Frederick National Laboratory
for Cancer Research

sponsored by the National Cancer Institute

Volume electron microscopy: advances and challenges in DL-based segmentation.

Kedar Narayan
Group Leader, Center for Molecular Microscopy
NCI & FNL

DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute

Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute

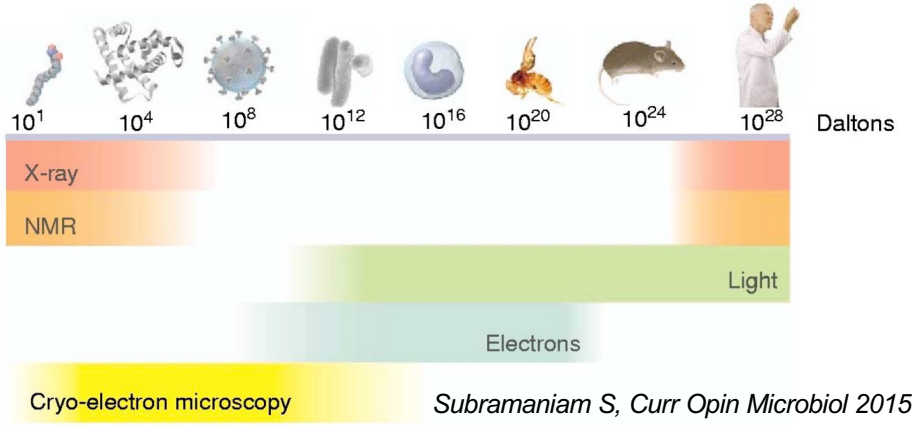
Outline

- What is volume electron microscopy (vEM), what's the data like
- Segmentation challenges in vEM
- CEM500K as a resource for the community
- Outlook

Take-aways

- An understanding of volume EM image data
- Our approach of tackling the segmentation bottleneck
- An exciting area for DL work!

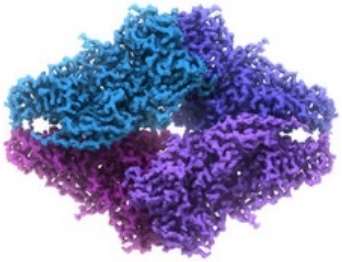
What is volume electron microscopy ? (Hint: It's NOT cryoEM)

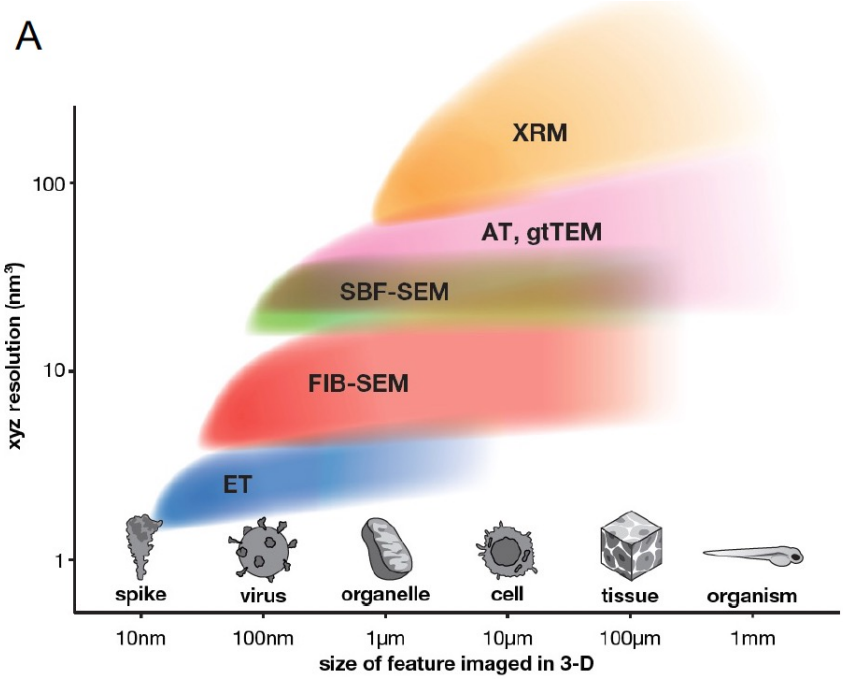


Subramaniam S, Curr Opin Microbiol 2015

Photons and electrons have widely varying resolving powers
Cell biology questions can be addressed by LM and EM
Correlative microscopy leverages their orthogonal advantages
LM and EM imaging (CLEM) can be combined in 2D and 3D

"cryoEM" = Structural studies of soluble and membrane protein complexes at near-atomic resolution.

