

Quantifying the Impact of Segmentation Inaccuracies on Cell Phenotyping in Imaging Mass Cytometry

Johannes Schuiki¹ Markus Steiner^{2,3} Heinz Hofbauer¹ Stephan Drothler^{2,3,4}
Giulia Pessina^{2,3} Richard Greil^{2,3} Nadja Zaborsky^{2,3} Andreas Uhl¹

¹Dept. of Artificial Intelligence and Human Interfaces, Paris-Lodron-University Salzburg, Austria

²Cancer Cluster Salzburg, Austria

³Dept. of Internal Medicine III, Paracelsus Medical University, Austria

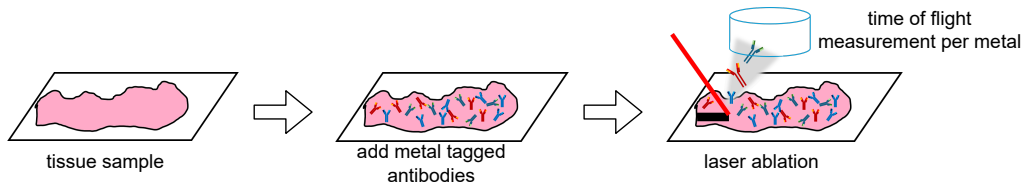
⁴Dept. of Biosciences, Paris-Lodron-University Salzburg, Austria

Aug 22, 2026



Imaging mass cytometry

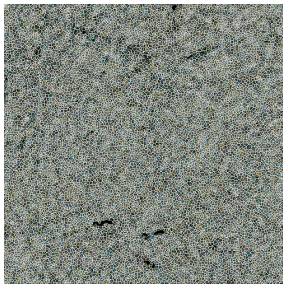
- use rare earth metal - antibody mixture that binds on proteins
- vaporize the tissue, accelerate the "smoke" and do time of flight measurement



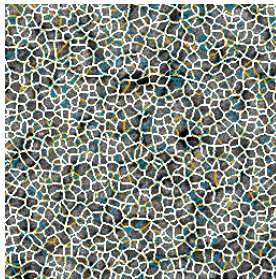
- commercialized system named "Hyperion" by Fluidigm
- fixed-size resolution: $1\mu\text{m}$ per pixel
- many channels (our data has 50)
- tissue sample gets destroyed in the process

Motivation through previous work

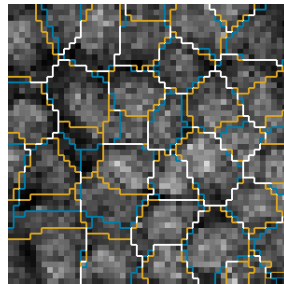
- Experts annotated patches manually using Ilastik¹ to generate pixel probability maps (background, nuclei, membrane)
- Probability maps are expanded to the whole mage using CellProfiler²



full 1000 x 1000 image



256 x 256 crop



50 x 50 crop (upscaled)

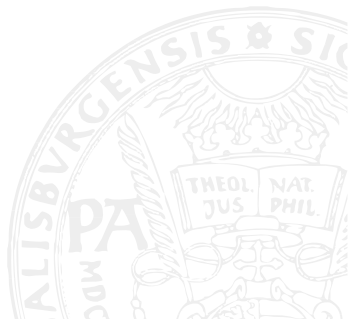
¹ S. Berg et al. (2019). "ilastik: interactive machine learning for (bio)image analysis". In: *Nature Methods*

² D. R. Stirling et al. (2021). "CellProfiler 4: improvements in speed, utility and usability". In: *BMC Bioinformatics*

