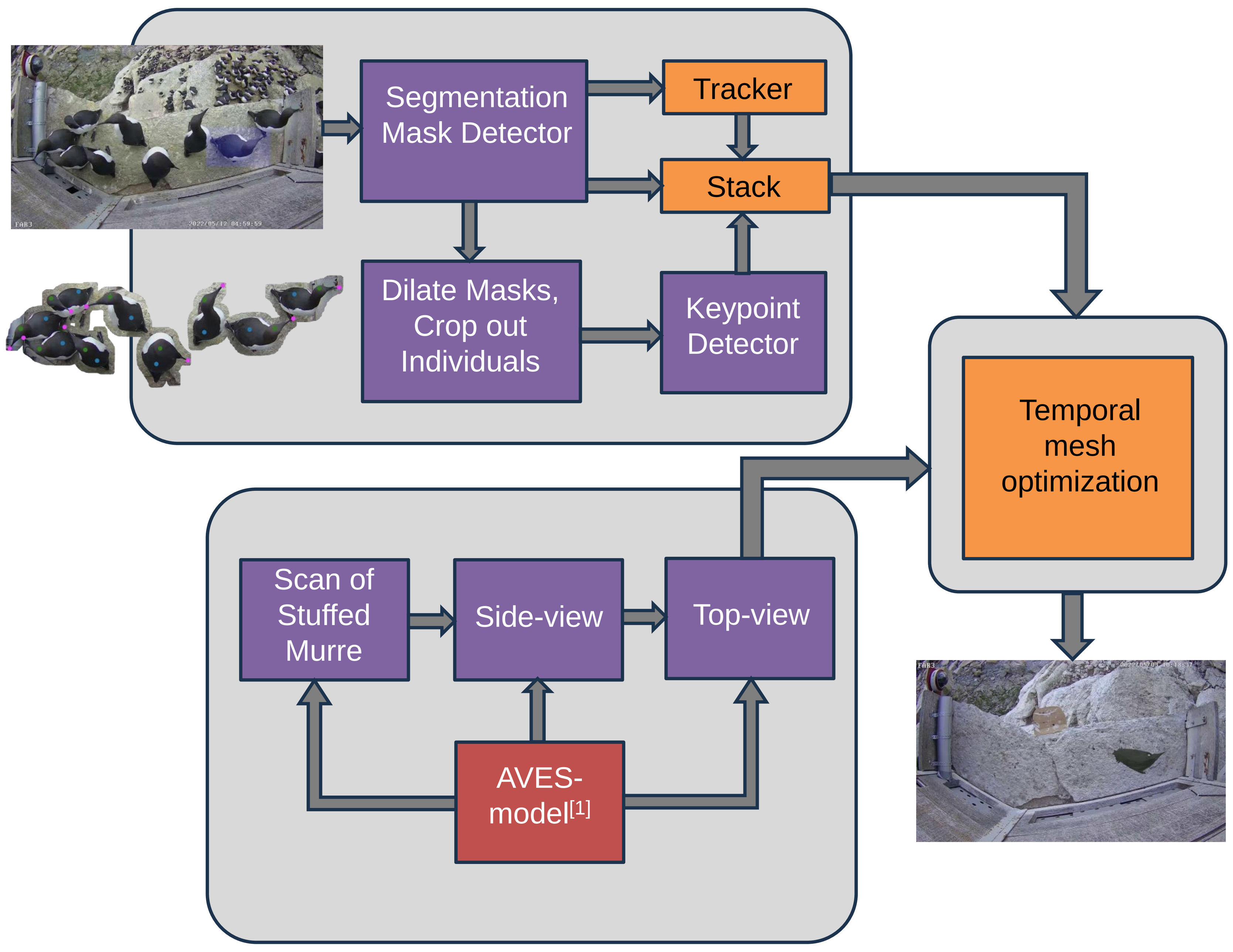
Temporally-consistent 3D Reconstruction of Seabirds

Johannes Hägerlind[1], Jonas Hentati Sundberg[2], Bastian Wandt[1]
[1]Linköping University, Sweden [2]Swedish University of Agricultural Sciences

Contribution

- We reconstruct the 3D pose and shape from monocular videos of the common murre.
  - A full pipeline of detection, tracking, segmentation, and temporally consistent 3D reconstruction (using a parameterised mesh-model [1]).
  - A temporal loss that extends current single-image 3D bird pose and shape estimators [1, 2, 3] to the temporal domain.
- A dataset of 10K frames, including a smaller test set with bird-specific keypoint labels.