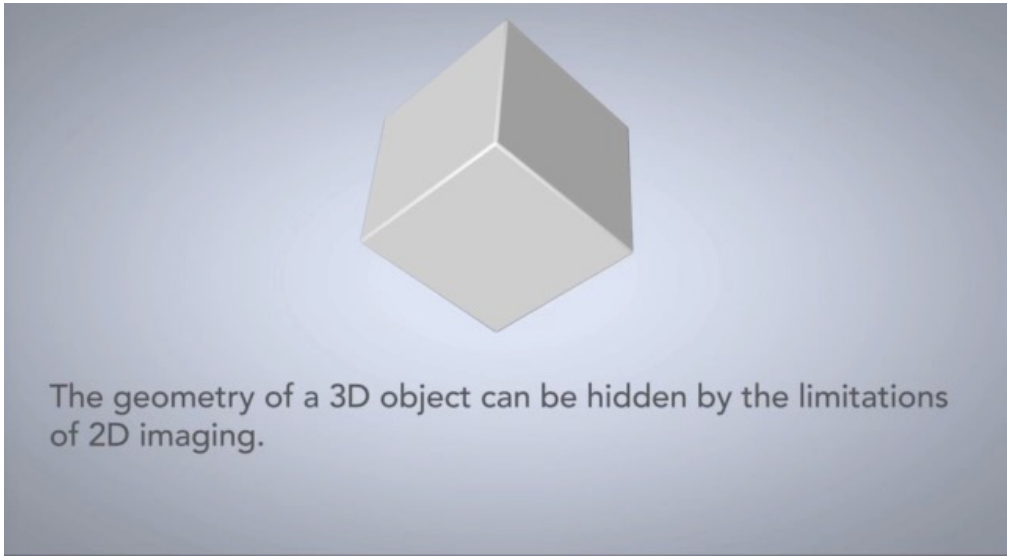




"**volume EM**" = A group of imaging approaches to study cells and tissue ultrastructure in 3D at nanoscale resolutions.

Baena V et al, Viruses 2021

Why volume electron microscopy ?



Typical vEM “pipeline”

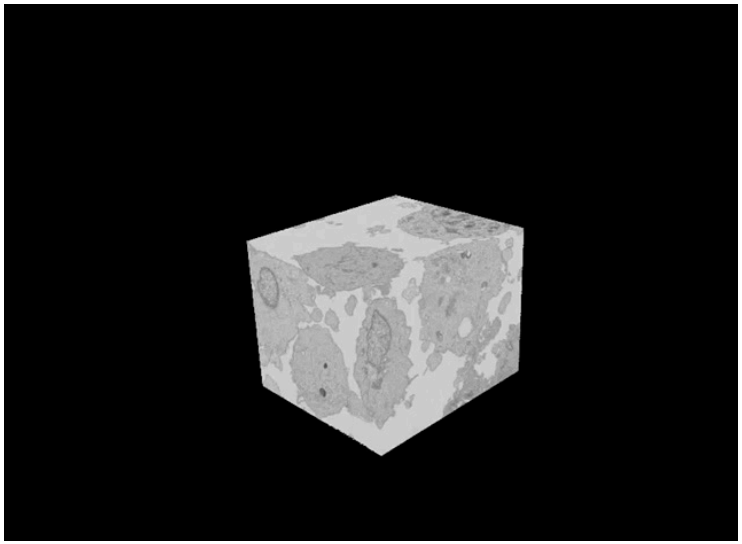
Baena V et al, Viruses 2021
Peddie C and Collinson, L Micron 2014
Narayan K and Subramaniam S, Nat Methods 2015