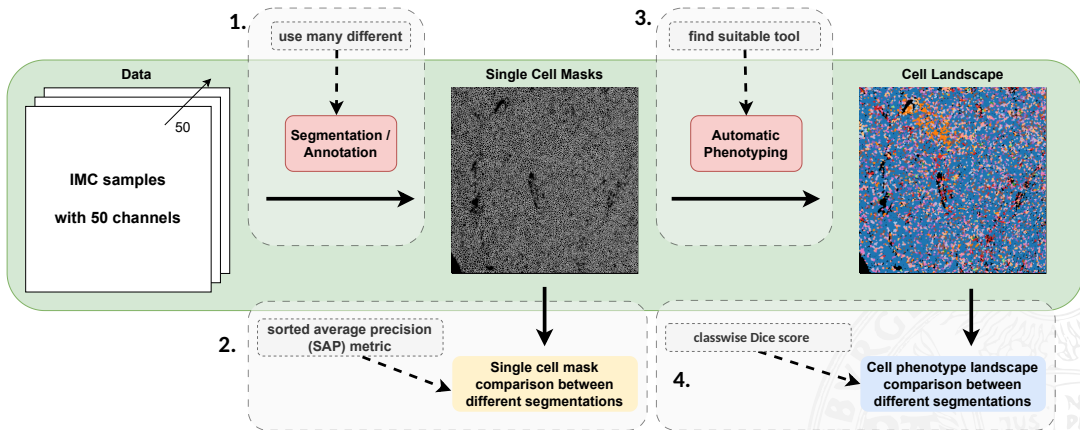
Aim of this research

Evaluate effects of segmentation errors for downstream task of phenotyping

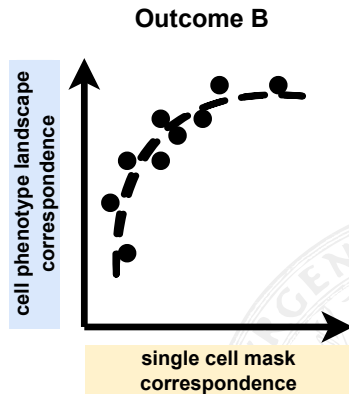
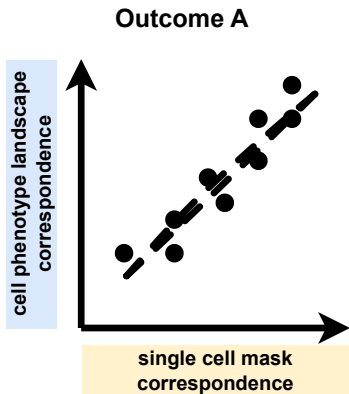
- One common downstream task is cell phenotyping (assign celltype label to each single cell)
- Segmentation is a prior step for this
- Do masks need to be "perfect" or just "good enough"?



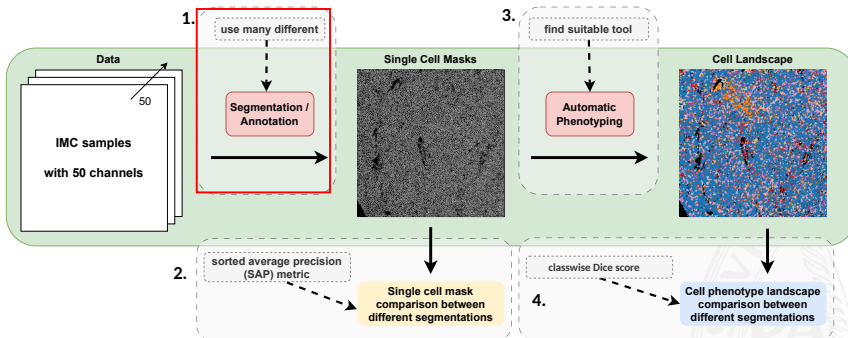
Study design



Expected possible outcomes



Block 1: Single Cell Mask Sources



Single Cell Mask Sources

Single Cell Mask Source	Abbreviation
Manual Annotator 1	M1
Manual Annotator 2	M2
Manual Annotator 3	M3
Manual Annotator 4	M4
DeepCell Mesmer ³ in steinbock framework	DC-F
DeepCell Mesmer ² standalone	DC
DeepCell Mesmer ² standalone stitched	DC-S
CellSAM ⁴ stitched	CS-S
Cellpose ⁵ v3	CP
Cellpose ² v3 stitched	CP-S
Vista2D ⁶ stitched	V2D-S

