining a transformer encoder with a lightweight decoder (Xie et al., 2021). The encoder uses a hierarchical transformer backbone (ResNet34) with multi-head self-attention to capture long-range dependencies. The decoder aggregates multi-level features via a simple MLP. Terms: Self-attention: mechanism allowing each pixel to attend to all others; MLP: multi-layer perceptron; Feature map: spatial representation at each encoder stage." **(see Fig. 7)**.

Benchmark metrics (p. 13) How these are applied to the datasets need to be better outlined

Clarify: The subsequent paragraphs in Section 2.6 explain in detail how these metrics are applied in this study.

- "We include the background class …" **(see L274)**.
- "We also include ROC-AUC, which evaluates the model's ability to distinguish between fracture and background pixels across thresholds …" **(see L288)**.
- **Table 5**

(p. 15) what is the field of view of this image? And from what dataset?

Agree: Scale bar added to the figure, and figure caption now end with: *" Image from the Ovaskainen22, KL5 dataset. Abbreviations are defined in Table 5."* **(see Fig. 8)**.

(p. 15) You need to define there in the figure caption

Agree: We added to the caption this sentence: "The abbreviations are already presented in table 5" **(see Fig. 8)**.

"discarding the first row" (p. 15) unclear why/what you mean here

Agree: We have clarified this. The caption now reads: "Values in bold show the best scores, excluding scenarios a, b and c." **(see Fig. 8)**.

"FracSim: a fracture-level similarity metric" (p. 16) I like this test, it adds considerable value to your tests.

We appreciate the positive feedback.

"§2.6 and §2.7" (p. 16)

Agree: Section symbols have been replaced with written section/subsection references.

"2.9 Deployment As a first step toward practical deployment, we developed a lightweight web-based prototype that enables users to upload 270 paired RGB images and DEMs and obtain deep-learning–based fracture predictions via the hassle-free Gradio interface (Abid et al.,

2019). The tool, hosted on Hugging Face Spaces https://huggingface.co/spaces/ayoubft/fractex2D_tuto, illustrates how the trained model can be made accessible beyond a purely research-oriented setting. While still preliminary, this prototype highlights the potential of interactive platforms to broaden the adoption of automated fracture detection within the geoscience community and offers a concrete example of integrating deep-learning outputs into field and laboratory workflows. The inter275 face additionally allows users to explore a range of computer-vision filters and metrics calculation." (p. 17) This feels more like a discussion point than your methods, consider moving down. When following the link, i got stuck on "building" for ~15 minuets so could not assess the viability of the claims.
(p. 17) Eventually it worked, it is quite a useful tool, however, i also think it highlights the limitations of automated methods in that it does not consider geological connections.

Agree: We have moved Section 2.9 to the Discussion (Section 4.3).

The delay was a cold start: Hugging Face free Spaces "sleep" when idle, so the first visit must wake the server, pull the container, and download the models. We have since implemented a workaround that keeps our Space live.

We have also developed a simple experimental QGIS plugin (as recommended by the second reviewer) https://plugins.qgis.org/plugins/fraXtex/ that allows users to apply the trained models.

Yes: a critical limitation of the current semantic segmentation approach is that it only predicts per-pixel confidence that a location is fracture versus background; this is why we emphasise the need for post-processing.

"(CV)" (p. 17) Given CV usually refers to Coefficient of Variance, consider changing this.

Agree: We have replaced "CV" with "computer vision" written in full throughout the manuscript.

Results

"Figure 8. Radar chart summarizing the performance of the classical filters across all evaluation metrics. All metrics originally lie between 0 and 1, except PSNR⊙ and FracSim⊗. And metrics where lower values indicate better performance (AE⊖, MSE⊖, Loss⊖ and FracSim⊗) were flipped by computing (1 – value) so that the chart consistently displays better performance toward the outer radius. FracSim⊗ was calculated on the modified test set." (p. 18) Add the evaluation metrics to a key or the figure caption to aid the reader

Agree: We added to the caption this sentence: "Abbreviations are defined in table 5" **(see Fig. 9)**.