Biological experiment → Sample preparation → vEM imaging → Image processing → **segmentation** → analysis

vEM has turbocharged connectomics research
(many refs, read *Kubota Y et al, Front Neural Circuits 2018)*

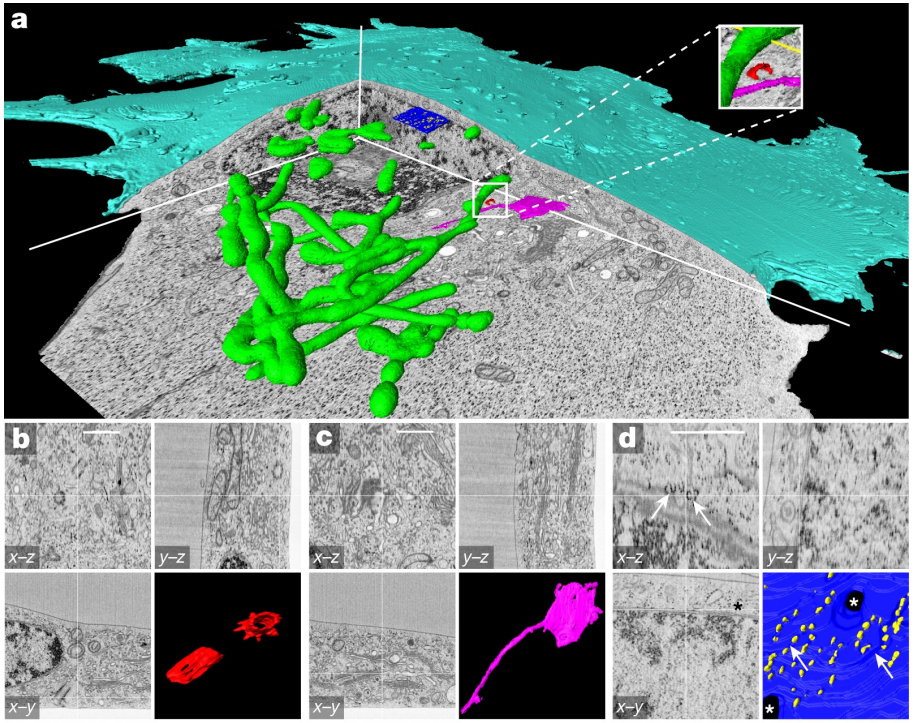
The vEM segmentation challenge

The conversion of large, information-rich, high-resolution low SNR, grayscale, “non-specific” 2D micrographs (stack) into accurate and precise binary label maps and 3D meshes for downstream analysis



vEM dataset sizes are mostly 1-100GB, now easily entering TB range (1 dataset published at 0.5 PB)

There have been significant advances made in this area in the recent past



Xu CS et al, Nature 2021

So… what's the problem?

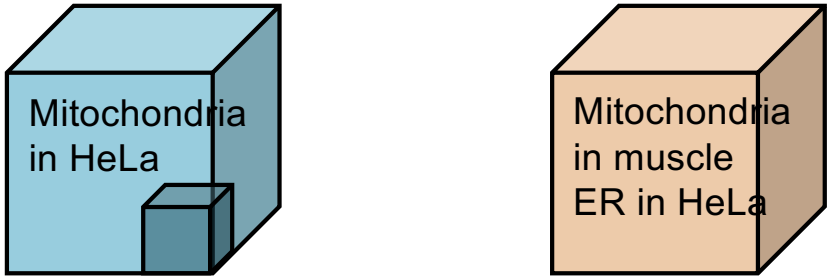
- Manual segmentation will never catch up with speed of acquisition
- Current segmentation efforts by DL approaches are improving throughput
- BUT transfer learning is a big problem

Typical vEM "automatic segmentation" pipeline:

- Acquire vEM data → Manually segment sub-volume → Train fancy model → Infer on full vEM dataset

Inadequate training and context can have bad consequences!
https://en.wikipedia.org/wiki/Ecce_Homo_(Mart%C3%ADnez_and_Gim%C3%A9nez)

