Revision Notes, egusphere-2025-3146

Dear Armin,

We would like to express our sincere gratitude for your time and feedback on our manuscript, " **Rapid Flood Mapping from Aerial Imagery Using Fine-Tuned SAM and ResNet-Backboned U-Net**."

The author's reply to the comment:

Comment	Response
Dear Hadi, Thank you—please don't get me wrong—you are one of our remote sensing community, and your work is good. I don't want to dwell on the similarity aspect, as we've already explored comparisons between SAM and UNet50 (with ResNet backbone) in the context of water segmentation using close-range remote sensing images (from UAVs, smartphones, and handheld cameras, within a 1–300 meter range) (https://doi.org/10.1109/ACCESS.2024.3385425). However, I'd like to suggest that you consider referencing our results, as they align with your findings and could strengthen your discussion section. For instance, we observed that SAM performs very well in general. However, when segmenting images from the same area, UNet actually produced even better results. This nuance might enrich your discussion, especially when highlighting the practical performance differences between models.	Dear Armin, Thank you once again for your valuable feedback. We truly appreciate the time and effort you have taken to engage with our work. In our work, we not only compared U-NET and SAM, but also evaluated two types of input prompts in SAM (points and bounding boxes) and two types of backbones for U-NET (ResNet-50 and ResNet-100). We agree that including a direct comparison with your findings will make our paper more comprehensive, and we will add this comparison to the results section of our revised manuscript. Regarding the ViT backbone, we also evaluated multiple variants and observed that their effectiveness for flood-affected area segmentation was comparable. To optimize computational resources, we selected ViT-Base, as it provides a favorable balance between accuracy and efficiency.

Another point is related to the computational aspects: SAM typically operates on 1024×1024 patches, and when fine-tuning with a frozen ViT backbone, it still requires significant computational resources. The choice of ViT backbone also matters—ViT-H is quite heavy and not ideal for fine-tuning, whereas the smaller variants (like Tiny ViT and Medium ViT) tend to perform better with fewer resources. Lastly, regarding datasets: One could argue that with SAM, we might not need large annotated datasets anymore. While SAM reduces the need for manual annotation, I would still say that datasets are necessary. The real question is: how many do we actually need? To help address this, you might consider referencing this recent paper by Professor Anette Eltner from Dresden University: https://doi.org/10.1080/01431161.2025.2457131 It discusses the sensitivity of model performance to dataset size and could help you frame this as a potential advantage of SAM over UNet. I'd love to hear your thoughts and see how you might incorporate some of these ideas into your discussion. Warm regards, Armin	We also appreciate your recommendation of Professor Anette Eltner's recent article. It provides an excellent perspective on dataset size requirements. The number of labeled images needed depends on the complexity of the task. For instance, in our previous study on erosion and deposition segmentation (https://doi.org/10.1016/j.catena.2025.108954) we worked with about 400 labeled images and observed clear performance gains with increasing dataset size, an effect that was particularly relevant given the higher complexity of that task compared to flood mapping. This increased complexity was because eroded and non-eroded soil often have very similar visual characteristics, whereas flooded areas are usually more distinct from their surroundings. Thank you again for your thoughtful suggestions. They will certainly help us improve the clarity and impact of our paper.