"Using grid-searched parameters across all test sites, the classical filters produced somewhat useful but generally sparse outputs (Figure 9)." (p. 18) You need to better lead the reader through these results by pulling on examples in the text.

Agree: We have added specific examples referencing figure panels. The revised text now reads: *"Using grid-searched parameters across all test sites, the classical filters produced somewhat useful but generally sparse outputs (Fig. 9). For example, in the Ovaskainen22 test patch (Fig. 9a–f), Phase Congruency (panel b) detected several fracture segments but introduced substantial noise, while Sato (panel e) captured only the most prominent ridges. In the Samsu19 patch (Fig. 9m–x), all filters struggled with the coarser resolution and sedimentary texture. Phase Congruency and Canny often highlighted many true fracture segments, but their responses included substantial noise and many thin lines."* **(see L359)**.

"Sobel and Canny captured more pronounced edges along major traces, while Sato produced thicker and more continuous lines." (p. 19) as above, link directly to the figure

Agree: The text now links directly to the figure. The qualitative results section has been revised with specific panel references **(see Section 3.1.2, from L359)**.

"These tuned parameters, although effective in isolated cases, often fail when applied to heterogeneous datasets." (p. 19) for example, xxxx

Agree: This conclusion is drawn from the comparison between parameters chosen individually on each site and those automatically optimized across all sites. The grid-searched universal parameters, while effective on individual sites during manual tuning, produce noisy or incomplete maps when applied uniformly. The revised text reads: *"While manually tuned parameters can produce visually plausible outputs on individual test sites, this performance does not generalise well to other locations, or even to different regions within the same image. Traditional filters are sensitive to local contrast, scale, and texture, requiring site-specific adjustments to capture relevant features."* **(see L370)**.

"Deep networks can, in principle, generalise beyond specific sites by capturing shared structures across datasets."(p. 20) But what could it also not degrade the interpretation where there are significant differences?

Agree: We agree that learning shared representations across datasets may also reduce sensitivity to site-specific characteristics if geological differences are sufficiently large. Generalisation is therefore not guaranteed and depends on the diversity and representativeness of the training data. This is now noted in the Discussion: *"Our cross-dataset experiments show that diverse training data improves model performance. A general model trained on all sites consistently outperformed models trained on individual datasets, especially on overlap- and detection-sensitive metrics such as F1 score and IoU."* **(see L495)**.

And the limitation is acknowledged in Data Section 2.1.1: *"We acknowledge that three datasets covering two granite types and one clastic sedimentary setting, while providing meaningful lithological contrast for a first benchmark, do not represent the full diversity of geological settings encountered in practice."* **(see L145)**.

"Figure 11 shows a representative training run for both architectures using the fixed hyperparameters mentioned in the Methods section and a single random seed (42), which

ensures that weight initialization and any other stochastic processes are reproducible and consistent across runs.” (p. 21) please describe what this means in relating to training fracture sets

Agree: The use of a fixed random seed (42) ensures that another researcher running the same code with the same data and seed would obtain identical loss curves and model weights, providing full reproducibility. The text now reads: *"Fig. 11 shows a representative training run for both architectures using the fixed hyperparameters described in the Methods section and a single random seed (42), ensuring that weight initialisation and other stochastic processes are reproducible across runs. The training curves illustrate the model learning process on the fracture datasets, showing how the segmentation error changes as the models learn to identify fracture patterns from the annotated RGB+DEM patches over successive epochs."* **(see L381)**.

“the predicted pixels presenting fractures corresponded to true positives, while recall remained moderate (0.45),”(p. 22) Where fracture mapping is used for analysis such as resource estimations, missing ~50% of fractures mapped at a scale that we know will not capture everything is a significant loss!

Agree: This is an important point. We wish to clarify that the recall of 0.45 refers to pixel-level recall, not fracture-level recall. A fracture that is, for example, 4 pixels wide in the ground truth but predicted as 2 pixels wide would contribute 50% false negatives at the pixel level, yet the fracture itself is still detected. This is precisely why we propose the FracSim metric, which operates at the fracture-segment level and better reflects the practical impact on downstream analysis. The revised FracSim section now includes: *"Pixel-wise metrics are widely used for segmentation evaluation but have limitations when applied to thin, linear features such as fracture traces. Small differences in line thickness or slight spatial offsets can disproportionately affect precision, recall, and F1-score. For example, a correctly detected fracture predicted slightly thinner than the annotation will incur false negatives along its edges, substantially reducing recall despite successful detection."* **(see L308)**.