Infer on slightly different dataset → poor results



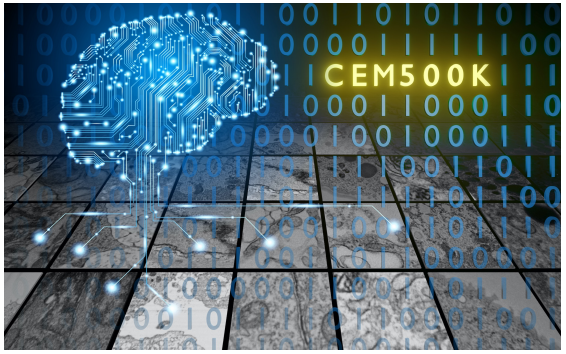
The limited data distribution during supervised training methods narrow the range of contexts available to a model.

CEM500K: a resource for DL-based segmentation of vEM data

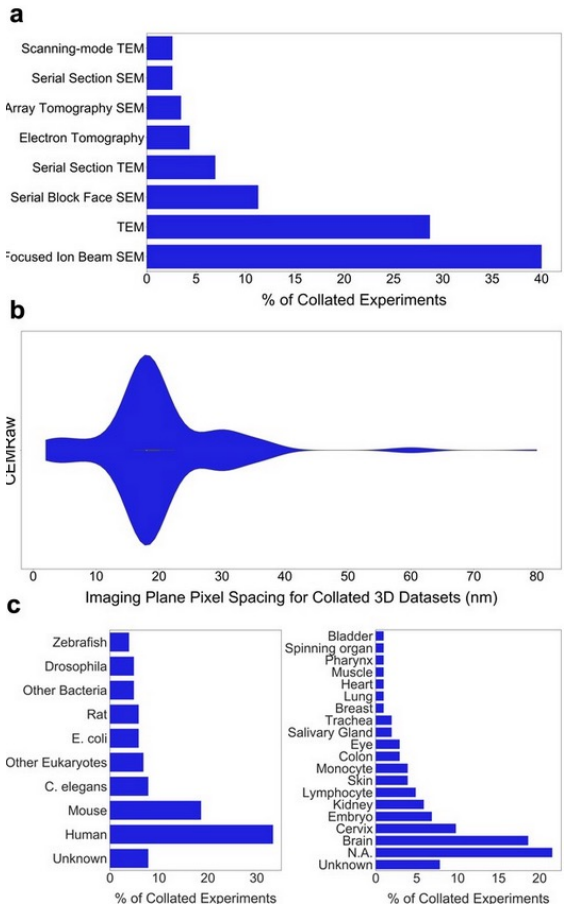
- **Insight:**
  - Provide the model data in more contexts, and remove constraints of supervision
- **Idea:**
  - Pre-train a model on a general task (generic feature recognition in EM images)
  - Then use the parameters for specific downstream tasks (organelle segmentation)
- **Approach:**
  - Curate a **relevant, heterogenous, information-rich, non-redundant** EM dataset
    - Cellular Electron Microscopy 0.5 x 10^6 images = CEM500K
  - Unsupervised model pre-trained on CEM500K = no need for up-front segmentation
    - Momentum Contrast algorithm (MoCoV2) for pre-training *He K et al, 2019. https://arxiv.org/abs/1911.05722*
  - Train and test against publicly available vEM benchmarks

