Reviewer #1

- What are the major claims of the paper?

The authors claim to develop an integrated strategy package for reducing the memory footprint in 3D dense prediction of large volume. Their approach integrates operator spatial tiling, operator fusion, normalization statistic aggregation, and on-demand feature re-computation to reduce memory usage and accelerate runtime. They demonstrated the efficacy and superb performance of the approach using several well-known geological interpretation ML models.

- Are they novel and will they be of interest to others in the community and the wider field?

Yes. The proposed methods are new and I foresee that they will be used by a lot of relevant fields that rely on 3D dense prediction, e.g., medical imaging, physics field prediction, etc.

- Is the work convincing, and if not, what further evidence would be required to strengthen the conclusions?

I am impressed by the efficacy and performance of the proposed approach. The work is very convincing as it provides sufficient technical details, including pseudocodes, of their approach. This makes the approach reproducible and validated by other users. There are several minor issues I think the authors can improve: (1) Labels in Figure 3/S1 charts are too small, and need to be enlarged to ensure publication quality. (2) Please briefly describe why for ChannelSeg3D there is no obvious memory reduction compared with the original implementation, as I can see that this contradicts your claim that the new strategy will reduce memory footprint. (3) Discuss the potential of the proposed strategies for training phase usage. My experience with dense 3D volume prediction is that at least for some tasks, like RGT prediction (which is like monocular depth estimation), people may still some properly large data in the training phase to ensure inference quality, not simply like training on $128 \times 128 \times 128$ volume and simply predicting on $1024 \times 1024 \times 1024$ volume. (4) I would recommend adding some definitive mathematical equations for the described strategies in the Methods section. This will make people better understand the logic behind the strategies and improve the scientific value of the strategies. Software infrastructure change over time, so it is important to properly define the procedure with clear equations/algorithms so to pertain long-lasting value and enable implementations with other software architectures.

- On a more subjective note, do you feel that the paper will influence thinking in the field?

Absolutely. The work is the first that I know to specifically address a realistic and important issue for 3D dense prediction – memory usage. This will enable the practical usage and improve the performance of many ML models in the field of seismic interpretation and fields alike. This will enable people to focus more on science rather than worrying about memory bottleneck, and potentially accelerate scientific and engineering discoveries in many fields. It has both promising scientific value and engineering significance.

Reviewer #2

Dear authors,

I have been applying different deep learning models on reflection seismic data as large as 1.5 TB in size, and **the problem that your submitted manuscript addresses has been one of the fundamental challenges in this area.** Providing a solution to reduce memory consumption during model inference—without changing model training parameters-**is certainly a research area that deserves attention.** As you have clearly mentioned in the manuscript, several strategies have already been proposed, each with its pros and cons.

The pace of new network development is very fast, which makes it difficult to build memory-efficient inference approaches for all of them. Moreover, memory efficiency is usually not the primary goal when new networks are developed; instead, the focus is on improving prediction quality. However, **your grouping of the three main types of models currently used (Figure 2 in your manuscript) is well structured and clear.**

Overall, you have explained the challenge and your proposed four strategies **very well**, and you have supported the approach with at least two examples: faults and RGT cases. These two examples represent two rather opposite features in subsurface data—discontinuity for faults and continuity for RGT. The success of your proposed approach in improving both types is a strong indication of its efficiency.

After these positive comments, I have included a few additional remarks in the attached PDF file. They should be relatively straightforward to address.

One minor comment concerns the verification of results, for example in Figure 4. The common objective in fault and RGT interpretation is to improve continuity; however, this is not always geologically correct. In some cases, discontinuity is actually the right interpretation. I discussed this in detail in Alaei and Torabi (2024). Since subsurface data interpretation is inherently non-unique, it would be wise to mention in the discussion that improvements achieved through the four strategies should always be geologically sense-checked, as pitfalls can remain (except in synthetic label cases).

I enjoyed reading your manuscript, and I wish you the best of luck moving forward.

Reviewer #3

Dear Editor of Nature Communications Engineering and Authors,

I have read and reviewed the paper "Memory-Efficient Full-Volume Inference for Large-Scale 3D Dense Prediction without Performance Degradation" by Authors Jintao Li and Xinming Wu.

General comments:

This paper addresses an important and commonly encountered issue, being the mandatory partitioning or 'tiling' of multidimensional data input to Artificial Neural Networks (ANNs). This

partitioning, usually a result of CPU or GPU memory limitations, indeed results in boundary and edge effects or scaling issues when large spatial structures are to be segmented in a full 3D volume. These boundary/edge/scaling effects can be remediated by predicting with overlapping tiles, downscaled tiles and post-processing filters to removed edge artefacts. I myself am a practitioner of ANN segmentation, I recognize the problem and developed or adapted remediation solutions.

It is however much more desirable to attack the problem in its core: by manipulating the operators and kernels and preventing sub-optimal computation paths. **An added advantage of the current new approach is that the trained ANN models and input data need no retraining or conditioning, this is of great benefit for the AI community that uses ANNs for segmentation.**

The Authors demonstrate their four main improvements on some well-known ANN segmentation problems: 1) spatially chunked implementations, 2) operator chunked variants, 3) operator fusion and unified chunking, 4) on-demand re-computation. The results are very convincing and will help the AI community substantially. **This paper has therefore a large impact.**

The Authors claim that they achieve reduced memory usage and accelerated runtime at no performance degradation. They do raise some limitations and drawbacks in the section 3 Discussion, this is fine. The lack of performance degradation can certainly be the case, but the classic principle of 'there's no such thing as a free lunch' cannot be ruled out. The comparison Figure 4 looks compelling in terms of hardly any difference between the classic chunked strategy and the full-volume strategy. But I would like to see either 1) a difference section (chunked cube minus full-volume cube) or 2) an extreme zoom-in of the two strategies or 3) a table with some difference metrics/attributes like signal-to-noise, correlation, coherency etc. Or, as the Authors state, if there is a fundamental reasoning that can demonstrate their point, for example equivalent chunked and full convolutions, then please state this with a math formula, schematic or graph.

Specific comments:

In line 3 of page 2, the authors state that "These tasks typically require continuous estimation across the entire 3D volume, ..." This is certainly not always the case, as many 3D volumes of data contain small enough structures to be segmented by tiled input, and sometimes smaller tiles are even advantageous. It's all a matter of scale and resolution of the input data. I would rephrase this sentence to "These tasks sometimes require continuous estimation across the entire 3D volume, ..."

All in all, **this paper is a welcome contribution to the AI community and brings substantial improvements to the field**. Given that the Authors will publicly publish their code, this work and its analyses will be reproducible by peers in the field. I recommend minor revisions as mentioned above for publication.