“the ground truth in terms of image similarity, but the segmentation metrics demonstrate reliable detection of fracture structures.” (p. 22) very unclear what you mean by this, please show in an image

Agree: This sentence means that for the task of detecting thin elongated features, the image similarity metrics are very sensitive to small spatial noise and thus report large mismatches between ground truth and predictions, even when the fractures are correctly detected. Now the sentence reads: *“SSIM (0.49) and PSNR (14.7 dB) confirm that the reconstructed fracture masks deviate notably from the ground truth in terms of image similarity, yet the segmentation metrics demonstrate **decent** detection of fracture structures.”* **(see L394)**.

“emphasising the challenging nature of this task and relatively limited available training data” (p. 22)split into a new sentence, this is a very key point

Agree: This has been split into a new sentence. The revised text now reads: *"Although the U-Net and SegFormer models clearly outperform filter-based methods, their absolute performance remains well below standard deep learning studies (F1 > 0.8, IoU > 0.6) (Guo et al., 2023; Kus et al., 2024; Sumi et al., 2024). This emphasises the challenging nature of fracture segmentation as a task and the relatively limited training data currently available."* **(see L399)**.

"This demonstrates that DL predictions not only overlap ground truth spatially (IoU, F1) but also better reproduce fracture segment length statistics." (p. 22) Have you considered showing two of the histograms? you rely heavily on the metric, but it would be good to see the data.

Clarify: We agree that visualising the fracture length distributions could provide additional insight. However, FracSim is computed directly from these histograms, and plotting them for every experiment would add a large number of figures with limited additional information beyond the reported metric. We therefore retain FracSim as a compact quantitative summary of the distributional similarity.

"The comparison of models within each site (Table 6) shows that the combined model (M_all) provides clear advantages over dataset-specific models, particularly in terms of F1 and IoU." (p. 22) A limitation of this work is that the workflow is not tested on a non-interpreted dataset, so you would expect the performance to be fairly good in comparison to a blind test.

Agree: We note that the Lapiés di Bou and Bingie Bingie patches included in this study were not part of the training set and serve as qualitative blind tests (Fig. 14). However, we agree that a fully quantitative blind test on a non-interpreted dataset is needed.

"This represents more than a twofold increase in overlap metrics, suggesting that the increased variability in the combined training set enables better generalization to the more complex fracture patterns present in the Samsu19 dataset." (p. 23) For these sorts of statements, it would be very helpful to have access to the original interpretations without having to go to the source publications. At a minimum, you need to include a new figure that shows these, but ideally, expand on the geological setting section to describe key aspects of the networks used to train the data.

Agree: New figures were added to the appendix **(see Figs A1-A3)**.

"Error Case Analysis" (p. 24) I found this section difficult to link to parts of the figure, please ensure you better link to your figures throughout, maybe using i, ii, iii for each panel.

Agree: We have referenced specific figure panels throughout the Error Case Analysis text. The revised text now includes panel-level references: *"High-confidence errors. In several cases, the CNN produced confident predictions that did not match the annotations. …, e.g., the over-simplified traces and omissions visible in panels (h) and (i) of Fig. 14. …*

Another high-confidence error type…data in panel (e) ….

Low-confidence errors. .. in panels (a–c). …, or the black-and-white scale panel in (d))…, while panel (f),...." **(see L440)**.

"annotation quality, artifacts in inputs, and domain heterogeneity" (p. 25) But all data will have these

Agree: We agree that these issues are inherent to all geological datasets, and this is precisely the challenge that motivates the need for robust, generalisable models. The revised text acknowledges this: *"Taken together, this analysis suggests that many errors stemmed not from the CNN architecture itself but from data limitations: annotation quality, input artifacts, and domain heterogeneity."* **(see L457)**.