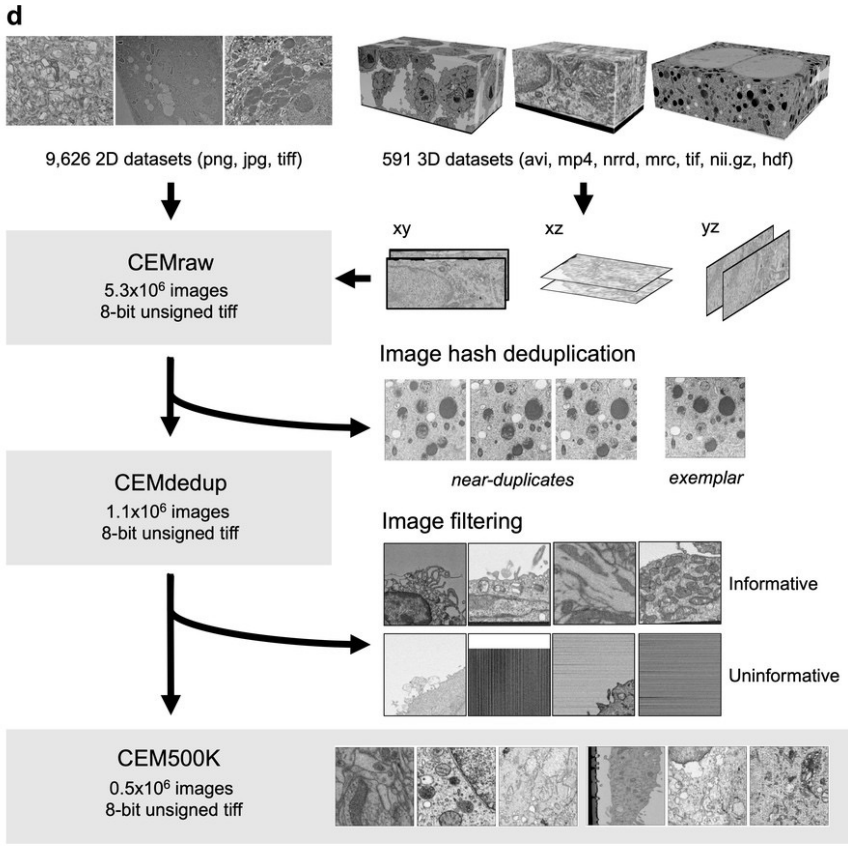
NOTE: the CEM500k dataset and pre-training approach is agnostic to the architecture of the models.

CEM500K: a resource for DL-based segmentation of vEM data



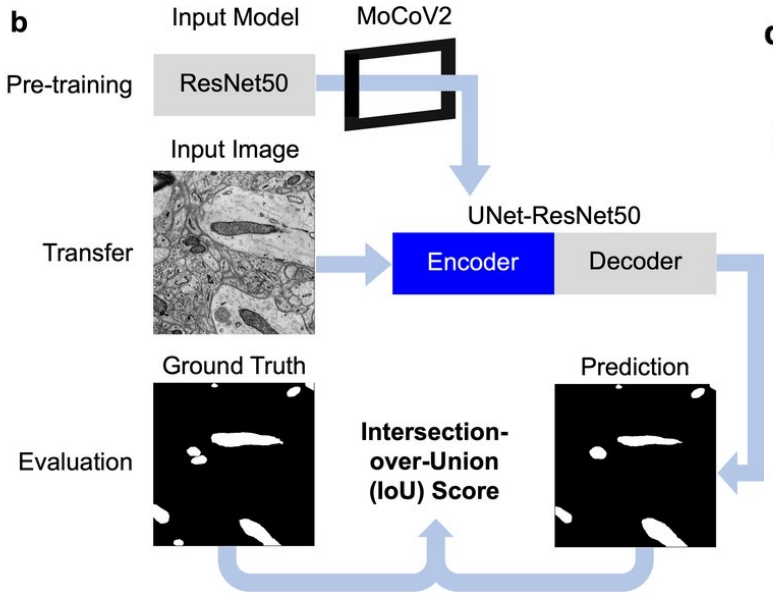
Ryan C and Narayan K, eLife 2021
https://elifesciences.org/articles/65894
https://github.com/volume-em/cellemnet