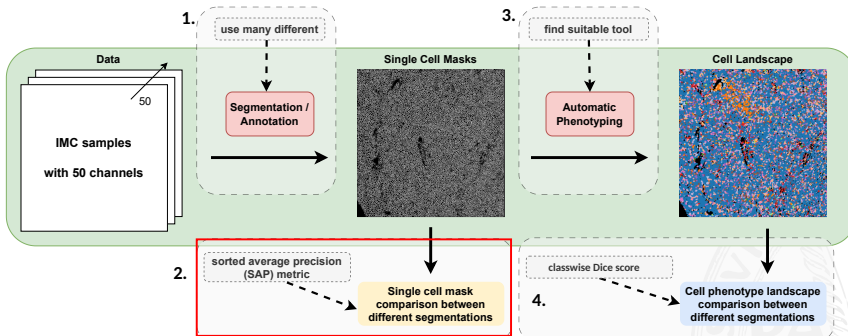
³ N. F. Greenwald et al. (2021). "Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning". In: *Nat Biotechnol*

⁴ U. Israel et al. (2023). *A Foundation Model for Cell Segmentation*. Preprint: biorxiv

⁵ C. Stringer et al. (2025). "Cellpose3: one-click image restoration for improved cellular segmentation". In: *Nature Methods*

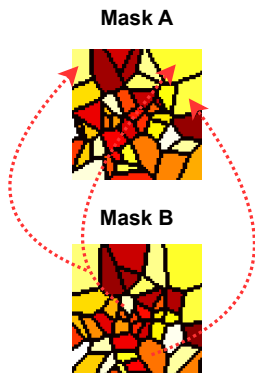
⁶ NVIDIA (2024). *VISTA-2D: A foundational model for cell segmentation in spatial omics workflows*. <https://github.com/Project-MONAI/VISTA/tree/main/vista2d>. Version 0.3.0

Block 2: Pairwise mask comparison using sortedAP



Evaluation Metric: SAP

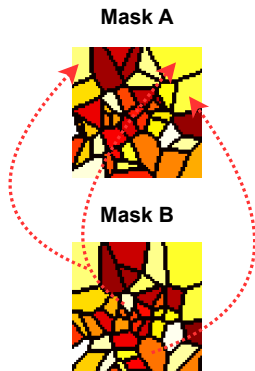
- 1 Calc IoU between all corresponding cell instances of two images



	B_0	B_1	B_2	B_3	B_4	B_5	...
A_0	.34	.12	.22	0	0	0	
A_1	.01	.15	.04	0	0	0	
A_2	0	.21	.16	0	0	0	
A_3	0	.07	.01	.05	.07	0	...
A_4	0	0	.16	.27	.03	0	
A_5	0	0	0	0	.60	0	
...				...			

Evaluation Metric: SAP

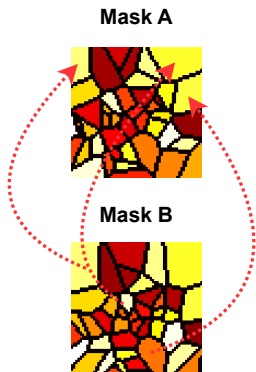
- 1 Calc IoU between all corresponding cell instances of two images



	B_0	B_1	B_2	B_3	B_4	B_5	...
A_0	.34	.12	.22	0	0	0	
A_1	.01	.15	.04	0	0	0	
A_2	0	.21	.16	0	0	0	
A_3	0	.07	.01	.05	.07	0	...
A_4	0	0	.16	.27	.03	0	
A_5	0	0	0	0	.60	0	
...				...			

Evaluation Metric: SAP

- 2 Determine matching objects by treating this as an max-weight assignment problem



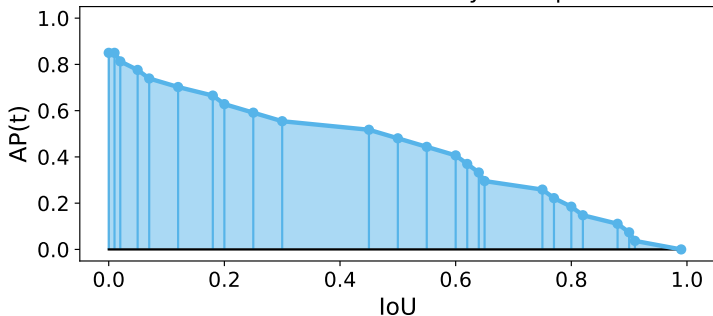
	B_0	B_1	B_2	B_3	B_4	B_5	...
A_0	.34	.12	.22	0	0	0	
A_1	.01	.15	.04	0	0	0	
A_2	0	.21	.16	0	0	0	
A_3	0	.07	.01	.05	.07	0	...
A_4	0	0	.16	.27	.03	0	
A_5	0	0	0	0	.60	0	
...				...			