"In contrast, our automatic 395 deep learning approach generated probability maps in 6 seconds (on free tier huggingface space (2vCPU and 16GB RAM)). While post-processing is still required to derive a fracture map (instance segmentation and vectorisation), this speed-up could enable a substantial reduction in effort and time during large-scale fracture interpretation." (p. 25) How long did the post-processing take? I think there is a key aspect of this that there will be a learning curve to apply any of these methods correctly - stating 6 seconds per map i think is vastly exaggerating what is realistic.

Clarify: The reported 6-second timing refers only to probability map inference; post-processing (vectorisation) is outside the scope of this study and was not timed. We agree that the 6-second figure should not be interpreted as the total time required to produce a final usable fracture map. There is no learning curve involved in generating the probability maps themselves, as users simply input their imagery and receive a probability map in return.

", automatically with the U-Net model (g, h) and with SegFormer model (i, j)" (p. 26) Although the automatic methods have done a reasonable job of capturing the main geometries of features, the connections are poorly captured based on the above images.

Agree: We agree that a critical limitation of the current semantic segmentation approach is its inability to explicitly model topological relationships between fractures. Addressing this will require post-processing algorithms that are aware of connectivity, or models that first detect fracture nodes. This is now discussed in Section 4.2 ("The Need for Post-processing"): *"Semantic segmentation models do not explicitly preserve fracture connectivity; intersections and network topology are therefore often poorly reconstructed even when the main fracture geometries are correctly detected. Without these steps, downstream fracture statistics remain noisy and difficult to interpret."* **(see L504)**.

Discussion

"4 Discussion" (p. 26) This discussion does not fit your work into any of the wider work in this field, and alongside the introduction, is currently unfortunately not sufficient for publication. Many of these effects are discussed within the image processing literature, and should be linked back

to why your work matters. The results are in general good, if difficult to follow at times, but without fitting these into context, the work will not be picked up by the community.

Agree: The Discussion has been substantially expanded to situate our work within the broader literature. A new paragraph comparing our results with previous studies has been added (Section 4.1): *"Our results are consistent with previous studies that applied U-Net architectures to fracture detection.Chudasama et al. 2024 reported F1 scores of 0.78 - 0.926 on granite outcrops in Finland using U-Net, though their evaluation did not include the background class, which inflates metrics relative to our methodology. Mattéo et al. 2021 achieved a Tversky Index (similar to IoU) of 0.35 - 0.68 using a similar architecture but on a single lithology. Our M_all model achieves F1 = 0.48 and IoU = 0.31 across heterogeneous datasets with the background class included, suggesting that cross-site generalisation comes at a performance cost that is expected to decrease as more training data become available."* **(see L476)**.

"Unsurprisingly, traditional edge and ridge filters proved inadequate for robust fracture mapping across heterogeneous datasets." (p. 26) link to your data

Agree: We now link directly to specific results. The revised text reads: *"Traditional edge and ridge filters proved inadequate for robust fracture mapping across heterogeneous datasets (Figs. 10, 11 and Table E1)."* **(see L472)**.

(p. 26) how does this compare to previous studies?

Agree: See the expanded Discussion text (Section 4.1), quoted in the response to the general Discussion 2 comments above.

"Nevertheless, absolute performance remained modest compared with segmentation benchmarks in other domains. This is likely due to the highly unusual nature of this semantic segmentation task - the selection of thin (high-aspect-ratio) pixel groupings (edges) in a highly class-imbalanced 410 dataset - and the variety of geological and environmental factors that change fracture appearance or introduce obfuscating features. Label noise and misalignment are also likely contributors, although practical limitations around data collection and manual labelling make it challenging to acquire better training data." (p. 26) link back to the fracture network literature, and why geological heterogeneity will impact this. Training data will be inherently bias, based on a range of factors and the scale it is collected at.

Agree: The Discussion now links these points to the fracture network characterisation literature. We cite Peacock and Sanderson (2026) on the influence of map resolution and quality on fault network topology, and discuss how geological heterogeneity and annotation scale affect training data quality. The revised text in Section 4.1 reads: *"This likely reflects the unusual nature of the task—selecting thin (high-aspect-ratio) pixel groupings in a strongly class-imbalanced dataset—and geological and environmental factors that alter fracture appearance or introduce confounding features. Label noise and misalignment are also likely contributors, although practical limits on data collection and manual labelling make better training data hard to acquire."* **(see L484)**.