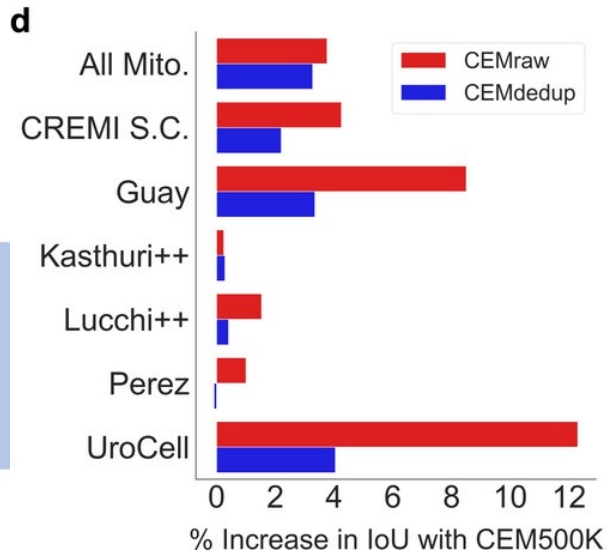


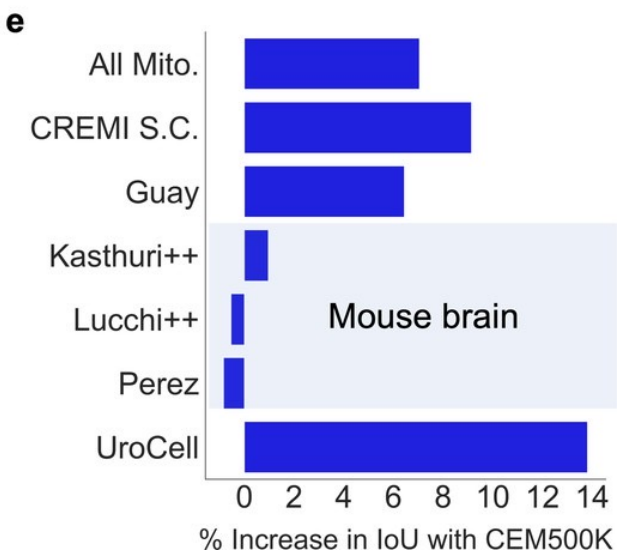


Frederick National Laboratory for Cancer Research

Pre-training by CEM500K improves transfer learning







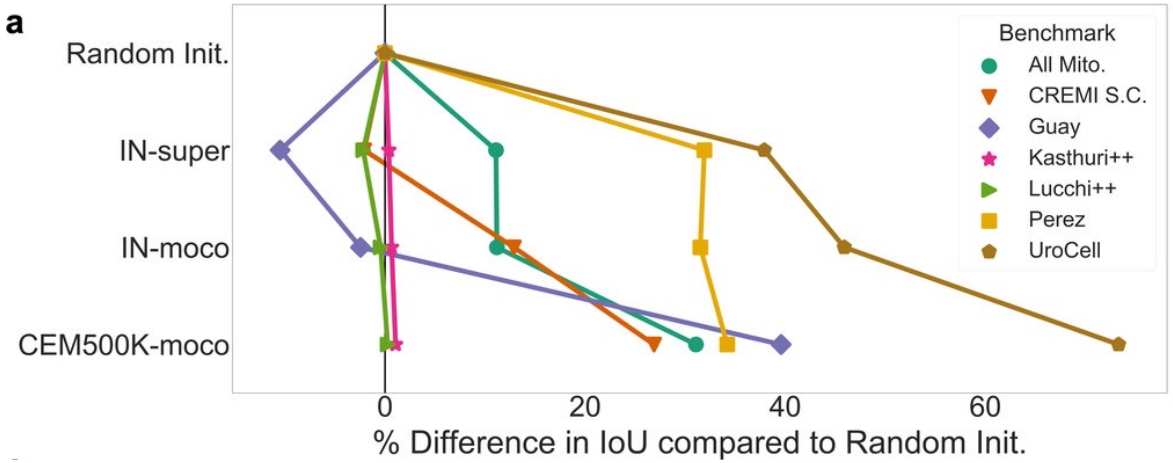
Overall strategy – nothing too fancy

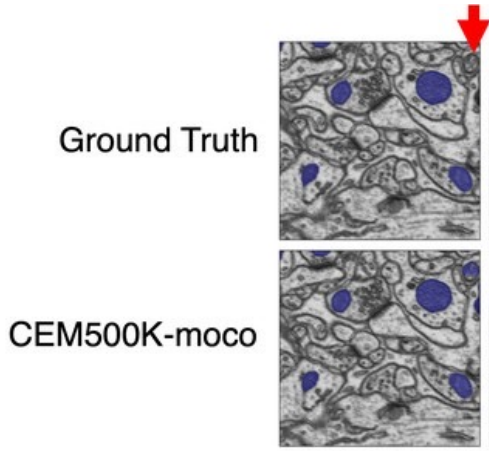
Trimming the CEM dataset improves performance