References

Hägerlind, J., Hentati-Sundberg, J., & Wandt, B. (2024). Temporally-consistent 3D Reconstruction of Birds. *arXiv preprint arXiv:2408.13629*.

[1] Wang, Y., Kolotouros, N., Daniilidis, K., & Badger, M. (2021). Birds of a feather: Capturing avian shape models from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14739-14749).
[2] Badger, M., Wang, Y., Modh, A., Perkes, A., Kolotouros, N., Pfrommer, B. G., ... & Daniilidis, K. (2020, August). 3D bird reconstruction: a dataset, model, and shape recovery from a single view. In *European Conference on Computer Vision* (pp. 1-17). Cham: Springer International Publishing.
[3] Hägerlind, J. (2023). 3D-Reconstruction of the Common Murre (master's thesis).

Temporal optimization

- Total loss is:

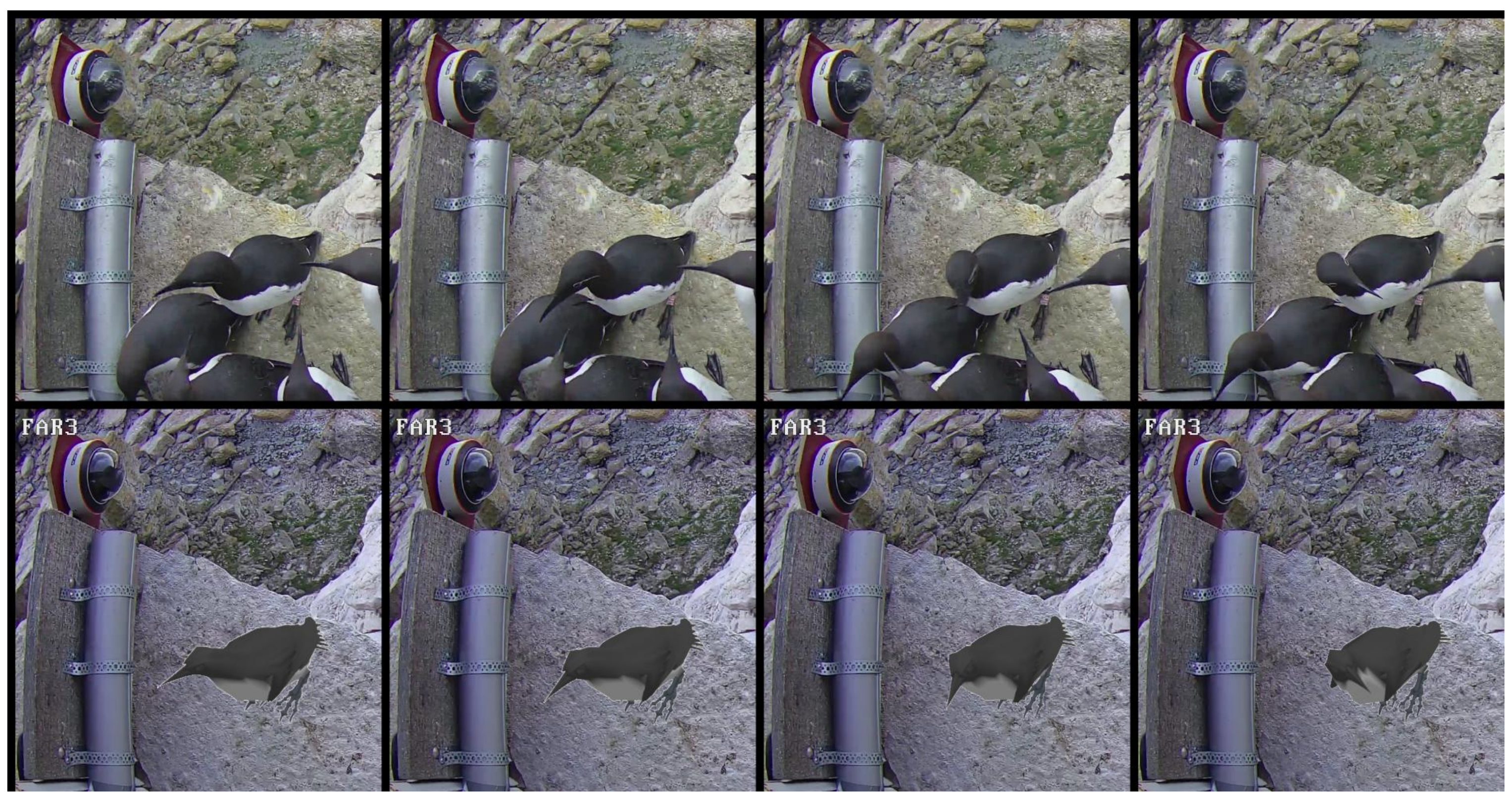
$$E = \lambda_{kpt}E_{kpt} + \lambda_{msk}E_{msk} + \lambda_{pp}E_{pp} + \lambda_{bp}E_{bp} + \lambda_{vel}E_{vel} + \lambda_{acc}E_{acc}$$