Evaluation Metric: SAP

- Sort matched values, compute $AP(t)$ at every step and calc area under curve⁷

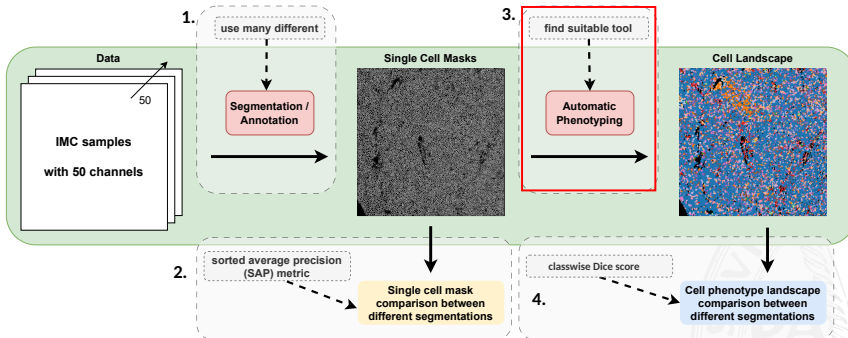
$$AP(t_{IoU}) = \frac{TP(t_{IoU})}{TP(t_{IoU}) + FN(t_{IoU}) + FP(t_{IoU})}$$

Metric visualized on toy example



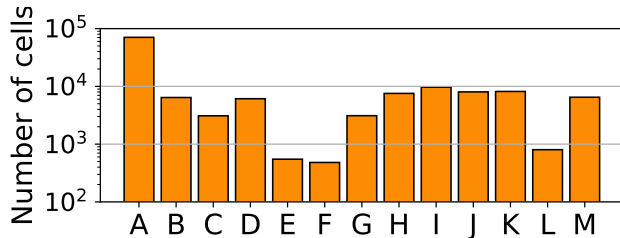
⁷ L. Chen et al. (2023). "SortedAP: Rethinking Evaluation Metrics for Instance Segmentation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*

Block 3: Automatic phenotyping approaches



Phenotyping ground truth

- A "ground truth" (expert phenotyping) was created using the steinbock framework: Deepcell Mesmer for Segmentation + FlowSOM for assigning classes to cells.



A: naive B cells
B: Dendritic Cells
C: Germinal Center B cells
D: Macrophages
E: Natural Killer cells
F: Plasmablasts
G: Smooth Muscle Cells

Legend

H: CD4 memory T cells
I: CD4 terminal effector memory T cells
J: CD8 effector memory T cells
K: CD8 terminal effector memory T cells
L: regulatory T cells
M: artefacts

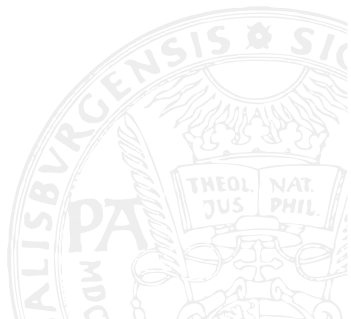
Automatic phenotyping approaches

Four approaches were considered (Cross Validation!)

1 Phenotyping foundation model: KRONOS⁸

⁸

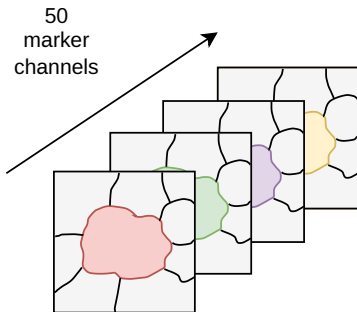
M. Shaban et al. (2025). *A Foundation Model for Spatial Proteomics*.



Automatic phenotyping approaches

Four approaches were considered (Cross Validation!)