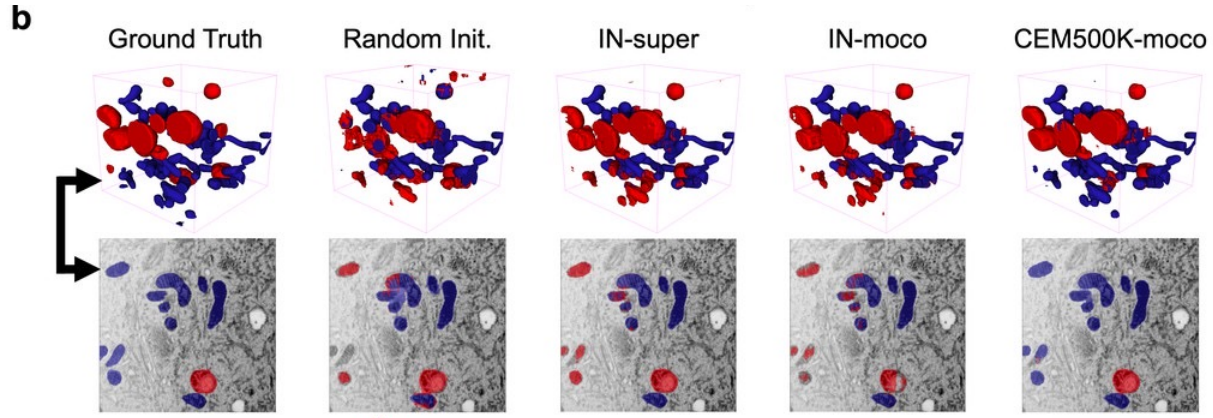
For diverse data, pre-training on CEM is better than on a mouse brain dataset

Cool result (Fig. 3): without *a priori* knowledge, the model recognizes organelles as relevant features in these images! The model also performs better with image variations (contrast, noise etc) expected from variable data acquisition