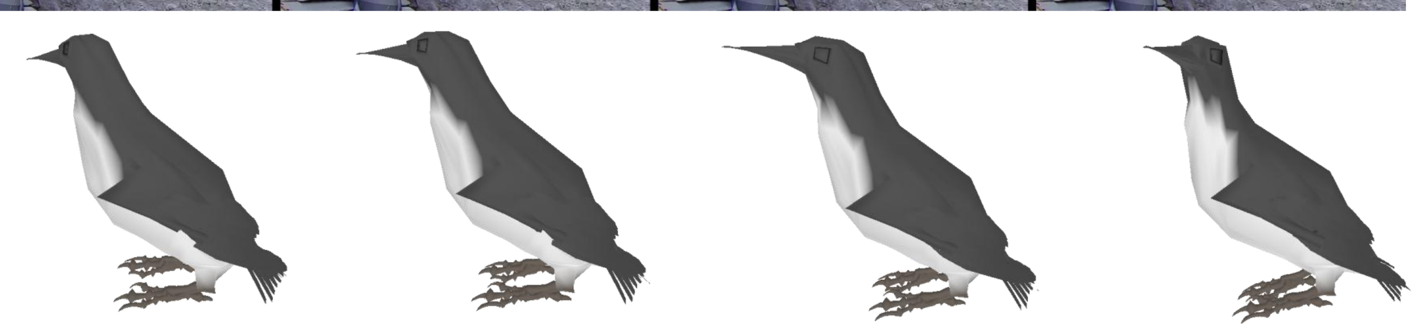
  - E_{msk} : mask loss, squared L2 norm [3],
  - E_{kpt} : keypoint reprojection loss, using weighted Geman-McClure [2],
  - E_{pp} : pose prior loss [2], using Mahalanobis distance,
  - E_{bp} : bone lengths prior loss [2, 3], using Mahalanobis distance,
  - λ_{kpt}, λ_{msk}, λ_{pp} and λ_{bp} from [3].
- We introduce a temporal loss:

$$E_{vel} = \sum_{k\in\{g,p\}} \beta_k \sum_{i=1}^{N} \left\|\theta_{k,i+1} - \theta_{k,i}\right\|_2$$

$$E_{acc} = \sum_{k\in\{g,p\}} \beta_k \sum_{j=2}^{N} \left\|\theta'_{k,j+1} - \theta'_{k,j}\right\|_2,$$

where θ_g and θ_p represent the global orientation and bone poses, respectively.





	λ_{vel}	λ_{acc}	med	size	me_p
Baseline[3]	0	0	False	False	0.0824
Ours	0	0	True	False	0.0809
	100	100	False	True	0.0791
	100	*100*	*True*	*False*	*0.0758*
Ours (best)	100	100	True	True	0.0756

- ***size*** – optimize a single size in the temporal window
- ***med*** – use a weighted median filter on the keypoint predictions.
- me_p – the mean reprojection error for the keypoints, with respect to the longest side of the bounding box around a bird.

Paper **Dataset**