- 1 Phenotyping foundation model: KRONOS⁸
- 2 Average expression per channel



Average expression:

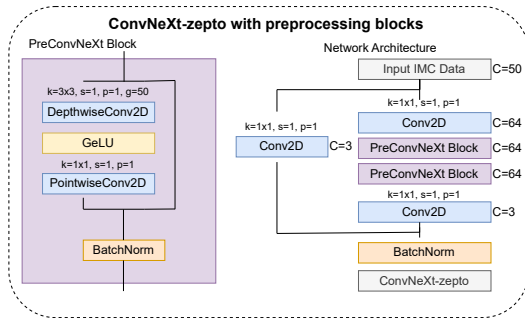
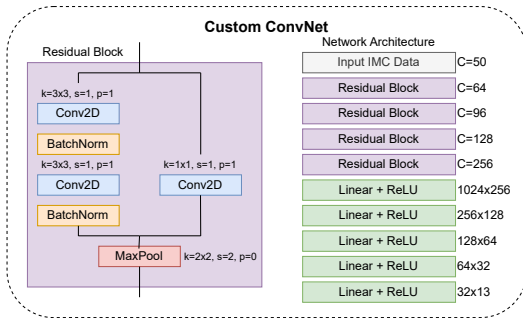
Take average per channel into feature vector

$$v \in \mathbb{R}^{50}$$

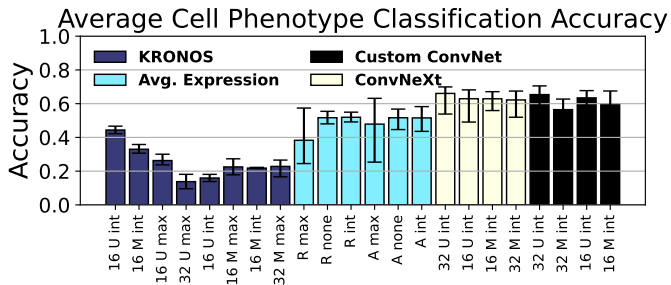
Automatic phenotyping approaches

Four approaches were considered (Cross Validation!)

- 1 Phenotyping foundation model: KRONOS⁸
- 2 Average expression per channel
- 3 ConvNets:

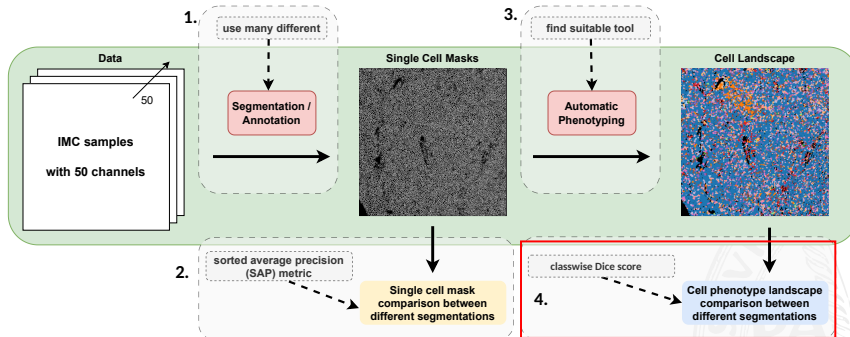


Automatic phenotyping approach: Classification results

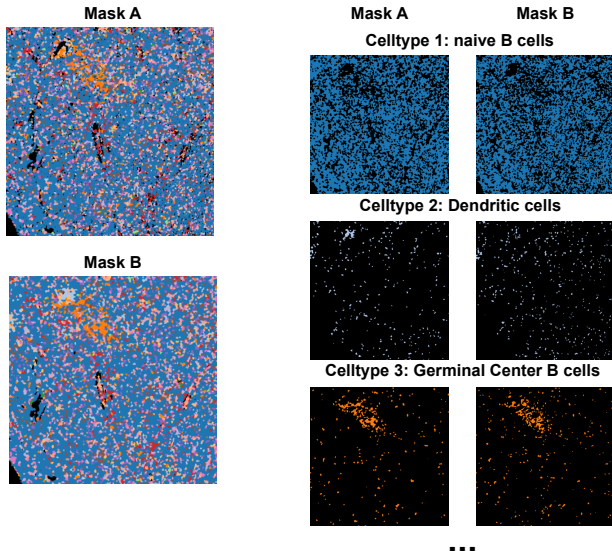


- U/M: unmasked/masked
- 16/32: pixel size around masks
- R/A: reduced marker set (KRONOS compatible) / all 50 markers
- int/max/none: normalization of channels (int= $\frac{1}{2^{16}}$)