CEM500K beats current vEM benchmarks – and uncover human errors





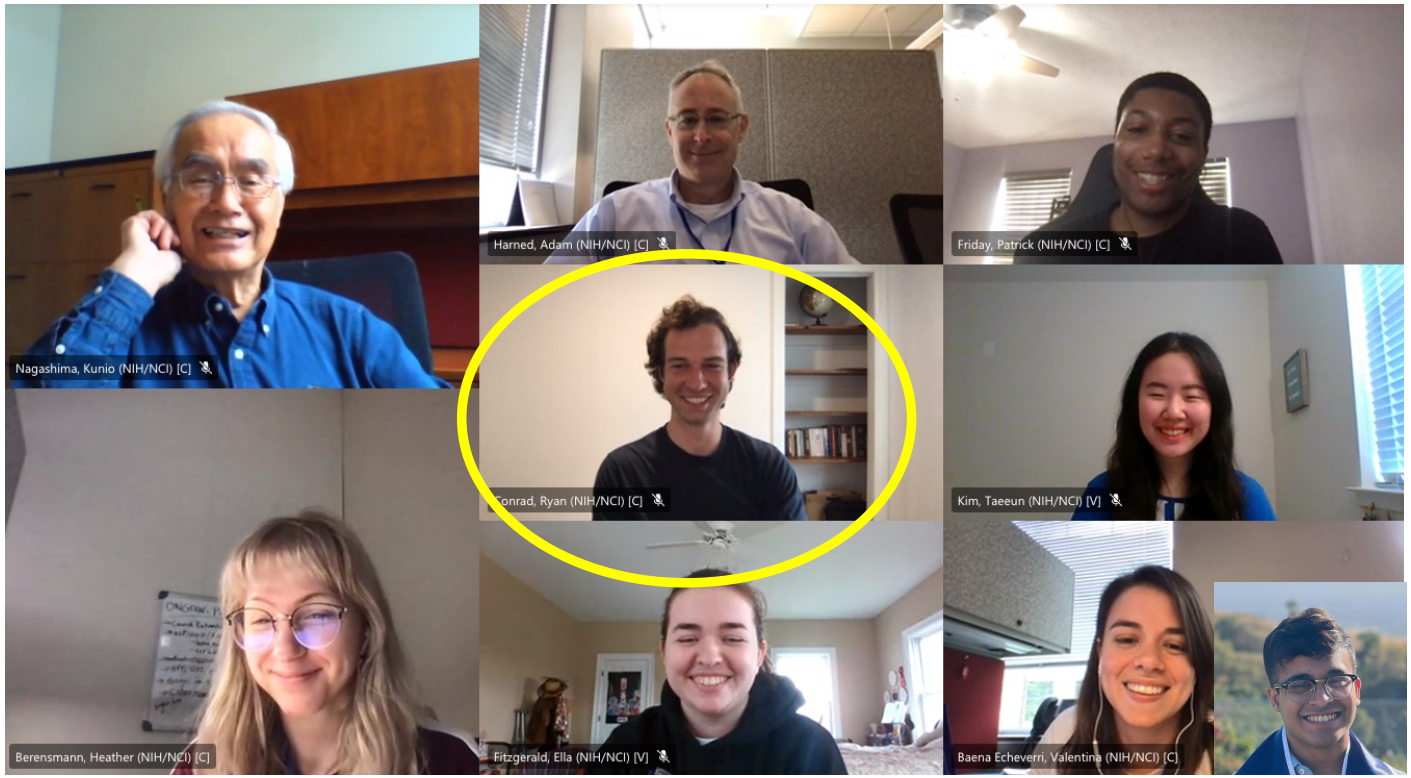


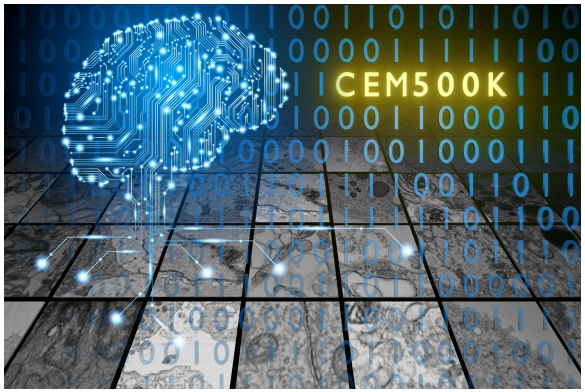
Benchmark	Training Iterations	Random Init.	IN-super	IN-moco	CEM500K-moco	Reported
All Mitochondria	10000	0.587	0.653	0.653	**0.770**	–
CREMI Synaptic Clefts	5000	0.000	0.196	0.226	**0.254**	–
Guay (Guay et al., 2020)	1000	0.308	0.275	0.300	**0.429**	0.417
Kasthuri++ (Casser et al., 2018)	10000	0.905	0.908	0.911	**0.915**	0.845
Lucchi++ (Casser et al., 2018)	10000	0.894	0.865	0.892	**0.895**	0.888
Perez (Perez et al., 2014)	2500	0.672	0.886	0.883	**0.901**	0.821

Outlook

- These are good results, but universal vEM segmentation models is the ultimate aim
- Need better/challenging benchmarks and community agreement on robustness metrics
- Better “DL + clean-up” pipelines, better communication with biologists
- Transition from “pretty pictures” to quantitative data
- Newcomers: Be wary of going down the rabbit hole with models and parameters
  - **Data (not model architecture) is key!**

Acknowledgments





https://insite.cancer.gov/community/corporate-communications/blog/2021/04/14/data-set-lets-ai-teach-itself-to-analyze-microscopy-better

Paper: https://elifesciences.org/articles/65894
Code: https://github.com/volume-em/cellemnet
Dataset: EMPIAR-10592