“Interestingly,” (p. 27) subjective, remove

Agree: Replaced with "Notably," **(see L403)**.

“(see subsection 3.2.” (p. 27) link to figs & close bracket

Agree: Corrected to "(see subsection 3.2 and Figure 12)."

“Their” (p. 27) ?

Agree: "Their" has been replaced with "SegFormer's" for clarity.

“This suggests that exposure to varied lithologies, lighting, and fracture patterns helps the model learn 425 more transferable features and reduces overfitting to site-specific conditions. These results highlight the importance of multisite datasets for robust fracture segmentation.” (p. 27) I am not convinced that this would be the case across a wide range of data outside the training datasets - without testing the method on a dataset without training, could it not be that these sites all show similar arrangements?

Clarify: We agree that generalisation to entirely new datasets cannot be guaranteed from three training sites alone. The revised text in the Discussion now acknowledges this: *"Exposure to varied lithologies, lighting, and fracture patterns appears to yield more transferable features and reduce overfitting to site-specific conditions. These results underscore the value of multi-site datasets for robust fracture segmentation."* **(see L497)**.

And the limitation paragraph in Section 2.1.1 adds: *"We acknowledge that three datasets covering two granite types and one clastic sedimentary setting, while providing meaningful lithological contrast for a first benchmark, do not represent the full diversity of geological settings encountered in practice; notably absent are carbonate, metamorphic, and volcanic lithologies. FraXet is intended as a living benchmark to be expanded as further annotated public datasets become available; including these other lithological settings is a priority for future work."* **(see L145)**.

“fracture connectivity).” (p. 27) From the presented maps, fracture connectivity is very poorly captured, manual analysis will consider connections and test the geological realism of the interpretation during digitalization. THe knowledge of the network gained in this phase should not be disregarded, and care will need to be take using datasets derived using these methods without due diligence of the geological processes than underpin them.

Agree: We agree that a critical limitation of the current semantic segmentation approach is its inability to explicitly model topological relationships between fractures. This is now discussed in Section 4.2 ("The Need for Post-processing"): *"Semantic segmentation models do not explicitly preserve fracture connectivity, so intersections and network topology are often poorly reconstructed even when the main fracture geometries are correctly detected*. Without these steps, downstream fracture statistics remain noisy and hard to interpret." **(see L504)**.

And we emphasise the importance of incorporating domain knowledge in post-processing: *"Post-processing should also incorporate domain knowledge (for example, weak orientation priors) to reduce spurious short segments and improve the geological plausibility of extracted traces."* **(see L507)**.

"Furthermore, we suggest that it might be possible to develop post-processing algorithms that can leverage the rich information in probabilistic model outputs and local knowledge on e.g., likely fracture orientations to (1) identify and remove false-positive fracture detections, and (2) extrapolate detected fractures across small gaps (based on their orientation), as is commonly done by human interpreters." (p. 27) This has not been tested however, and it is unclear to me how much post-processing time was required within the vectorization phase.

Clarify: In this study, we do not perform post-processing or vectorisation; our objective is limited to generating and benchmarking fracture probability maps. The paragraph in question is intended to outline future research directions rather than describe methods evaluated here. This is clarified in the revised Discussion (Section 4.2): *"Furthermore, we suggest that it might be possible to develop post-processing algorithms that can leverage the rich information in probabilistic model outputs and local knowledge on, for example, likely fracture orientations to (1) identify and remove false-positive fracture detections, and (2) extrapolate detected fractures across small gaps (based on their orientation), as is commonly done by human interpreters."* **(see L508)**.

(p. 32) Reference list is not sufficient to support the work and clearly identify the gaps that it aims to fill. See earlier comments regarding scope.

Agree: The reference list has been substantially expanded. New references added throughout the revised manuscript include: Aydin (2000), Berkowitz (2002), Bemis et al. (2014), Villarreal et al. (2022), Palamakumbura et al. (2020), Potts (2002), Bonnet (2001), Sanderson (2015), Bond (2007), Scheiber (2015), Peacock (2019), Andrews (2019), Peacock and Sanderson (2026), and others across both the classical methods and deep learning paragraphs.