Block 4: Landscape overlap metric - weighted DSC



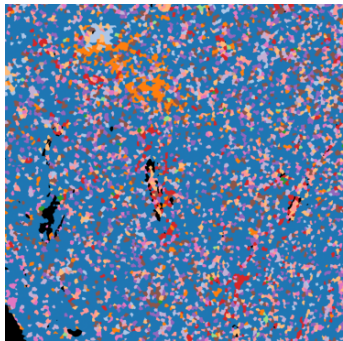
Block 4: Landscape overlap metric - weighted DSC



$$\text{DSC} = \frac{2 \cdot \text{Area}_{\text{overlap}}}{\text{Area}_{\text{mask A}} + \text{Area}_{\text{mask B}}}$$

Comparison against random Voronoi masks as sanity check

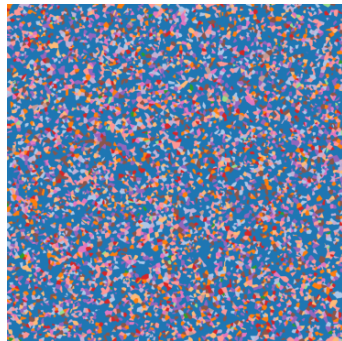
Real Mask



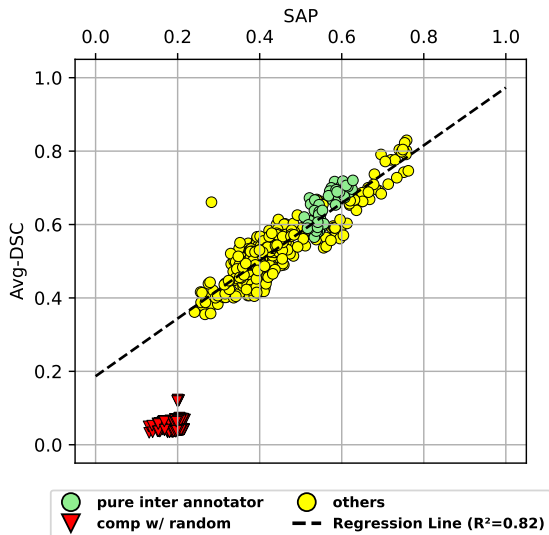
Random Voronoi mask
with known celltype
distribution



Random Voronoi mask
with known celltype
distribution



SAP vs Dice

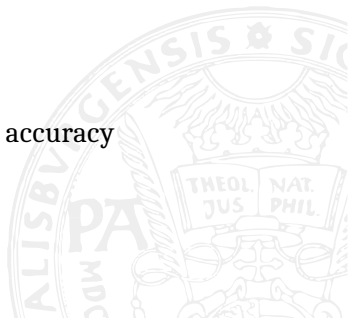


This study did

- We investigated how changes in segmentation masks translate to automatic cell phenotyping
- Results suggest that masks "good enough" are not sufficient and masks should be as good as possible

Limitations

- Phenotyping / Cell classification model did not achieve 100% accuracy



Q & A

Find resources here:

Data



Code (not yet uploaded)



Extra Slide: Qualitative example of overlap per celltype

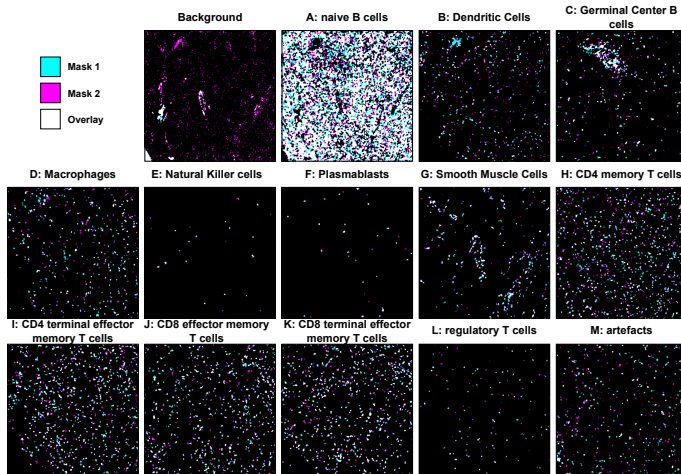


Figure: Exemplary phenotype landscape visualization using results from two segmentation masks on the same sample. Source mask 1: Annot1, Source mask 2: deepcell; sample id: 50. SAP: 0.44, avg-DSC: 0.44.