Dear Dr Thomas Dewez,

We thank you for the positive and insightful review. We are particularly grateful for the detailed technical comments on preprocessing, which have helped us improve the clarity and reproducibility of our methodology.

Please note that to facilitate the evaluation of our revision, the page and line numbers of the reviewer's comments refer to the originally submitted manuscript while page and line numbers of our responses refer to our revised manuscript.

Best regards,
Ayoub Fatihi, Jefter Caldeira, Tom Beucler, Samuel T. Thiele, and Anindita Samsu

Reply to RC2: 'Comment on egusphere-2026-1097', Thomas Dewez, 23 Jun 2026

The article "Towards robust fracture mapping: benchmarking automatic fracture mapping in 2D outcrop imagery" addresses a key element in collating quantitative structural geology : fracture trace extraction.

In this contribution, the authors compiled a benchmark of three published empirical datasets named FraXet which combine colour RGB orthoimages, corresponding digital elevation models (DEM) and human-interpreted fracture traces. FraXet was processed in a coherent fashion with traditional image edge detection filters on the one hand, and with two deep learning approaches on the other hand in order to reproduce the manual fracture traces.

To maximize transposabiity of their findings, effort was made to render image contents comparable by standardizing image digital counts to zero mean and unit variance over patches of 256x256 pixels. Issues like departure between annotated fracture trace serving as ground truth and predicted fracture location is effectively taken into account in the procedure. I found this particlarly relevant as reference traces are never easy to draw and are subject of debate. A reference fracture data sets is always faulty in some way given locally ill-defined surface geometry or weak radiometric contrast. In accounting for this effect, the authors are not trying to hide this difficulty and provide a clever means of mitigating the issue.

The principle and merits of each processing are well presented. The methodology is clear and well explained even for those unfamiliar with deep learning methods nitty gritties. I enjoyed reading this work and feel more informed now than before.

Coming to manuscript improvements, I found that an object definition was lacking. Talking about "fractures" and their occurrence in rock faces may seem obvious and common-sense. Yet, a clear definition of them, particularly in term of their appearance in images and DEM would improve the purpose of the work and imply a clearer fitness-for-purpose evaluation. A descriptive paragraph explaining which raster properties fractures do display in images and DEM would be useful. Underlying my remark is the need to recall that a fracture trace is the piece-wise linear pattern (sometimes very shortly linear) drawn by the intersection of a possibly, but not always, planar rock surface discontinuity with the topography. Both surfaces needn't be absolutely planar and the projection of their intersection on an arbitrary plane (that of the orthomosaic) will affect their final appearance geometry.

Agree: A definition paragraph has been added at the start of the Introduction. The revised text now reads: *"A fracture trace is the linear to curvilinear intersection of a fracture with an exposed rock surface, commonly mapped as a piecewise-linear polyline (Potts, 2022; Bonnet et al., 2001; Sanderson and Nixon, 2015). In RGB orthophotographs it appears as a dark, high-contrast linear feature produced by differential weathering and illumination (Vasuki et al., 2014; Thiele et al., 2017b; Ren and Malik, 2003), whereas in the DEM it shows as a subtle linear relief or step (Koike et al., 1995; Raghavan et al., 1993). Its apparent geometry therefore depends on the projection of the discontinuity-surface intersection onto the orthorectification plane (Potts, 2022; Bemis et al., 2014)."* **(see L26)**.

Not only is the shape of a fracture traces not strictly linear, but they also have different appearances dependant on scale. Fracture lengths (and widths) cover a very broad set of sizes. Currently the text addresses this size definition question by talking about raster patches of 256 x 256 pixels. An actual metric dimension is lacking to relate image space to object space. Table 3 is not quite sufficient to grasp the descriptive scale of each data set.

Agree: We now state the physical dimensions represented by image patches for each dataset. The revised text reads: *"Given the spatial resolutions of the datasets, a 256 × 256 pixel patch represents a square region with side lengths ranging from approximately 0.13–0.33 m (Matteo21), 1.28–1.54 m (Ovaskainen22), and 7.4–8.2 m (Samsu19)."* **(see L172)**.

Further, the appearance of fractures in an image is a function of the fracture set orientation with respect to the outcrop surface and with respect to the orthorectification plane. I perhaps missed this information but representing the domain of relative orientation might be a good way to make your paper last through time. Learning has occurred on three ortho

mosaics cross-cutting fracture sets at such range of angle only. One guesses that the fracture sets are at steep angle with the projection plane so that traces are mostly narrow, but specifying parameter range explictly may help further down the line when the benchmark data will be used for other learning experiments.

Agree: In the current benchmark, the outcrops are predominantly sub-horizontal, so fracture orientation relative to the orthorectification plane is not a major source of variation. While we agree that orientation statistics would be valuable for future applications of the benchmark, extracting precise orientation distributions for each dataset would require additional analysis beyond the scope of this revision. We note this as a recommendation for future work in the Discussion (Section 4.4): *"Our results highlight several concrete next steps toward automated and objective fracture mapping: improving annotation precision (multi-scale labels, consensus labelling) in training data, expanding dataset diversity by including more sites as additional public data become available, and adopting architectures and loss functions tailored to the unusual thin geometry of fractures."* **(see L521)**.

To summarize, this paper is a very valuable contribution to 2D image-based structural geology analysis.

Implementation for the broader public through a QGIS plugin would certainly expand the user base and maximize this contribution. In there, offering a function for extracting segment lengths and intersection types for apparent fracture persistence and hierarchical connection would be excellent.

Agree: *We have developed a simple experimental QGIS plugin (https://plugins.qgis.org/plugins/fraXtex/) that allows users to apply the trained models (see Section 4.3). The revised Discussion now reads:* "To facilitate integration into existing geological workflows, we additionally developed a simple QGIS plugin (https://plugins.qgis.org/plugins/fraXtex/) that enables inference directly within QGIS. Together, these tools demonstrate how the proposed workflow can be readily adopted by the geoscience community." **(see L517)**.

We agree that extracting individual segments and properties such as segment length, intersection topology, and hierarchical connection would be very valuable. However, this is beyond the scope of the present contribution and is part of future planned work, as noted in Section 4.2 ("The Need for Post-processing") **(see L500)**.

Specific comments

line 135 about wide fracture expansion. The maximum filtering is clever but :

* what neighbourhood did it consider for the kernel? Kernel size of 2.

* what threshold qualifies as "dark pixels" ? Threshold of 30 (on a 0 to 255 scale).

* what justifies filtering only the green image band ? Wouldn't one of the principal component analysis components fit the purpose of capturing the most contrasted image better? Is it implictly evoking the Bayer matrix of digital cameras where there are twice more green-sensitive pixels than reds and blues? It wouldn't hurt to make an explanation sentence. But my processing preference would go for the richest PCA band. The green band was selected because most digital cameras use a Bayer filter with twice as many green-sensitive pixels as red and blue, resulting in higher spatial contrast. Additionally, vegetation is most prominent in the green band, making it useful for distinguishing vegetated areas from fractures. We note that the first principal component from PCA or other bands could alternatively be used.

The wide-fracture expansion now specifies all parameters: *"This algorithm (Fig. 4c) fills dark linear regions (in the green channel) resulting from a thresholding operation (value of 30 on a 0 to 255 scale) by applying a maximum filter with a kernel size of 2 pixels followed by morphological dilation and then merging the resulting (dark and connected) regions with the original binary mask. Only expanded areas connected to existing labelled pixels are retained to prevent false positives."* **(see L201)**.

line 138+ about the multi-scale label smoothing. Again, I really liked this clever approach, but:

* what kernel size did you use for buffering the reference fracture trace? An expansion radius of 2 pixels was used for the initial buffering, followed by additional dilation kernels of 2, 5, and 7 pixels weighted respectively *1/3, 1/5, and 1/9*.

* how does this kernel size relate to the digitization scale (or actual metric distance)? The kernel sizes are defined in pixels to take into consideration annotation uncertainty, not to represent a specific metric distance.

* does a single kernel size apply to all data sets irrespective of the operator and the metric size of the images? Yes, a single pre-processing is applied to all datasets.

A single kernel size seems inappropriate given the different scale of images and objects.

Clarify: The purpose of the label smoothing is to account for annotation uncertainty, i.e., the small spatial offset between the manually digitized trace and the true fracture location, rather than to represent fracture width. For this reason, the smoothing is defined in pixel space and is independent of the image GSD. Since fractures are resolved at a minimum of one pixel regardless of resolution, annotation uncertainty is expected to be similar across datasets in pixel

terms, with interpreters typically placing traces within a few pixels of the true location. Therefore, using a single kernel size for all datasets is appropriate.

The multi-scale label smoothing now specifies the kernel sizes and weights: *"First, labels are expanded with a distance of 2 pixels to fill small gaps and buffer the reference traces. This is followed by three successive morphological dilations applied to the expanded mask using disk-shaped structuring elements with radii of 2, 5, and 7 pixels. These bands are then combined into a graded representation using decreasing weights of 1/3, 1/5, and 1/9 respectively. This produces a smooth boundary around fractures (Fig. 4d)."* **(see L206)**.

line 145+ data augmentation

This is a classic approach in deep learning to augment the domain of variation of the source images.

How do you justify the gaussian blur dimension implemented? This is a simulation of out-of-focus photographs. How realistic/unrealistic is it regarding your practice of outcrop photography? Is the smoothing kernel size in range with encoutered out of focus image portions? Wouldn't it be wise to also implement the second source of image blur : c? The augmentation set fails to capture motion blur, arising from slow shutter speed. This image defect will probably enter in resonnance with fracture orientation. Fracture-normal motion blur will broaden the fracture trace and reduce local contrast. Fracture-colinear motion blur will likely not affect the appearance of the fracture. In turn, motion blur will affect both fracture trace location and sharpness (or probability, since you introduced this notion as metric accounting for prediction precision). A comment on the matter will suffice, no need to re-run the learning process.

Clarify: A Gaussian blur with a kernel size of 5 pixels (sigma $\in$ [0.1, 2.0]) was applied. This was used primarily to introduce texture variation (smoothing) in the target rock rather than to simulate out-of-focus blur specifically. We did not include motion blur because our benchmark consists of orthophotographs generated from multiple overlapping images, where blurred frames generally have limited influence on the final orthomosaic. More generally, motion blur primarily affects the quality of the input data rather than the fracture-mapping model itself. If the fracture location is displaced or obscured in the input image, the model can only learn and predict from the available visual information, not recover the underlying true fracture location. This is a well-recognised limitation of supervised machine learning and reflects the dependence of model performance on the quality of the training data. The augmentation parameters are listed in Table C1 in the Appendix.

How do you justify the over/underexposure values applied? Are they arbitrary? Do they correspond to one f-stop of under/over-exposure relating to possible camera's on-the-fly

mis-calculation? Could you say why and when this situation should occur in field photographs?

Clarify: Brightness and contrast adjustments (factors of 0.7–1.7 for brightness, 0.7–1.5 for contrast) are applied as random photometric augmentations rather than explicit f-stop simulations. They approximate realistic exposure variations caused by automatic camera settings, changing weather conditions, cloud cover, and time of acquisition. The ranges were chosen empirically to introduce moderate variation without degrading image quality beyond realistic field conditions. The full parameter ranges are listed in Table C1.

line 158 you suggest that only one channel was considered for traditional image processing filters. Which PCA channel? The text is not explicit. Is it always the first component that contains the best fracture signal?

Agree: The first principal component (PC1) was used for all traditional image processing filters. This has now been stated explicitly in the manuscript (Section 2.2, "Channel selection"): *"PCA (principal component analysis) was then applied to the normalised four-band data to obtain a single-band representation containing a maximal amount of variance."* **(see L195)**.

And in the filter section: *"All filters were applied to the first principal component (channel) resulting from the PCA reduction."* **(see L227)**.

The motivation was not the assumption that PC1 always contains the best fracture signal, but rather to provide a consistent single-channel input for the Python implementations of the traditional computer vision algorithms, which operate on grayscale images. PC1 preserves the maximum variance in the data. We note that the green band could also serve as a single-channel alternative.